

Explaining Collaborative Filtering Recommendations

Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl

Dept. of Computer Science and Engineering

University of Minnesota

Minneapolis, MN 55112 USA

{herlocke, konstan, riedl}@cs.umn.edu

<http://www.grouplens.org/>

ABSTRACT

Automated collaborative filtering (ACF) systems predict a person's affinity for items or information by connecting that person's recorded interests with the recorded interests of a community of people and sharing ratings between like-minded persons. However, current recommender systems are black boxes, providing no transparency into the working of the recommendation. Explanations provide that transparency, exposing the reasoning and data behind a recommendation. In this paper, we address explanation interfaces for ACF systems – *how* they should be implemented and *why* they should be implemented. To explore *how*, we present a model for explanations based on the user's conceptual model of the recommendation process. We then present experimental results demonstrating what components of an explanation are the most compelling. To address *why*, we present experimental evidence that shows that providing explanations can improve the acceptance of ACF systems. We also describe some initial explorations into measuring how explanations can improve the filtering performance of users.

Keywords

Explanations, collaborative filtering, recommender systems, MovieLens, GroupLens

INTRODUCTION

Automated collaborative filtering (ACF) systems predict a user's affinity for items or information. Unlike traditional content-based information filtering system, such as those developed using information retrieval or artificial intelligence technology, filtering decisions in ACF are based on human and not machine analysis of content. Each user of an ACF system rates items that they have experienced, in order to establish a profile of interests. The ACF system then matches together that user with people of similar interests or tastes.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CSCW'00, December 2-6, 2000, Philadelphia, PA.

Copyright 2000 ACM 1-58113-222-0/00/0012...\$5.00.

Then ratings from those similar people are used to generate recommendations for the user.

ACF has many significant advantages over traditional content-based filtering, primarily because it does not depend on error-prone machine analysis of content. The advantages include the ability to filter any type of content, e.g. text, art work, music, mutual funds; the ability to filter based on complex and hard to represent concepts, such as taste and quality; and the ability to make serendipitous recommendations.

It is important to note that ACF technologies do not necessarily compete with content-based filtering. In most cases, they can be integrated to provide a powerful hybrid filtering solution.

ACF systems have been successful in research, with projects such as GroupLens [9,12], Ringo [13], Video Recommender [6], and MovieLens [4] gaining large followings on the Internet. Commercially, some of the highest profile web sites like Amazon.com, CDNow.com, MovieFinder.com, and Launch.com have made successful use of ACF technology.

While automated collaborative filtering systems have proven to be accurate enough for entertainment domains[6,9,12,13], they have yet to be successful in content domains where higher risk is associated with accepting a filtering recommendation. While a user may be willing to risk purchasing a music CD based on the recommendation of an ACF system, he will probably not risk choosing a honeymoon vacation spot or a mutual fund based on such a recommendation. There are several reasons why ACF systems are not trusted for high-risk content domains.

First, ACF systems are stochastic processes that compute predictions based on models that are heuristic approximations of human processes. Second, and probably most important, ACF systems base their computations on extremely sparse and incomplete data. These two conditions lead to recommendations that are often correct, but also occasionally very wrong. ACF systems today are black boxes, computerized oracles that give advice, but cannot be questioned. A user is given no indicators to consult to determine when to trust a recommendation and when to doubt one. These problems have prevented acceptance of ACF systems in all but the low-risk content domains.

Explanation capabilities provide a solution to building trust and may improve the filtering performance of people using ACF systems. An explanation behind the reasoning of a ACF recommendation provides transparency into the workings of the ACF system. Users will be more likely to trust a recommendation when they know the reasons behind that recommendation. Explanations will help users understand the process of ACF, and know where its strengths and weaknesses are.

SOURCES OF ERROR

Explanations help either detect or estimate the likelihood of errors in the recommendation. Let us examine the different sources for errors. Errors in recommendations by automated collaborative filtering (ACF) systems can be roughly grouped into two categories: model/process errors and data errors.

Model/Process Errors

Model or process errors occur when the ACF system uses a process to compute recommendations that does not match the user's requirements. For example, suppose Nathan establishes a rating profile containing positive ratings for both movie adaptations of Shakespeare and Star Trek movies. Whether he prefers Shakespeare or Star Trek depends on his context (primarily whether or not he is taking his lady-friend to the movies). The ACF system he is using however does not have a computational model that is capable of recognizing the two distinct interests represented in his rating profile. As a result, the ACF system may match Nathan up with hard-core Star Trek fans, resulting in a continuous stream of recommendations for movies that could only be loved by a Star Trek fan in spite of whether Nathan is with his lady-friend or not.

Data Errors

Data errors result from inadequacies of the data used in the computation of recommendations. Data errors usually fall into three classes: not enough data, poor or bad data, and high variance data.

