If you liked Herlocker et al.'s explanations paper, then you might like this paper too

Derek Bridge
Insight Centre for Data Analytics
University College Cork, Ireland
derek.bridge@insight-centre.org

Kevin Dunleavy
School of Computer Science and IT
University College Cork, Ireland
kevdunleavy@gmail.com

ABSTRACT

We present *explanation rules*, which provide explanations of user-based collaborative recommendations but in a form that is familiar from item-based collaborative recommendations; for example, "People who liked *Toy Story* also like *Finding Nemo*". We present an algorithm for computing explanation rules. We report the results of a web-based user trial that gives a preliminary evaluation of the perceived effectiveness of explanation rules. In particular, we find that nearly 50% of participants found this style of explanation to be helpful, and nearly 80% of participants who expressed a preference found explanation rules to be more helpful than similar rules that were closely-related but partly-random.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information Filtering*

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Recommender Systems, Explanations

1. INTRODUCTION

An *explanation* of a recommendation is any content, additional to the recommendation itself, that is presented to the user with one or more of the following goals: to reveal how the system works (transparency), to reveal the data it has used (scrutability), to increase confidence in the system (trust), to convince the user to accept the recommendation (persuasion), to help the user make a good decision (effectiveness), to help the user make a decision more quickly (efficiency), or to increase enjoyment in use of the system (satisfaction) [11, 14]. The focus in this paper is effectiveness: explanations that help users to decide which item to consume.

IntRS 2014, October 6, 2014, Silicon Valley, CA, USA.

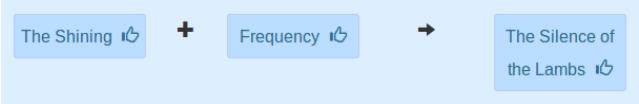

Figure 1: An explanation rule

The problem that we examine in this paper is how to produce effective explanations of user-based collaborative recommendations. It is relatively easy to explain the recommendations of *content-based recommenders*, e.g. by displaying meta-descriptions (such as features or tags) that the active user's profile and the recommended item have in common [10, 13]. *Item-based collaborative recommendations* are also amenable to explanation, e.g. by displaying items in the user's profile that are similar to the recommended item [8, 6]. *User-based collaborative recommendations*, on the other hand, are harder to explain. Displaying the identities of the active user's neighbours is unlikely to be effective, since the user will in general not know the neighbours; displaying their profiles is unlikely to be effective, since even the parts of their profiles they have in common with the active user will be too large to be readily comprehended.

It is possible to explain a recommendation using data other than that which the recommender used to generate the recommendation [2]. For example, a system could explain a user-based collaborative recommendation using the kind of data that a content-based recommender uses (features and tags), e.g. [9]. In our work, however, we try to preserve a greater degree of fidelity between the explanation and the operation of the recommender. Specifically, we generate the explanation from co-rated items on which the active user and her nearest-neighbour agree.

We propose an algorithm for making item-based explanations, also referred to as influence-style explanations [1]; for example, "People who liked *Toy Story* also like *Finding Nemo*". This style of explanation is familiar to users of amazon.com [6], for example. These are the kind of explanation most commonly produced by item-based collaborative recommenders. But we will show how to produce them in the case of user-based collaborative recommenders. The algorithm is adapted from one recently proposed to explain case-based classifiers [7]. It produces explanations in the form of *explanation rules*. The antecedent of an explanation rule characterizes a subset of the active user's tastes that are predictive of the recommended item, which appears in the consequent of the rule; see the example in Figure 1.

	Alien	*Brazil*	*Crash*	*Dumbo*	*E.T.*	*Fargo*
Ann	2	4	1	2		4
Bob	5	4		1		5

Table 1: A ratings matrix

2. EXPLANATION ALGORITHM

We use a conventional user-based collaborative recommender of the kind described in [4]. Like theirs, our recommender finds the active user's 50 nearest neighbours using significance-weighted Pearson correlation; for each item that the neighbours have rated but the active user has not, it predicts a rating as the similarity-weighted average of deviations of neighbours' ratings from their means; it recommends the items with the highest predicted ratings.

