Interfaces for Eliciting New User Preferences in Recommender Systems

Sean M. McNee, Shyong K. Lam, Joseph A. Konstan, and John Riedl

GroupLens Research Project
Department of Computer Science and Engineering
University of Minnesota–Twin Cities
Minneapolis, MN, 55455 USA
{mcnee, lam, konstan, riedl}@cs.umn.edu

Abstract. Recommender systems build user models to help users find the items they will find most interesting from among many available items. One way to build such a model is to ask the user to rate a selection of items. The choice of items selected affects the quality of the user model generated. In this paper, we explore the effects of letting the user participate in choosing the items that are used to develop the model. We compared three interfaces to elicit information from new users: having the system choose items for users to rate, asking the users to choose items themselves, and a mixed-initiative interface that combines the other two methods. We found that the two pure interfaces both produced accurate user models, but that directly asking users for items to rate increases user loyalty in the system. Ironically, this increased loyalty comes despite a lengthier signup process. The mixed-initiative interface is not a reasonable compromise as it created less accurate user models with no increase in loyalty.

1 Introduction

Recommender systems help users sort through the vast quantities of information deluging them every day. These systems build user models based on users' stated preferences for items in a domain (e.g., ratings of movies) and use these models to generate personalized recommendations for the user.

All of these systems share a common problem: generating recommendations for new users of the system. One solution is to make non-personalized recommendations at first, dynamically learning the user's preferences and incrementally adjusting the user model. Brusilovsky, however, claims that this approach will not learn accurate enough models, and suggests that the system will still need to elicit information from the user [3].

In this paper, we explore the question of how to elicit user preference information in the context of recommender systems. We contrast system-controlled approaches, where the system decides which items the user can rate, with methods that allow the user to specify some or all of the items to be rated.

2 Eliciting New User Preferences

There are several effective strategies for creating user models in recommender systems. One common strategy is to use the text of items a user likes to build a keyword profile and then recommend new items that match the profile. These content-based filtering systems work well when the content of items is amenable to machine processing. Other effective strategies for building models include the use of Bayesian networks [2], adaptive decision trees [13], and rule-based systems [5].

An alternative family of strategies, called collaborative filtering (CF), works directly with users' preferences to build a user model [11]. Typically, CF systems generate recommendations for a target user by generating a "neighborhood" of other like-minded users. The preferences found in the neighborhood are then used to create personalized recommendations for the target user.

CF systems can make recommendations in any domain and are more likely than content-based systems to recommend serendipitous items, items to which users may never have been exposed but may like. However, CF systems are particularly afflicted by the new user problem [11] as they typically require several ratings from a new user before they can find similar users with which to make personalized recommendations.

The obvious solution: ask new users to rate items. But how should the system select items to ask the users about that are both useful to the system and will not bore or alienate new users? Asking a user to rate the movie "The Godfather" doesn't help a movie recommender, since most people have seen "The Godfather" and enjoyed it. Asking the user to rate the movie "Grand Exit" from 1935 is equally useless, since the odds that the user has seen the movie are quite low–regardless of how much the system would gain from knowing the user's rating of that movie. Therefore, a solution to this problem should attempt to balance cost to the new user, in terms of frustration, time, and effort spent, with the amount of information the system would gain from each rating.

Pennock et al. proposed that information-theoretic calculations might prove useful for solving this new user problem in a CF-based recommender [9]. Rashid et al. showed that choosing items with a high score on a metric that combined an entropy calculation with a measure of the item's popularity reduced users' rating effort while maintaining high recommendation quality [10]. This metric is simple to calculate and was more effective than other metrics that personalized the items presented to the users.

Brusilovsky discusses three approaches to collaborative user modeling in building personalized hypermedia systems [3]. The first approach automatically updates the user model based on user-provided information. This is implicit in collaborative filtering since the user model is based on user-provided ratings. The entropy calculations from Rashid et al. employ the second of these approaches: to directly ask the user for data the system needs to update the model. Brusilovsky's last approach allows users to make direct changes to their model. In the context of the new user problem, this translates into allowing the users to directly specify what items they are interested in, instead of having the system question them.

3 New Interfaces: Who's In Charge?

In order to effectively test how these different methods of creating user models compare to each other in a recommender system, we created user interfaces to reflect these different model-building approaches. We then compared these alternative interfaces for new users to enter preference information in a CF-based recommender system.

