Chapter 15
Designing and Evaluating Explanations for Recommender Systems

Nava Tintarev and Judith Masthoff

Abstract This chapter gives an overview of the area of explanations in recommender systems. We approach the literature from the angle of evaluation: that is, we are interested in what makes an explanation "good", and suggest guidelines as how to best evaluate this. We identify seven benefits that explanations may contribute to a recommender system, and relate them to criteria used in evaluations of explanations in existing systems, and how these relate to evaluations with live recommender systems. We also discuss how explanations can be affected by how recommendations are presented, and the role the interaction with the recommender system plays w.r.t. explanations. Finally, we describe a number of explanation styles, and how they may be related to the underlying algorithms. Examples of explanations in existing systems are mentioned throughout.

15.1 Introduction

In recent years, there has been an increased interest in more user-centered evaluation metrics for recommender systems such as those mentioned in [42]. It has also been recognized that many recommender systems functioned as black boxes, providing no transparency into the working of the recommendation process, nor offering any additional information to accompany the recommendations beyond the recommendations themselves [29].

Explanations can provide that transparency, exposing the reasoning and data behind a recommendation. This is the case with some of the explanations hosted on Amazon, such as: *"Customers Who Bought This Item Also Bought ... "*. Expla-

Nava Tintarev
University of Aberdeen, Aberdeen, U.K, e-mail: n.tintarev@abdn.ac.uk

Judith Masthoff
University of Aberdeen, Aberdeen, U.K. e-mail: j.masthoff@abdn.ac.uk

F. Ricci et al. (eds.), *Recommender Systems Handbook*,
DOI 10.1007/978-0-387-85820-3_15, © Springer Science+Business Media, LLC 2011

nations can also serve other aims such as helping to 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. In this way, we distinguish between different explanation such as e.g. explaining the way the recommendation engine works (transparency), and explaining why the user may or may not want to try an item (effectiveness). An effective explanation may be formulated along the lines of *"You might (not) like Item A because..."*. In contrast to the Amazon example above, this explanation does not *necessarily* describe how the recommendation was selected - in which case it is not transparent.

This chapter offers guidelines for designing and evaluating explanations in recommender systems as summarized in Section 15.2. Expert systems can be said to be the predecessors of recommender systems. In Section 15.3 we therefore briefly relate research on evaluating explanations in expert systems to evaluations of explanations in recommender systems. We also identify the developments in recommender systems which may have caused a revived interest in explanation research since the days of expert systems.

Up until now there has been little consensus as to how to evaluate explanations, or why to explain at all. In Section 15.4, we list seven explanatory criteria, and describe how these have been measured in previous systems. These criteria can also be understood as advantages that explanations may offer to recommender systems, answering the question of *why* to explain. In the examples for effective and transparent explanations above, we saw that the two evaluation criteria could be mutually exclusive.

In Section 15.5, we consider that the underlying recommender system affects the evaluation of explanations, and discuss this in terms of the evaluation metrics normally used for recommender systems (e.g. accuracy and coverage). We mention and illustrate examples of explanations throughout the chapter, and offer an aggregated list of examples in commercial and academic recommender systems in Table 15.6. We will see that explanations have been presented in various forms, using both text and graphics.

Additionally, explanations are not decoupled from recommendations themselves or the way in which users can interact with the recommender system: both factors influence each other and the explanations that can be generated, which in turn affects the degree to which explanatory goals are achieved. We discuss these types of design choices in Section 15.6 – in Section 15.6.1 we mention different ways of presenting recommendations, and Section 15.6.2 how users can interact and give input to a recommender system.

Moreover, the underlying algorithm of a recommender engine may influence the types of explanations that can be generated, although it is also possible that the explanations selected by the system developer do *not* reflect the underlying algorithm. This is particularly the case for computationally complex algorithms for which explanations may be more difficult to generate, such as collaborative filtering [29, 31]. In this case, the developer must consider the trade-offs between different explanatory goals such as satisfaction (as an extension of understandability) and transparency. In Section 15.7, we relate the most common explanation styles and how they may relate to the underlying algorithms. Finally, we conclude with a summary and future directions in Section 15.8.

15.2 Guidelines

The content of this chapter is divided into sections which each elaborate on the following design guidelines for explanations in recommender systems.

- Consider the benefit(s) you would like to obtain from the explanations, and the best metric to evaluate on the associated criteria (Section 15.4).
- Be aware that the evaluation of explanations is related to, and may be confounded with, the functioning of the underlying recommendation engine, as measured by criteria commonly used for evaluating recommender systems (Section 15.5).
- Think about how the way that you present the recommendations themselves, and the the interaction model, affect each other and the explanations (Section 15.6). These factors in turn affect the degree to which different explanatory goals can be achieved.
- Last, but certainly not least, consider the relationship between the underlying algorithm and the type of explanations you choose to generate (Section 15.7). Do the explanations that you generate help you achieve your explanatory goals?

15.3 Explanations in Expert Systems

Explanations in intelligent systems are not a new idea: explanations have often been considered as part of the research in the area of expert systems [8, 32, 38, 27, 66]. This research has largely been focused on what kind of explanations can be generated and how these have been implemented in real world systems [8, 32, 38, 66]. Overall, *there are few evaluations of the explanations in these systems*. When they did occur evaluations of explanations have largely focused on *user acceptance* of the system such as [15] or acceptance of the systems' conclusions [67]. An exception is an evaluation in MYCIN which considered the decision support of the system as a whole [27]. In contrast, the commercial intent behind recommender systems targeting a wide user base was previously unseen in expert systems, has extended the evaluation goals for explanations beyond acceptance.

Also, developments in recommender systems have revived explanation research, after a decline of studies in expert systems in the 90's. One such development is the increase in data: due to the growth of the Web, there are now more users using the average (recommender) system. Systems are also no longer developed in isolation of each other, making the best possible reuse of code (open source projects) and datasets such as the MovieLens [2] and Netflix dataset [3]. In addition, new algorithms, in particular in the domain of collaborative filtering, have been adapted and developed (see also Chapter 4 on neighborhood based approaches, and Chapter 5 on advances in collaborative filtering). These approaches mitigate domain dependence, and allow for greater generalizability, and are more suitable for large and often sparse datasets. One sign of the revived interest in explanation research is the success of a recent series of workshops on explanation aware computing (see e.g. [53, 54]).

For further reading, see the following reviews on expert systems with explanatory capabilities for three of the most common inference methods: heuristic-based methods [36], Bayesian networks [35], and case-based reasoning [22].

15.4 Defining Goals

Guideline 1: Consider the benefit(s) you would like to obtain from the explanations, and the best metric to evaluate on the associated criteria.

Surveying the literature on explanations in recommender systems, we see that recommender systems with explanatory capabilities have been evaluated according to different criteria, and identify seven different goals for explanations. Here we mention goals that are applicable to single item recommendations, i.e. when a single recommendation is being offered. When recommendations are made for multiple items, such as in a list, the criteria may be different and consider other factors such as diversity (e.g. are the items in the list sufficiently varied).

Table 15.1 states these goals, which are similar to those desired (but not evaluated on) in expert systems, c.f. MYCIN [10]. In Table 15.2, we summarize previous evaluations of explanations in recommender systems, and the criteria by which they have been evaluated. Works that have no clear criteria stated, or have not evaluated the system on the explanation criteria which they state, are omitted from this table.

For example, in Section 15.3 we mentioned that expert systems were commonly evaluated in terms of user acceptance and the decision support of the system as a whole. User acceptance can be defined in terms of our goals of satisfaction or persuasion. If the evaluation measures acceptance with the system as whole, such as [15] who asked questions such as *"Did you like the program?"*, then this reflects user satisfaction. If rather the evaluation measures user acceptance of advice or explanations, as in [67], the criterion can be said to be persuasion.

It is important to identify these goals as distinct, even if they may interact, or require certain trade-offs. Indeed, it would be hard to create explanations that do well on all criteria, in reality it is a trade-off. For instance, in our work we have found that while personalized explanations may lead to greater user satisfaction, they do not necessarily increase effectiveness [61]. Other times, goals that seem to be inherently related are not necessarily so, for example it has been found that transparency does not necessarily aid trust [20]. For these reasons, while an explanation in Table 15.2 may have been evaluated for several criteria, it may not have achieved them all.

The type of explanation that is given to a user is likely to depend on the criteria of the designer of a recommender system. For instance, when building a system that sells books one might decide that user trust is the most important aspect, as it leads to user loyalty and increases sales. For selecting tv-shows, user satisfaction could be more important than effectiveness. That is, it is more important that a user enjoys the service, than that they are presented the best available shows.

In addition, some attributes of explanations may contribute toward achieving multiple goals. For instance, one can measure how *understandable* an explanation is, which can contribute to e.g. user trust, as well as satisfaction.

In this section we describe seven criteria for explanations, and suggest evaluation metrics based on previous evaluations of explanation facilities, or offer suggestions of how existing measures could be adapted to evaluate the explanation facility in a recommender system.

