Tagsplanations: Explaining Recommendations Using Tags

Jesse Vig
Grouplens Research
University of Minnesota
jvig@cs.umn.edu

Shilad Sen
Grouplens Research
University of Minnesota
ssen@cs.umn.edu

John Riedl
Grouplens Research
University of Minnesota
riedl@cs.umn.edu

ABSTRACT
While recommender systems tell users what items they might like, *explanations* of recommendations reveal *why* they might like them. Explanations provide many benefits, from improving user satisfaction to helping users make better decisions. This paper introduces *tagsplanations*, which are explanations based on community tags. Tagsplanations have two key components: *tag relevance*, the degree to which a tag describes an item, and *tag preference*, the user's sentiment toward a tag. We develop novel algorithms for estimating tag relevance and tag preference, and we conduct a user study exploring the roles of tag relevance and tag preference in promoting effective tagsplanations. We also examine which types of tags are most useful for tagsplanations.

ACM Classification Keywords
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

Author Keywords
Explanations, tagging, recommender systems

INTRODUCTION
While much of the research in recommender systems has focused on improving the accuracy of recommendations, recent work suggests a broader set of goals including trust, user satisfaction, and transparency [16, 1, 13, 6]. A key to achieving this broader set of goals is to *explain* recommendations to users. While recommendations tell users what items they might like, explanations reveal *why* they might like them. An example is the "Why this was recommended" feature on Netflix[1]. Netflix explains movie recommendations by showing users similar movies they have rated highly in the past.

[1]http://www.netflix.com

IUI'09, February 8 - 11, 2009, Sanibel Island, Florida, USA.

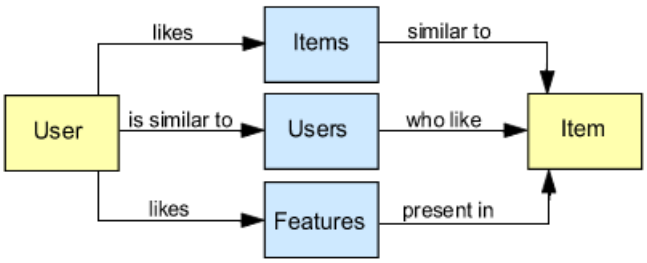

Figure 1: Intermediary entities (center) relate user to recommended item.

Research shows that explanations help users make more accurate decisions [1], improve user acceptance of recommendations [6], and increase trust in the recommender system [13]. Moreover, studies indicate that users *want* explanations of their recommendations – a survey of users of one movie recommender site showed that 86% of those surveyed wanted an explanation feature added to the site [6].

While many different types of explanation facilities exist, they all seek to show how a recommended item relates to a user's preferences. As Figure 1 illustrates, a common technique for establishing the relationship between user and recommended item is to use an *intermediary entity*. An intermediary entity is needed because the direct relationship between user and item is unknown, assuming that the user has not yet tried the item. In the Netflix example above, the intermediary entity is the list of previously rated movies shown in the explanation. The relationship between the user and these movies is that he or she has rated them positively. The relationship between these movies and the recommended movie is that other users who liked these movies also liked the recommended movie.

Explanations of recommendations fall into one of three categories: *item-based*, *user-based*, and *feature-based*, depending on the type of intermediary entity used to relate the user to the recommended item. In item-based explanations like the Netflix example, a set of items serves as the intermediary entity. User-based explanations utilize other users as intermediary entities. For example, Herlocker et al. designed a explanation that shows a user how other users with similar taste rated the recommended item [6]. Feature-based approaches use features or characteristics of the recommended item as intermediary entities. For example, one movie recommender prototype uses movie features including genre, director, and cast to justify recommendations [14].

We present a new type of explanation that uses tags as features, which we call a *tagsplanation*. The intermediary entity for tagsplanations is a tag or a set of tags. For example:

"We recommend the movie Fargo because it is tagged with ***quirky*** *and you have enjoyed other movies tagged with* ***quirky****."*

Tags have become increasingly popular on websites such as Delicious[2], Flickr[3], and Amazon[4], and they have many qualities that make them useful for explanations. As described in [5], tags may describe what an item is, what it is about, or what its characteristics are – all of which may be useful for explaining a recommendation. Another advantage is that no experts are needed to create and maintain tags, since tags are applied by the users themselves. Furthermore, tags provide both factual and subjective descriptions [11]. However, tags present unique challenges, including issues of tag quality [10] and tag redundancy [5].

We study two aspects of tag-based explanations: the relationship of the tag to the recommended item, which we call *tag relevance*, and the relationship of the user to the tag, which we call *tag preference*. Tag relevance represents the degree to which a tag describes a given item. For example, consider a tagsplanation for the movie *Pulp Fiction* that uses the tag "dark comedy". In this example, tag relevance would measure how well "dark comedy" describes *Pulp Fiction*. Tag preference, on the other hand, measures the user's sentiment to the given tag, for example how much the user likes or dislikes dark comedies. Tag relevance and tag preference are orthogonal to one another: the former is item-specific and the latter is user-specific.

Our design of tagsplanations is motivated by three goals: justification, effectiveness, and mood compatibility. *Justification* is the ability of the system to help the user understand why an item was recommended [6]. Justification differs from transparency [16] because justifications may not reveal the actual mechanisms of the recommender algorithm. Tintarev et al. define *effectiveness* as the ability of the explanation to help users make good decisions. *Mood compatibility* is the ability of the explanation to convey whether or not an item matches a user's mood. A recent study showed that users are interested in explanations with mood-related features [15].

