Explaining Recommendations

Nava Tintarev
ntintare@csd.abdn.ac.uk

Department of Computing Science, University of Aberdeen, UK

Abstract. This thesis investigates the properties of a good explanation in a movie recommender system. Beginning with a summarized literature review, we suggest seven criteria for evaluation of explanations in recommender systems. This is followed by an attempt to define the properties of a useful explanation, using a movie review corpus and focus groups. We conclude with planned experiments and evaluation.

1 Research Area

Explanations in intelligent systems began with expert systems which were predominantly based on heuristics [1], but also on case-based reasoning (CBR) [2], and model based approaches [3]. In recent years more commercial or entertainment inclined expert systems called recommender systems have begun to offer explanations as well [4–6]. These systems represent user preferences for the purpose of suggesting items to purchase or examine, i.e. recommendations. An explanation in this type of system is formulated along the lines of *"Item A is recommended to you because..."*. The justification following may depend on the underlying recommendation algorithm (e.g. content-based, collaborative-based). Explanations are also intrinsically linked to the way recommendations are presented and the degree of interactivity, see [7] for an in-depth discussion.

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 [8]. User satisfaction, and derivatives thereof such as serendipity [8], diversity [9] and trust [10] are increasingly seen as important. The definition of a *good* explanation is still largely open, and the ways in which explanations can contribute to a recommender system will be the topic of my thesis.

2 Aims and Objectives

The aim of our research is to provide explanations that are optimal for a given user and given criterion (e.g. Trust, see Section 3.1). Our objectives are, for a selection of criteria, to:

- decide upon metrics; *e.g. Trust - increased usage, users return to system* .
- investigate what constitutes optimal content; *what content optimizes Trust?*

Table 1. Criteria

Criteria	**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 (also called conversion)
Efficiency	Help users make decisions faster
Satisfaction	Increase the ease of usability or enjoyment

- investigate what constitutes optimal length; *which length optimizes Trust?*
- build and evaluate an explanation generation system.

We believe that explanations should take into consideration which properties are important for each user. For instance, [5] showed poor acceptance for explanations using information about the user's favorite actor or actress. It would seem plausible that this property (actor/actress) is more important to some users than others. In fact, this is likely to be the case, given that the variance in acceptance for this type of explanations was exceptionally high. Also, we would like to follow in the footsteps of [11, 10] who suggest that concise explanations may be more persuasive and trust inducing respectively.

In later stages of our work we plan to evaluate our conclusions by incorporating explanations into a movie recommender system, using the Duine toolkit [1].

3 Work done so far

3.1 Criteria

To determine what makes a good explanation, it is first necessary to consider the ways in which explanations can be evaluated. In a literature survey (see [7] for details) we have identified seven different criteria by which explanations for single recommendations have been evaluated with users in the past: transparency [12], scrutability [13], trust [10, 14], effectiveness [1], persuasiveness [5], efficiency [15], and satisfaction [16]. We describe each criteria briefly in Table 1. A tentative definition of metrics can be found in [7].

3.2 Analysis of review corpus

Having determined the possible advantages of explanations as criteria, we chose to investigate if there is a difference between explanations that were considered useful for deciding whether or not to watch a movie, i.e. *Effective* explanations.

[1] Telematica Instituut: http://duine.sf.net

Table 2. Properties with frequency counts

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

For this purpose we analyzed 74 user reviews [2] of DVD movies on the British Amazon website [3]. Amazon's reviews are particularly suitable for analysis. The reviews themselves are rated by other users as useful or not. This function may reflect not only what kind of reviews people write, but also what kind of reviews people like to *read*. The corpus referred to 37 movies, each with one useful and one non-useful review. Each review was voted useful/non-useful by at least half, but not less than five of the voters.

In a parallel study of 49 reviews, for 49 different movies, we investigated which properties were mentioned the most often (see Table 2). These properties were based on an informal exploration of reviews on the MovieLens [4] website . For each review, we recorded the frequency of mentioned properties. A property was awarded a point for each mention, regardless of whether it was in favor or disfavor of the movie.

Results: Table 3 summarizes the general properties of useful and non-useful reviews. Useful reviews were longer ($p < 0.01$), and included (a longer) synopsis ($p<0.01$). We also found that useful reviews were more linguistically complex, with a higher Flesch-Kincaid Grade Level ($p<0.01$). The difference for the percentage of passive sentences was not significant however.

Table 3. Mean values for amazon reviews

	total length (words)	**synopsis length (words)**	**% Passive**	**Grade level**
Useful	294.3	87.6	10.6	9.9
Non-Useful	102	3.0	6.1	8.0

We found that reviewers referenced a particular character, rather than the actor or actress. Often, users mentioned that the type of movie was what they would or would not expect in the genre, such as *"the best comedy that I have*

[2] Although reviews are not identical to explanations within a recommender system, we believe that they are sufficiently similar to deduce properties of a useful explanation.