Missing and sparse data are two inherent factors of ACF computation. If the data were complete, there would be no need for ACF systems to predict the missing data points. Items and users that have newly entered the ACF system are particularly prone to error. When a new movie is first released, very few people have rated the movie, so the ACF must base predictions for that movie on a small number of ratings. Because there are only a small number of people who have rated the movie, the ACF system may have to base recommendations on ratings from people who do not share the user's interests very closely. Likewise, when new users begin using the system, they are unwilling to spend excessive amounts of time entering ratings before seeing some results, forcing the ACF system to produce recommendations based on a small and incomplete profile of the user's interests. The result is that new users may be matched with other people who share their interests on a small subset of items, but in actuality don't share much more in common.

Even in cases where considerable amounts of data are available about the users and the items, some of the data may contain errors. For example, suppose Nathan accidentally visits a pornography site because the site is located at a URL very similar to that of the White House – the residence of the US president. Nathan is using an ACF system that considers web page visits as implicit preference ratings. Because of his accidental visit to the wrong web site, he may soon be surprised by the type of movies that are recommended to him. However, if he has visited a large number of web sites, he may not be aware of the offending rating.

High variance data is not necessarily considered bad data, but can be the cause of recommendation errors. For example, of all the people selected who have rating profiles similar to Nathan's, half rated the movie "Dune" high and half rated it low. As a movie that polarizes interests, the proper prediction is probably not the average rating (indicating ambivalence), although this is probably what will be predicted by the ACF system.

EXPLANATIONS

Explanations provide us with a mechanism for handling errors that come with a recommendation. Consider how we as humans handle suggestions as they are given to us by other humans. We recognize that other humans are imperfect recommenders. In the process of deciding to accept a recommendation from a friend, we might consider the quality of previous recommendations by the friend or we may compare how that friend's general interests compare to ours in the domain of the suggestion. However, if there is any doubt, we will ask "why?" and let the friend explain their reasoning behind a suggestion. Then we can analyze the logic of the suggestion and determine for ourselves if the evidence is strong enough.

It seems sensible to provide explanation facilities for recommender systems such as automated collaborative filtering systems. Previous work with another type of decision aide – expert systems – has shown that explanations can provide considerable benefit. The same benefits seem possible for automated collaborative filtering systems. Most expert systems that provided explanation facilities, such as MYCIN[1], used rule-based reasoning to arrive at conclusions. MYCIN provided explanations by translating traces of rules followed from LISP to English. A user could ask both why a conclusion was arrived at and how much was known about a certain concept. Other work describing explanation facilities in expert systems includes Hovitz, Breese, and Henrion[7]; and Miller and Larson[10]. Since collaborative filtering does not generally use rule-based reasoning, the problems of explanation there will require different approaches and different solutions.

Work related to explanations can be found in cognitive science, psychology, and philosophy. Johnson & Johnson[8] have begun research into the components of a unified theory of explanation in human-computer interfaces. To support

their theories, they performed empirical experiments to help determine the logical components of an explanation. There has also been considerable study into the psychology of questioning and question answering with humans and how it can be applied to human-computer interfaces[2,3]. Philosophers have studied the rules and logic of human discourse – such as in the book “The Uses of Argument” by Toulmin[14].

Building an explanation facility into a recommender system can benefit the user in many ways. It removes the black box from around the recommender system, and provides transparency. Some of the benefits provided are:

- (1) Justification. User understanding of the reasoning behind a recommendation, so that he may decide how much confidence to place in that recommendation.
- (2) User Involvement. User involvement in the recommendation process, allowing the user to add his knowledge and inference skills to the complete decision process.
- (3) Education. Education of the user as to the processes used in generating a recommendation, so that he may better understand the strengths and limitations of the system.
- (4) Acceptance. Greater acceptance of the recommender system as a decision aide, since its limits and strengths are fully visible and its suggestions are justified.

Together, the potential for increasing the impact of automated collaborative filtering systems is great.

RESEARCH QUESTIONS

There are three key research questions that we are interested in answering about the use of explanations with automated collaborative filtering (ACF) systems.

What models and techniques are effective in supporting explanation in an ACF system?

An ACF system's computational model can be complex. What is the right amount of detail to expose? How much information is too much? What type of data do we present? Can we develop a theoretical model to guide our design decisions? Many such issues can be explored through theoretical design and experimentation with users.

Can explanation facilities increase the acceptance of automated collaborative filtering systems?

We believe that by providing transparency into the workings of the ACF process, we will build users' confidence in the system, and increase their willingness to use the ACF system as a decision aid.

Can explanation facilities increase the filtering performance of ACF system users?

The goal of an ACF system is to reduce information overload by helping the user to separate the good and valuable from

that which is not. The information filter helps users to make decisions about which items to consume, such as what books to read or what movies to watch. We want ACF systems that result in more of the correct decisions, as well as filters that improve the rate at which we can process information, without missing anything important. But we also want systems that can reduce stress by making us more confident about our decisions. Can explanation interfaces strengthen these effects?

BUILDING A MODEL OF EXPLANATIONS

There are many different ways that we can explain a recommendation from an automated collaborative filtering (ACF) system. What kinds of data summaries are the most important or the most effective? What format of presentation is the most understandable and effective? What are the most compelling forms of explanation that we can give for a collaborative filtering recommendation?

