Effective Explanations of Recommendations: User-Centered Design

Nava Tintarev
University of Aberdeen
Department of Computing Science
Scotland, U.K., AB24 3UE
+44 1224 272839
ntintare@csd.abdn.ac.uk

Judith Masthoff
University of Aberdeen
Department of Computing Science
Scotland, U.K., AB24 3UE
+44 1224 272299
jmasthoff@csd.abdn.ac.uk

ABSTRACT
This paper characterizes general properties of useful, or *Effective*, explanations of recommendations. It describes a methodology based on focus groups, in which we elicit what helps moviegoers decide whether or not they would like a movie. Our results highlight the importance of personalizing explanations to the individual user, as well as considering the source of recommendations, user mood, the effects of group viewing, and the effect of explanations on user expectations.

Categories and Subject Descriptors
H.5.2 [**User Interfaces**]: *User-centered design, Evaluation/Methodology*

General Terms
Design, Experimentation, Human Factors

Keywords
Recommender systems, explanations

1. INTRODUCTION
The recommender systems community is reaching a consensus that accuracy metrics such as mean average error (MAE), precision and recall, can only partially evaluate a recommender system [9]. User satisfaction and derivatives thereof such as serendipity [8], diversity [12] and trust [3] are increasingly seen as important. Explanations of recommendations can play an important role in improving the user experience. However, the definition of a *good* explanation is still largely open and depends on the general aim of the recommender system. Previous recommender systems with explanation facilities have been evaluated in a number of ways, reviewed and discussed in-depth in [11]. Among other things, good explanations could help inspire user trust and loyalty, increase satisfaction, make it quicker and easier for users to find what they want, and persuade them to try or purchase a recommended item. Table 1 defines seven possible aims of explanation facilities in recommender systems. In this paper, we investigate the general properties of an explanation that helps a movie recommender system fulfill the criterion of Effectiveness, i.e. helps users to make good decisions.

RecSys'07, October 19–20, 2007, Minneapolis, Minnesota, USA.

Table 1. Possible aims for explanations

Aim	Definition
Transparency	Explain how the system works
Scrutability	Allow users to tell the system it is wrong
Trust	Increase users' confidence in the system
Effectiveness	Help users make good decisions
Persuasiveness	Convince users to try or buy
Efficiency	Help users make decisions faster
Satisfaction	Increase the ease of usability or enjoyment

The paper is organized as follows. In Section 2 we discuss the motivation for this work. In Section 3, we discuss the methodology and results from our focus groups. Therein we survey how moviegoers discuss their favourite and other movies in *dialogue,* and what they consider when making a choice. We conclude in Section 4 with plans for future work.

2. MOTIVATION
Herlocker et al. [6] found that out of twenty-one explanation interfaces participants were most likely to see a movie if they saw a histogram of how similar users had rated the item, with the "good" ratings clustered together and the "bad" ratings clustered together. However, a limitation of this experiment in terms of Effectivness is a bias toward positive ratings in the MovieLens dataset. Using another dataset, Bilgic and Mooney [2] have shown that using this type of histogram causes items to be overestimated, and suggest this is due to the skew towards high ratings. That is, participants *think* they like an item more than they really would. This means that the histogram based explanation may be more Persuasive than it is Effective.

A second limitation of the experiment of Herlocker et al. is that the explanations based on movie properties such as favorite actor/actress were not personalized for the participants, but rather for the main author of the paper [6]. This may have resulted in the *relatively* poor, yet significant, acceptance for explanations using this type of information. It would seem plausible that a movie feature such as favorite actor/actress is more important to some users than others, and that it would depend on each user's disposition toward the particular actor/actress. The high variance

in acceptance for this type of explanations in the Herlocker et al experiment suggests that this is likely.

We consider the role of mentioning item features to users in explanations. The rationale behind studying user's utilization of features is that simply stating that two items are similar does not always help users see the commonality between items, while an explanation using feature-based information may better help a user understand how two items are related. For example, Hingston [7] who studied the perceived Effectiveness of explanations found that participants requested information about why items were judged to be similar to one another in an explanation interface which compared the recommended item to similar items the user had liked in the past. Similarly, Bilgic and Mooney [2] failed to show a significant effect on Effectiveness for an explanation interface which used information about previously rated items, but where the explicit relations between these previously rated items and the current recommendation are not clear.

Therefore, in our study we set out to find out what helps moviegoers decide whether or not they would like a movie. If personalization has any merit, we wanted know what and how to personalize. In addition, we hoped that natural dialogue would help us discover how to best present this information to moviegoers.