In this paper, we investigate the roles of tag relevance and tag preference in tagsplanations. Specifically, we consider the following research questions:

RQ-Justification: What is the role of *tag preference* and *tag relevance* in helping users understand their recommendation?

Explanations help users understand why a given item was recommended. Tagsplanations promote understanding by demonstrating how a tag relates to the item and how the user relates to the tag. We investigate the role of these two components in helping users understand the recommendation overall.

RQ-Effectiveness: What is the role of *tag preference* and *tag relevance* in helping users determine if they will like the item?

We investigate whether users prefer information about the relationship between themselves and the tag (tag preference) or between the tag and the item (tag relevance) to make good decisions.

RQ-Mood: What is the role of *tag preference* and *tag relevance* in helping users decide if an item fits their current mood?

Recommender systems typically do not consider the user's mood when generating recommendations. Explanations provide users with additional information that can help them decide if an item fits their current mood. We investigate the relative importance of tag preference and tag relevance for revealing mood compatibility.

RQ-Tag-Type: What types of tags are most useful in tagsplanations?

A wide variety of tags may be applied to an item, some of which may be more suitable for explanations than others. As discussed in [10], *factual* tags identify facts about an item such as people, places, or concepts, while *subjective* tags express users' opinions of an item. We investigate the relative usefulness of factual tags versus subjective tags, and analyze which specific tags perform best.

To answer these research questions, we designed a tagsplanation feature for the MovieLens movie recommender site[5] and conducted an online user study. Participants in the study viewed 4 types of tagsplanations, each of which handles tag preference and tag relevance in a different way. Participants evaluated each tagsplanation based on how well it achieved the goals of justification, effectiveness, and mood compatibility. We then analyze the results to determine the roles of tag relevance and tag preference in promoting these 3 goals. Subjects also evaluate specific tags, and we compare the results for subjective tags versus factual tags.

RELATED WORK

Item-based explanations. Item-based approaches use items as intermediary entities to relate the user to the recommended item. An example is the "Why this was recommended" feature on Netflix, which shows users their past ratings for a set of related movies. Similarly, Amazon shows users their past purchases that motivated a recommendation of a new item. Studies show that item-based explanations improve users'

[2]http://del.icio.us
[3]http://www.flickr.com
[4]http://www.amazon.com
[5]http://www.movielens.org

acceptance of recommendations [6] and help users make accurate decisions [1]. As discussed in [15], a shortcoming of item-based explanations is that users may not understand the connection between the explaining items and the recommended item.

Item-based approaches have several properties that we also use in our tag-based approach. First, item-based approaches present the relationship between the user and the set of related items in a way that users can easily interpret. In the Netflix example, the relationship is defined by the rating the user has given to each movie shown in the explanation. We use a similar approach, by expressing the relationship between user and tag as a 1-to-5 star inferred rating. Second, item-based explanations often use ratings correlation as a criteria for choosing the related items. We use a similar approach for tagsplanations by selecting tags with preference values that strongly correlate with ratings for the item.

User-based explanations. User-based explanations utilize other users as intermediary entities. The relationship between the main user and the explaining users is typically that they share similar tastes, and the relationship between the explaining users and the recommended item is that they have rated the item positively. For example, Herlocker et al. developed an explanation facility that displays a histogram of ratings of the recommended item by other users with similar taste [6]. This approach was successful at persuading users to try an item, but less effective at helping users make accurate decisions [1]. Motivated by these results, we study how well tagsplanations help users make accurate decisions.

Feature-based explanations. Feature-based approaches use qualities or characteristics of the recommended item as intermediary entities. Two types of features have been used in recommendation explanations:

- **Predefined categories.** Tintarev et al. developed a movie recommender prototype that explains recommendations by using categories such as genre, director, and cast [14]. While these features have tremendous potential for use in recommendation explanations, using predefined categories like these also presents several challenges. Shirky describes several such issues: (1) experts or system designers are needed to create the categorization scheme, (2) the categorization scheme might not be sufficiently descriptive, and (3) the categories may be too static to accommodate changing needs [12]. A tag-based approach addresses these issues by putting control in the hands of the users and allowing free-form metadata.
- **Keywords.** Many types of items, such as books, articles, or websites, contain textual content that may be mined for keywords [7, 2]. Keyword-based approaches use these words as features in the explanation. The Libra book recommender, for example, extracts keywords from the book text based on their predictive power in a naive Bayesian classifier and uses them to explain the recommendation [7]. Bilgic et al. showed that these explanations helps users make more accurate decisions [1].

  Explanations using item content face several challenges, many of which may be addressed by using tags. One limitation is that items such as music or pictures may not have readily available textual content. In contrast, tags may be applied to virtually any type of item. A second issue is that keywords extracted from content represent data rather than metadata, and therefore may be too low-level. In one example described in [1], the keywords used to explain a book recommendation are *Heart*, *Beautiful*, *Mother*, *Read*, *Story*. While these words do suggest qualities of the book, meta-level tags such as *fiction*, *mother-daughter relationship*, or *touching* might be more suitable.

  We considered adapting keyword-style approaches to generate tagsplanations. Existing keyword-style approaches [7, 1] only require that an item have a set of words and corresponding frequencies. If tags were used rather than words, the same algorithm could be utilized. We chose to develop a new approach for several reasons. First, existing keyword-style approaches require that users rate items using binary categories, such as {*like*, *dislike*}. In the MovieLens recommender system used in our study, users rate items on a 5-star scale. Second, existing approaches do not account for two issues that tagging systems face: low quality tags [10] and redundant tags [5]. Third, we wished to represent the relationships between tag and item and between user and tag in a more intuitive way than in existing keyword-style approaches. For example, the Libra book recommender display the *strength* of each keyword, which equals the count of the word multiplied by the weight assigned by a naive Bayesian classifier. While these strength values may reflect a user's preferences towards the keywords, users may have difficulty interpreting the values.