To address the first research question, we work “outside-in.” That is to say, that we start with the user-perceived conceptual model[11] of the ACF system, and then from that we generate the key components of explanation. We discuss the white-box model and the black-box conceptual models as well as misinformed conceptual models.

White Box Conceptual Model

One of the strengths of ACF is that it has an easily communicated and understood conceptual model of operation. The operation of an ACF system is analogous to the human word-of-mouth recommendation. Users of an ACF system are provided with the following three-step conceptual model of the operation of the ACF system.

- (1) User enters profile of ratings
- (2) ACF system locates people with similar profiles (neighbors)
- (3) Neighbors' ratings are combined to form recommendations

At the implementation level, these steps are broken up into more detailed steps, but the user is generally not aware of such details. A user's perception of the performance of the above listed three tasks will affect her perception of the performance of the overall ACF system.

From this model, we can derive potential means of explanation of an ACF recommendation. We can focus on techniques to justify that the ACF system is indeed performing each of the above steps to the satisfaction of the user and her current context. Let us examine each of the steps in more detail, focusing on two components that we need to explain: the process and the data.

(1) User enters profile of ratings

Explaining step (1) may seem relatively straightforward, but consideration of this step brings light to important information that can be very important to the user. Consider process information: exactly how was the profile information

collected? We can collect an immense amount of preference information from the user, both implicit, such as page-views, and explicit, such as numeric ratings. Any interactions the user has with a system can possibly affect the outcome of the recommendations. The user can benefit from knowing how her actions have effected her recommendations. For example, a user, upon learning that web-page visits are considered weak ratings, may determine to provide more explicit preference ratings in order to have greater control over her recommendations.

An explanation might need to explain what kinds of preference information were used in a given explanation. What kinds of data does the profile consist of? Was the movie recommendation for Titanic based purely on ratings of previously seen movies, or on the fact that the user spent time reading plot summaries of romance movies? In addition, has the user provided enough profile information to allow for high-quality recommendations? Perhaps the user has not rated a large enough or diverse enough set of movies to allow the ACF system to provide accurate recommendations with confidence. An explanation interface might be required to give the user feedback on the quality of her profile. For example, we have designed an explanation interface that identifies movie ratings in a user profile that had the most significant effect on a prediction. Ratings in the profile that have an exceptionally significant affect on the recommendation are a sign that the profile may not be diverse enough, as well as an indication of potential similarities in content or taste between the significant item and the item being recommended.

(2) ACF system locates people with similar profiles (neighbors)

It is in performing step (2) that ACF systems show their true value over normal human word-of-mouth recommendations, with ACF systems being able to examine thousands of potential soulmates, and choose the most similar of the bunch. What do we have to do to assure the user that the ACF system has identified the correct set of neighbors for the user's current context of need? The process that is used to locate other people with similar profiles is one key to the success of the collaborative filtering technology. If the neighbors selected by the system are the best predictors for the user's current information need, then the resulting recommendations will be the best possible. This is especially important for higher risk domains, where the user will often want to know when approximations and shortcuts are taken. For example, most ACF systems have huge numbers of profiles; their user community often numbers in the millions. These same ACF systems must also provide thousands of predictions per second. Supporting this large numbers of users at that level of performance requires many approximations. In most cases, the neighbors selected are not necessarily the "best" neighbors, but rather are the most similar profiles that could be found by sampling the profiles that are available in high-speed memory.

The similarity metric that is used to judge potential neighbors can also be important in evaluating a prediction. Does the "closeness" measured by the given similarity metric match the users current content of information need?

Providing descriptions of the data processed in locating a neighbor can be important to explaining a prediction. How many potential neighbors were examined (i.e. what was the sample size?) From the neighbors that were selected – what do their profiles look like? Do their interests match the users current context of information need? When measuring similarity between users, most ACF systems will give equal weight to all items that occur in both profiles. However, the user will often have strong constraints that are not captured by the system. For example, a user may feel that anyone who rated Star Wars low has no right giving ratings for science fiction movies. An explanation could give the user the ability to examine the ratings of the chosen neighbors and when the user discovers the offending neighbor, he can disregard the prediction, or perhaps the system will allow him to manually remove that neighbor from consideration.

(3) Neighbors ratings are combined to form recommendations

The final step is explaining the data and the process of taking the ratings of the neighbors and aggregating them into a final prediction. It is at this level that many of the symptoms of weak predictions can be discovered with good explanations.

The data are the most important in explaining this step. Users can benefit greatly from knowing exactly how each of their neighbors rated the item being explained, or if there are large numbers of ratings, the distribution of ratings for the item being recommended. They can combine this information with information from step (2), such as knowing how "good" or "close" their neighbors are. Users can detect instances where the prediction is based on a small amount of data, and investigate further to determine if a recommendation is an error, or just a sleeper item. For example, imagine that Jane has received a movie recommendation from an ACF-based movie recommender. She requests an explanation for the movie. She finds that the recommendation is based only on the ratings of five of her neighbors. From this Jane knows that the movie is either very new, not well known, or has received bad publicity. Of the five ratings from her neighbors, three are exceptionally high, and two are ambivalent or slightly negative. She then looks closely at the profiles of the neighbors. The three who liked the movie seem to share her interests in eclectic art films. The two who did not rate the movie higher seemed to only share Jane's interest in popular Hollywood films. From this information, Jane determines that the movie is probably a not-well-known art film and decides to trust the recommendation.

