Explanations of Recommendations

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

ABSTRACT

This thesis focuses on explanations of recommendations. Explanations can have many advantages, from inspiring user trust to helping users make good decisions. We have identified seven different aims of explanations, and in this thesis we will consider how explanations can be optimized for some of these aims. We will consider both an explanation's content and its presentation. As a domain, we are currently investigating explanations for a movie recommender, and developing a prototype system. This paper summarizes the goals of the thesis, the methodology we are using, the work done so far and our intended future work.

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 [15]. User satisfaction and derivatives thereof such as serendipity [15], diversity [21] and trust [4] 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, which we have reviewed and discussed in-depth in [19]. 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. When choosing and comparing explanation techniques, it is very important to agree on what the explanation is trying to achieve. For example, although the study by [6] measured user appraisal of *recommendations* (persuasion), this is not the same as measuring appraisal of actual *items* (effectiveness) [1]. In this thesis, we investigate the general properties of explanations that help a movie recommender system fulfill some of the criteria. In our work so far, we have focused on Effectiveness (i.e. helping users to make good decisions), Trust (i.e. increasing users' confidence in the system), and Satisfaction. We are looking both at the content of explanations and the way they are presented.

Table 1. Possible aims for explanations

Aim	**Definition**
Transparency	Explain how the system works
Scrutability	Allow users to tell the system it is wrong
Trustworthiness	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

2. RELATED WORK

Since there is no consensus on how to evaluate an explanation facility in a recommender system, work to date strongly reflects the fact that explanations in different recommender systems try to achieve different aims.

The aim of transparency stems back to the days of expert systems such as MYCIN [2]. More recent studies such as [17] investigated how explanations could contribute to *Transparency* in recommender systems. Other work (e.g. [5]) thoroughly evaluates an adaptive hypertext according to its ability to grant users the power to control the personalization process (*Scrutability*). Chen and Pu [4] have made a substantial step toward understanding how a number of presentational choices affect user *Trust* and intent to return to a system. Bilgic and Mooney [1] also considered presentational choices and compared three ways of presenting explanations for book recommendations. The explanations were evaluated according to how much they helped users find books they really ended up liking (*Effectiveness*). In contrast, Herlocker et al. [6] evaluated twenty-one different explanation interfaces for a movie recommendation, but measured them according to how likely a user thought they would be to see this movie at the cinema (*Persuasiveness*). Carenini and Moore [3] also focused on persuasion, and found concise and personalized explanations to be more persuasive in the house domain. Other systems aimed at decreasing the time it takes a user to find a good item, i.e. increasing recommendation *Efficiency.* This has been done by helping the user understand the relation between competing options [13, 14].

It is unlikely for an explanation facility in a recommender system to be optimized for all seven aims, and thus all the more important that this is a conscious choice. Occasionally, the choice of aim is not conscious and leads to trade-offs the creators of the system initially did not consider. For example consider the work of [9], which found that while medical staff in a neonatal unit *preferred* graphics (*Satisfaction*), they actually made *better* choices (Effectiveness) as well as were able to make decisions *quicker* (Efficiency) for the textual version.

3. WORK DONE SOFAR

3.1 Survey of Explanations and Metrics

As discussed above, we have conducted a comprehensive review of explanations [19]. In our review, we argue that explanations are intrinsically linked with the way recommendations are presented (for instance, top item, top N-items, etc), and with the degree of interactivity offered. The tables therein offer an overview of explanation facilities in existing commercial and academic recommender systems respectively. In addition we considered how to measure the 'goodness' of explanations for each of the aims in Table 1. For instance, one way to evaluate how effective explanations are would be to measure the liking of the recommended item prior to and after consumption. Persuasiveness can be measured as the effect on the likelihood of selecting an item. We refer back to [19] for an in detail discussion for each of the aims.