Table 15.1: Explanatory criteria and their definitions

Aim	**Definition**
Transparency (Tra.)	Explain how the system works
Scrutability (Scr.)	Allow users to tell the system it is wrong
Trust	Increase users' confidence in the system
Effectiveness (Efk.)	Help users make good decisions
Persuasiveness (Pers.)	Convince users to try or buy
Efficiency (Efc.)	Help users make decisions faster
Satisfaction (Sat.)	Increase the ease of use or enjoyment

15.4.1 Explain How the System Works: Transparency

An anecdotal article in the Wall Street Journal titled "*If TiVo Thinks You Are Gay, Here's How to Set It Straight*" describes users' frustration with irrelevant choices made by a video recorder that records programs it assumes its owner will like, based on shows the viewer has recorded in the past[69]. For example, one user, Mr. Iwanyk, suspected that his TiVo thought he was gay since it inexplicably kept recording programs with gay themes. This user clearly deserved an explanation.

An explanation may clarify *how* a recommendation was chosen. In expert systems, such as in the domain of medical decision making, the importance of trans-

Table 15.2: The criteria by which explanations in recommender systems have been evaluated. System names are mentioned if given, otherwise we only note the type of recommended items. Works that have no clear criteria stated, or have not *evaluated* the system on the explanation criteria which they state, are omitted from this table. Note that while a system may have been evaluated for several criteria, it may not have achieved all of them. Also, for the sake of completeness we have distinguished between multiple studies using the same system.

System (type of items)	**Tra.**	**Scr.**	**Trust**	**Efk.**	**Per.**	**Efc.**	**Sat.**
(Internet providers) [23]			X		X		X
(Digital cameras, notebooks computers) [49]			X				
(Digital cameras, notebooks computers) [50]			X	X			
(Music) [55]			X				
(Movies) [61]				X	X		X
Adaptive Place Advisor (restaurants) [59]				X		X	
ACORN (movies) [65]							X
CHIP (cultural heritage artifacts) [19]	X		X	X			
CHIP (cultural heritage artifacts) [20]	X		X				X
iSuggest-Usability (music) [30]	X			X			
LIBRA (books) [11]				X			
MovieLens (movies) [29]					X		X
Moviexplain (movies) [58]				X			X
myCameraAdvisor [63]			X				
Qwikshop (digital cameras) [39]				X		X	
SASY (e.g. holidays) [21]	X	X					X
Tagsplanations (movies) [62]	X			X			

parency has also been recognized [10]. Transparency or the heuristic of "Visibility of System Status" is also an established usability principle [44], and its importance has also been highlighted in user studies of recommender systems [55].

Vig et al. differentiate between transparency and justification [62]. While transparency should give an honest account of how the recommendations are selected and how the system works, justification can be descriptive and decoupled from the recommendation algorithm. The authors cite several reasons for opting for justification rather than genuine transparency. For example some algorithms that are difficult to explain (e.g. latent semantic analysis where the distinguishing factors are latent and may not have a clear interpretation), protection of trade secrets by system designers, and the desire for greater freedom in designing the explanations.

Cramer et al. have investigated the effects of transparency on other evaluation criteria such as trust, persuasion (acceptance of items) and satisfaction (acceptance) in an art recommender [19, 20]. Transparency itself was evaluated in terms of its effect on actual and perceived understanding of how the system works [20]. While actual understanding was based on user answers to interview questions, perceived understanding was extracted from self-reports in questionnaires and interviews.

The evaluation of transparency has also been coupled with scrutability (Section 15.4.2) and trust (Section 15.4.3), but we will see in these sections that these criteria can be distinct from each other.

15.4.2 Allow Users to Tell the System it is Wrong: Scrutability

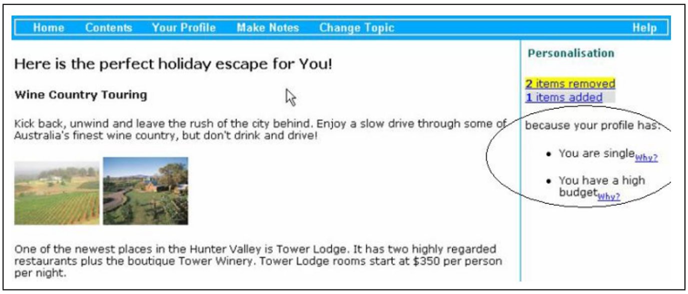
Explanations may help isolate and correct misguided assumptions or steps. When the system collects and interprets information in the background, as is the case with TiVo, it becomes all the more important to make the reasoning available to the user. Following transparency, a second step is to allow a user to correct reasoning, or make the system *scrutable* [21]. Explanations should be part of a cycle, where the user understands what is going on in the system and exerts control over the type of recommendations made, by correcting system assumptions where needed [56]. Scrutability is related to the established usability principle of User Control [44]. See Figure 15.1 for an example of a scrutable holiday recommender.

While scrutability is very closely tied to the criteria of transparency, it deserves to be uniquely identified. The explanations in Table 15.4 are scrutable, but not (fully) transparent even if they offer some form of justification. For example, there is nothing in this Table that suggests that the underlying recommendations are based on a Bayesian classifier. In such a case, we can imagine that a user attempts to scrutinize a recommender system, and manages to change their recommendations by modifying their ratings, but still does not understand exactly what happens within the system.

Czarkowski found that users were not likely to scrutinize on their own, and that extra effort was needed to make the scrutability tool more visible [21]. In addition, it was easier to get users perform a given scrutinization task such as changing the personalization (e.g. “Change the personalisation so that only Current Affairs programs are included in your 4:30-5:30 schedule.”) Their evaluation included metrics such as task correctness, and if users could express an understanding of what information was used to make recommendations for them. They understood that adaptation in the system was based on their personal attributes stored in their profile, that their profile contained information they volunteered about themselves, and that they could change their profile to control the personalization [21].

15.4.3 Increase Users’ Confidence in the System: Trust

Trust is sometimes linked with transparency: previous studies indicate that transparency and the possibility of interaction with recommender systems increases user trust [23, 55]. A user may also be more forgiving, and more confident in recommendations, if they understand why a bad recommendation has been made. Trust in the recommender system could also be dependent on the accuracy of the recommen-



Here is the perfect holiday escape for You!

Wine Country Touring

Kick back, unwind and leave the rush of the city behind. Enjoy a slow drive through some of Australia's finest wine country, but don't drink and drive!

One of the newest places in the Hunter Valley is Tower Lodge. It has two highly regarded restaurants plus the boutique Tower Winery. Tower Lodge rooms start at $350 per person per night.

because your profile has:

- You are single
- You have a high budget

Fig. 15.1: Scrutable holiday recommender [21]. The explanation is in the circled area, and the user profile can be accessed via the "why" links.

dation algorithm [41]. A study of users' trust (defined as perceived confidence in a recommender system's *competence*) suggests that users intend to return to recommender systems which they find trustworthy [16]. We note however, that there is a case where transparency and trust were not found to be related [20].

We do not claim that explanations can fully compensate for poor recommendations, but good explanations may help users make better decisions (see Section 15.4.5 on effectiveness). A user may also appreciate when a system is "frank" and admits that it is not confident about a particular recommendation.

In addition, the interface design of a recommender system may affect its credibility. In a study of factors determining web page credibility, the largest proportion of users' comments (46.1%) referred to the appeal of the overall visual design of a site, including layout, typography, font size and color schemes [25]. Likewise the perceived credibility of a Web article was significantly affected by the presence of a photograph of the author [24]. So, while recommendation accuracy, and the criteria of transparency are often linked to the evaluation of trust, design is also a factor that needs to be considered as part of the evaluation.

Questionnaires can be used to determine the degree of trust a user places in a system. An overview of trust questionnaires can be found in [45] which also suggests and validates a five dimensional scale of trust. Note that this validation was done with the aim of using celebrities to endorse products, but was not conducted for a particular domain. Additional validation may be required to adapt this scale to a particular recommendation domain.

A model of trust in recommender systems is proposed in [16, 50], and the questionnaires in these studies consider factors such as intent to return to the system, and intent to save effort. Also [63] query users about trust, but focus on trust related beliefs such as the perceived competence, benevolence and integrity of a virtual adviser. Although questionnaires can be very focused, they suffer from the fact that self-reports may not be consistent with user behavior. In these cases, implicit measures (although less focused) may reveal factors that explicit measures do not.

One such implicit measure could be loyalty, a desirable bi-product of trust. One study compared different interfaces for eliciting user preferences in terms of how they affected factors such as loyalty [41]. Loyalty was measured in terms of the number of logins and interactions with the system. Among other things, the study found that allowing users to independently choose which items to rate affected user loyalty. It has also been thought that Amazon's conservative use of recommendations, mainly recommending familiar items, enhances user trust and has led to increased sales [57]. We encourage readers who would like to learn more about trust in recommender systems to read Chapter 20 which is dedicated to this topic.

15.4.4 Convince Users to Try or Buy: Persuasiveness

Explanations may increase user acceptance of the system or the given recommendations [29]. Both definitions qualify as persuasion, as they are attempts to gain benefit for the system rather than for the user.