3. FOCUS GROUPS

We conducted two focus groups to gain an intuition if particular movie features such as e.g. actors, awards etc determine whether or not participants will see and like a movie. For this purpose, an initial list of features was obtained in an exploratory analysis of online reviews from Amazon.co.uk (see Table 2, the numbers indicate how many times each feature was mentioned across 48 reviews). In particular we aimed to find out how participants would like to be recommended, or dissuaded, from watching a movie. As we intend to create a system using explanations, we hoped that in the informal setting of a focus group we would find particular formulations and keywords moviegoers use and prefer in justifications of recommendations.

Table 2. Common features in corpus analysis

Cast (28)	Good in its genre (26)	Initial expectations (22)
Script (19)	Visuals (18)	Suites mood (18)
Realistic (15)	Director (12)	Subject matter (12)
Easy viewing (8)	Good for kids (7)	Repulsive/ violent (7)
Dialogs (6)	Pace (5)	Soundtrack (5)
Original (5)	Movie Studio (2)	Sex (1)

A limitation of the focus groups is that what users like to say may differ from what they like to hear. We will not know for sure how helpful a participant's justification would be to a potential user, although we can watch the reactions of other participants.

3.1 Design

3.1.1 Procedure

A total of eleven participants were spread over two focus groups, with the same facilitator. Audio recordings were made for later analysis. Participants were welcomed, and explained the purpose of the focus group.

The focus group began with each participant telling what their favorite movie was and why they liked it. Next, the facilitator asked participants to suggest movies, selecting one that the majority had seen. Participants were informally asked about their initial expectations for this movie, and if something in particular made them consider watching it. They were also asked about their impression after watching the movie, and which features helped form this impression. This was augmented with a final question about how they would like to be recommended or dissuaded from watching the discussed movie. Care was taken to phrase this question so that the recommendation would be directed to them by someone who knew their tastes, e.g. "*If a friend would recommend, or tell you not to see, this movie, how would they motivate it?"*

The facilitator aimed primarily to let the participants themselves suggest movies. However, they could not always think of suggestions, in which case a movie from a prepared list (Appendix A[1]) was used, based on the most rated movies in MovieLens balanced with a handful of the most popular movies in each genre on IMDB[2]. The list was printed out and shown openly to participants. After the introductions, each focus group discussed an average of five movies in detail. Participants also sometimes referred to a movie to illustrate a point, although this movie had not been seen by a majority of participants. Including these discussions, but not the introductions, the average number of movies mentioned in each focus group is ten.

We concluded the focus groups with a summary of what had been said so far. The participants were asked for feedback on this summary. For completeness, they were also asked if any type of movie or deciding feature had been neglected in the discussion. At the end, we orally went through the list of features mentioned in Table 2, and asked participants to note their importance. In total each session took between 1 and 1 1/2 hours.

3.1.2 Participants

Eleven participants interested in movies were recruited from the staff and student population of Aberdeen University (eight males and three females, aged 24-33). Participants varied in nationality (Irish, Israeli, French (3), Scottish (2), Spanish, South African, Swiss/Bolivian and Vietnamese). Using academic and multi-cultural participants may lead to a bias in taste, such as an increased preference for independent cinema. We did however find a great divergence in taste, both in terms of the types of movies participants liked and the weight different features had in determining if a movie was worth viewing or not. For example, the genres discussed varied greatly and included: action, children, animated, comedy, crime/gangster, documentary, horror, fantasy, musical, romance, science fiction, thriller and western.

3.2 Results and Discussion

3.2.1 What Features are Important to Mention

Participants' introductions of their favorite movies show which features they intuitively considered important to mention

[1] http://www.csd.abdn.ac.uk/~ntintare/appendix/recSys07.rtf

[2] http://www.imdb.com: retrieved November 2006

(Citations of introductions are available in Appendix B[1]). Table 3 shows the features mentioned and how often they were mentioned across all participants. The features mentioned varied largely between subjects, the most commonly mentioned feature was "good in its genre" (e.g. this movie is funny, when talking about a comedy) followed by "script complexity" and "mood".

Table 3. Number of times each feature was mentioned

Good in its genre (6)	Script complexity (4)	Mood (4)
Subject matter (2)	Initial expectations (2)	Cast (2)
Director (1)	Visuals (1)	Realistic (1)
Original (1)		

Note that "mood" may relate to several sub-features in turn such as affect (Appendix B, quote 4), genre preferences (quote 8), and atmosphere (quote 11).

