The Effectiveness of Personalized Movie Explanations: An Experiment Using Commercial Meta-data

Nava Tintarev and Judith Masthoff

University of Aberdeen,
Aberdeen, U.K.

Abstract. This paper studies the properties of a helpful and trustworthy explanation in a movie recommender system. It discuss the results of an experiment based on a natural language explanation prototype. The explanations were varied according to three factors: degree of personalization, polarity and expression of unknown movie features. Personalized explanations were not found to be significantly more Effective than non-personalized, or baseline explanations. Rather, explanations in all three conditions performed surprisingly well. We also found that participants evaluated the explanations themselves most highly in the personalized, feature-based condition.

1 Introduction

Recommender systems represent user preferences for the purpose of suggesting items to purchase or examine, i.e. recommendations. Our work focuses on explanations of recommended items [1,2,3], explaining how a user might relate to an item unknown to them. More concretely, we investigate explanations in the movie domain with the aim of helping users make qualified decisions, i.e. Effective explanations. An explanation may be formulated along the lines of *"You might (not) like Item A because..."*. The justification following may depend on the underlying recommendation algorithm (e.g. content-based, collaborative-based), but could also be independent. Our approach is algorithm independent, as it aims to explain a randomly selected item rather than the recommendation. In this way we implicitly differentiate between explaining the way the recommendation engine works (Transparency), and explaining why the user may or may not want to try an item (Effectiveness). In addition, since items are selected randomly the explanation can vary in polarity: being positive, neutral or negative.

The experiment described in this paper measures how different explanations affect the likelihood of trying an item (Persuasion) versus making informed decisions (Effectiveness) and inspiring user trust (Trust). We investigate the effects different types of explanations have on these three explanation aims.

As in a study by Bilgic and Mooney [1], we define an Effective explanation as one which helps the user make a correct estimate of their valuation of an item. Persuasion only reflects a user's initial rating of item, but not their final rating

W. Nejdl et al. (Eds.): AH 2008, LNCS 5149, pp. 204–213, 2008.
© Springer-Verlag Berlin Heidelberg 2008

after trying it. While the user might initially be satisfied or dissatisfied, their opinion may change after exposure. As in [1], Effectiveness can be measured by (1) the user rating the item on the basis of the explanation, (2) the user trying the item, (3) the user re-rating the item. While it would be preferable if users could actually try the item, in an experimental setting step 2 may be approximated by e.g. allowing users to read item reviews written by other users. The metric suggested by [1] is optimized when the mean difference between the two ratings (step 1 - step 3) is close to zero, has a low standard deviation, and there is a strong positive correlation between the two ratings. If an explanation helps users make good decisions, getting more (accurate and balanced) information or trying the item should not change their valuation of the item greatly.

Although [1] did not explicitly consider the direction of skew, the difference between the two ratings may be either positive (over-estimation of the item) or negative (under-estimation). Over-estimation may result in false positives; users trying items they do not end up liking. Particularly in a high investment recommendation domain such as real-estate, a false positive is likely to result in a large blow to trust in the system. Under-estimation may on the other hand lead to false negatives; users missing items they might have appreciated. If a user recognizes an under-estimation due to previous knowledge or subsequent exposure, this may lead to a loss of trust as well. Likewise, an under-estimation may needlessly decrease an e-commerce site's revenue.

2 Factors That May Impact Explanation Effectiveness

2.1 Features and Personalization

Users like to know what it is about a particular item that makes it worthy (or not) of recommendation. Bilgic and Mooney [1] did not find a significant result for Effectiveness for an 'Influence based explanation' which listed other books previously rated highly by the user as influential for the recommendation. Other work surveying a similar type of interface suggests that users would like to know the explicit relationship between a recommended item and similar items used to form the recommendation [4]. An explanation based on item features may be one way to do this, e.g. "You have rated books with the *same author* highly in the past."