[20] evaluated the acceptance of recommended items in terms of how many recommended items were present in a final selection of six favorites. In a study of a collaborative filtering- and rating-based recommender system for movies, participants were given different explanation interfaces (e.g. Figure 15.2)[29]. This study directly inquired how likely users were to see a movie (with identifying features such as title omitted) for 21 different explanation interfaces. Persuasion was thus a numerical rating on a 7-point Likert scale.

In addition, it is possible to measure if the evaluation of an item has changed, i.e. if the user rates an item differently after receiving an explanation. Indeed, it has been shown that users can be manipulated to give a rating closer to the system's prediction [18]. This study was in the low investment domain of movie rental, and it is possible that users may be less influenced by incorrect predictions in high(er) cost domains such as cameras[1]. It is also important to consider that too much persuasion may backfire once users realize that they have tried or bought items that they do not really want.

Persuasiveness can be measured in a number of ways. For example, it can be measured as the difference between two ratings: the first being a previous rating, and the second a re-rating for the same item but with an explanation interface [18]. Another possibility would be to measure how much users actually try or buy items compared to users in a system without an explanation facility. These metrics can also be understood in terms of the concept of "conversion rate" commonly used in e-Commerce, operationally defined as the percentage of visitors who take a desired action. For a more in-depth discussion of persuasion in recommender systems the reader may consult Chapter 14.

[1] In [60] participants reported that they found incorrect overestimation less useful in high cost domains compared to low cost domains.

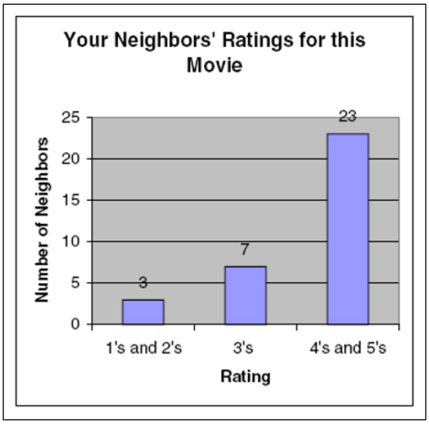


Fig. 15.2: One out of twenty-one interfaces evaluated for persuasiveness - a histogram summarizing the ratings of similar users (neighbors) for the recommended item grouped by good (5's and 4's), neutral (3's), and bad (2's and 1's), on a scale from 1 to 5 [29].

15.4.5 Help Users Make Good Decisions: Effectiveness

Rather than simply persuading users to try or buy an item, an explanation may also assist users to make *better* decisions. Effectiveness is by definition highly dependent on the accuracy of the recommendation algorithm. An effective explanation would help the user evaluate the quality of suggested items according to their own preferences. This would increase the likelihood that the user discards irrelevant options while helping them to recognize useful ones. For example, a book recommender system with effective explanations would help a user to buy books they actually end up liking. Bilgic and Mooney emphasize the importance of measuring the ability of a system to assist the user in making accurate decisions about recommendations based on explanations such as those in Figure 15.3 and Tables 15.3, 15.4 and 15.5 [11]. Effective explanations could also serve the purpose of introducing a new domain, or the range of products, to a novice user, thereby helping them to understand the full range of options [23, 49].

Vig et al. measure perceived effectiveness: *"This explanation helps me determine how well I will like this movie."* [62]. Effectiveness of explanations can also be calculated as the *absence of a difference* between the liking of the recommended item prior to, and after, consumption. For example, in a previous study, users rated a book twice, once after receiving an explanation, and a second time after reading the book [11]. If their opinion on the book did not change much, the system was considered effective. This study explored the effect of the whole recommendation process, explanation inclusive, on effectiveness. The same metric was also used to evaluate whether personalization of explanations (in isolation of a recommender system) increased their effectiveness in the movie domain [61].

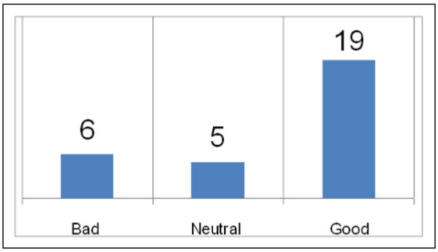


Fig. 15.3: The Neighbor Style Explanation - a histogram summarizing the ratings of similar users (neighbors) for the recommended item grouped by good (5's and 4's), neutral (3's), and bad (2's and 1's), on a scale from 1 to 5. The similarity to Figure 15.2 in this study was intentional, and was used to highlight the difference between persuasive and effective explanations [11].

Table 15.3: The keyword style explanation by [11]. This recommendation is explained in terms of keywords that were used in the description of the item, and that have previously been associated with highly rated items. "Count" identifies the number of times the keyword occurs in the item's description, and "strength" identifies how influential this keyword is for predicting liking of an item.

Word	**Count**	**Strength**	**Explain**
HEART	2	96.14	*Explain*
BEAUTIFUL	1	17.07	*Explain*
MOTHER	3	11.55	*Explain*
READ	14	10.63	*Explain*
STORY	16	9.12	*Explain*

While this metric considers the difference between the before and after ratings, it does not discuss the effects of over- contra underestimation. If a user's evaluation of an item decreases after exposure to an item, their initial rating was an overestimation. Likewise, if their evaluation increases after exposure to the item, the initial rating was an underestimation. In our work we found that users considered overestimation to be less effective than underestimation, and that this varied between domains. Specifically, overestimation was considered more severely in high investment domains compared to low investment domains. In addition, the strength of the effect on perceived effectiveness varied depending on where on the scale the prediction error occurred [60].

Another way of measuring the effectiveness of explanations has been to test the same system with and without an explanation facility, and evaluate if subjects who receive explanations end up with items more suited to their personal tastes [19].

Other work evaluated explanation effectiveness using a metric from marketing [28], with the aim of finding the single *best* possible item (rather than "good enough items" as above) [17]. Participants interacted with the system until they found the

Table 15.4: A more detailed explanation for the "strength" of a keyword which shows after clicking on *"Explain"* in Table 15.3. In practice "strength" probabilistically measures how much more likely a keyword is to appear in a positively rated item than a negatively rated one. It is based on the user's previous positive ratings of items ("rating"), and the number of times the keyword occurs in the description of these items ("count") [11].

Title	**Author**	**Rating**	**Count**
Hunchback of Notre Dame	Victor Hugo, Walter J. Cobb	10	11
Till We Have Faces: A Myth Retold	C.S. Lewis, Fritz Eichenberg	10	10
The Picture of Dorian Gray	Oscar Wilde, Isobel Murray	8	5

item they would buy. They were then given the opportunity to survey the entire catalog and to change their choice of item. Effectiveness was then measured by the fraction of participants who found a better item when comparing with the complete selection of alternatives in the database. So, using this metric, a low fraction represents high effectiveness.

Effectiveness is the criterion that is most closely related to accuracy measures such as precision and recall [19, 58, 59]. In systems where items are easily consumed, e.g. internet news, these can be translated into recognizing relevant items and discarding irrelevant options respectively. For example, there have been suggestions for an alternative metric of "precision" based on the number of profile concepts matching with user interests, divided by the number of concepts in their profile [19].

15.4.6 Help Users Make Decisions Faster: Efficiency

Explanations may make it *faster* for users to decide which recommended item is best for them. Efficiency is another established usability principle, i.e. how quickly a task can be performed [44]. This criterion is one of the most commonly addressed in the recommender systems literature given that the task of recommender systems is to find needles in haystacks of information.

Efficiency may be improved by allowing the user to understand the relation between competing options. [39, 43, 49] use so called critiquing, a sub-class of knowledge-based algorithms based on trade-offs between item properties, which lends itself well to the generation of explanations. In the domain of digital cameras, competing options may for example be viewed by selecting *"Less Memory and Lower Resolution and Cheaper"* [39]. This way users are quickly able to use this query revision to find a cheaper camera if they are willing to settle for less memory and lower resolution. More details on critiquing-based recommender systems can also be found in Chapter 13 of this handbook.

Efficiency is often used in the evaluation of so-called conversational recommender systems, where users continually interact with a recommender system, refining their preferences (see also Section 15.6.2). In these systems, the explanations can be seen to be implicit in the dialog. Efficiency in these systems can be measured by the total amount of interaction time, and number of interactions needed to find a satisfactory item [59]. Evaluations of explanations based on improvements in efficiency are not limited to conversational systems however. Pu and Chen for example, compared completion time for two explanatory interfaces, and measured completion time as the amount of time it took a participant to locate a desired product in the interface [49].

Other metrics for efficiency also include the number of inspected explanations, and number of activations of repair actions when no satisfactory items are found [23, 52]. Normally, it is not sensible to expose users to all possible recommendations and their explanations, and so users can choose to inspect (or scrutinize) a given recommendation by asking for an explanation. In a more efficient system, the users would need to inspect *fewer* explanations. Repair actions consist of feedback from the user which changes the type of recommendation they receive, as outlined in the sections on scrutability (Section 15.4.2). Examples of user feedback/repair actions can be found in Section 15.6.2.