We confirmed that participants differ in the features they use when describing their favorite movies. For instance, consider these examples:

P1: *"..normally I don't have a favorite actor or actress, but Jet Li is probably one of my favorite actors. Anything from him is good…"*

Facilitator: *"What made you watch it?"*

P2: *"The director I think, Scorsese"*

Participant P1 differentiated the movie according to one particular actor; while P2 mentioned the director as an important factor in choosing a movie and was in fact consist in this preference throughout the focus group. We note that in both cases, it was not only enough to mention director or actor *in general*; each participant found it important to refer specifically to their *favorite*.

In a similar manner, some participants cared more about the overall movie aesthetics and musical score, while others did not notice or consider these features particularly important.

3.2.2 Mood Influences Features

Participants believed that their mood is likely to influence the genre they choose to see, and as a secondary effect, what features they consider important; such as script complexity, affect (e.g. feel good movie). These factors were often situational*; "I mean for a musical I don't really need a great script, a great plot at least, uh or for uh what I call a pre-exam uh film the night before I mean. Bruce Willis saving the world is just what I need. Uh you know you don't want something, you just want to use two neurons and that's it, just relax."*.

3.2.3 Social Viewing Influences Features

In both groups, for most of the participants there was a clear distinction between movies viewed in larger, more casual groups of friends, and movies seen alone or in more intimate circumstances such as with a partner. Movies seen with groups of friends were often light or easy viewing. Other movies, such as Schindler's list are better viewed in more intimate company or even alone; *"I think I watched it on my own or something, I'm kind of thinking it's not the kind of thing you watch […] in a group"*. The reason behind this seems two-fold. In larger gatherings the aim is often light-hearted entertainment, the viewers aim to enjoy themselves rather than conduct a mental activity. More serious or dramatic movies on the other hand may invoke strong emotions and tension. Secondly, in large gatherings there is often a lot of simultaneous activity, someone is always speaking, going to get a tea or coffee etc. which obstructs the viewers from following a complex plot.

3.2.4 Who Should Give the Explanation

Participants listened to their friends' recommendations, in particular when they had time to spare. Whether or not participants listen to a recommendation depends on how it was given: *"It probably depends on the way they describe the movie rather than who they are."* Participants in both groups also agreed that the same advice coming from different people wouldn't have the same impact on them. It depended on whether or not this person had similar taste, i.e. agreed on movies in the past: *"But it depends on the style of the movie; because if it's like a romantic comedy and my sister tells me its brilliant then I'll go and see it. If it's an action then I'll listen to what my brother thought of it. […] if I know they have similar tastes in that kind of film to me then I'll listen to them."*

3.2.5 Explanations and Satisfaction

Explanations may help users enjoy movies more, rather than serve merely as decision aids. Participants believed that correcting faulty expectations for sequels or adaptations of a movie would not influence whether or not they saw it. Rather both groups unanimously felt that it could increase their acceptance upon viewing, and save potential disappointment. One participant stated that he liked musicals, but had to know what to expect in advance: *"If I go to see a musical I have to know it's a musical before watching it"*.

3.2.6 Dissuading Users

In retrospect, none of the participants felt that they would have wanted to be dissuaded from watching movies they had disliked. Participants even watched popular movies which they expected to be disappointed by. They wanted to form their own opinion, and they did not want to reject social invitations, or refute the general consensus without strong warrant: *"I wouldn't rush to watch certain genres, but if I was with somebody that was into that then yeah. I always think you try and take everything for what it is and try and look for the good parts"*. We suggest that this is mainly due to the social nature of movie viewing, and may be weaker for less social types of recommendations such as books.

3.2.7 Modifications to Features

The initial features suggested by our exploratory analysis of online reviews were moderately modified in scope by the results of these focus groups. Firstly, we considered "realistic" to be a feature of a movie. Participants in both groups strongly differentiate between the terms realistic and believable. One participant explicitly stated: *"…you used these two words and I think they are really important; realistic and believable. I don't care about it being realistic; I care about it being believable"*. Particularly in genres such as Action and Science Fiction, realism seemed to be watered down to "believable" which is important in the negative sense, e.g. flaws in coherence make a movie less attractive.