Using item features also makes it possible to personalize explanations, as different users may place different importance on different feature, and have individual tastes with regard to these features (i.e. not everyone has the same favorite actor). The seminal study by Herlocker et al. [2] on explanation interfaces shows a strong *persuasive* effect for an explanation interface referring to a particular movie feature, namely "favorite actor or actress". This feature (favorite actor/actress) may be more important to some users than others since a high variance in acceptance for this type of explanation was found. Qualitative feedback from focus groups also shows that users vary with regard to which movie features they find important[5,6].

If it is the case that some features are more important for particular users, it would seem plausible that explanations that tailor which features to describe would be more Persuasive and Effective than explanations with randomly selected features, and non-feature based explanations. In the real-estate domain Carenini and Moore have shown that user-tailored evaluative arguments (such as *"the house has a good location"* for a user who cares about location) increase users' likelihood to adopt a particular house compared to non-tailored arguments [7].

While similar, our work differs from the studies in [7] and [2], which primarily considered the *Persuasive* power of arguments and explanations, but did not study Effectiveness. Arguably [7] varied the polarity of the evaluative arguments, but given the domain (real-estate) it was difficult for them to consider the final valuation of the items. Our aim is therefore to consider how user-tailoring of item features can affect explanation Effectiveness, Persuasion as well as Trust.

2.2 Polarity

An explanation may contain both positive and negative information, and in that sense may have a polarity in a similar way as a numerical rating of an item. [8] showed that manipulating a rating prediction can alter the user valuation of a movie, causing either an over- or underestimation. Modifying the polarity of an explanation is likely to lead to a similar skew in Effectiveness. In the study by Herlocker et al [2] participants were most likely to see a movie if they saw an explanation interface consisting of a barchart of how similar users had rated the item. This bar chart had one bar for "good", a second for "ok" and a third for "bad" ratings. A weakness of this result is a bias toward positive ratings in the used dataset[1]. Bilgic and Mooney [1] later showed that using this type of histogram causes users to overestimate their valuation of the items (books).

We have analyzed online movie reviews mined from the Amazon website, to see if we could distinguish the properties of reviews that are considered helpful [6]. We found that users were more prone to write positive reviews and that negative reviews were considered significantly less helpful by other users than positive ones. Similar correlations between item rating and review helpfulness were found in other domains such as digital cameras and mobile phones [9]. All of this makes us consider whether negative explanations are likely to be found less helpful by users, or may instead help mitigate users' overly optimistic beliefs about items.

2.3 Certainty

The Herlocker et al study [2] considered an interface which looked at recommendation confidence, which portrays to which degree the system has sufficient information to make a strong recommendation. Their study did not find that confidence displays had a significant effect on how likely a participant was to see a movie. McNee et al [10] also studied the effect of confidence displays on user acceptance. They found that users that were already familiar with their

[1] MovieLens - http://www.grouplens.org/node/12#attachments

recommender system (MovieLens) were less satisfied with the system overall after being exposed to confidence displays. On the other hand, more experienced users also found the confidence display more valuable than new users.

As part of a larger study comparing different types of explanation interfaces we held three focus groups (comprising of 23 participants) discussing the confidence display used in [2]. We found that many participants found information about confidence displays confusing. They did not understand what to do with the confidence rating or felt that the system should not make predictions if it was not confident. This raised the question of how lack of confidence would affect explanation Effectiveness, Persuasion as well as Trust. In particular, we were curious how users would react to missing information. In real data-sets as our data retrieved from Amazon's e-Commerce Service (ECS), detailed feature meta-data is sometimes missing. Is it better to refrain from presenting these items to users altogether, to talk about another feature which might not be as important to the user, or candidly state that the system is missing certain information?

2.4 Other Factors

Effectiveness of explanations can also be affected by a number of other factors. If the quality of the information used to form the recommendation or recommendation accuracy are compromised this is likely to lead to poor Effectiveness. Likewise, the nature of the recommended object and presentation of the recommended items are likely to be contributing factors. While these are highly relevant topics, they will not be discussed further in this paper. We conduct a study where no recommendation engine is used, in a single domain (movies), with all items presented in the same manner (one stand-alone item).