15.4.7 Make the use of the system enjoyable: Satisfaction

Explanations have been found to increase user satisfaction with, or acceptance of, the overall recommender system [23, 29, 55]. The presence of longer descriptions of individual items has been found to be positively correlated with both the *perceived* usefulness [60], and ease of use of the recommender system [55]. Also, many commercial recommender systems such as those seen in Table 15.6 are primarily sources of entertainment. In these cases, any extra facility should take notice of the effect on user satisfaction. Figure 15.4 gives an example of an explanation evaluated on the criterion of satisfaction.

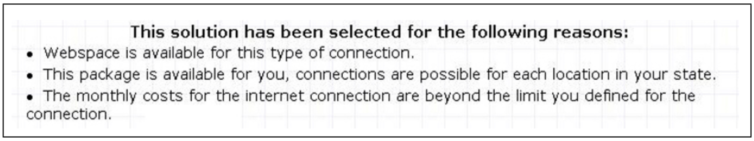
When measuring satisfaction, one can directly ask users whether the system is enjoyable to use [15], or if users like the explanations themselves [60]. Satisfaction can also be measured indirectly by measuring user loyalty [41, 23] (see also Section 15.4.3), and likelihood of using the system for a search task [20].

In measuring explanation satisfaction, it is important to differentiate between satisfaction with the recommendation process[2], and the recommended products (persuasion) [20, 23]. One (qualitative) way to measure satisfaction with the process would be to conduct usability testing methods such as record a think-aloud protocol for a user conducting a task [37].

[2] Here we mean the entire recommendation process, inclusive of the explanations. However, in Section 15.5 we highlight that evaluation of explanations in recommender systems are seldom fully independent of the underlying recommendation process.

In this case, the participants describe their entire experience using the system: what they are looking at, thinking, doing and feeling, as they go about a task such as finding a satisfactory item. Objective notes of everything that users say are taken, without interpretation or influencing the users in any way. Video and voice recordings can also be used to revisit the session and to serve as a memory aid. In such a case, it is possible to identify usability issues and even apply quantitative metrics such as the ratio of positive to negative comments; the number of times the evaluator was frustrated; the number of times the evaluator was delighted; the number of times and where the evaluator worked around a usability problem etc.

It is also arguable that users would be satisfied with a system that offers effective explanations, confounding the two criteria. However, a system that aids users in making good decisions, may have other disadvantages that decrease the overall satisfaction (e.g. requiring a large cognitive effort on the part of the user). Fortunately, these two criteria can be measured by distinct metrics.



This solution has been selected for the following reasons:

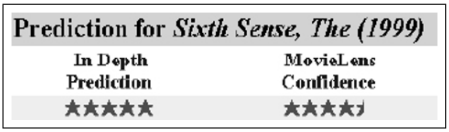
- Webspace is available for this type of connection.
- This package is available for you, connections are possible for each location in your state.
- The monthly costs for the internet connection are beyond the limit you defined for the connection.

Fig. 15.4: An explanations for an internet provider, describing the provider in terms of user requirements: “This solution has been selected for the following reasons . . . ” [23].

15.5 Evaluating the Impact of Explanations on the Recommender System

Guideline 2: Be aware that the evaluation of explanations is related to, and may be confounded with, the functioning of the underlying recommendation engine, as measured by criteria commonly used for evaluating recommender systems.

We have now identified seven criteria by which explanations in recommender systems can be evaluated, and given suggestions of how such evaluations can be performed. To some extent, these criteria assume that we are evaluating only the explanation component. It also seems reasonable to evaluate the system as a *whole*. In that case we might measure the general system usability and accuracy, which will depend on both the recommendation algorithm as well as the impact of the explana-



Prediction for *Sixth Sense, The (1999)*

In Depth Prediction	MovieLens Confidence
★★★★★	★★★★½

Fig. 15.5: Confidence display for a recommendation, [29] - the movie is strongly recommended (5/5), and there is a large amount of information to support the recommendation (4.5/5).

tion component. Therefore, in this section, we describe the interaction between the recommender engine and our explanation criteria, organized by the evaluation metrics commonly used in recommender system evaluations: accuracy, learning rate, coverage, novelty/serendipity and acceptance.

15.5.1 Accuracy Metrics

Accuracy metrics regard the ability of the recommendation engine to predict correctly, but accuracy is likely to interact with explanations too. For example, with respect to the relationship between transparency and accuracy: Cramer et al. found that transparency led to changes in user behavior that ultimately decreased recommendation accuracy [19].

The system's own confidence in its recommendations is also related to accuracy and can be reflected in explanations. An example of an explanation aimed to help users understand (lack of) accuracy, can be found in confidence displays such as Figure 15.5. These can be used to explain e.g. poor recommendations in terms of insufficient information used for forming the recommendation. For further work on confidence displays see also [40].

Explanations can also help users understand how they would relate to a particular item, possibly supplying additional information that helps the user make more informed decisions (effectiveness). In the case of poor accuracy, the risk of missing good items, or trying bad ones increases while explanations can help decrease this risk. By helping users to correctly identify items as good or bad, the accuracy of the recommender system as a whole may also increase.

15.5.2 Learning Rate

The learning rate represents how quickly a recommender system learns a user's preferences, and how sensitive it is to changes in preferences. Learning rate is likely to affect user satisfaction as users would like a recommender system to quickly learn their preferences, and be sensitive to short term as well as long term interests. Explanations can increase satisfaction by clarifying or hinting that the system considers changes in the user's preferences. For example, the system can flag that the value for a given variable is getting close to its threshold for incurring a change, but that it has not reached it yet. A system can also go a step further, and allow the user to see just how it is learning and changing preferences (transparency), or make it possible for a user to delete old preferences (scrutability). For example, the explanation facility can request information that would help it learn/change quicker, such as asking if a user's favorite movie genre has changed from action to comedy.

15.5.3 Coverage

Coverage regards the range of items which the recommender system is able to recommend. Explanations can help users understand where they are in the search space. By directing the user to rate informative items in under-explored parts of the search space, explanations may increase the overlap between certain items or features (compared to sparsity). Ultimately, this may increase the overall coverage for potential recommendations. Understanding the remaining search options is related to the criterion of transparency: a recommender system can explain why certain items are not recommended. It may be impossible or difficult to retrieve an item (e.g. for items that have a very particular set of properties in a knowledge-based system, or the item does not have many ratings in a collaborative-filtering system). Alternatively, the recommender system may function under the assumption that the user is not interested in the item (e.g. if their requirements are too narrow in a knowledge-based system, or if they belong to a very small niche in a collaborative-based system). An explanation can explain why an item is not available for recommendation, and even how to remedy this and allow the user to change their preferences (scrutability).

Coverage may also affect evaluations of the explanatory criteria of effectiveness. For example, if a user's task is not only to find a "good enough" item, but the best item for them, then the coverage needs to be sufficient to ensure that "best" items are included in the recommendations. Depending on how much time retrieving these items takes, coverage may also affect efficiency.

15.5.4 Acceptance

It is possible to confound acceptance, or satisfaction with a system with other types of satisfaction. If users are satisfied with a system with an explanation component, it remains unclear whether this is due to: satisfaction with the *explanation component*, satisfaction with *recommendations*, or general design and visual appeal. Satisfaction with the system due to the recommendations is connected to accuracy metrics, or even novelty and diversity, in the sense that sufficiently good recommendations need to be given to a user in order to keep them satisfied. Although explanations may help increase satisfaction, or tolerance toward the system, they cannot function as a substitute for e.g. good accuracy. Indeed, this is true for all the mentioned explanatory criteria. An example of an explanation striving toward the criterion of satisfaction may be: *"Please bare with me, I still need to learn more about your preferences before I can make an accurate recommendation."*

15.6 Designing the Presentation and Interaction with Recommendations

Guideline 3: Think about how the way that you present the recommendations themselves, and the the interaction model, affect each other and the explanations. These factors affect the degree to which explanatory goals are achieved.

The way recommendations are presented are likely to affect the interaction model that can be used for eliciting users preferences. Likewise, both factors can affect the types of explanations that can be generated. In turn, some of the explanations that can be generated may be more suitable for particular explanatory criteria. Chapter 16 of this handbook, also discusses a complementary evaluation framework for preference-based (such as critiquing which is described in Chapter 13) recommender systems and focuses on the design of both presentation of recommendations and interaction model. For example one guideline states: *"Showing one search result or recommending one item at a time allows for a simple display strategy which can be easily adapted to small display devices; however, it is likely to engage users in longer interaction sessions or only allow them to achieve relatively low decision accuracy."* (Guideline 9).