3.2 Content of Effective Explanations

Using a user-centered design approach, we have investigated what characterizes general properties of useful, or Effective, explanations of recommendations. We have used a methodology of corpus analyses and focus groups [20] to elicit important features in our movie domain, as well as to obtain heuristics such as the optimal number of features to mention. The rationale behind studying user's utilization of item 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 [1] 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.

We conducted two corpus analyses based on online movie reviews. Although reviews are not the same as explanations, we argue that an explanation aimed at Effectiveness is different from an explanation aimed at e.g. Transparency. That is, an explanation aiming to help users make better decisions is more likely to be based on evaluative arguments (such as reviews) than an explanation aiming to explain how an item was selected.

The first study of movie reviews suggests that helpful reviewers mention roughly 4-5 features. Together, the two studies of movie reviews suggest that the optimal length for reviews lies around 200 words, describing each feature in some detail. However, explanations differ from reviews, and in a real context, explanations cannot be this long. In particular a recommender suggesting several items at once could suffer from severe space limitations. In this case it is important to make the choice between mentioning more features, or fewer but in greater detail.

There was a general consensus across focus groups and corpus analysis for a (short) list of features (such as good in its genre, script complexity, and mood). However, participants also weighed features differently, suggesting that personalization of feature selection is needed. In addition, we found that this personalization should take the viewing context and mood into account (see [20] for more results from our focus groups).

As an explanation in a recommender system may need to be presented in a limited space, we have also empirically investigated how to best strike a balance between the number of features to mention, and the degree of detail in which they are described.

Our experiment suggests that when forced to choose, people are more likely to prefer fewer features, but described in more detail. People however differ in their degree of interest in features and details, and in the relative proportion of words spent on the "plot" (descriptive) and "opinions" (justifying features). Therefore, while few features in detail may be a suitable default for an explanation in a recommender system, our results also suggest that there may be merit in personalizing the number of features, and degree of detail as well.

3.3 Explanations that Increase Trust

We have started empirical work on explanations that increase user trust. In a pilot experiment, we have investigated whether the balance between more features versus more detail has an impact on user trust. Each participant was given one explanation out of set of three (varying in features and detail). We used questions from two validated trust questionnaires developed by [8, 16], to determine how much participants trusted the recommender system based on the explanation. The results of this pilot study are not conclusive, clearly only one explanation is too little to impact user trust. There may also be personalization issues, with some users' trust being impacted positively by more features, and others by more detail. We are in the process of planning an improved experiment.

3.4 Presentation of Explanations

We have also begun to compare different methods of presentation of explanations. We have conducted seven focus groups, with a total of sixty-seven participants. Our intention was to receive qualitative feedback on interfaces similar to those used in the [6] study of explanations in collaborative recommender systems.

Firstly, we perceived that the study often did not control for content, i.e. although a graphical interface gained the most acceptance, it did not have a textual counterpart. In this study we therefore often compared graphical and textual equivalents. In fact, participants initially preferred an equivalent textual description to the bar chart preferred by most participants in [6], until we changed the graphic to a pie chart! We received a great deal of detailed qualitative feedback, including the relevance of the content. For example, most of the participants felt that information about recommendation confidence should be omitted altogether, or at least not be shown as a justification for a recommended item (in any medium). Rather, the participants felt that uncertain recommendations should simply be omitted. In addition, participants suggested more detailed possible improvements to interfaces. As participants told us what they liked and preferred, this study focused on *perceived* Effectiveness of explanations.