Before presenting the explanation algorithm, we define some terms:

Explanation partner: The *explanation partner* is the member of the set of nearest neighbours who is most similar to the active user and who likes the recommended item. Often this will be the user who is most similar to the active user — but not always. In some cases, the most similar user may not have liked the recommended item: the recommendation may be due to the votes of other neighbours. In these cases, one of these other neighbours will be the explanation partner. It may appear that *recommendations* exploit the opinions of a set of neighbours (for accuracy), but *explanations* exploit the opinions of just one of these neighbours, the explanation partner. But this is not completely true. As we will explain below, the items included in the explanation are always members of the explanation partner's profile, but they are also validated by looking at the opinions of *all* other users (see the notions of coverage and accuracy below).

Candidate explanation conditions: Let u be the active user and v be the explanation partner; let j be a co-rated item; and let r_{uj} and r_{vj} be their ratings for j. We define *candidate explanation conditions* as co-rated items j on which the two users agree.

In the case of numeric ratings, we do not insist on rating equality for there to be agreement. Rather, we define agreement in terms of liking, indifference and disliking. For a 5-point rating scale, the candidate explanation conditions would be defined as follows:

$$\begin{aligned}\text{candidates}(u,v) = \\ &\{\text{likes}(j) : r_{uj} > 3 \wedge r_{vj} > 3\} \cup \\ &\{\text{indiff}(j) : r_{uj} = 3 \wedge r_{vj} = 3\} \cup \\ &\{\text{dislikes}(j) : r_{uj} < 3 \wedge r_{vj} < 3\}\end{aligned}$$

For example, the candidate explanation conditions for users Ann and Bob in Table 1 are

$$\{\text{likes}(\textit{Brazil}), \text{dislikes}(\textit{Dumbo}), \text{likes}(\textit{Fargo})\}$$

Alien does not appear in a candidate condition because Ann's and Bob's ratings for it disagree; *Crash* and *E.T.* do not appear in candidate conditions because neither of them is co-rated by Ann and Bob.

Input: user profiles U, recommended item i, active user u, explanation partner v
Output: an explanation rule for i
$R \leftarrow$ if _ then i;
$Cs \leftarrow$ candidates(u, v);
while accuracy$(R) < 100 \wedge Cs \neq \{\ \}$ **do**
  $Rs \leftarrow$ the set of all new rules formed by adding singly each candidate condition in Cs to the antecedent of R;
  $R^* \leftarrow$ most accurate rule in Rs, using rule coverage to break ties between equally accurate rules;
  if accuracy$(R^*) \leq$ accuracy(R) **then**
    return R;
  $R \leftarrow R^*$;
  Remove from Cs the candidate condition that was used to create R;
return R;

Algorithm 1: Creating an explanation rule

Rule coverage: A rule *covers* a user if and only if the rule antecedent is satisfied by the user's profile. For example, the rule in Figure 1 covers any user u whose profile contains ratings $r_{u,\textit{TheShining}} > 3$ and $r_{u,\textit{Frequency}} > 3$, irrespective of what else it contains. Rule *coverage* is then the percentage of users that the rule covers.

Rule accuracy: A rule is *accurate* for a user if and only if the rule covers the user and the rule consequent is also satisfied by the user's profile. For example, the rule in Figure 1 is accurate for any user u whose profile additionally contains $r_{u,\textit{TheSilenceoftheLambs}} > 3$. Rule *accuracy* is then the percentage of covered users other than the active user for whom the rule is accurate.

The algorithm for building an explanation rule works incrementally and in a greedy fashion; see Algorithm 1 for pseudocode. Initially, the rule has an empty antecedent, and a consequent that contains the recommended item i, written as 'if _ then i' in Algorithm 1. On each iteration, the antecedent is refined by conjoining one of the candidate explanation conditions, specifically the one that leads to the most accurate new rule, resolving ties in favour of coverage. This continues until either the rule is 100% accurate or no candidate explanation conditions remain.

3. EXPERIMENTS

We tested three hypotheses, the first using an offline experiment, the other two using a web-based user trial.

3.1 Practicability of explanation rules

The number of candidate explanation conditions can be quite large. If explanation rules are to be practicable, then the number of conditions that the algorithm includes in the antecedent of each explanation rule needs to be quite small.

Hypothesis 1: that explanation rules will be short enough to be practicable.