15.6.1 Presenting Recommendations

We summarize the ways of presenting recommendations that we have seen for the systems summarized in this paper. While there are a number of possibilities for the *appearance* of the graphical user interface, the actual *structure* of offering recommendations can also vary. We identify the following categories for structuring the presentation of recommendations:

- **Top item.** Perhaps the simplest way to present a recommendation is by offering the user the best item for them. E.g. *"You have been watching a lot of sports, and football in particular. This is the most popular and recent item from the world cup."*

- **Top N-items.** The system may also present several items at once.*"You have watched a lot of football and technology items. You might like to see the local football results and the gadget of the day."* Note that while this system could be able to explain the relation between chosen items, it could also explain the rational behind each single item.
- **Similar to top item(s).** Once a user shows a preference for one or more items, the recommender system can offer *similar* items. E.g. *"You might also like...Oliver Twist by Charles Dickens"*.
- **Predicted ratings for all items.** Rather than forcing selections on the user, a system may allow its users to browse all the available options. Recommendations are then presented as predicted ratings on a scale (say from 0 to 5) for each item. A user might query why a certain item, for example local hockey results, is predicted to have a low rating. The recommender system might then generate an explanation like: *"While this is a sports it is about hockey, which you do not seem to like!"*.
- **Structured overview.** The recommender system can give a structure which displays trade-offs between items [49, 68]. The advantage of a structured overview is that the user can see how items compare, and what other items are still available if the current recommendation should not meet their requirements.

15.6.2 Interacting with the Recommender System

There are different ways in which a user can give input to the recommender system. This interaction is what distinguishes conversational systems from "single-shot" recommendations. They allow users to elaborate their requirements over the course of an extended dialog [51] rather than each user interaction being treated independently of previous history.

We expand on the four ways suggested by [26], supplying examples of current applications[3]. Note that although there are more unobtrusive ways to elicit user preferences, e.g. via usage data [46] or demographics [6], this section focuses on *explicit* feedback from users.

- **The user specifies their requirements.** The user can specify their requirements through a dialog about their preferences in plain English [43, 64]. Such a dialog does not make use of the user's previous interests, nor does it explain *directly*. That is, there is no sentence that claims to be a justification of the recommendation. It does however do so indirectly, by reiterating (and satisfying) the user's *requirements*.
- **The user asks for an alternation.** A more direct approach is to allow users to explicitly critique recommended items (see also Chapter 13 on the evolution of critiquing), for instance using a structured overview (see Section 15.6.1).

[3] A fifth section on mixed interaction interfaces is appended to the end of this original list.

One such system explains the difference between a selected item and remaining items [39].

- **The user rates items.** To change the type of recommendations they receive, the user may want to correct predicted ratings, or modify a rating they made in the past. The *influence based explanation* in Table 15.5 shows which rated titles influenced the recommended book the most [11].
- **The user gives their opinion.** A common usability principle is that it is easier for humans to recognize items, than to draw them from memory. For example, a user could specify whether they think an item is interesting or not, if they would like to see more similar items, or if they have already seen the item previously [12, 57].
- **Mixed interaction interfaces.** Recommender systems can also combine different types of interactions [17, 41].

Table 15.5: The *influence based explanation* showed which rated titles influenced the recommended book the most. Although this particular system did not allow the user to modify previous ratings, or degree of influence, in the explanation interface, it can be imagined that users could directly change their rating here. Note however, that it would be much harder to modify the degree of influence, as it is computed: any modification is likely to interfere with the regular functioning of the recommendation algorithm [11].

BOOK	**YOUR RATING Out of 5**	**INFLUENCE Out of 100**
Of Mice and Men	4	54
1984	4	50
Till We Have Faces: A Myth Retold	5	50
Crime and Punishment	4	46
The Gambler	5	11

15.7 Explanation Styles

Guideline 4: Consider the relationship between the underlying algorithm and the type of explanations you choose to generate. Do the explanations that you generate help you achieve your explanatory goals?

In this section we describe explanations inspired by a particular underlying algorithm, or different "explanation styles". We caution that explanations may follow the "style" of a particular algorithm irrespective of whether or not this is how the recommendations have been retrieved or computed. In other words, the explanation style for a given explanation *may, or may not,* reflect the underlying algorithm by which the recommendations are computed. There often is a divergence between how the recommendations are retrieved and the style of the given explanations. Consequently, this type of explanation would not be consistent with the goal of transparency, but may support other explanatory goals.

Table 15.6: Examples of explanations in commercial and academic systems, ordered by explanation style (case, collaborative, content, conversational, demographic and knowledge/utility-based).

System	Example explanation	Explanation style
iSuggest-Usability [30]	See e.g. Figure 15.8	Case-based
LoveFilm.com	*"Because you have selected or highly rated: Movie A"*	Case-based
LibraryThing.com	"Recommended By User X for Book A"	Case-based
Netflix.com	A list of similar movies the user has rated highly in the past	Case-based
Amazon.com	*"Customers Who Bought This Item Also Bought . . . "*	Collaborative
LIBRA [11]	Keyword style (Tables 15.3 and 15.4); Neighbor style (Figure 15.3); Influence style (Figure 15.5)	Collaborative
MovieLens [29]	Histogram of neighbors (Figure 15.2) and Confidence display (Figure 15.5)	Collaborative
Amazon.com	*"Recommended because you said you owned Book A"*	Content-based
CHIP [20]	*"Why is 'The Tailor's Workshop recommended to you'? Because it has the following themes in common with artworks that you like: * Everyday Life * Clothes . . . "*	Content-based
Moviexplain [58]	See Table 15.7	Content-based
MovieLens: *"Tagsplanations"* [62]	Tags ordered by relevance or preference (see Figure 15.7)	Content-based
News Dude [12]	*"This story received a [high/low] relevance score, because it contains the words f1, f2, and f3."*	Content-based

Continued on next page

Table 15.6 – continued from previous page

System	Example explanation	Explanation style
OkCupid.com	Graphs comparing two users according to dimensions such as “more introverted”; comparison of how users have answered different questions	Content-based
Pandora.com	*“Based on what you’ve told us so far, we’re playing this track because it features a leisurely tempo . . . ”*	Content-based
Adaptive place Advisor [59]	Dialog e.g. “Where would you like to eat?” “Oh, maybe a cheap Indian place.”	Conversational
ACORN [65]	Dialog e.g. “What kind of movie do you feel like?” “I feel like watching a thriller.”	Conversational
INTRIGUE [6]	*“For children it is much eye-catching, it requires low background knowledge, it requires a few seriousness and the visit is quite short. For yourself it is much eye-catching and it has high historical value. For impaired it is much eye-catching and it has high historical value.”*	Demographic
Qwikshop [39]	*“Less Memory and Lower Resolution and Cheaper”*	Knowledge/utility-based
SASY [21]	*“. . . because your profile has: *You are single; *You have a high budget”* (Figure 15.1)	Knowledge/utility-based
Top Case [43]	“Case 574 differs from your query only in price and is the best case no matter what transport, duration, or accommodation you prefer”	Knowledge/utility-based
(Internet Provider) [23]	*“This solution has been selected for the following reasons: *Webspace is available for this type of connection . . . ”* (Figure 15.4)	Knowledge/utility-based
”Organizational Structure” [49]	Structured overview: *“We also recommend the following products because: *they are cheaper and lighter, but have lower processor speed.”* (Figure **??**)	Knowledge/utility-based

Continued on next page

Table 15.6 – continued from previous page

System	Example explanation	Explanation style
myCameraAdvisor [63]	e.g "…cameras capable of taking pictures from very far away will be more expensive …"	Knowledge/utility-based

Transparency is not the only explanatory goal to consider when deciding upon explanation style. For example, for a given system one might find that users are more satisfied with content-based style explanations even though critique-based style explanations are more efficient. As of yet, there is little comparison between explanation styles with regard to their performance on explanatory goals. Only Hingston [30] has compared the understandability and scrutability of different explanation styles inspired by algorithm, although in these cases, the explanations were directly influenced by different underlying algorithms as well. Other studies have however considered the effects of different explanation interfaces on different explanatory goals [20, 29, 61].

Notwithstanding, the underlying algorithm of a recommender engine will to a certain degree influence the types of explanations that can be generated. Table 15.6 summarizes the most commonly used explanation styles (case-based, content-based, collaborative-based, demographic-based, knowledge and utility-based) with examples of each. In this section we describe each style: their *assumed* inputs, processes and generated explanations. For commercial systems where this information is not public, we offer educated guesses. While conversational systems are included in the Table, we consider conversational systems as more of an interaction style than a specific algorithm.

In the following sections we will give further examples of how explanation styles can be inspired by common algorithms as classified by Burke [13]. For each example we also mention how the recommendations are presented, and the interaction model that was chosen.

For describing the interface between the recommender system and explanation component we use the notation used in [13]: **U** is the set of users whose preferences are known, and $\mathbf{u} \in U$ is the user for whom recommendations need to be generated. **I** is the set of items that can be recommended, and $\mathbf{i} \in \mathbf{I}$ is an item for which we would like to predict u's preferences.

15.7.1 Collaborative-Based Style Explanations

For collaborative-based style explanations the assumed input to the recommender engine are user **u**'s ratings of items in **I**. These ratings are used to identify users that are similar in ratings to **u**. These similar users are often called "neighbors" as nearest-neighbors approaches are commonly used to compute similarity. Then, a prediction for the recommended item is extrapolated from the neighbors' ratings of **i**.