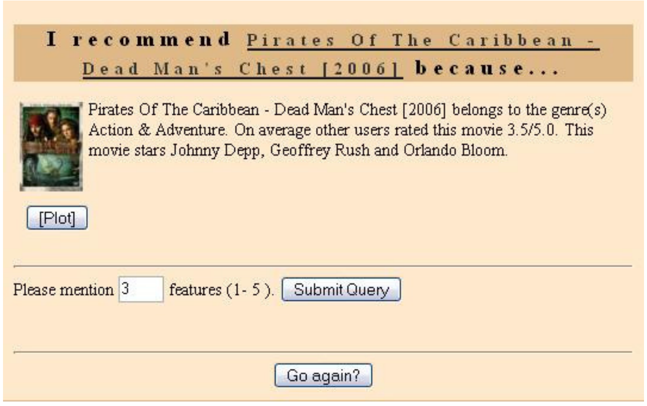
3.5 Semantic Similarity

This thesis started in the domain of News recommendations, but changed domain after feedback in the Recommender Systems summer school in Bilbao last year. We had already explored different ways to measure similarity for news headlines [18]. We compared human judgements of similarity with Lin's taxonomy-based measure [12] and the WASP measure that uses annotated corpus data [11]. The main aim of this work was to better understand similarity, so that it can be used to explain recommendations to users. We found that both the Lin and WASP measures were feasible options for calculating similarity in the context of real-world news headlines. WASP has the advantage of not needing word senses. However, the Lin measure is better suited for constructing explanations that can be understood by users. We proposed a hybrid approach, in which we calculate both the WASP and the Lin measures, using word sense 1 for all words (to avoid manual annotation). We can then use the WASP measure as the similarity measure to decide on recommendations, but whenever the measures give similar results, use the least common subsumer as used by Lin to explain the similarity to the user. Though this work was done in the news domain, we believe that it will be applicable to the movie domain as well, and expect to use this in future versions of our prototype.

3.6 Prototype

Currently, we are developing a prototype which generates explanations for movie recommendations. Figure 1 shows an explanation for a single movie, based on the features that a user found most important features (other people's ratings, and actors), and specific genre preference (action and adventure). Our findings suggest that plot and genre information is important to most if not all users, so these are made available by default. However, to cater for potential space restrictions, the full plot description only extends when explicitly requested by the user.

Figure 2 illustrates the basic structure of this prototype. The user model used in this system weighs the movies features elicited by our studies, according to specified user utility.



I recommend Pirates Of The Caribbean - Dead Man's Chest [2006] **because...**

Pirates Of The Caribbean - Dead Man's Chest [2006] belongs to the genre(s) Action & Adventure. On average other users rated this movie 3.5/5.0. This movie stars Johnny Depp, Geoffrey Rush and Orlando Bloom.

Please mention 3 features (1- 5).

Figure 1. Single item explanation

The flexibility of a natural language generation system allows users to change text preferences such as the number of movies to show, the explanations' degree of detail, and the number of features mentioned. The prototype can then be used as a further test-bed for explanations in recommendations. For example, we could vary degree of detail and number of features for a number of movies, and study their effect on explanation aims such as Trust and Effectiveness.

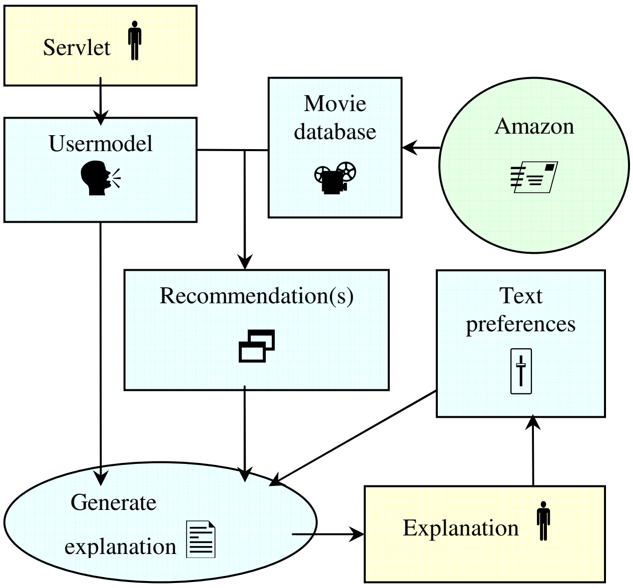


Figure 2. Prototype structure

4. PLANNED WORK AND CONCLUSIONS

In the next year we consider pursuing one of the following directions. Firstly, we would like to study the effect of the explanation *explicitly* stating that it does not have any information about a particular feature (e.g. the director information is unknown), vs. selecting movies that do describe all the features that are important for a particular user. This line of research is inspired by the negative comments by participants of the focus groups surveying presentation interfaces based on recommendation confidence. It would also be interesting to see if non-recommendations (e.g. "don't see this movie because...") inspire more or less trust than positive recommendation (e.g. "see this movie because…"). More specifically, we would like to find out whether or not this can increase the Trustworthiness of the system, or help the user make a correct decision whether to watch a movie or not (increase Effectiveness).