We ran the user-based collaborative recommender that we described at the start of the previous section on the MovieLens 100k dataset, and obtained its top recommendation for each user in the dataset. We then ran the explanation algorithm to produce an explanation rule that would explain

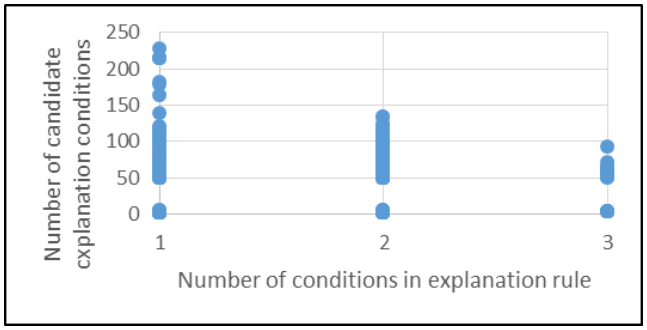

Figure 2: Rule length

the recommended item to that user. In Figure 2, we plot the number of candidate explanation conditions (vertical axis) against the number of these conditions that the algorithm includes in the rule (horizontal axis).

From the Figure, we see that the longest rules contained only three items in their antecedents. Not only that, but actually only 4% of the rules had three items in their antecedents; the other 96% were split nearly evenly between those having one and those having two items. We also see that the more candidates there are, the shorter the explanation rule tends to be. We have not investigated the exact reasons for this.

We repeated this experiment using a dataset with unary ratings to see what difference this might make. We took a LastFM dataset that contains artist play counts for 360 thousand users and 190 thousand artists.[1]. We converted play counts to unary ratings, i.e. recording 1 if and only if a user has played something by an artist. The results were very similar to those in Figure 2 (which is why we do not show them here), again with no rule having more than three items in its antecedent.

These are encouraging results for the practicability of explanation rules.

3.2 Effectiveness of this style of explanation

We designed a web-based user trial, partly inspired by the experiment reported in [5], drawing data from the MovieLens 1M dataset. Trial participants visited a web site where they progressed through a series of web pages, answering just three questions. An initial page established a context, essentially identical to the one in [5]:

> Imagine you want to go to the cinema but only if there is a movie worth seeing. You use an online movie recommender to help you decide. The movie recommender recommends one movie and provides an explanation.

First, we sought to elicit the perceived effectiveness of this *style* of explanation with the following hypothesis:

Hypothesis 2: that users would find explanation rules to be an effective style of explanation.

We showed participants an explanation rule for a recommendation and we asked them to rate its helpfulness on a 5-point scale. Specifically, we asked "Would this style of explanation help you make a decision?" with options Very unhelpful, Unhelpful, Neutral, Helpful, and Very helpful. Our

[1] `mtg.upf.edu/node/1671`

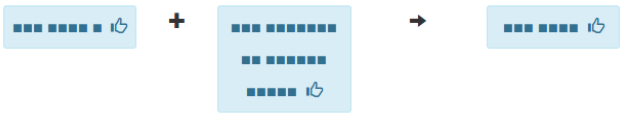

Figure 3: A redacted explanation rule

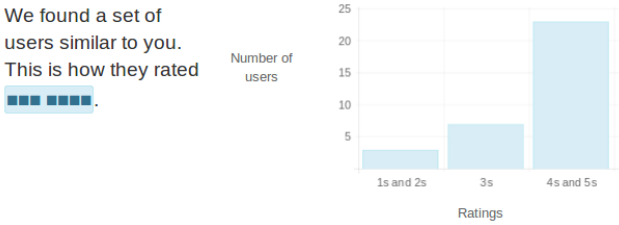

Figure 4: A redacted explanation in the style of [5]

wording differs from that used by [5]. They asked how likely the user would be to go and see the movie, with answers on a 7-point scale. Our wording focuses on explanation effectiveness (helpfulness in making a decision), whereas theirs focuses on persuasiveness.[2]

To encourage participants to focus on explanation *style*, we followed [5] in redacting the identity of the recommended movie. A participant's feedback is then not a function of the quality of the recommendation itself. For the same reasons, we obscured the identities of the movies in the antecedent of the explanation rule; see the example in Figure 3.