Commercially, the most well known usage of collaborative-style explanations are the ones used by Amazon.com: *"Customers Who Bought This Item Also Bought ... "*. This explanation assumes that the user is viewing an item which they are already interested in. It implies that the system finds similar users (who bought this item), and retrieves and recommends items that these similar users bought. The recommendations are presented in the format of similar to top item. In addition, this explanation assumes an interaction model, whereby ratings are implicitly inferred through purchase behavior.

Herlocker et al. suggested 21 explanation interfaces using text as well as graphics [29]. These interfaces varied with regard to content and style, but a number of these explanations directly referred to the concept of neighbors. Figure 15.2 for example, shows how neighbors rated a given (recommended) movie, a bar chart with "good", "ok" and "bad" ratings clustered into distinct columns. Again, we see that this explanation is given for a specific way of recommending items, and a particular interaction model: this is a single recommendation (either top item or one item out of a top-N list), and assumes that the users are supplying rating information for items.

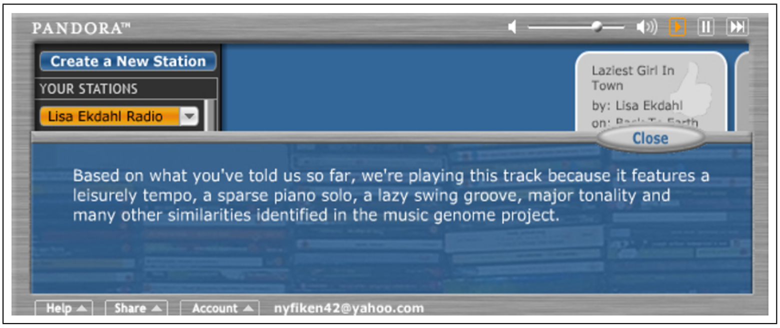
15.7.2 Content-Based Style Explanation

For content-based style explanations the assumed input to the recommender engine are user **u**'s ratings (for a sub-set) of items in **I**. These ratings are then used to generate a classifier that fits **u**'s rating behavior and use it on **i**. A prediction for the recommended item is based on how well it fits into this classifier. E.g. if it is similar to other highly rated items.

If we simplify this further, we could say that content-based algorithms consider similarity between items, based on user ratings but considering item properties. In the same spirit, content-based style explanations are based on the items' properties. For example, [58] justifies a movie recommendation according to what they infer is the user's favorite actor (see Table 15.7). While the underlying approach is in fact a hybrid of collaborative and content-based approaches, the explanation style suggests that they compute the similarity between movies according to the presence of features in highly rated movies. They elected to present users with several recommendations and explanations (top-N) which may be more suitable if the user would like to make a selection between movies depending on the information given in the explanations (e.g. feeling more like watching a movie with Harrison Ford over one starring Bruce Willis). The interaction model is based on ratings of items.

A more domain independent approach is suggested by [62] who suggest a similarity measure based on user specified keywords, or tags. The explanations used in this study use the relationship between keywords and items (tag relevance), and the relationship between tags and users (tag preference) to make recommendations (see Figure 15.7). Tag preference, or how relevant a tag is for a given user, can be seen as a form of content-based explanation, as it is a weighted average of a given user's ratings of movies with that tag. Tag relevance, or how relevant a keyword is for recommending an item, on the other hand is the correlation between (aggregate) users' preference for the tag, and their preference for a movie with which the tag is associated. In this example, showing recommendations as a single top item allows the user to view many of the tags that are related to the item. The interaction model is again based on numerical ratings.

The commercial system Pandora, explains its recommendations of songs according to musical properties such as tempo and tonality. These features are inferred from users ratings of songs. Figure 15.6 shows an example of this [1]. Here, the user is offered one song at a time (top item) and gives their opinion as "thumbs-up" or "thumbs-down" which also can be considered as numerical ratings.

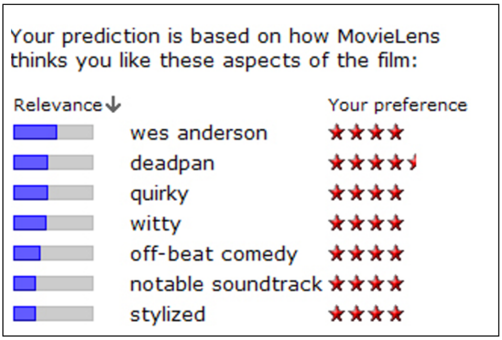


Based on what you've told us so far, we're playing this track because it features a leisurely tempo, a sparse piano solo, a lazy swing groove, major tonality and many other similarities identified in the music genome project.

Fig. 15.6: Pandora explanation: *"Based on what you've told us so far, we're playing this track because it features a leisurely tempo . . . "*

Table 15.7: Example of an explanation in Moviexplain, using features such as actors, which occur for movies previously rated highly by this user, to justify a recommendation [58].

Recommended movie title	**The reason is the participant**	**who appears in**
Indiana Jones and the Last Crusade (1989)	Ford, Harrison	5 movies you have rated
Die Hard 2 (1990)	Willis, Bruce	2 movies you have rated



Your prediction is based on how MovieLens thinks you like these aspects of the film:

Relevance ↓		Your preference
▬	wes anderson	★★★★
▬	deadpan	★★★★½
▬	quirky	★★★★
▬	witty	★★★★
▬	off-beat comedy	★★★★
▬	notable soundtrack	★★★★
▬	stylized	★★★★

Fig. 15.7: Tagsplanation with both tag preference and relevance, but sorted by tag relevance

15.7.3 Case-Based Reasoning (CBR) Style Explanations

Explanations can also omit mention of significant properties and focus primarily on the similar items used to make the recommendation. The items used are thus considered cases for comparison, resulting in case-based style explanations. We note that CBR systems greatly vary with regard to the recommendation algorithm. For example, the FINDME recommender [14] is based on critiquing, and the ranking of items in [5] is based on their presence in travel plans of users who expressed similar interests. While these CBR systems have also used different methods to present their explanations, we recall that this section, and the sections describing the other explanation styles, are focused on the *style* of the explanation rather than the actual underlying algorithm. As such, each of these systems could in theory have had a case-based style explanation.

In fact, in this chapter we have already seen a type of case-based style explanation, the "influence based style explanation" of [11] in Figure 15.5. Here, the influence of an item on the recommendation is computed by looking at the difference in the score of the recommendation with and without that item. In this case, recommendations were presented as top item, assuming a rating based interaction. Another study computed the similarity between recommended items[4], and used these similar items as justification for a top item recommendation in the "learn by example" explanations (see Figure 15.8) [30].

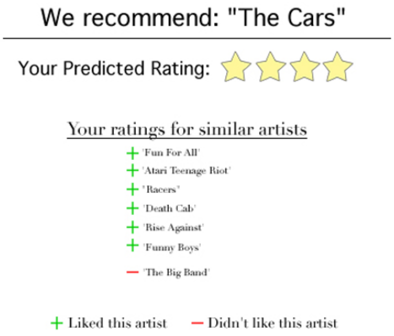
[4] The author does not specify which similarity metric was used, though it is likely to be a form of rating based similarity measure such as cosine similarity.

15.7.4 Knowledge and Utility-Based Style Explanations

For knowledge and utility-based style explanations the assumed input to the recommender engine are description of user **u**'s needs or interests. The recommender engine then infers a match between the item **i** and **u**'s needs. One knowledge-based recommender system takes into consideration how camera properties such as memory, resolution and price reflect the available options as well as a user's preferences [39]. Their system may explain a camera recommendation in the following manner: *"Less Memory and Lower Resolution and Cheaper"*. Here recommendations are presented as a form of structured overview describing the competing options, and the interaction model assumes that users ask for alterations in the recommended items.

Similarly, in the system described in [43] users gradually specify (and modify) their preferences until a top recommendation is reached. This system can generate explanations such as the following for a recommended holiday titled "Case 574": *"Top Case: Case 574 differs from your query only in price and is the best case no matter what transport, duration, or accommodation you prefer"*.

It is arguable that there is a certain degree of overlap between knowledge-based, content-based style (Section 15.7.2) and case-based style explanations (Section 15.7.3) which can be derived from either type of algorithm depending on the details of the implementation.


We recommend: "The Cars"

Your Predicted Rating: ★★★★

Your ratings for similar artists

+ 'Fun For All'
+ 'Atari Teenage Riot'
+ 'Racers'
+ 'Death Cab'
+ 'Rise Against'
+ 'Funny Boys'
— 'The Big Band'

+ Liked this artist — Didn't like this artist

Fig. 15.8: Learn by example, or case based reasoning [30].

15.7.5 Demographic Style Explanations

For demographic-based style explanations, the assumed input to the recommender engine is demographic information about user **u**. From this, the recommendation algorithm identifies users that are demographically similar to **u**. A prediction for the recommended item **i** is extrapolated from how the similar users rated this item, and how similar they are to **u**.