The process used to aggregate neighbor ratings into a prediction may also be of interest to the user. However, in most cases, the prediction is simply a weighted average of the neighbor's ratings.

Black Box Model

Often, there is not the opportunity or possibly the desire to convey the conceptual model of the system to each user of the system. In such cases, the ACF system becomes a black box recommender, producing recommendations like those of an oracle. The user may not even be aware that the ACF system is collecting implicit ratings, such as time-spent-reading[9] to fuel the recommendations. For example, a video store could use past rental history as rating profiles, and produce personalized recommendations for users based on ACF technology. For fear that other video stores will copy their technique, they do not wish to reveal the process by which they compute recommendations, yet they would like to provide some sort of explanation, justification, or reason to trust the recommendation.

In these situations, the forms of explanations generated the white box model are not appropriate. We must focus on ways to justify recommendation that are independent of the mechanics that occur within the black box recommender. One technique is to use the past performance of the recommender as justification. For example, an explanation might include the statement "This system has been correct for you 80% of the time when recommending items similar to this one." Another technique might be to bring in supporting evidence that may not have been used during the computation of the recommendation. For example, even though the video store recommendation was based only on the purchase records of customers, the video store could justify its predictions by displaying correlating recommendations from local film critics.

Any white box can be viewed as a black box by focusing only on the inputs and outputs. Because of this, forms of explanation for black box recommenders should also be useful in providing explanations for white box recommenders. For example, even if information about the process and data used in the computation is available to the user, knowing the system's past overall performance can be very useful.

Misinformed Conceptual Models

It is inevitable that some users will form incorrect conceptual models of the ACF systems that they are using to filter information. One common misconception that users acquire is that an ACF system is making decisions based on content characteristics. For example, several users of MovieLens have written us with comments that make it clear they believe we are recommending based on movie content characteristics such as director, actors/actresses, and genre. Here, the educational aspect of explanations comes into play. Users with conflicting conceptual models will quickly realize that the explanations do not match their expectations, and through the process of examining explanation, learn the proper conceptual model.

A related issue occurs when users are intentionally lead to believe in an incorrect conceptual model. This might happen if the computational model is believed to be too complex to

explain, so users are lead to believe that a simpler, more understandable process is being used. There could even be instances where the recommender is using what could be considered subversive methods by the user, such as claiming to provide personalized recommendations, while pushing high inventory or high margin items. All these issues greatly complicate explanations, and we do not focus on them in this paper.

EXPERIMENT 1 – INVESTIGATING THE MODEL

The cognitive models described act as a guide that can indicate potential key areas of explanation. However, there are huge amounts of information that could be explained in a prediction. Automated collaborative filtering (ACF) tools evolved to combat information overload, and we should avoid creating a new kind of information overload by presenting too much or too confusing data.

When we design an explanation interface to an ACF system, we are faced with the initial problem: what exactly do we explain and in what manner? The model we have described solves this problem to some extent by suggesting information that is key in the user's cognitive model of the system. However, even with the model, we are left with a huge number of features to potentially explain.

What makes an explanation interface successful will vary from situation to situation. In some cases, a successful explanation is one that successfully convinces you to purchase a recommended product. In other cases, a successful explanation interface is one that helps you to identify predictions that have weak justification. In all cases, a successful explanation interface will be one that users perceive to be useful or interesting, and will continue to use.

To explore this issue, we have performed an experiment that measures how users of an ACF system respond to different explanations, each derived from a different components of the explanation models described in the previous section.

Design

The study was performed as a survey; test subjects were volunteer users of the MovieLens web-based movie recommender. MovieLens uses ACF technology to produce personalized recommendations for movies and videos. The MovieLens database currently contains 4.6 million ratings from 74,000 users on 3500 movies, currently averaging approximately 1000 active users per week.

Study participants were presented with the following hypothetical situation:

Imagine that you have \$7 and a free evening coming up. You are considering going to the theater to see a movie, but only if there is a movie worth seeing. To determine if there is a movie worth seeing, you consult MovieLens for a personalized movie recommendation. MovieLens recommends one movie, and provides some justification.

Each user was then provided with 21 individual movie recommendations, each with a different explanation component, and asked to rate on a scale of 1 – 7 how likely they would be to go and see the movie. An example of one explanation interface is shown in Figure 1.



Figure 1. One of the twenty-one different explanation interfaces given shown in the user survey. Notice that the title has been encoded, so that it does not influence a user’s decision to try a movie.

To ensure that the response to each stimulus could be compared fairly, the 21 different explanation interfaces all describe the same movie recommendation. The explanations are based on data from an observed recommendation on MovieLens.¹ The recommendation chosen for the survey was one that we, as experienced experts with the system, recognized as having good justification. That is to say that, had we been presented with the explanation data, we would believe that its probability of being correct was high.