3 Experiment

This experiment is based on a prototype system which dynamically generates natural language explanations[2] for movie items based on meta-data retrieved from Amazon (ECS)[3]. The aim of this experiment was to see if using movie features (e.g. lead actors/actresses), and personalization could affect the Effectiveness of explanations. We studied if explanation polarity, and clearly stating that some information is missing could affect Effectiveness. We also wanted to know whether the effect was the same for Persuasion. When we help users make decisions that are good for them (Effectiveness), will they end up buying/trying fewer items (Persuasion)? Likewise, we are interested in the effects these factors have on user Trust.

[2] Realized with simpleNLG, a simple and flexible natural language generation system created by Ehud Reiter. See also http://www.csd.abdn.ac.uk/∼ ereiter/simplenlg

[3] The used meta-data considers the finding of focus groups and analysis of online movie reviews [5,11,6] as well as which features are readily available via Amazon's ECS e.g. actors, directors, genre, average rating, and certification (e.g. rated PG).

3.1 Design

First, participants entered their movie preferences: which genres they were in the mood for, which they would not like to see, how important they found movie features (elicited in previous studies [6]), and their favourite actors/directors. The user model in our prototype can weigh the movies' features, according to feature utility as well as the participant's genre preferences.

Fifty-nine movies were pre-selected as potential recommendations to participants. Thirty are present in the top 100 list in the Internet Movie Database (IMDB[4]) and the other twenty-nine were selected at random, but all were present in both the MovieLens 100.000 ratings dataset[5] and Amazon.com.

Each participant evaluated *ten* recommendations and explanations for movies selected at random from the pre-selected set. Note that the explanations tell the user what they might think about the item, rather than how the item was selected. Moreover, these explanations differ from explanations of recommendations as they may be negative, positive, or neutral, as the movies shown to the user are selected at random. Since we did not want the users to have any pre-existing knowledge of the movies they rated, we prompted them to request a new recommendation and explanation if they felt they might have seen the movie. Next, we followed the experimental design of [1] for each movie:

1. Participants were shown the title and cover image of the movie and explanation, and answered the following questions:
   - *How much do you think you would like this movie?*
   - *How good do you think the explanation is?*
2. Participants read movie reviews on Amazon.com, care was taken to differentiate between our explanation facility and Amazon.
3. They re-rated the movie, the explanation and their trust of our system: *"Given everything you've seen* ***so far*** *how much do you now trust the explanation facility in this system?"*

On all questions, participants selected a value on a Likert scale from 1 (bad) to 7 (good), or opted out by saying they had "no opinion". They could give qualitative comments to justify their response. In a between subjects design, participants were assigned to one of three degrees of personalization:

1. **Baseline:** The explanation is neither personalized, nor describes item features. This is a generic explanation that could apply to anyone, e.g. *"This movie is one of the top 100 movies in the Internet Movie Database (IMDB)."* or *"This movie is not one of the top 100 movies in the Internet Movie Database (IMDB)."* No additional information is supplied about the movie.
2. **Random choice, feature based:** The explanation describes item features, but the movie feature is selected at random, e.g. *"This movie belongs to your preferred genre(s): Action & Adventure. On average other users rated this movie 4/5.0"*. The feature 'average rating' may not be particularly important to the user.

[4] http://www.imdb.com

[5] http://www.grouplens.org/node/12#attachments

3. **Personalize choice, feature based:** The explanation describes the item feature that is most important to the participant, e.g. *"Although this movie does not belong to any of your preferred genre(s), it belongs to the genre(s): Documentary. This movie stars Liam Neeson your favorite actor(s)"*. For this user, the most important feature is leading actors.