Surveying a number of systems which use a demographic-based filter e.g. [6, 34, 48], we could only find one which offers an explanation facility: *"For children it is much eye-catching, it requires low background knowledge, it requires a few seriousness and the visit is quite short. For yourself it is much eye-catching and it has high historical value. For impaired it is much eye-catching and it has high historical value."*[6]. In this system recommendations were offered as a structured overview, categorizing places to visit according to their suitability to different types of travelers (e.g. children, impaired). Users can then add these items to their itinerary, but there is no interaction model that modifies subsequent recommendations

To our knowledge, there are no other systems that make use of demographic style explanations. It is possible that this is due to the sensitivity of demographic information; anecdotally we can imagine that many users would not want to be recommended an item based on their gender, age or ethnicity (e.g. *"We recommend you the movie Sex in the City because you are a female aged 20-40."*).

15.8 Summary and future directions

In this chapter, we offer guidelines for the designers of explanations in recommender systems. Firstly, the designer should consider what benefit the explanations offer, and thus which criteria they are evaluating the explanations for (e.g. *transparency, scrutability, trust, efficiency, effectiveness, persuasion or satisfaction*). The developer may select several criteria which may be related to each other, but may also be conflicting. In the latter case, it is particularly important to distinguish between these evaluation criteria. It is only in more recent work that these trade-offs are being shown and becoming more apparent [20, 61].

In addition, the system designer should consider the *metrics* they are going to use when evaluating the explanations, and the dependencies the explanations may have with different parts of the system, such as the way recommendations are presented (e.g. top item, top N-items, similar to top item(s), predicted ratings for all items, structured overview), the way users interact with the explanations (e.g. the user specifies their requirements, asks for an alteration, rates items, gives their opinion, or uses a hybrid interaction interface) and the underlying recommender engine.

To offer a single example of the relation between explanations and other recommender system factors, we can imagine a recommender engine with low recommendation accuracy. This may affect all measurements of effectiveness in the system, as users do not really like the items they end up being recommend. These measurements do not however reflect the effectiveness of the *explanations* themselves. In this case, a layered approach to evaluation [47], where explanations are considered in isolation from the recommendation algorithm as seen in [61], may be warranted. Similarly, thought should be given to how the method of presenting recommendations, and the method of interaction may affect the (evaluation of) explanations.

We offered examples of explanation styles influenced by the most common algorithms (e.g. content-based, collaborative, demographic, or knowledge/utility-based), and how they have been used in existing systems. To a certain extent these types of explanations can be reused (likely at the cost of transparency) for hybrid recommendations, and other complex recommendation methods such as latent semantic analysis, but these areas of research remain largely open. Preliminary works for some of these areas can be found in e.g. [33] (explaining Markov decision processes) and [31] (explaining latent semantic analysis models).

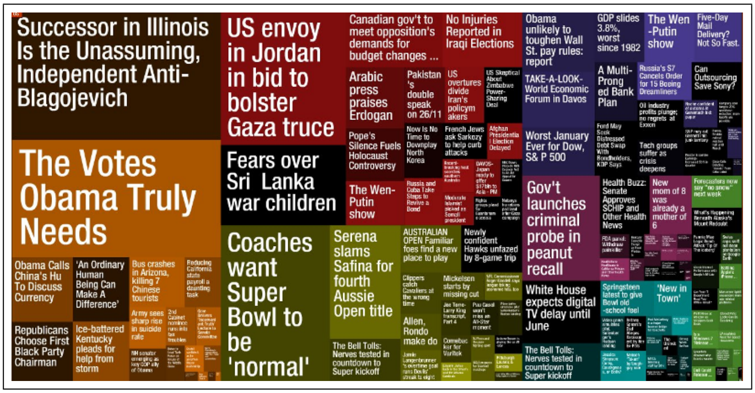

Successor in Illinois Is the Unassuming, Independent Anti-Blagojevich
The Votes Obama Truly Needs
Obama Calls China's Hu To Discuss Currency
'An Ordinary Human Being Can Make A Difference'
Bus crashes in Arizona, killing 7 Chinese tourists
Republicans Choose First Black Party Chairman
Ice-battered Kentucky pleads for help from storm
US envoy in Jordan in bid to bolster Gaza truce
Fears over Sri Lanka war children
Coaches want Super Bowl to be 'normal'
Canadian gov't to meet opposition's demands for budget changes ...
No Injuries Reported in Iraqi Elections
Arabic press praises Erdogan
Pakistan's double speak on 26/11
US overturns divide Iran's policym akers
Pope's Silence Fuels Holocaust Controversy
The Wen-Putin show
Serena slams Safina for fourth Aussie Open title
AUSTRALIAN OPEN Familiar foes find a new place to play
Newly confident Hawks unfazed by 8-game trip
Allen, Rondo make do
The Bell Tolls: Nerves tested in countdown to Super kickoff
Obama unlikely to toughen Wall St. pay rules: report
TAKE-A-LOOK-World Economic Forum in Davos
Worst January Ever for Dow, S&P 500
Gov't launches criminal probe in peanut recall
White House expects digital TV delay until June
The Bell Tolls: Nerves tested in countdown to Super kickoff
GDP slides 3.8%, worst since 1982
The Wen-Putin show
Five-Day Mail Delivery? Not So Fast.
A Multi-Prong ed Bank Plan
Can Outsourcing Save Sony?
Tech groups suffer as crisis deepens
Health Buzz: Senate Approves SCHIP and Other Health News
New mom of 8 was already a mother of 6
Springsteen latest to give Bowl old-school feel
'New in Town'

Fig. 15.9: Newsmap - a treemap visualization of news. Different colors represent topic areas, square and font size to represent importance to the current user, and shades of each topic color to represent recency.

As of yet, there has been little comparison between explanation styles with regard to their performance on explanatory goals. This is an avenue of research in which we hope to see further progress in the near future. Also, future work will likely involve more advanced interfaces for explanations. For example, the "treemap" structure (see Figure 15.9 [4]) offers an overview of the search space [9]. This type of overview may also be used for explanation. Assume for example, that a user is being recommended the piece "The Votes Obama Truly Needs", and that this rectangle is highlighted. This interface "explains" that this item is being recommended because the user is interested in current US news (orange color), it is popular (big square), and that it is recent (bright color).

Last, but certainly not least, researchers are starting to find that explanations are part of a cyclical process. The explanations affect a user's mental model of the recommender system, and in turn the way they interact with the explanations. In fact this may also impact the recommendation accuracy negatively [7, 20]. For example [7] saw that recommendation accuracy decreased as users removed keywords from their profile for a news recommender system. Understanding this cycle will likely be one of the future strands of research.

References

1. Pandora (2006). `http://www.pandora.com`
2. Movielens dataset (2009). `http://www.grouplens.org/node/73`
3. Netflix dataset (2009). `http://www.netflixprize.com/`
4. Newsmap (2009). `http://www.marumushi.com/apps/newsmap/index.cfm`
5. Nutking (2010). `http://nutking.ectrldev.com/nutking/jsp/language.do?action=english`
6. Adrissono, L., Goy, A., Petrone, G., Segnan, M., Torasso, P.: Intrigue: Personalized recommendation of tourist attractions for desktop and handheld devices. Applied Artificial Intelligence **17**, 687–714 (2003)
7. Ahn, J.W., Brusilovsky, P., Grady, J., He, D., Syn, S.Y.: Open user profiles for adaptive news systems: help or harm? In: WWW '07: Proceedings of the 16th international conference on World Wide Web, pp. 11–20. ACM Press, New York, NY, USA (2007).
8. Andersen, S.K., Olesen, K.G., Jensen, F.V.: HUGIN—a shell for building Bayesian belief universes for expert systems. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1990)
9. Bederson, B., Shneiderman, B., Wattenberg, M.: Ordered and quantum treemaps: Making effective use of 2d space to display hierarchies. ACM Transactions on Graphics **21(4)**, 833–854. (2002)
10. Bennett, S.W., Scott., A.C.: The Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project, chap. 19 - Specialized Explanations for Dosage Selection, pp. 363–370. Addison-Wesley Publishing Company (1985)
11. Bilgic, M., Mooney, R.J.: Explaining recommendations: Satisfaction vs. promotion. In: Proceedings of the Wokshop Beyond Personalization, in conjunction with the International Conference on Intelligent User Interfaces, pp. 13–18 (2005)
12. Billsus, D., Pazzani, M.J.: A personal news agent that talks, learns, and explains. In: Proceedings of the Third International Conference on Autonomous Agents, pp. 268–275 (1999)
13. Burke, R.: Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction **12(4)**, 331–370 (2002)
14. Burke, R.D., Hammond, K.J., Young, B.C.: Knowledge-based navigation of complex information spaces. In: AAAI/IAAI, Vol. 1, pp. 462–468 (1996)
15. Carenini, G., Mittal, V., Moore, J.: Generating patient-specific interactive natural language explanations. Proc Annu Symp Comput Appl Med Care pp. 5–9 (1994)
16. Chen, L., Pu, P.: Trust building in recommender agents. In: WPRSIUI in conjunction with Intelligent User Interfaces, pp. 93–100 (2002)
17. Chen, L., Pu, P.: Hybrid critiquing-based recommender systems. In: Intelligent User Interfaces, pp. 22–31 (2007)
18. Cosley, D., Lam, S.K., Albert, I., Konstan, J.A., Riedl, J.: Is seeing believing?: how recommender system interfaces affect users' opinions. In: CHI, *Recommender systems and social computing*, vol. 1, pp. 585–592 (2003).
19. Cramer, H., Evers, V., Someren, M.V., Ramlal, S., Rutledge, L., Stash, N., Aroyo, L., Wielinga, B.: The effects of transparency on perceived and actual competence of a content-based recommender. In: Semantic Web User Interaction Workshop, CHI (2008)