To obtain a 'yardstick', we also showed participants another explanation and asked them whether it too was helpful. For this purpose, we used the most persuasive explanation style from [5]. This explanation takes the form of a histogram that summarizes the opinions of the nearest neighbours. Figure 4 contains an example of this style of explanation (again with the recommended item redacted).

In the experiment, the software randomly decides the order in which it shows the two explanation styles. Approximately 50% of participants see and rate the explanation rule before seeing and rating the histogram, and the remainder see and rate them in the opposite order.

Prior to asking them to rate either style of explanation, users saw a web page that told them that we had obscured the movie titles, and we showed them an explicit example of a redacted movie title. We conducted a pilot run of the experiment with a handful of users before launching the real experiment. Participants in the pilot run did not report and difficulty in understanding the redacted movie titles or the redacted explanation rules.

We had 264 participants who completed all parts of the experiment. We did not collect demographic data about the participants but, since they were reached through our own contact lists, the majority will be undergraduate and postgraduate students in Irish universities.

Figure 5 shows how the participants rated explanation rules for helpfulness. Encouragingly, nearly 50% of participants found explanation rules to be a helpful or very helpful style of explanation (100 and 16 participants out of the 264,

[2] This is an observation made by Joseph A. Konstan in lecture 4-4 of the Coursera course *Introduction to Recommender Systems*, `www.coursera.org`.

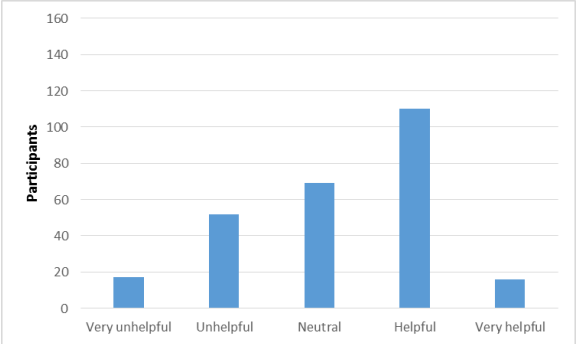

Figure 5: Helpfulness of redacted explanation rules

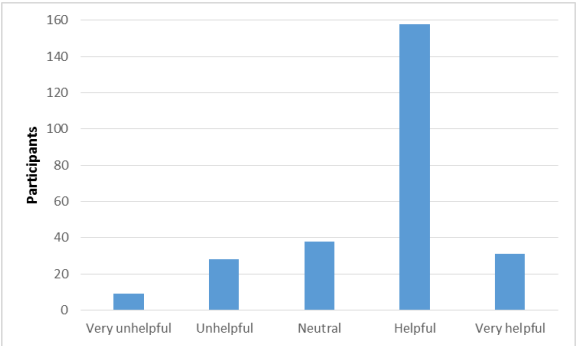

Figure 6: Helpfulness of redacted histograms

resp.); but about a quarter of participants found them neutral (69 participants), and a quarter found them unhelpful or very unhelpful (52 and 17, resp.). Figure 6 shows the same for the other style of explanation. Just over 70% of participants found this style of explanation to be helpful or very helpful (158 and 31 participants, resp.).

Note that we did not ask participants to *compare* the two styles of explanation. They are not in competition. It is conceivable that a real recommender would use both, either side-by-side or showing one of the two explanations by default and only showing the other to users who click through to a more detailed explanation page.

Furthermore, as the reader can judge by comparing Figures 3 and 4, any direct comparison of the results is unfair to the explanation rules since they have two levels of redaction (the recommended movie and the antecedents in the rules) whereas the histogram has just one (the recommended movie). As far as we can tell, there is no explanation style in [5] that would give comparable levels of redaction for a fair experiment.

For some readers, this may raise the question of why we showed participants the redacted histograms at all. The reason is to give a 'yardstick'. If we simply reported that nearly 50% of participants found explanation rules to be helpful or very helpful, readers would not know whether this was a good outcome or not.

From the results, we cannot confidently conclude that the hypothesis holds: results are not in the same ball-park as the 'yardstick'.[3] But we can conclude that explanation rules are a promising style of explanation: many users perceive them to be a helpful style of explanation, and they are therefore deserving of further study in a more realistic setting.

We note as a final comment in this subsection that the experiment reported in [1], which uses a very different methodology and no redaction of movie titles, found item-based explanations (there referred to as influence style explanations) to be better than neighbourhood style explanations.