Our previous findings [11,6] suggest that genre information is important to most if not all users, so both the second and third condition contain a sentence regarding the movie genre in a personalized way. This sentence notes that the movie belongs to some of the user's disliked genres (negative polarity), preferred genres (positive polarity), or lists the genres it belongs to though they are neither disliked nor preferred (neutral polarity). In negative explanations, the movie belongs to a genre the user dislikes. We do not explicitly state what the user may think of the item, e.g. "You might like/dislike this movie" as this is likely to bias their rating. Also, there are times when Amazon is missing information. An example explanation for a negative explanation with unknown information is: *"Unfortunately this movie belongs to at least one genre you do not want to see: Horror. Director information is unknown."*. Seventeen movies lack director information and their explanations explicitly state that this is missing.

Also, a movie may star one of the user's favorite actors or director in which case this will also be mentioned as a *"favorite"*, e.g. "This movie starts Ben Kingsley, Ralph Fiennes and Liam Neeson your *favorite* actor(s)."

Fifty-one students and university staff participated in the experiment. Of these, five were removed based on users' comments suggesting that they had either rated movies for which they had a pre-existing opinion, or Amazon's reviews instead of our explanations. Of the remaining, 25 were male, 21 female and the average age was 26.5. Participants were roughly equally distributed among the three conditions (14, 17 and 15 respectively).

We hypothesize that personalized feature based explanations will be more Effective than random choice feature based explanations and the baseline explanations.

3.2 Results and Discussion

Table 1 summarizes the means of all the recorded values.

Table 1. Means (and StDev) of user ratings and percentage "no opinions". First ratings are given after viewing the explanation, second ratings after viewing the Amazon reviews.

Condition	**Movie rating1**	**Movie rating2**	**Explanation rating1**	**Explanation rating2**	**Trust**
Baseline	3.45 (1.26) 8.8%	4.11 (1.85) 0%	2.38 (1.54) 2.2%	2.85 (1.85) 0%	2.69 (1.94) 0.7%
Random choice	3.85 (1.87) 7.2%	4.43 (2.02) 3.6%	2.50 (1.62) 3.0%	2.66 (1.89) 3.0%	2.56 (1.74) 3.6%
Personalized	3.61 (1.65) 3.1%	4.37 (1.93) 0.6%	3.09 (1.70) 0.6%	3.14 (1.99) 0%	2.91 (1.60) 1.3%

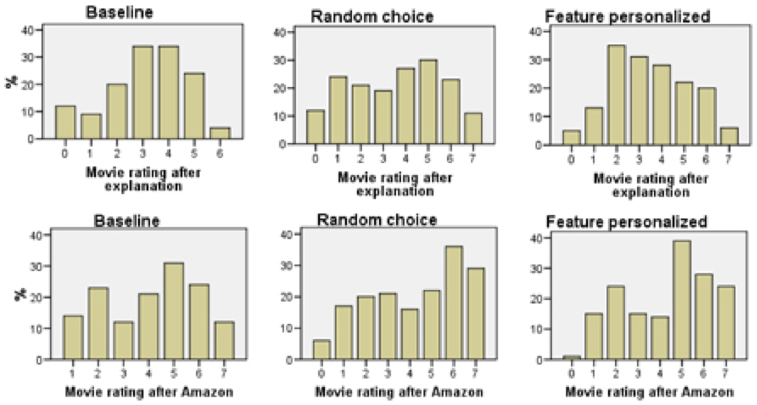

Fig. 1. First and second movie ratings - the distribution is considered with regard to *percentage* of ratings in each condition

Enough to Form an Opinion? Since our explanations are very short we first considered whether they were sufficient for the user to form an opinion of the movie. In Table 1 we note the percentage of no-opinions in each condition. We see that this is small though perhaps not negligible. The percentage for the first movie as well as for the first explanation is smallest in the personalized condition. In Figure 1 we consider the actual ratings of the movie. We see that the first and second rating of the movie are distributed beyond the mean rating of 4, suggesting that participants are able to form polarized opinions.

Are Personalized Explanations More Effective? Next we considered Effectiveness. Similar to the metric described by [1] we consider the mean of the difference between the two movie ratings. Unlike [1] (who considered the signed values) we consider the *absolute*, or unsigned, difference between the two ratings in Table 2. *Independent samples t-tests show no significant difference between the means of the three conditions.* This suggests that the degree of personalization or using item features does not increase explanation Effectiveness.