The study performed was organized as a randomized block design, with the blocks being users and the treatment being the different explanation interfaces. The 21 different interfaces were presented in a random order for each user to account for learning effects. The survey was presented to 78 users of the MovieLens site. A list of the different explanation interfaces provided is shown in Table 1, along with the accompanying results.

Results

The number of times each question was asked differed slightly because if any users clicked on the reload or refresh button of their browser, no result was recorded and the survey moved to the next question.

Explanations 11 and 12 are the base case. They represent no additional explanation data, beyond the simple knowledge of the white-box cognitive model. Therefore, explanations 13

and greater can be seen as negatively contributing to the acceptance of the recommendation.

#		N	Mean Response	Std Dev
1	Histogram with grouping	76	5.25	1.29
2	Past performance	77	5.19	1.16
3	Neighbor ratings histogram	78	5.09	1.22
4	Table of neighbors ratings	78	4.97	1.29
5	Similarity to other movies rated	77	4.97	1.50
6	Favorite actor or actress	76	4.92	1.73
7	MovieLens percent confidence in prediction	77	4.71	1.02
8	Won awards	76	4.67	1.49
9	Detailed process description	77	4.64	1.40
10	# neighbors	75	4.60	1.29
11	No extra data – focus on system	75	4.53	1.20
12	No extra data – focus on users	78	4.51	1.35
13	MovieLens confidence in prediction	77	4.51	1.20
14	Good profile	77	4.45	1.53
15	Overall percent rated 4+	75	4.37	1.26
16	Complex graph: count, ratings, similarity	74	4.36	1.47
17	Recommended by movie critics	76	4.21	1.47
18	Rating and %agreement of closest neighbor	77	4.21	1.20
19	# neighbors with std. deviation	78	4.19	1.45
20	# neighbors with avg correlation	76	4.08	1.46
21	Overall average rating	77	3.94	1.22

Table 1. Mean response of users to each explanation interface, based on a scale of one to seven. Explanations 11 and 12 represent the base case of no additional information. Shaded rows indicate explanations with a mean response significantly different from the base cases (two-tailed $\alpha = 0.05$).

Analysis

First, it is important to recognize the big winners: histograms of the neighbors’ ratings, past performance, similarity to other items in the user’s profile, and favorite actor or actress.

There were three rating different histograms. The best performing histogram (explanation 1) is shown in Figure 2. Due to space considerations, not all explanation screens are depicted here, but they can be found online at <http://www.cs.umn.edu/~herlocke/>. Explanation 3 was a standard bar chart histogram, with one bar for each category of rating (1-5). Explanation 4 presented the same data as Explanation 3, but in numerical tabular format instead of a bar chart (see Figure 1). Explanation 1 performed better than a basic bar chart because it reduced the dimensionality of the data to a point where only one binary comparison is necessary (the good versus the bad). The hypothesis that simple graphs are more compelling is supported by observing the poor performance of Explanation 16, which presents a superset of the data shown in histograms, adding information about how close each neighbor is to the user.

¹ The explanation interfaces were based on the MovieLens recommendation of “Titanic” for the primary author.

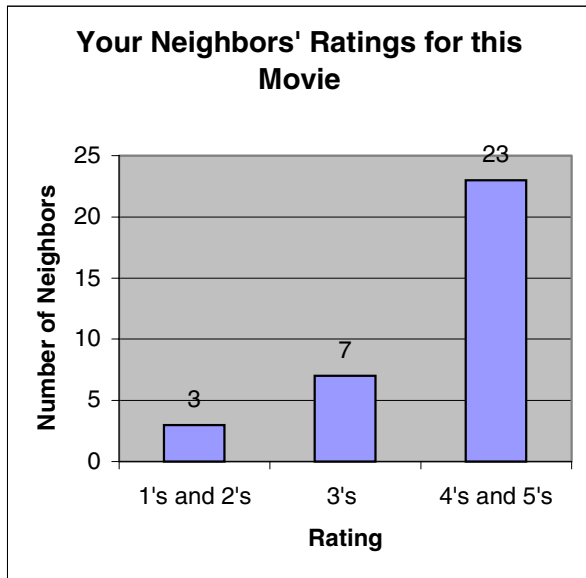


Figure 2. A histogram of neighbors' ratings for the recommended item, with the "good" ratings clustered together and the "bad" ratings clustered together, and the ambivalent ratings separated out. The result is that the user has to do only a single binary visual comparison to understand the consequence of the explanation. This was the best performing explanation.

Stating positive past performance of the ACF system was just compelling as demonstrating the ratings evidence behind the recommender. The exact explanation was "MovieLens has predicted correctly for you 80% of the time in the past."² This highlights the fact that in many cases, if a recommendation source is accurate enough, we may not really care on how the recommendation was formed. While this form of explanation is useful for setting the context, it is not valuable for distinguishing between different recommendations from the same system. However, recommendation-specific explanation information could be introduced by providing explanations such as "MovieLens has been 80% on movies similar to this one."