DESIGN SPACE

The goal of explanations of recommendations is to relate the recommended item to the user's tastes and interests. As discussed in the introduction, a common approach for establishing this relationship is to use an intermediary entity that relates to both user and recommended item. For tagsplanations, the intermediary entity is a tag. Tagsplanations must clearly identify the relationship between user and tag and between tag and recommended item.

A recommendation explanation may be one of two types: description or justification. *Descriptions* reveal the actual mechanism that generates the recommendation. For example, k-nearest-neighbor item-item algorithms generate recommendations based on a user's ratings for the k items most similar to the recommended item [8]. In this case, an item-based explanation could be used to accurately depict the algorithm. *Justifications*, on the other hand, convey a conceptual model that may differ from that of the underlying algorithm. For example, a book recommender might use a k-nearest-neighbor item-item algorithm to recommend books, but may justify a recommendation based on the fact that the book was written by a user's favorite author.

While descriptions provide more transparency than justifications, there are several reasons why justifications might be

preferred. First, the algorithm may be too complex or unintuitive to be described in terms that a user can understand. Dimension reduction models, for example, generate recommendations based on a set of latent factors in the data that may not have a clear interpretation [4]. Second, the system designers may want to keep aspects of the algorithm hidden, for example to protect trade secrets. Third, using justification allows a greater freedom in designing explanations since they are not constrained by the recommender algorithm.

Relationship between item and tag: tag relevance

We use the term *tag relevance* to describe the relationship between a tag and an item. One possible measure of relevance is tag popularity. That is, have many users applied a given tag to an item or only a few users? A tag applied by many users is probably more relevant to the given item. Another potential measure of relevance is the correlation between item preference and tag preference. That is, do users who like a given tag also tend to like the associated item? A strong correlation may suggest that the tag is highly relevant. While both tag popularity and preference correlation may indicate tag relevance, the two measures may not agree with one another. On MovieLens, the tag "Bruce Willis" is unpopular for the film *Predator* (no one has applied the tag to this movie); we suspect this is because Bruce Willis is not in *Predator*. However, a strong correlation may exist between users' preference for the tag "Bruce Willis" and their preference for *Predator* because *Predator* is the type of movie Bruce Willis is often in.

Tag relevance may be represented as either a binary relationship (*relevant*, *not relevant*) or as a value on a continuous scale. While a binary relationship might convey a simpler conceptual model, it lacks the precision of a continuous scale. Users may wish to know *how* relevant a given tag is, rather than just whether it is relevant or not.

Relationship between user and tag: tag preference

We define tag preference to be the user's sentiment towards a tag. The key design choices concern how to represent this relationship and how to compute the value for a given user and tag. Tag preference may be modeled as a binary relationship (*like*, *dislike*) or as a continuous variable representing the degree to which the user likes or dislikes a tag. A binary approach provides a simpler conceptual model to users, but is less precise than a continuous approach. A user's preference towards a tag may be computed in one of two ways. Preference may be assessed directly, by asking a user his or her opinion of a tag, or it may be inferred based on a user's actions. For example, if a user tends to give high ratings to items that have a particular tag, the system could infer that he or she has a positive preference toward that tag. An advantage of inferring tag preference over asking users directly is that no additional work is needed from the users.

DESIGN DECISIONS

Platform

We used the MovieLens movie recommender website (Fig. 2) as a platform for implementing tagsplanations. MovieLens members rate movies on 5-star scale and receive recommendations based on a k-nearest-neighbor item-item recommender algorithm. Over 164,000 members have rated at least one movie, and members have contributed over 15 million movie ratings in total. In January 2006, MovieLens introduced tagging [11, 10]. 3,758 users have created 14,675 distinct tags, which have been applied to 7,268 different movies. Users rate the quality of tags by clicking on a thumbs-up or thumbs-down icon next to each tag.

Figure 2: A sample screen on MovieLens

MovieLens does not take tags into account when generating movie recommendations. Therefore, tag-based explanations serve as a justification rather than a description of the underlying recommender algorithm.

Tag Preference

In this section we first give an overview of our approach, followed by formal definitions. As previously discussed, tag preference may be computed in one of two ways. Tag preference may be measured directly, by asking a user his or her opinion of a tag, or it may be inferred based on user behavior. We use the latter approach, in order to spare users from having to explicitly evaluate many different tags. Specifically, we infer users' tag preferences based on their movie ratings. We use movie ratings because they are the central mechanism on MovieLens by which users express preferences and because they are in abundant supply (the average MovieLens user has rated 75 movies).

To estimate a user's preference for a tag, we compute a weighted average of the user's ratings of movies with that tag. For example, suppose we are estimating Alice's preference for the tag "violence" and Alice has rated the following movies tagged with "violence": *Planet of the Apes*, *Reservoir Dogs*, *Sin City*, and *Spider-Man 2*. Her inferred preference toward "violence" is a weighted average of those four movie ratings. We chose to use a weighted average for two reasons. First, it provides a simple and computationally efficient algorithm for computing tag preference. Second, it guarantees that the computed tag preference values will lie in the same 0.5 - 5.0 range as the movie ratings, since the output value of a weighted average is bounded by the minimum and maximum input values. This allows us to represent tag preference values using the familiar 5-star paradigm used for movie ratings. Although we considered a binary representation (*like*, *dislike*), we chose the 5-star scale because it provides a fine-grained level of information yet is easy to interpret since it follows the standard of the movie rating scale. Previous studies have shown that users prefer fine-grained rating scales [3].