Secondly, we would like to tackle a few presentational issues of explanations. We would like to answer questions such as, in regards to e.g. Trust and Effectiveness, when is text preferable to graphics and vice versa? When do the two media compliment each other? Does the order in which features are mentioned have an effect? We also plan to consider how different types of explanation content (e.g. content-based, collaborative-based, and case-based) affect these types of preferences (e.g. choice of media).

Last, but not least, remains the question of the underlying recommender engine. The focus of this thesis lies with the optimal presentation and content for explanations rather than designing a strong recommendation systems engine. For the sake of feasibility it would however be beneficial if our prototype simulates or uses a pre-existing engine. A possible solution would be to extend the prototype to use Amazon's recommendation engine, via a similarity lookup on Amazon's ECS. However, while our explanations are based on content-based preferences, Amazon's recommendations are computed according to a collaborative algorithm. One of the main challenges for the coming year will be to bridge this gap, possibly by augmenting limited meta-data with information from customer and editorial reviews, extracted with the help of semantic similarity measures adapted from [18].

4. REFERENCES

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

[2] Buchanan, B.G. & Shortliffe, E.H. *(ed.)* The Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project 30-35 *Addison-Wesley Publishing Company*, 571-665, 1985

[3] Carenini, G. and Moore, J. An empirical study of the influence of argument conciseness on argument effectiveness. *ACL*, 2000.

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

[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] Hong, T. Contributing Factors to the Use of Health-Related Websites. *Journal of Health Communication*, 11:2, 149-165, 2006

[9] Law, A.S., Freer, Y. Hunter, J., Logie, R.; N, N.M. & Quinn, J. A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. *J Clin Monit Comput., 19:3*, 183-94, 2005

[10] Kincaid, J.P., Fishburne Jr., R.P., Rogers, R.L. and Chissom, B.S. Derivation of new readability formulas. Technical report, TN: Naval Technical Training, U. S. Naval Air Station, Memphis, TN, 1975.

[11] Kilgarriff, A., and Tugwell, D. (2001). Word sketch: Extraction and display of significant collocations for lexicography. *Collocations Workshop* in association with ACL, 2001.

[12] Lin, D. An information-theoretic definition of similarity. Proceedings of the Fifteenth International Conference on Machine Learning, 296 – 304, 1998.

[13] McCarthy, K., Reilly, J., McGinty, L. & Smyth, B. Thinking Positively - Explanatory Feedback for Conversational Recommender Systems. *European Conference on Case-Based Reasoning Explanation Workshop,* 2004

[14] McSherry, D.. Explanation in recommender systems. *Artificial Intelligence Review*, 24(2):179 – 197, 2005.

[15] 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*

[16] Ohanian, R. Construction and Validation of a Scale to Measure Celebrity Endorsers' Perceived Expertise, Trustworthiness, and Attractiveness *Journal of Advertising*, 19:3, 39, 1990.

[17] Sinha, R. & Swearingen, K. The role of transparency in recommender systems
Conference on Human Factors in Computing Systems, 2002

[18] Tintarev, N. & Masthoff, J. Similarity for News Recommender Systems. *WPRSIUI*, in association with AH, 2006.

[19] Tintarev, N. and Masthoff, J. Survey of explanations in recommender systems. *WPRSIUI*, in association with ICDE, 2007.

[20] Tintarev, N. & Masthoff, J. Effective Explanations of Recommendations: User-Centered Design. In *ACM Recommender System*, 2007.

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