Explanation 5 – movie similarity – was "This movie is similar to 4 other movies that you rated 4 stars or higher." This kind of explanation can either be determined using content analysis or identifying movies that have correlated rating patterns.

The success of explanation 6 – favorite actor/actress - shows that domain specific content features are an important element of any explanation. However, notice the unusually high variance. Clearly there is a division between those who evaluate movies based on actors/actresses and those who don't.

² It is important to note that this is the only explanation that is not based on actual data. We extrapolated this number from our experiences with observation of the accuracy of our prediction algorithm.

Some explanations (18–21) had significantly lower mean response than the base cases. Poorly designed explanations can actually decrease the effectiveness of a recommender system. This stresses the importance of good design of explanations interfaces.

One of the key parameters in ACF systems is the similarity between the user and the neighbors. It is often the case that some of the neighbors chosen do not really share that much in common, so indicating the similarity can be important. Yet explaining similarity is tricky, since the statistical similarity metrics that have been demonstrated as the most accurate such as correlation[5] are hard to understand for the average user. For example, in the recommendation explained in this study the average correlation was 0.4, which we recognize from experience as being very strong for movie rating data. However, users are not aware that correlations greater than 0.4 are rare; they perceive 0.4 to be less than half on the scale of 0 to 1. This highlights the need to recode the similarity metric into a scale is perceptually balanced. In this specific case, we might recode the correlations into three classes: good, average, and weak.

It is interesting to note that external "official" rating services such as awards or critics did not fare particularly well (explanations 8 and 17). This indicates that users believe personalized recommendation to be more accurate than ratings from critics, a fact that has been shown by previous work[6].

Prior to the main study, we performed a small pilot study where we had the opportunity to interview participants after they took the survey. From these interviews, we learned that many users perceived each "recommendation" as having been generated using a different model – which was then explained. Each explanation was changing the user's internal conceptual model of how the recommender computed predictions. In the primary study, we attempted to control for this effect by clearly stating to study participants up front that the model was going to be the same in each case.

EXPERIMENT 2 – ACCEPTANCE AND FILTERING PERFORMANCE

In the previous section, we addressed the first research question. In this section, we present an experiment that addresses the remaining two research questions: (2) can explanations improve acceptance of automated collaborative filtering (ACF) systems and (3) can explanations improve the filtering performance of users?

Hypotheses

The goal of this experiment was to test two central hypotheses related to the research questions:

Hypothesis 1: adding explanation interfaces to an ACF system will improve the acceptance of that system among users.

Hypothesis 2: adding explanation interfaces to an ACF system will improve the performance of filtering decisions made by users of the ACF system.

Design

The experimental subjects were users of the MovieLens web-based ACF movie recommender. A link inviting users to try experimental interfaces was placed on the front page, and users volunteered to participate in approximately a month-long study. Experimental subjects were assigned randomly, either to a control group or to a group that was presented with a new experimental explanation interface. Members of control groups either saw the standard MovieLens interface or saw the standard interface with aesthetic changes to encourage them to believe they were seeing a significantly different system.

Figures 3 and 4 depict two of the explanations interfaces shown. Figure 3 presents a simple discrete confidence metric, while Figure 4 presents graphically the distribution of ratings for a movie within the user’s neighborhood, based on the similarity of each neighbor.

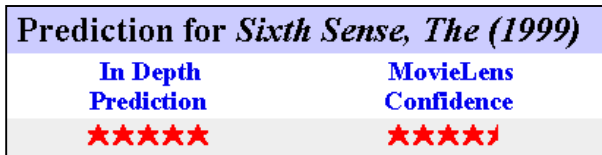


Figure 3. A simple confidence interval as an explanation for a movie recommendation.

Each experimental subject was given a survey upon entering and leaving the experiment regarding their impressions of the MovieLens site. These surveys were used to assess how explanation interfaces might affect the acceptance of ACF systems.

Each experimental subject was asked to return to use MovieLens how they normally would for recommendations, but to return to MovieLens whenever they saw a new movie and fill out a mini-survey with the following questions:

- Which movie did you see?
- Did you go because you thought you would enjoy the movie or did you go for other reasons (such as other viewers)?
- Did you consult MovieLens before going?
- If you consulted MovieLens, what did MovieLens predict?
- How much did MovieLens influence your decision?
- Was the movie worth seeing?
- What would you now rate the movie?

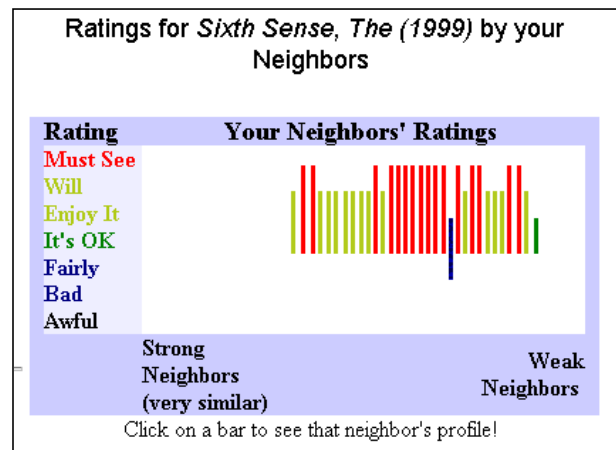


Figure 4. A screen explaining the recommendation for the movie “The Sixth Sense.” Each bar represents a rating of a neighbor. Upwardly trending bars are positive ratings, while downward trending ones are negative. The x-axis represents similarity to the user.