We weight the average because some movie ratings may suggest tag preference more strongly than others. For example, *Reservoir Dogs* is much more violent than *Planet of the Apes*, so a user's rating for *Reservoir Dogs* is probably a stronger indicator of their preference toward "violence" than their rating for *Planet of the Apes*, even though both movies have been tagged with "violence". We use tag frequency to measure the relative importance of a movie in determining tag preference. For example, MovieLens users have tagged *Reservoir Dogs* with "violence" 7 times, while *Planet of the Apes* has only been tagged once. Therefore we assign a higher weight to *Reservoir Dogs* when computing a user's preference for "violence".

We now provide formal definitions for the concepts described above. First, we define *tagshare*[6], a measure of tag frequency that is used to assign weights to movies. The *tagshare* of a tag t applied to an item i is the number of times t has been applied to i, divided by the number of times *any* tag has been applied to i. We denote this value as $\text{tag\_share}(t, i)$. For example, on MovieLens the tag "Bruce Willis" has been applied to the movie *Die Hard* 8 times and the number of applications of all tags to *Die Hard* is 56. Therefore $\text{tag\_share}(\text{"Bruce Willis"}, \textit{Die Hard})$ equals 8/56 = 0.14. We add a constant smoothing factor of 15 to the denominator when computing tagshare, in order to reduce the value when an item has a small number of tags[7]. This smoothing reflects the possibility that applications of a tag may be due to chance.

We now formally define the measure we use to estimate tag preference, which we call *tag_pref*. Let I_u to be the set of items that user u has rated, let $r_{u,i}$ to be the rating user u has given to item i, and let $\bar{r}_u$ be user u's average rating across all items. User u's tag preference for tag t is computed as follows:

$$\text{tag\_pref}(u,t) = \frac{\left(\sum_{i \in I_u} r_{u,i} \cdot \text{tag\_share}(t,i)\right) + \bar{r}_u \cdot k}{\left(\sum_{i \in I_u} \text{tag\_share}(t,i)\right) + k}$$

$\text{tag\_pref}(u,t)$ is undefined if user u has not rated any items with tag t. k is a smoothing constant that we assign a value of 0.05[7]. The smoothing constant k accounts for users who have rated few items with a given tag. This smoothing serves to bring the computed tag preference closer to the user's average rating, because ratings of a small number of items may not properly reflect a user's tag preference.

Tag Relevance

As discussed earlier, two possible measures of tag relevance are tag popularity and preference correlation. Tag popularity reflects the number of users who have applied the tag to the movie, while preference correlation is the correlation between users' preference for the tag and their preference for the movie. We chose to use preference correlation to represent tag relevance, because it directly relates tag preference (the relationship between user and tag) with item preference (the relationship between user and item). To determine the preference correlation between item i and tag t, we compute the Pearson correlation between the sequence of ratings for i across all users and the sequence of inferred tag preference values for t across all users. Tag popularity is used implicitly as a filtering criteria; we assign a relevance of zero to tags that have not been applied at least once to a given item. We represent tag relevance on a continuous scale rather a binary one (*relevant*, *not relevant*), because this allows users to discern the relative importance of one tag versus another when both tags are relevant to the item.

We now formally define our measure of tag relevance, which we call *tag_rel*. For a given tag t and item i, we define U_{ti} to be the subset of users who have rated i and have a well-defined tag preference for t. (Users must have rated at least one item with tag t in order to have a well-defined tag preference for t.) We define X to be the set of ratings for item i across all users in U_{ti}, adjusted by each user's average rating. That is, $X = \{r_{u,i} - \bar{r}_u : u \in U_{ti}\}$. We subtract each user's average rating to account for individual differences in rating behavior. We define Y to be the set of inferred tag preference values[8] toward tag t for all users in U_{ti}, adjusted by each user's average rating. That is, $Y = \{\text{tag\_pref}(u,t) - \bar{r}_u : u \in U_{ti}\}$. Tag_rel is defined using Pearson correlation, as follows:

$$\text{tag\_rel}(t,i) = \begin{cases} \text{pearson}(X,Y), & \text{if } t \text{ has been applied to } i; \\ 0, & \text{otherwise.} \end{cases}$$

Tag Filtering

We filtered tags based on three criteria:

- **Tag quality.** One study showed that only 21% of tags on MovieLens were of high enough quality to display to the community [10]. We filtered tags based on implicit and explicit measures of tag quality. First, we require that a tag has been applied by at least 5 different users and to at least 2 different items. Second, we require that the average thumb rating of a tag across all items satisfy a minimum threshold. As discussed earlier, MovieLens members use thumb ratings to give explicit feedback about tag quality. We used a smoothed average of thumb ratings as described in [10] and retained the top 30% of tags. However, we chose not to filter tags representing movie genres or names of directors and actors. MovieLens members tended to rate these tags poorly, but we suspect this is due to the fact that genre, cast, and director are displayed next to each film, and tags containing the same information might appear redundant. For tagsplanations, we do not display this information and therefore such tags would not be redundant.
- **Tag redundancy.** Different tags may have very similar meanings [5]. These similarities arise when two tags are synonyms (*film*, *movie*), different forms of the same

[6]The term *tagshare* was first used by Tim Spalding from LibraryThing in a blog post on February 20, 2007: http://www.librarything.com/thingology/2007/02/when-tags-works-and-when-they-dont.php

[7]We chose smoothing constants based on qualitative analysis over a series of test cases.