Figure 2 graphically depicts the *signed* distribution of Effectiveness. We see here that under-estimation is more frequent than overestimation in all three conditions. We also note the peak at zero in the random choice, feature based

Table 2. Effectiveness over absolute values with "no-opinions" omitted, and Pearson's correlations between the two movie ratings

Condition	$\overline{m}$ **(StDev)**	**Correlation**	**p**
Baseline	1.38 (1.20)	0.427	0.000
Random choice	1.14 (1.30)	0.650	0.000
Personalized	1.40 (1.21)	0.575	0.000

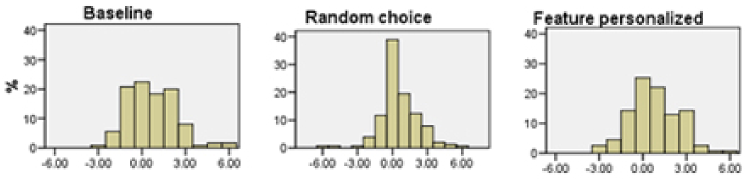

Fig. 2. Distribution of (signed) Effectiveness - "no opinions" omitted

condition. Around 40% of explanations in this condition are perfectly Effective, i.e. the difference between the two ratings is zero.

We investigated this further and found that the random choice condition has significantly higher initial ratings than the other two conditions. We compared this condition with the personalized condition to see if there was any factor that could cause this. The percentage of shown movies that were in the top 100 in IMDB was comparable, and the distribution of movie titles did not show an evident skew. In the personalized condition most participants chose "actors" as their most preferred movie feature (75%) while the participants in the random choice condition received explanations describing the four movie features in fairly equal proportions. The explanations in the random choice condition have *fewer* mentions of favorite actors and directors, *more* explanations with unknown information, and *fewer* movies in the participants' preferred genres than in the personalized condition. All of this, except the difference in features mentioned, would be expected to lead to a *lower* initial rating rather than the found higher rating. With regards to features, we speculated that the difference may be due to the feature "average rating by other users" being mentioned more often, as we observed a positive bias of average ratings on the Amazon website. However, we found that mentioning this feature correlated more with *low* ratings of movies. So, we have not yet found a satisfactory explanation.

Since [1] did not consider the sign of the difference between the two ratings, their metric of Effectiveness also requires that the two ratings are correlated. This correlation is still interesting for our purposes. Table 2 shows a significant and positive correlation between these two ratings for all three conditions. *That is, explanations in all three conditions perform surprisingly well.*

Explanations and User Satisfaction. In Table 1 we see that the second set of explanation ratings are higher than the first. This may be partly due to some participants confounding our explanations with the Amazon reviews, thus rating our explanation facility higher. The mean rating for trust and explanations is low overall, but users rate the first explanation rating significantly highest in the personalized condition (independent sample t-tests, $p<0.001$). This suggests that while the personalized explanations may not help users make better decisions, users may still be more satisfied. This is confirmed by the qualitative comments. Participants in the personalized condition appreciated when their preferred feature was mentioned: *"...explanation lists main stars, which attracts me a little to watch the movie..."*. Participants felt that vital information was missing in particular in the random choice condition: *"...I indicated that Stanley Kubrick is one of my favorite directors in one of the initial menus but the explanation didn't tell me he directed this. That would have piqued my interest. The explanation didn't have this important detail so a loss of trust happened here..."* Another participant in the random choice condition had set actors as the most important feature and left the following comment for an explanation with information about the director: *"...not much useful information in the explanation - I do not know many directors, so do not really care who directs a movie."*. In contrast, participants in the baseline condition expressed that they were dissatisfied with the explanation: *"Not very helpful explanation even if it is top 100..."*