[3] http://amazon.co.uk

[4] http://movielens.umn.edu/

ever seen". Initial expectations were often influenced by adaptations from books, previous releases, awards, and previous reviews.

3.3 Focus groups

To investigate how these properties could be applied to explanations, two focus groups with a total of 11 participants were conducted. In these focus groups the participants described movies they had seen; their initial expectations, their reactions after seeing the movie, what it was that formed their opinion of the movie and what kind of explanation they would like to receive. A limited summary follows below.

- The decisive properties for seeing a movie varied between users e.g. director, script complexity, dialogs, genre, and subject matter.

- Movies seen with groups of friends were often "light, easy viewing", with simpler scripts than those viewed in more intimate company or alone. Light movies were also preferred before a mentally demanding activity such as an exam.

- Participants did not want to be dissuaded from watching movies, even if it would have helped them avoid watching a movie they had not enjoyed in the past. Social effects such as movie popularity and an outing with friends were often in play.

- Some properties were more descriptive, such as cast, filming location, and black and white. This became particularly clear when participants attempted to clarify the identity of a movie.

- Reviews may help users enjoy movies more, rather than serve merely as decision aids. Participants believed that correcting faulty expectations of a movie would not influence whether or not they saw it. Rather, it could increase their acceptance upon viewing, and save potential disappointment.

4 Planned Work and conclusion

We plan to conduct a number of experiments in order to refine our idea of how explanations should be presented in natural language. In one study participants will be asked to compare sets of reviews controlled for scenario, and type of movie properties. They will be asked to edit these reviews as well as to specify which of the properties from a list should be mentioned in a review of this movie. A second study will target the question of balancing the number of properties to mention versus the amount of detail. We will also compare different interfaces, and user preferences for text versus graphics.

The final model of explanations will be implemented in a movie recommender system. Our aim is to evaluate the system with users, according to several of the criteria mentioned in Section 3.1. For example a likely metric for *Effectiveness* is the difference in rating for a movie upon recommendation, and after viewing [4]. We plan to compare the system with and without explanations for each criterion, and also measure criteria against each other (e.g. longer explanations inspire *Trust* but decrease *Efficiency*).

We hope that our work on explanations will contribute to the field of recommender systems, via an understanding of how to personalize explanations, and how much content to present.

References

1. Buchanan, B.G., Shortliffe, E.H., eds.: 30-35. In: The Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Addison-Wesley Publishing Company (1985) 571–665
2. Doyle, D., Tsymbal, A., Cunningham, P.: A review of explanation and explanation in case-based reasoning. Technical report, Department of Computer Science, Trinity College, Dublin (2003)
3. Druzdzel, M.J.: Qualitative verbal explanations in bayesian belief networks. Artificial Intelligence and Simulation of Behaviour Quarterly, special issue on Bayesian networks (1996) 43–54
4. Bilgic, M., Mooney, R.J.: Explaining recommendations: Satisfaction vs. promotion. In: Beyond Personalization Workshop, IUI. (2005)
5. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Computer Supported Cooperative Work. (2000)
6. Mcsherry, D.: Explanation in recommender systems. Artificial Intelligence Review **24(2)** (2005) 179 – 197
7. Tintarev, N., Masthoff, J.: A survey of explanations in recommender systems. In: WPRSIUI associated with ICDE. (2007)
8. McNee, S.M., Riedl, J., Konstan, J.A.: Being accurate is not enough: How accuracy metrics have hurt recommender systems. In: Extended Abstracts of CHI. (2006)
9. Ziegler, C.N., McNee, S.M., Konstan, J.A., Lausen, G.: Improving recommendation lists through topic diversification. In: WWW'05. (2005)
10. Chen, L., Pu, P.: Trust building in recommender agents. In: WPRSIU'02. (2002)
11. Carenini, G., J. Moore, J.: An empirical study of the influence of argument conciseness on argument effectiveness. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics. (2000)
12. Sinha, R., Swearingen, K.: The role of transparency in recommender systems. In: Conference on Human Factors in Computing Systems. (2002)
13. Czarkowski, M.: Evaluating scrutable adaptive hypertext. In: 10th International Conference on User Modeling, Workshop 3: Evaluation of Adaptive Systems. (2005)
14. Swearingen, K., Sinha, R.: Interaction design for recommender systems. In: Designing Interactive Systems. (2002)
15. Thompson, C.A., Göker, M.H., Langley, P.: A personalized system for conversational recommendations. J. Artif. Intell. Res. (JAIR) **21** (2004) 393–428
16. Sinha, R., Swearingen, K.: Comparing recommendations made by online systems and friends. In: DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries. (2001)