[8]When computing tag preference values for t, we excluded item i in order to avoid spurious correlations with ratings for i.

word (*violence*, *violent*), or at different levels of specificity (*comedy*, *dark comedy*). Our process for filtering redundant tags consists of three steps: preprocessing, redundancy detection, and winner selection. In the preprocessing step, we stem[9] the words in each tag and remove stopwords[10]. In the redundancy detection step, we use a simple heuristic: if two tags in the same explanation contain any of the same words or differ from one another by only 1 character, they are classified as redundant. The winning tag is the one with higher relevance to the given item.

- **Usefulness for explanation.** We removed all tags with an undefined value for tag preference (which occurs when a user has not rated any items with that tag) or with a tag relevance score less than 0.05. We also limited the number of tags we show in each tagsplanation to 15, in order to conserve screen space and to avoid overloading users with too much information.

Interface

Figure 3 shows an example of the tagsplanation interface. Tag relevance is represented as a bar of varying length, while tag preference is depicted as a number of stars rounded to the nearest half. An arrow indicates sort order of the tags. We designed four variations of the interface, which differ in the data displayed (tag relevance, tag preference, or both) and the sorting order (tag relevance or tag preference):

EXPERIMENT

We conducted a within-subjects study of each of the four interfaces: RelSort, PrefSort, RelOnly, and PrefOnly. Subjects completed an online survey in which they evaluated each interfaces on how well it helped them (1) understand why an item was recommended (*justification*), (2) decide if they would like the recommended item (*effectiveness*), and (3) determine if the recommended item matched their mood (*mood compatibility*). Based on survey responses, we draw conclusions about the role of tag preference and tag relevance in promoting justification, effectiveness, and mood compatibility.

- **Interface 1: RelSort.** Shows relevance and preference, sorts tags by relevance (Fig. 3)
- **Interface 2: PrefSort.** Shows relevance and preference, sorts tags by preference (Fig. 4)
- **Interface 3: RelOnly.** Shows relevance only, sorts tags by relevance (Fig. 5)
- **Interface 4: PrefOnly.** Shows preference only, sorts tags by preference (Fig. 6)

[9]We used Porter's stemming algorithm as implemented at http://nltk.sourceforge.net

[10]We used a standard list of stopwords from http://nltk.sourceforge.net plus one domain-specific stopword: "movie"

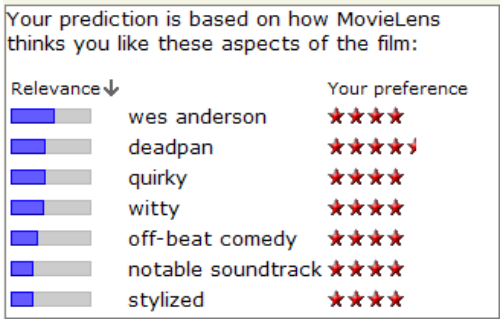

Figure 3: RelSort interface, for movie *Rushmore*. (The list of tags shown in each of these 4 figures is truncated).

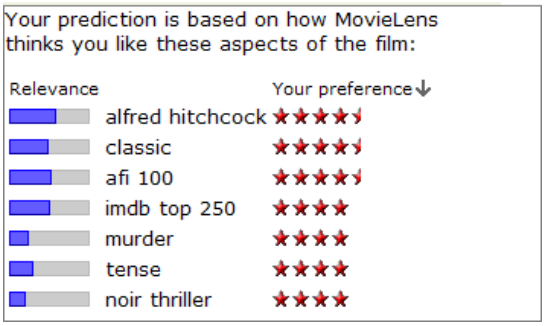

Figure 4: PrefSort interface, for movie *Rear Window*.

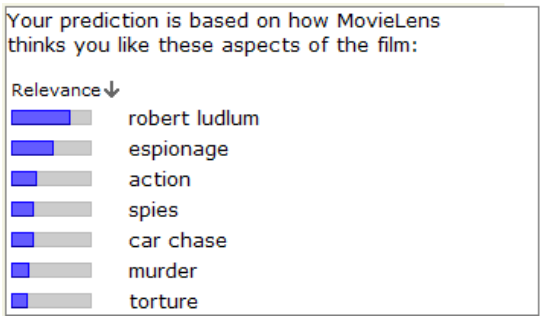

Figure 5: RelOnly interface, for movie *The Bourne Ultimatum*.

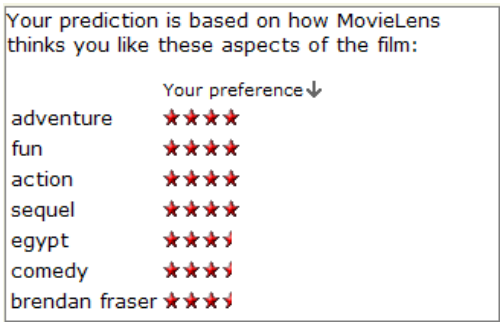

Figure 6: PrefOnly interface, for movie *The Mummy Returns*.