Results

210 users participated in this study, filling out 743 mini-surveys. In 315 of those mini-surveys, users consulted MovieLens before seeing the movie. In 257 of those mini-surveys, MovieLens had some effect on the user’s decision to see the movie. In 213 (83%) of the cases where MovieLens had an effect on the decision, the MovieLens recommendation was not the sole reason for choosing a movie.

Figure 5 shows the filtering performance of each experimental group. There was no statistically significant difference between any two experimental groups (based on a one-way ANOVA with $\alpha = 0.05$).

In exit surveys given at the end of the study, users in non-control groups were asked if they would like to see the explanation interface they had experienced added to the main MovieLens interface. 97 experimental subjects filled out the exit survey. 86% of these users said that they would like to see their explanation interface added to the system.

As part of the exit surveys, users were given the opportunity to provide qualitative feedback in the form of text comments. They were asked what they liked most about the explanation interfaces, what they like least about the explanation interfaces, and given the chance to provide suggestions and comments. The qualitative feedback from all those who responded (60 users) was almost entirely positive. Comments ranged from

“It made sense of what seemed to be somewhat senseless numerical ratings”

to

“I could see the variety of responses to a film which corresponds to what I do with my friends. It helps me see how strongly they felt and the power or range of that diversity which always helps me be prepared for a film which evokes powerful response.”

Some users were particularly excited with the ability to view the ratings of their neighbors. The viewable ratings profiles gave some more substance and reality to the previously invisible “neighbors.” Several users asked for features to explore their interests with neighbors further (i.e. just show me the movies we agreed on), while others wanted to meet and converse with their neighbors.

Another user wanted to be able to bookmark certain users, so she could return and see what movies they were going to see.

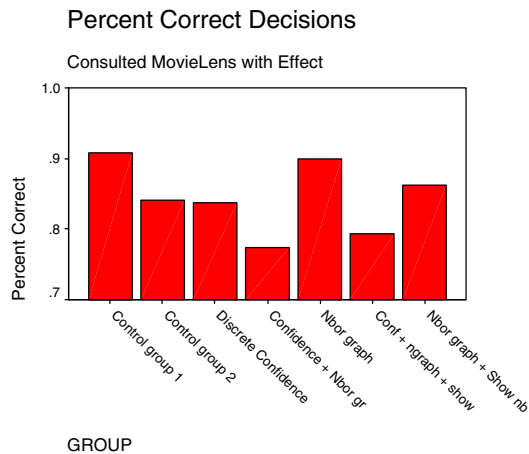


Figure 5. Percentage of correct movie decisions users made while using different versions of the explanation system. The first two bars represent control groups where no explanation interface was seen. The data was of extremely high variance, with none of the differences being statistically significant.

The majority of the negative comments stemmed from the question “What did you like least about the explanation interfaces?” and were related to inadequacies in the prediction algorithm and not in the explanation interface. By using the explanation interfaces, users discovered many predictions that were based on small amounts of data or neighbors who weren’t that similar. It was this that they complained about, not the quality of the explanation interfaces.

Analysis

The overwhelming request to see the explain feature added surveys to the system and the supporting positive remarks from the text comments indicates that users see explanation as a valuable component of an ACF system. The experimental subjects successfully used the system to identify predictions based on small amounts of data or neighbors that weren’t that similar.

The filtering performance measurements performed during this study were inconclusive. The results were confounded primarily by lack of good data. Most of the filtering decisions reported by the study participants were made without consulting MovieLens first, even in the groups that received explanations. There was also a considerably large amount of uncontrolled variance, especially between users. A more

controlled study would undoubtedly reveal the true effect of explanations on decision performance.

One of the key components to the explanations that we built was the graph of neighbors’ neighbors ratings shown in Figure 4. We believed this graph to be exceptionally effective at conveying a large amount of data about the prediction in a small amount of space. Most “experts” to whom we demonstrated this graph were impressed by its utility. However, though the process of performing this experiment and experiment 1, we have learned that while this graph is preferred by expert users, it is too complex for ordinary users. The confusion factor related to this graph may have affected people decisions not to use the explanation facilities more frequently. However, experiment 1 has demonstrated what the most effective explanation components are, which will be useful in designing the next generation of explanation interface.

SUMMARY AND CONCLUSIONS

Explanations have shown themselves to be very successful in previous work with expert systems. From this knowledge, it seems intuitive that they will prove to be successful in interfaces to automated collaborative filtering systems. The challenges will be to extract meaningful explanations from computational models that are more ad hoc than rule-based expert systems, and to provide a usable interface to the explanations. The result will be filtering systems that are more accepted, more effective, more understandable, and which give greater control to the user.