3.3 Effectiveness of the selection mechanism

Next, we sought to elicit the perceived effectiveness of our algorithm's way of building explanation rules:

Hypothesis 3: that users would find the algorithm's selection of conditions in the antecedents of the rules (based on accuracy and coverage) to be better than random.

In the same web-based user trial, we showed the participants two rules side-by-side (the ordering again being determined at random). One rule was constructed by Algorithm 1. The other rule was constructed so as to have the same number of conditions in its antecedent, but these were selected at random from among the candidate explanation conditions. Note they are not wholly random: they are still candidate explanation conditions (hence they are co-rated items on which the user and explanation partner agree) but they are not selected using accuracy and coverage.

We asked participants to compare the two rules. They selected one of four options: the first rule was more helpful than the second; the second was more helpful than the first; the two rules were equally helpful; and they were unable to tell which was the more helpful ("don't know").

There was no redaction in this part of the experiment. It was important that participants judged whether the movie preferences described in the antecedents of the rules did support the recommended movie. Prior to asking users to rate the two explanation rules, users saw a web page that told them: that they would see a recommendation; that they should pretend that the recommended movie was one that they would like; that they would see two explanations; that movie titles would no longer be obscured; and that they should compare the two explanations for helpfulness. There are, of course, the risks that measuring effectiveness before consumption like this may result in judgements that overlap with persuasiveness, and that measuring perceived effectiveness is not as reliable as measuring something more objective [12].

Figure 7 shows the outcomes of this part of the experiment. We see that 32% found the explanation rule to be more helpful (85 participants) and only 10% (27 participants) found the partly-random rules to be more helpful. This means that, of those who expressed a preference (85 plus 27 participants), 76% preferred the explanation rules and only 24% preferred the partly-random rules. Furthermore, a two-tailed z-test shows the difference to be significant at the 0.01 level. This suggests that the algorithm does select candidate explanation conditions in a meaningful way.

However, 36% of participants found the rules to be equally helpful and 22% could not make a decision (95 and 57 participants resp.). This means, for example, that (again using

[3]For readers who insist on a comparison: using Very Unhelpful = 1, Unhelpful = 2, etc., the mean rating for the redacted explanation rules is 3.21 (st.dev. 1.03), the mean rating for the redacted histograms is 3.66 (st.dev. 0.94); and, using Welch's t-test, we reject at the 0.01 level the null hypothesis that there is no difference in the means.

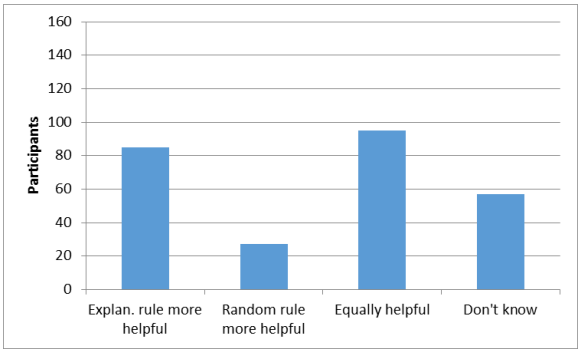

Figure 7: Helpfulness of explanation rules compared with partly-random rules

Explanation rule		Partly-random rule	
Accuracy	Coverage	Accuracy	Coverage
91%	2%	56%	15%
83%	1%	68%	4%
76%	11%	42%	33%
25%	3%	20%	13%

Table 2: Accuracy and coverage of pairs of rules

a two-tailed z-test), there is no significant difference between the proportion who found explanation rules to be more helpful and the proportion who found the two rules to be equally helpful.

There are at least two reasons for this. The first is that the participant is required to put herself 'in the shoes' of another user. The recommendation and the rules are computed for a user in the MovieLens dataset, not for the person who is completing the experiment, who must pretend that she likes the recommendation. The person who completes the experiment may not know much, if anything, about the movies mentioned in the rules. This may be why the "don't know" option was selected so often.[4] The alternative was to require participants in the experiment to register with the recommender and to rate enough movies that it would be able to make genuine recommendations and build realistic explanation rules. We felt that this placed too great a burden on the participants, and would likely result in an experiment skewed towards users with relatively few ratings.