Methodology

We invited 2,392 users of MovieLens to participate in the study, which was launched on 08/19/2008. We only included

members who had logged in within the past 6 months and who had rated at least 50 movies. 556 users accepted the invitation. In the study, we included movies with at least 10 tags (after tag filtering) and between 1000 and 5000 movie ratings. 93 movies satisfied these conditions. We placed bounds on the number of ratings for each movie in order to find movies that were neither very popular nor obscure, because we wished to provide a mix of movies that some subjects had seen and others had not seen. Clearly a live system would not use such filters. However, we found there was no statistically significant difference in survey responses for movies with a higher number of ratings (4000 to 5000 ratings) versus movies with a lower number of ratings (1000 to 2000). We leave it to future work to determine the minimum amount of rating and tagging data needed to produce good-quality tagsplanations.

Each participant took a personalized online survey, consisting of three parts:

- **Part 1: Evaluation of tagsplanations**

  We showed each subject tagsplanations for 4 different movies, drawn randomly from the pool of movies that satisfied the filtering conditions described above. Further, we only showed users a tagsplanation if they had a well-defined tag preference for at least 7 of the tags (reducing the number of eligible movies per user from 93 to 83, on average). Subjects only saw tagsplanations for movies they *had not* seen, and they verified whether or not they had seen each movie. Each of the 4 tagsplanations utilized a different interface (RelSort, PrefSort, RelOnly, or PrefOnly) so that each subject would see all 4 interfaces. To account for order effects, we presented the 4 interfaces in a random order.

  For each tagsplanation, participants responded to the following statements using a 5-point Likert scale[11]:

  1. *This explanation helps me understand my predicted rating.* (Justification)
  2. *This explanation helps me determine how well I will like this movie.* (Effectiveness)
  3. *This explanation helps me decide if this movie is right for my current mood.* (Mood compatibility)

- **Part 2: Evaluation of particular tags**

  We showed users three randomly chosen tags from the last tagsplanation they saw in Part 1 of the study. For each tag, they responded to the following statements using a 5-point Likert scale:

  1. *This **element** helps me understand my predicted rating.*
  2. *This **element** helps me determine how well I will like this movie.*
  3. *This **element** helps me decide if this movie is right for my current mood.*

[11]The possible responses were: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree.

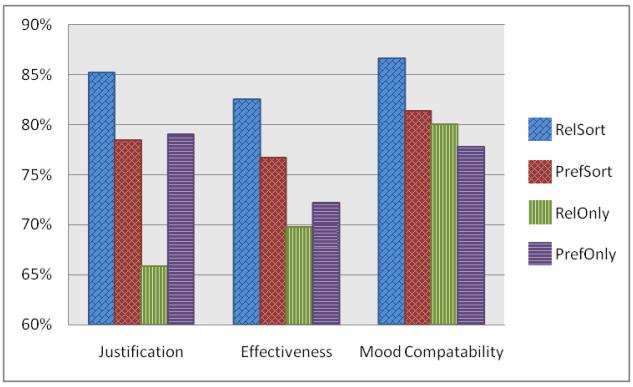

Figure 7: Percentage of responses that were *agree* or *strongly agree*, broken down type of interface (RelSort, PrefSort, RelOnly, PrefOnly). Neutral responses were excluded from the calculations.

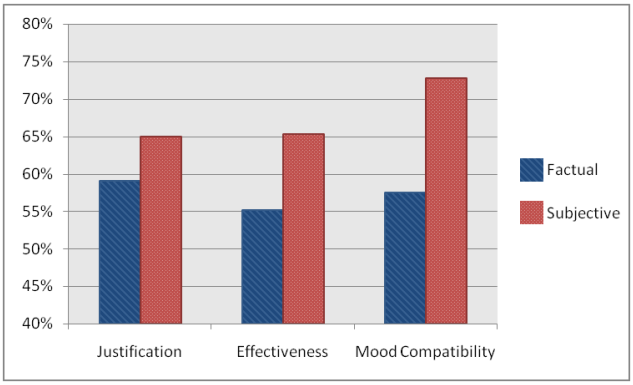

Figure 8: Percentage of responses that were *agree* or *strongly agree* for each type of statement, broken down by tag type (factual, subjective). Neutral responses were excluded from the calculations.

- **Part 3: Verification of tagsplanation accuracy**

  Subjects evaluated tagsplanations for two movies they had seen in the past, drawn from the same pool of movies described above. The purpose was to verify the accuracy of tagsplanations, since subjects had otherwise only evaluated tagsplanations for movies they *had not* seen. For each tagsplanation, subjects responded to the following statement using a 5-point Likert scale:

  Overall this is a good explanation given my knowledge of the movie.

Results

In Part 1 of the study, users responded to a series of statements about 4 different tagsplanation interfaces. Figure 7 shows the percentage of responses that were either *agree* or *strongly agree* for each type of statement (justification, effectiveness, mood compatibility), broken down by interface (RelSort, PrefSort, RelOnly, and PrefOnly). Neutral responses, which accounted for less than 30% of total responses in each category, were excluded from the calculations.

RelSort performed best in all three categories (justification, effectiveness, and mood-compatibility) by a statistically sig-

Top 10	score	Bottom 10	score
afi 100 (laughs)	100.0%	male nudity	0.0%
fantasy world	100.0%	narrated	7.7%
world war ii	100.0%	los angeles	12.5%
sci-fi	95.2%	directorial debut	16.7%
action	94.4%	childhood	17.6%
psychology	93.8%	matt damon	20.0%
disney	91.7%	movie business	25.0%
satirical	88.5%	afi 100	25.5%
drama	87.5%	new york city	26.7%
satire	86.4%	amnesia	28.6%

Table 1: 10 best and worst factual tags, based on percentage of responses that were *agree* or *strongly agree*. Includes tags with at least 10 responses.