During the course of the focus groups, we also realized that script complexity was strongly tied to mood. In addition, we realized that a simple script was pretty much synonymous to easy viewing. Some participants were happy to see movies as entertainment and did not place too much weight on the complexity of a story; others liked movies that presented a challenge, or were unpredictable: *"it depends a lot on how you come to the movies… [Participant X] would like a movie that challenges him, do a bit of thinking. Personally, I pretty much think of a movie as a form of entertainment – two hours of fun!"* This definition of mood differs from mood defined as a preference for certain genres, i.e. "I feel like seeing an *action* movie tonight", as well as in terms of affect, i.e. "I'd like to see a *feel good* movie".

Subject matter and how realistic a movie is were found to be very relevant for a documentary or historical movie, but not otherwise.

4. CONCLUSIONS AND FUTURE WORK

From the focus groups, we draw the following conclusions for recommender system explanations:

- *Feature selection in explanations needs to be tailored to the user.* We saw that users mentioned different movie features both when describing their favorite movie as well as throughout the discussion. Carenini and Moore [4] studied the effect of tailoring evaluative arguments in the domain of houses, and found that tailored arguments were significantly more persuasive than non-tailored arguments. Based on this, we believe that in any domain different users would weight features differently, even if there is a consensus about which features are generally important.
- *Feature selection in explanations needs to be tailored to the context.* We learned that users' context (such as social setting and mood) influences the importance they adhere to features. We suggest that recommender systems allow users to specify their current priorities, and save a number of such profiles.
- *Features can be selected from a relatively short list.* Though participants varied in which features they found important, a relatively short list provided good coverage. With some modifications, the features found in our analysis of on-line reviews were adequate. We believe that this finding and the method used to obtain the list are generalizable to many other recommender domains. Therefore, a short list of features should suffice for users to set their preferences and it should be possible for users to do this explicitly (in line with previous user studies on recommender systems that suggested that personalization should be easy and quick [1] and that users are willing to spend extra effort if they felt it resulted in better quality recommendations ([10], [8]). It may be possible to use some default settings (e.g. in general users like a synopsis), but these should be accessible and modifiable by users (i.e. Scrutable [5]).
- *Explanation source matters*. Explanations can be presented as coming from a credible source (e.g. different style for action movie recommendations). Collaborative algorithm may play a role in this. For example, we envision explanations of the type: "User X likes the same type of thrillers you do, such as 'Silence of the Lamb's. User X liked 'The Usual Suspects' too."

Currently, we are developing a prototype which generates explanations for movie recommendations. The user model used in this system weighs the movies features elicited by our exploratory corpus analysis and focus groups according to specified user utility. Relevant meta-data is extracted from the Amazon e-Commerce Service (ECS) for this purpose. Textual recommendations are generated by a flexible natural language generation system. This flexibility allows us to modify parameters such as which features to mention, and how to describe them. Further, the results of the focus group suggest that the optimal explanation is dependant on factors such as mood, and source, which allow us further control for different scenarios.

6. REFERENCES

[1] Van Barneveld, J. and Van Setten, M. *Personalized digital television*, chapter 10, 259–285. Kluwer, 2004.

[2] Bilgic, M. and Mooney, R.J. Explaining recommendations: Satisfaction vs. promotion. *Beyond Personalization Workshop, IUI*, 2005.

[3] Chen, L. and Pu, P. Trust building in recommender agents. *WPRSIUI workshop*, 2002.

[4] Carenini, G. & Moore, D.J. An Empirical Study of the Influence of User Tailoring on Evaluative Argument Effectiveness. IJCAI 2001.

[5] Czarkowski, M. *A Scrutable Adaptive Hypertext*. PhD thesis, University of Sydney, 2006.

[6] Herlocker, J. L., Konstan, J. A. and Riedl, J. Explaining collaborative filtering recommendations. In *CSCW*, 2000.

[7] Hingston, M. User friendly recommender systems. *Honours thesis*, University of Sydney, 2006.

[8] McNee, S.M., Lam, S.K., Konstan, J.A. and Riedl, J. Interfaces for eliciting new user preferences in recommender systems. *User Modeling*, 178–187, 2003.

[9] McNee, S.M., J. Riedl, and J. A. Konstan. Being accurate is not enough: How accuracy metrics have hurt recommender systems. *Extended Abstracts, CHI 2006*

[10] Swearingen, K. and Sinha, R. Interaction design for recommender systems. *Designing Interactive Systems*, 2002.

[11] Tintarev, N. and Masthoff, J. Survey of explanations in recommender systems. *WPRSIUI*, 2007.

[12] Ziegler, C., McNee, S.M., Konstan, J.A. and Lausen, G. Improving recommendation lists through topic diversification. *WWW 2005*