The second reason is that the partly-random rules are still quite good rules: they are considerably more meaningful than wholly-random rules. As Table 2 shows, one of the partly-random rules used in the experiment is nearly as accurate as its corresponding explanation rule. The partly-random rules also have high coverage because randomly selected movies are often popular movies. In our pilot run of the experiment, we had tried wholly-random rules, but they were so egregiously worse than their corresponding explanation rules that we felt that using them would prejudice the results of the real experiment. Ironically, the partly-random rules that we use instead perhaps include too many movies that are reasonable substitutes for the ones in their corresponding explanation rules, thus giving us much more equivocal results.

[4]An on-screen note told the participant that she was able to click on any title to get some information about the movie. If she did, we fetched and displayed IMDb genres and a one-line synopsis for the movie. But we did not record how many users exploited this feature.

4. CONCLUSIONS

We have presented an algorithm for building explanation rules, which are item-based explanations for user-based collaborative recommendations. We ran an offline experiment and web-based user trial to test three hypotheses. We conclude that explanation rules are a practicable form of explanation: on two datasets no rule antecedent ever contained more than three conditions. We conclude that explanation rules offer a promising style of explanation: nearly 50% of participants found them to be helpful or very helpful, but the amount of redaction used in the experiment makes it hard to make firm conclusions about their effectiveness. Finally, we conclude that users do find the algorithm's selection of conditions for the rule antecedent to be better than random: just under 80% of participants who expressed a preference preferred the explanation rule to a partly-random variant. But results here are also partly confounded by the conditions of the experiment, where a participant has to put herself 'in the shoes' of another user.

Given the caveats about the limitations of the experiments, our main conclusion is that explanation rules are promising enough that we should evaluate them further, perhaps in a comparative experiment such as the one reported in [3] or in A/B experiments in a real recommender.

5. ACKNOWLEDGMENTS

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289. We are grateful to Barry Smyth for discussions about the design of our experiment.

6. REFERENCES

[1] M. Bilgic and R. Mooney. Explaining recommendations: Satisfaction vs. promotion. In *Procs. of Beyond Personalization 2005: A Workshop on the Next Stage of Recommender Systems Research at the 2005 International Conference on Intelligent User Interfaces*, 2005.

[2] G. Friedrich and M. Zanker. A taxonomy for generating explanations in recommender systems. *AI Magazine*, 32(3):90–98, 2011.

[3] F. Gedikli, D. Jannach, and M. Ge. How should I explain? A comparison of different explanation types for recommender systems. *Int. J. Hum.-Comput. Stud.*, 72(4):367–382, 2014.

[4] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In F. Gey et al., editors, *Procs. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 230–237. ACM Press, 1999.

[5] J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In W. Kellogg and S. Whittaker, editors, *Procs. of the ACM Conference on Computer Supported Cooperative Work*, pages 241–250. ACM Press, 2000.

[6] G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[7] D. McSherry. A lazy learning approach to explaining case-based reasoning solutions. In B. Díaz-Agudo and I. Watson, editors, *Procs. of the 20th International Conference on Case-Based Reasoning*, LNCS 7466, pages 241–254. Springer, 2012.

[8] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Procs. of the 10th International Conference on World Wide Web*, pages 285–295, 2001.

[9] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos. MoviExplain: A recommender system with explanations. In *Procs. of the Third ACM Conference on Recommender Systems*, pages 317–320, 2009.

[10] N. Tintarev. Explanations of recommendations. In *Procs. of the First ACM Conference on Recommender Systems*, pages 203–206, 2007.

[11] N. Tintarev and J. Masthoff. Designing and evaluating explanations for recommender systems. In F. Ricci et al., editors, *Recommender Systems Handbook*, pages 479–510. Springer, 2011.

[12] N. Tintarev and J. Masthoff. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, 22(4–5):399–439, 2012.

[13] J. Vig, S. Sen, and J. Riedl. Tagsplanations: Explaining recommendations using tags. In *Procs. of the 14th International Conference on Intelligent User Interfaces*, pages 47–56, 2009.

[14] M. Zanker. The influence of knowledgeable explanations on users' perception of a recommender system. In *Procs. of the Sixth ACM Conference on Recommender Systems*, pages 269–272, 2012.