Top 10	score	Bottom 10	score
great soundtrack	90.9%	sexy	15.4%
fanciful	90.9%	intimate	25.0%
funny	90.0%	passionate	25.0%
poignant	88.9%	lyrical	30.8%
witty	88.0%	meditative	33.3%
dreamlike	87.5%	brilliant	41.2%
whimsical	87.5%	gritty	44.4%
dark	87.3%	understated	45.5%
surreal	86.7%	fun	50.0%
deadpan	84.2%	reflective	50.0%

Table 2: 10 best and worst subjective tags, based on percentage of users who agreed with statements about tag. Includes tags with at least 10 responses.

nificant margin[12] ($p \leq 0.005$, $p \leq 0.02$, and $p \leq 0.02$ respectively). RelOnly scored lowest in justification by a statistically significant margin ($p \leq 0.001$). PrefSort scored higher than RelOnly in effectiveness by a statistically significant margin ($p \leq 0.01$). None of the other differences were statistically significant.

In Part 2 of the study, users responded to a series of statements about individual tags. Figure 8 shows the percentage of responses that were *agree* or *strongly agree* for each type of statement (justification, effectiveness, and mood compatibility), summarized by type of tag (subjective or factual).[13] In all three categories, subjective tags outperformed factual tags by a statistically significant margin (justification: $p \leq 0.02$, effectiveness: $p \leq 0.001$, mood compatibility: $p \leq 0.001$). Tables 1 and 2 show the 10 best and worst performing factual and subjective tags, based on the percentage of responses that were *agree* or *strongly agree* across all types of statements. Neutral responses were excluded from the calculations.

Results from Part 3 of the study reveal that 81.7% of respondents agreed with the statement "*Overall this is a good explanation given my knowledge of the movie*". Neutral responses, which accounted for less than 20% of all responses, were excluded from this calculation. Broken down by interface, the percentage of agreement was 85.4% for RelSort, 80.3% for PrefSort, 75.7% for RelOnly, and 85.7% for PrefOnly.

[12]To determine statistical significance we used the Z-test of two proportions.

[13]In previous work [10], tags were manually coded as subjective or factual by two independent coders.

Discussion

RQ-Justification: What is the role of *tag preference* and *tag relevance* in helping users understand their recommendation?

The results in Figure 7 suggest that tag preference is more important than tag relevance for justifying recommendations. RelOnly, which only shows tag relevance, performed worst by a significant margin. PrefSort and PrefOnly received virtually the same score, even though PrefSort includes tag relevance and PrefOnly does not. Users may prefer seeing tag preference because they are skeptical that a recommender system can accurately infer their preferences. According to one user, "*The weights (relevance) don't do much good without the values (preferences) they're applied to when it comes to understanding the predicted rating. This is because what movielens thinks my preferences are might differ greatly from my actual preferences*".

However, users preferred *sorting* by tag relevance. (RelSort significantly outperformed PrefSort.) Thus tag relevance appears to serve best as an organizing principle for helping users understand recommendations. One subject commented: "*sorting by relevance ... feels more useful in terms of seeing how it actually comes up with the final score.*"

RQ-Effectiveness: What is the role of *tag preference* and *tag relevance* in helping users determine if they will like the item?

For effectiveness, tag preference and tag relevance appear to be of roughly equal importance, based on the fact that PrefOnly and RelOnly received similar scores. It is surprising that tag preference would be as important as tag relevance, since users should know their own preferences. One possible reason is that showing tag preference promotes *efficiency* [16], since users can more quickly spot the tags they might like or dislike. Another possibility is that users did not understand that tag preference is user-specific rather than item-specific, and they might have incorrectly thought that tag preference revealed information about the movie. One user commented: "*It is not clear if the preference ratings given are for the aspect in general, or for the value of the aspect exemplified by this movie.*" Again, subjects preferred tag relevance as a sort order for effectiveness. One subject commented: "*I like the relevance, it gives the breakdown of what elements the film has and you can gauge what you'll be experiencing when you watch it.*"

RQ-Mood: What is the role of *tag preference* and *tag relevance* in helping users decide if an item fits their current mood?

As Figure 7 shows, RelOnly performed significantly better

for mood compatibility than it did for effectiveness or justification. This suggests that relevance plays its most important role in helping reveal mood compatibility. One user commented: "*Exposing the supposedly relevant facets ... allows me to double check relevance against current interests and mood.*" Based on the superior performance of RelSort over PrefSort, tag relevance is again the preferred sort order.

RQ-Tag-Type: What types of tags are most useful in tagsplanations?

Figure 8 shows that subjective tags performed better than factual tags in all three categories. However, these results contradict prior work that showed users prefer factual tags over subjective tags [10]. One possible reason is that we filter out tags of low quality and low relevance, while the prior study did not. Subjective tags that survived the filtering step may compare more favorably to factual tags. Another possible reason for the discrepancy is context; subjects in our study evaluate tags in the context of specific tasks (understanding the recommendation, deciding if they will like the item, assessing mood compatibility), while subjects in the earlier study rated tags based on whether they should be shown in a general setting. For these three tasks, in particular assessing mood compatibility, users may prefer subjective tags, while factual tags may be preferred in a more general setting.