In this paper, we have explored the utility of explanations in automated collaborative filtering (ACF) systems. We have explored theoretically and experimentally three key research questions related to explanations in ACF systems.

What models and techniques are effective in supporting explanation in an ACF system?

ACF systems are built around complex mathematical models. Knowing exactly what and how to explain is not straightforward. We have presented an approach that develops the key process and data components of an explanation based on the user’s cognitive model of the explanation. Furthermore, we have performed an experiment to identify how compelling each of the identified explanation components is to the user. Rating histograms seem to be the most compelling ways to explain the data behind a prediction. In addition, indications of past performance; comparisons to similar rated items; and domain specific content features, such the actors and actresses in a movie are also compelling ways to justify a high recommendation.

Can explanation facilities increase the acceptance of automated collaborative filtering systems?

We hypothesized that adding explanation interfaces to ACF systems would increase their acceptance as filtering systems. Through an experiment with 210 users of the MovieLens web-based movie recommender, we have demonstrated that

most users value the explanations and would like to see them added to their ACF system (86% of survey respondents). These sentiments were validated by qualitative textual comments given by survey respondents.

Can explanation facilities increase the filtering performance of ACF system users?

We began an initial investigation into measuring the filtering performance of users both with and without the explanation interface. We believe that explanations can increase the filtering performance. Unfortunately, due to many factors, we were unable to prove or disprove our hypothesis. Users perform filtering based on many different channels of input, and attempting to isolate the affect of one filtering or decision aid requires well controlled studies, which are hard to perform through a web-site with users that you never meet.

FUTURE WORK

We believe that it is important to identify the characteristics of an explanation system that will result in improved filtering performance by the users. We should be able to use our results from this work and design more compelling explanation interfaces, which in turn will have a more dramatic effect on both usage and filtering performance. Then a more controlled study measuring filtering performance should reveal the true effect of explanations on filtering task performance.

Recommender systems with explanation interfaces exist in a dimension that varies from persuasive to informational. It would be interesting to explore how users respond to persuasive versus informational interfaces. In certain situations, a persuasive interface may result in improved filtering performance or vice versa.

EXPERIMENTAL NOTES

The two experiments described in this paper were actually performed in the reverse order that they are described. We have exchanged the order to communicate our ideas more effectively.

ACKNOWLEDGMENTS

We thank Tony Lam for the development of an experimental framework for MovieLens and Tim Getsch and Josh Chu for helping to build the explanation interfaces.

REFERENCES

1. Buchanan,B., Shortliffe,E., (Eds.) 1984. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, MA.
2. Lauer,T.W., Peacock E., Graesser,A.C., (Eds.) 1985. *Questions and Information Systems*. Lawrence Erlbaum and Associates.
3. Graesser,A.C., Black,J.B., (Eds.) 1985. *The Psychology of Questions*. Lawrence Erlbaum and Associates.
4. Dahlen,B.J., Konstan,J.A., Herlocker,J.L., Good,N., Borchers,A., Riedl,J., 1998. Jump-starting movielens: User benefits of starting a collaborative filtering system with "dead data". University of Minnesota TR 98-017.
5. Herlocker,J.L., Konstan,J.A., Borchers,A., Riedl,J., 1999. An algorithmic framework for performing collaborative filtering. *Proceedings of the 1999 Conference on Research and Development in Information Retrieval*.
6. Hill,W., Stead,L., Rosenstein,M., Furnas,G.W., 1995. Recommending and Evaluating Choices in a Virtual Community of Use. *Proceedings of ACM CHI'95 Conference on human factors in computing systems*. Denver, CO., (pp. 194-201).
7. Horvitz,E.J., Breese,J.S., Henrion,M., 1988. Decision Theory in Expert Systems and Artificial Intelligence. *International Journal of Approximate Reasoning* 2 (3), 247-302.
8. Johnson,H., Johnson,P., 1993. Explanation facilities and interactive systems. *Proceedings of Intelligent User Interfaces '93*. (pp. 159-166).
9. Konstan,J.A., Miller,B.N., Maltz,D., Herlocker,J.L., Gordon,L.R., Riedl,J., 1997. GroupLens: Applying collaborative filtering to Usenet news. *Communications of the ACM* 40 (3), 77-87.
10. Miller,C.A., Larson,R., 1992. An Explanatory and "Argumentative" Interface for a Model-Based Diagnostic System. *Proceedings of User Interface Software and Technology (UIST '92)*. (pp. 43-52).
11. Norman,D.A. 1989. *The Design of Everyday Things*. Currency-Doubleday, New York.
12. Resnick,P., Iacovou,N., Suchak,M., Bergstrom,P., Riedl,J., 1994. GroupLens: An open architecture for collaborative filtering of netnews. *Proceedings of 1994 Conference on Computer Supported Collaborative Work*. (pp. 175-186).
13. Shardanand,U., Maes,P., 1995. Social Information Filtering: Algorithms for Automating "Word of Mouth". *Proceedings of ACM CHI '95*. Denver, CO., (pp. 210-217).
14. Toulmin S.E. 1958. *The Uses of Argument*. Cambridge University Press