Trust, Classification and Completeness. In Table 1 we see that the mean trust is low in all three conditions, but seems best in the personalized feature based condition. Many participants felt that the genres were misclassified, and that this reduced their trust in the explanation facility. Although the genre classification is automatically retrieved from the Amazon ECS there are two things we could change in our explanations to mitigate these effects. In our prototype when a movie belongs to any of the users' favorite genres, only preferred genres are mentioned in the explanation even if the movie belongs to other genres as well. Similarly for disliked genres, only these are mentioned. A first improvement would be to mention all the genres a movie belongs to. Secondly, the genre explanations can be improved by considering even more detailed genre specification such as "Period Drama" rather than just "Drama" for costume dramas.

We received similar feedback for actors, where we only mention the user's favorite actor in the explanation: *"Benicio del Toro is in it, but so are others who aren't listed and who I really like..."*. That is, users might like to hear the names of all the leading actors even if only one is known to be their favorite.

Certainty and Polarity. None of the seven users for which director information was missing noted this, nor were there any explicit complaints about negative explanations where the movie belonged to a genre the user did not like.

4 Conclusions

In all three conditions participants largely have an opinion of the movie, and in all conditions there was more underestimation than overestimation. The mean Effectiveness deviated ca 1.5 from the optimum discrepancy of zero on a 7 point scale (StD < 1.5), regardless of the degree of personalization or whether or not the explanation used features such as actors. In light of this under-estimation we reconsider the fact that movie ratings in general, and their Amazon reviews in particular, tend to lean toward positive ratings. If Amazon reviews are overly positive, this may have affected our results.

Since there is no significant difference between conditions w.r.t. Effectiveness we consider the factors that the three conditions share, which is that they all expose the participant to the movie title and movie cover. A number of participants justify their ratings in terms of the image in their qualitative comments, in particular for the baseline explanation. So it is fair to assume that at least some participants use the image to form their judgment.

We are now planning a repeated experiment accounting for the factors discussed in this paper. Firstly, the experiment will consider explanations without images. Secondly, explanations regarding genre and actor will be more detailed and complete. Thirdly, a clearer distinction will be made between the personalized and random choice condition. Explanations in the random choice condition will describe all the genres of the movie, but not relate them to the user's preferences. Likewise it will list all the lead actors, and the director, but will not relate whether they are the user's favorites. We will consider alternative sources for approximating the user's true evaluation of the item, or repeat the experiment in a domain which does not have as strong a positive bias as Amazon. A final evaluation in which participants will *view* the movie they rate is also planned.

Acknowledgments. This research was partially supported by the EPSRC platform grant (EP/E011764/1).

References

1. Bilgic, M., Mooney, R.J.: Explaining recommendations: Satisfaction vs. promotion. In: Beyond Personalization Workshop, IUI (2005)
2. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Computer supported cooperative work (2000)
3. Mcsherry, D.: Explanation in recommender systems. Artificial Intelligence Review 24(2), 179–197 (2005)
4. Hingston, M.: User friendly recommender systems. Master's thesis, Sydney University (2006)
5. Tintarev, N., Masthoff, J.: Effective explanations of recommendations: User-centered design. In: Recommender Systems (2007)
6. Tintarev, N.: Explanations of recommendations. In: Recommender Systems (2007)
7. Carenini, G., Moore, D.J.: An empirical study of the influence of user tailoring on evaluative argument effectiveness. In: IJCAI (2001)
8. Cosley, D., Lam, S.K., Albert, I., Konstan, J.A., Riedl, J.: Is seeing believing?: how recommender system interfaces affect users' opinions. In: CHI (2003)
9. Kim, S.M., Pantel, P., Chklovski, T., Pennacchiotti, M.: Automatically assessing review helpfulness. In: EMNLP (2006)
10. McNee, S., Lam, S.K., Guetzlaff, C., Konstan, J.A., Riedl, J.: Confidence displays and training in recommender systems. In: INTERACT IFIP TC13 (2003)
11. Tintarev, N.: Explaining recommendations. In: User Modeling (2007)