While subjective tags generally outperformed factual ones, anecdotal evidence suggests that users prefer a factual tag over a subjective tag if both capture the same idea. For example, users preferred the factual tag *sexuality* (64.9% agreement over all categories) to the subjective tag *sexy* (15.4% agreement), and users preferred the factual tag *violence* (73.8% agreement) to the subjective tag *violent* (57.9% agreement).

Based on the top-rated factual tags in Table 1, users prefer factual tags that describe genres (*sci-fi*, *action*, *drama*) or general themes (*world war ii*, *psychology*). Based on the lowest-rated tags, subjects disliked factual tags that were highly specific (*los angeles*, *new york city*, *amnesia*, *movie business*) or tags that describe meta-level information (*directorial debut*, *narrated*, *afi 100*)

Based on the top-rated subjective tags in Table 2, subjects preferred subjective tags that are descriptive (*surreal*, *dreamlike*, *deadpan*) or suggest mood (*dark*, *whimsical*, *poignant*). Based on the lowest rated tags, users dislike subjective tags with sexual themes (*sexy*, *intimate*, *passionate*) or tags that express opinion without providing a description (*brilliant*, *fun*).

Beyond answering these specific research questions, the study also demonstrated the value of tagsplanations. Among survey respondents who expressed an opinion, over 80% agreed that the RelSort interface achieved the goals of justification, effectiveness, and mood compatibility. Over 85% agreed that the RelSort interface provided a good explanation given their prior knowledge of a movie. One user commented: "*This is as good a description of Waking Life as I could imagine being put together in a dozen words.*"

CONCLUSION

Our study showed that tag relevance and tag preference both play a key role in tagsplanations. Tag relevance serves best in an organizing role, and both tag relevance and tag preference help promote the goals of justification, effectiveness, and mood compatibility. The study also demonstrated the viability of tagsplanations. Over 80% of respondents who expressed an opinion agreed that the RelSort interface helped them (1) understand why an item was recommend, (2) decide if they would like the item, and (3) determine if an item fit their current mood.

One limitation of this study is that all data is self-reported. Rather than *asking* subjects how well tagsplanations promote effectiveness and mood compatibility, one could measure these objectives empirically. Bilgic et al. developed a method for quantifying how well explanations help users make accurate decisions [1]. One could use this same approach to test how well tagsplanations promote effectiveness.

Future research might also address a broader set of goals, including scrutability, transparency, and trust [16]. *Scrutability* allows the user to tell the recommender system it is wrong. Tagsplanations could incorporate this principle by letting users override their inferred tag preferences. Greater *transparency* could be provided in tagsplanations by showing users the items they rated that affected their computed tag preference. *Trust* reflects the confidence that users place in the recommender system. Future studies could gather feedback on how well tagsplanations foster trust in the recommender system.

Additionally, future work could explore more sophisticated methods for estimating tag preference. With a dataset of explicit tag ratings, one could use machine learning techniques to predict tag preference. SVD-based approaches, which have proven effective for predicting users' ratings of items [9], might also might be utilized to estimate users' preferences for tags.

ACKNOWLEDGMENTS

The authors thank S. Andrew Sheppard, Rich Davies, and the rest of GroupLens for their feedback and assistance with this paper. We thank the members of MovieLens for their ratings, feedback and suggestions. This paper is funded in part by National Science Foundation grants IIS 03-24851 and IIS 05-34420.

REFERENCES

1. M. Bilgic and R. J. Mooney. Explaining recommendations: Satisfaction vs. promotion. In *Proceedings of Beyond Personalization Workshop, IUI*, 2005.

2. D. Billsus and M. J. Pazzani. A personal news agent that talks, learns and explains. In *AGENTS '99:*

Proceedings of the third annual conference on Autonomous Agents, pages 268–275, New York, NY, USA, 1999. ACM.

3. 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.

4. J. Ellenberg. The psychologist might outsmart the math brains competing for the netflix prize. *Wired Magazine*, March 2008.

5. S. Golder and B. A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science*, 2006.

6. J. Herlocker, J. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2000. CHI Letters 5(1).

7. R. J. Mooney and L. Roy. Content-based book recommending using learning for text categorization. In *DL '00: Proceedings of the fifth ACM conference on Digital libraries*, pages 195–204, New York, NY, USA, 2000. ACM.

8. B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *WWW '01: Proceedings of the 10th International Conference on World Wide Web*, pages 285–295, Hong Kong, 2001. ACM Press.

9. B. M. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Application of dimensionality reduction in recommender systems - a case study. In *ACM WebKDD 00 (Web-mining for ECommerce Workshop)*, New York, NY, USA, 2000. ACM.

10. 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, New York, NY, USA, 2007. ACM.

11. 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.

12. C. Shirky. Ontology is overrated. http://www.shirky.com/writings/ontology_overrated.html, 2005. Retrieved on May 26, 2007.

13. R. Sinha and K. Swearingen. The role of transparency in recommender systems. In *CHI '02: CHI '02 extended abstracts on Human factors in computing systems*, pages 830–831, New York, NY, USA, 2002. ACM.

14. N. Tintarev. Explanations of recommendations. In *RecSys '07: Proceedings of the 2007 ACM conference on Recommender systems*, pages 203–206, New York, NY, USA, 2007. ACM.

15. N. Tintarev and J. Masthoff. Effective explanations of recommendations: user-centered design. In *RecSys '07: Proceedings of the 2007 ACM conference on Recommender systems*, pages 153–156, New York, NY, USA, 2007. ACM.

16. N. Tintarev and J. Masthoff. A survey of explanations in recommender systems. In *IEEE 23rd International Conference on Data Engineering Workshop*, pages 801–810, 2007.