20. Cramer, H.S.M., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., Aroyo, L., Wielinga, B.J.: The effects of transparency on trust in and acceptance of a content-based art recommender. User Model. User-Adapt. Interact **18**(5), 455–496 (2008).
21. Czarkowski, M.: A scrutable adaptive hypertext. Ph.D. thesis, University of Sydney (2006)
22. Doyle, D., Tsymbal, A., Cunningham, P.: A review of explanation and explanation in case-based reasoning. Tech. rep., Department of Computer Science, Trinity College, Dublin (2003)
23. Felfernig, A., Gula, B.: Consumer behavior in the interaction with knowledge-based recommender applications. In: ECAI 2006 Workshop on Recommender Systems, pp. 37–41 (2006)
24. Fogg, B., Marshall, J., Kameda, T., Solomon, J., Rangnekar, A., Boyd, J., Brown, B.: Web credibility research: A method for online experiments and early study results. In: CHI 2001, pp. 295–296 (2001)
25. Fogg, B.J., Soohoo, C., Danielson, D.R., Marable, L., Stanford, J., Tauber, E.R.: How do users evaluate the credibility of web sites?: a study with over 2,500 participants. In: Proceedings of DUX'03: Designing for User Experiences, no. 15 in Focusing on user-to-product relationships, pp. 1–15 (2003). URL `http://doi.acm.org/10.1145/997078.997097`
26. Ginty, L.M., Smyth, B.: Comparison-based recommendation. Lecture Notes in Computer Science **2416**, 731–737 (2002).
27. Hance, E., Buchanan, B.: Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Addison-Wesley (1984)
28. Häubl, G., Trifts, V.: Consumer decision making in online shopping environments: The effects of interactive decision aids. Marketing Science **19**, 4–21 (2000)
29. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: ACM conference on Computer supported cooperative work, pp. 241–250 (2000)
30. Hingston, M.: User friendly recommender systems. Master's thesis, Sydney University (2006)
31. Hu, Y., Koren, Y., Volinsky, C.: Collaborative filtering for implicit feedback datasets. In: ICDM (2008)
32. Hunt, J.E., Price, C.J.: Explaining qualitative diagnosis. Engineering Applications of Artificial Intelligence **1**(3), Pages 161–169 (1988)
33. Khan, O.Z., Poupart, P., Black, J.P.: Minimal sufficient explanations for mdps. In: Workshop on Explanation-Aware Computing associated with IJCAI (2009)
34. Krulwich, B.: The infofinder agent: Learning user interests through heuristic phrase extraction. IEEE Intelligent Systems **12**, 22–27 (1997)
35. Lacave, C., Diéz, F.J.: A review of explanation methods for bayesian networks. The Knowledge Engineering Review **17:2**, 107–127 (2002)
36. Lacave, C., Diéz, F.J.: A review of explanation methods for heuristic expert systems. The Knowledge Engineering Review **17:2**, 107–127 (2004)
37. Lewis, C., Rieman, J.: Task-centered user interface design: a practical introduction. University of Colorado (1994)
38. Lopez-Suarez, A., Kamel, M.: Dykor: a method for generating the content of explanations in knowledge systems. Knowledge-based Systems **7**(3), 177–188 (1994)
39. McCarthy, K., Reilly, J., McGinty, L., Smyth, B.: Thinking positively - explanatory feedback for conversational recommender systems. In: Proceedings of the European Conference on Case-Based Reasoning (ECCBR-04) Explanation Workshop,, pp. 115–124 (2004)
40. McNee, S., Lam S.K.and Guetzlaff, C., Konstan J.A.and Riedl, J.: Confidence displays and training in recommender systems. In: INTERACT IFIP TC13 International Conference on Human-Computer Interaction, pp. 176–183 (2003)
41. McNee, S.M., Lam, S.K., Konstan, J.A., Riedl, J.: Interfaces for eliciting new user preferences in recommender systems. User Modeling pp. pp. 178–187 (2003)
42. McNee, S.M., Riedl, J., Konstan, J.A.: Being accurate is not enough: How accuracy metrics have hurt recommender systems. In: Extended Abstracts of the 2006 ACM Conference on Human Factors in Computing Systems (CHI 2006) (2006)
43. McSherry, D.: Explanation in recommender systems. Artificial Intelligence Review **24(2)**, 179 – 197 (2005)

44. Nielsen, J., Molich, R.: Heuristic evaluation of user interfaces. In: ACM CHI'90, pp. 249–256 (1990)
45. Ohanian, R.: Construction and validation of a scale to measure celebrity endorsers' perceived expertise, trustworthiness, and attractiveness. Journal of Advertising **19:3**, 39–52 (1990)
46. O'Sullivan, D., Smyth, B., Wilson, D.C., McDonald, K., Smeaton, A.: Improving the quality of the personalized electronic program guide. User Modeling and User-Adapted Interaction **14**, pp. 5–36 (2004)
47. Paramythis, A., Totter, A., Stephanidis, C.: A modular approach to the evaluation of adaptive user interfaces. In: S. Weibelzahl, D.N. Chin, G. Weber (eds.) Evaluation of Adaptive Systems in conjunction with UM'01, pp. 9–24 (2001)
48. Pazzani, M.J.: A framework for collaborative, content-based and demographic filtering. Artificial Intelligence Review **13**, 393–408 (1999)
49. Pu, P., Chen, L.: Trust building with explanation interfaces. In: IUI'06, Recommendations I, pp. 93–100 (2006).
50. Pu, P., Chen, L.: Trust-inspiring explanation interfaces for recommender systems. Knowledge-based Systems **20**, 542–556 (2007)
51. Rafter, R., Smyth, B.: Conversational collaborative recommendation - an experimental analysis. Artif. Intell. Rev **24**(3-4), 301–318 (2005). URL `http://dx.doi.org/10.1007/s10462-005-9004-8`
52. Reilly, J., McCarthy, K., McGinty, L., Smyth, B.: Dynamic critiquing. In: P. Funk, P.A. González-Calero (eds.) ECCBR, *Lecture Notes in Computer Science*, vol. 3155, pp. 763–777. Springer (2004)
53. Roth-Berghofer, T., Schulz, S., Leake, D.B., Bahls, D.: Workshop on explanation-aware computing. In: ECAI (2008)
54. Roth-Berghofer, T., Tintarev, N., Leake, D.B.: Workshop on explanation-aware computing. In: IJCAI (2009)
55. Sinha, R., Swearingen, K.: The role of transparency in recommender systems. In: Conference on Human Factors in Computing Systems, pp. 830–831 (2002)
56. Sørmo, F., Cassens, J., Aamodt, A.: Explanation in case-based reasoning perspectives and goals. Artificial Intelligence Review **24(2)**, 109 – 143 (2005)
57. Swearingen, K., Sinha, R.: Interaction design for recommender systems. In: Designing Interactive Systems, pp. 25–28 (2002).
58. Symeonidis, P., Nanopoulos, A., Manolopoulos, Y.: Justified recommendations based on content and rating data. In: WebKDD Workshop on Web Mining and Web Usage Analysis (2008)
59. Thompson, C.A., Göker, M.H., Langley, P.: A personalized system for conversational recommendations. J. Artif. Intell. Res. (JAIR) **21**, 393–428 (2004).
60. Tintarev, N., Masthoff, J.: Over- and underestimation in different product domains. In: Workshop on Recommender Systems associated with ECAI (2008)
61. Tintarev, N., Masthoff, J.: Personalizing movie explanations using commercial meta-data. In: Adaptive Hypermedia (2008)
62. Vig, J., Sen, S., Riedl, J.: Tagsplanations: Explaining recommendations using tags. In: Intelligent User Interfaces (2009)
63. Wang, W., Benbasat, I.: Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. Journal of Managment Information Systems **23**, 217–246 (2007)
64. Wärnestål, P.: Modeling a dialogue strategy for personalized movie recommendations. In: Beyond Personalization Workshop, pp. 77–82 (2005)
65. Wärnestål, P.: User evaluation of a conversational recommender system. In: Proceedings of the 4th Workshop on Knowledge and Reasoning in Practical Dialogue Systems, pp. 32–39 (2005)
66. Wick, M.R., Thompson, W.B.: Reconstructive expert system explanation. Artif. Intell. **54**(1-2), 33–70 (1992).
67. Ye, L., Johnson, P., Ye, L.R., Johnson, P.E.: The impact of explanation facilities on user acceptance of expert systems advice. MIS Quarterly **19**(2), 157–172 (1995).

68. Yee, K.P., Swearingen, K., Li, K., Hearst, M.: Faceted metadata for image search and browsing. In: ACM Conference on Computer-Human Interaction (2003)
69. Zaslow, J.: Oh no! My TiVo thinks I'm gay (2002).