Conventional user interfaces fall into a spectrum of use from user-controlled interfaces (such as a word processor) to system-controlled interfaces (such as an ATM) [1, 12]. Splitting the difference between the two extremes are mixed-initiative interfaces [1, 5, 8, 14]. These interfaces are usually viewed as a dialogue between the system and the user with each side contributing to the interaction. Mixed-initiative interfaces can take many forms from augmented flight control systems where the computer monitors and maintains various aspects of controlling the plane reporting potential problems to the pilot, to 'smart' CAD programs that recommend modifications to various building structures for compliance with safety and fire codes [6].

The entropy-based calculations from Rashid et al. are an example of a system-controlled interface for soliciting new user preference information. Using interfaces of this type the system can efficiently gather information it needs to create high quality user models. Intelligently choosing which items to ask a new user is very important, but it limits the user to choose only from the items the system selects.

Instead of prompting the user to rate particular items, a user-controlled interface allows users to tell the system which items they want to rate. Interfaces of this type may be more difficult for new users since the users will be required to remember items with no context or prompting. At the same time, this interface will allow users to more fully express their interests across the entire universe of items and create a profile that they feel more accurately reflects who they are.

Lately, mixed-initiative interfaces have focused on the interaction of multiple system agents and one human user making decisions together [1, 8]. We instead will focus on the idea of a mixed-initiative interface as providing a choice to the user of a system. Thus, our mixed-initiative interface is a composite of the system-controlled and user-controlled interfaces (Figure 1). As such, it should provide the benefits of both interfaces by generating high quality user models that the users believe accurately reflect their tasks and opinions.

In summary, we hypothesize that human-controlled interfaces will have higher user satisfaction, whereas system-controlled interfaces will generate better user models. The mixed-initiative interface should perform better than both of the other interfaces, i.e., have the highest user satisfaction and generate the best user models.

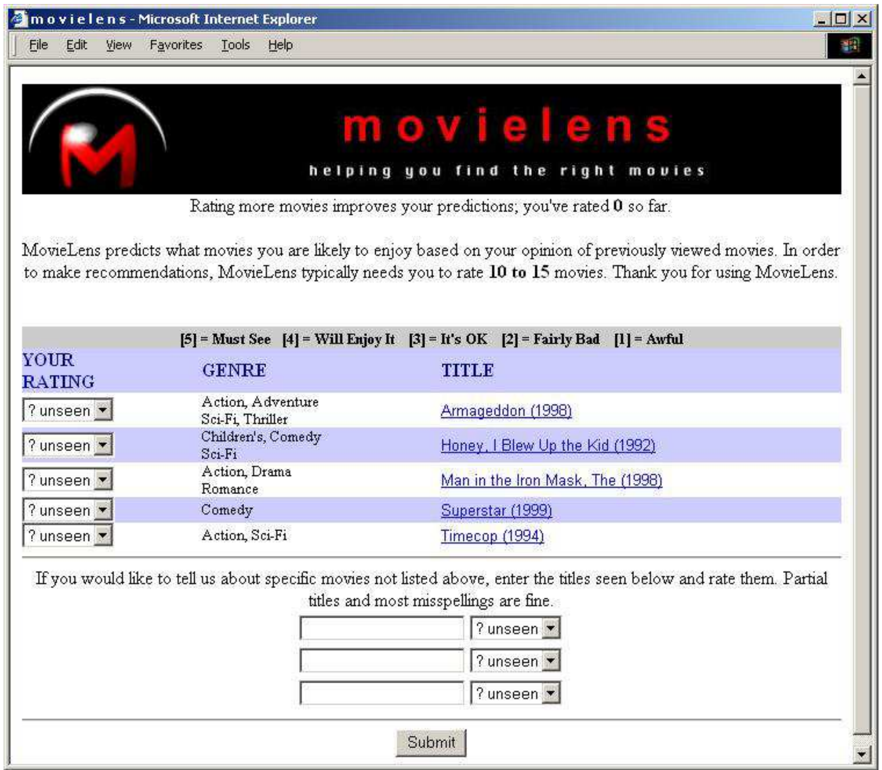

Fig. 1. The mixed-initiative interface includes elements from the system-controlled (top half) and user-controlled (bottom half) interfaces

4 Experiment

We performed a study using new users of MovieLens (www.movielens.org), an online collaborative filtering-based movie recommender system. New users registering for a MovieLens account were asked to volunteer for an experiment concerning alternative interfaces for rating movies. The signup process requires new users to rate a minimum number of movies, typically 10 to 15, on a 5-star scale before being allowed to receive personalized recommendations.

Users agreeing to participate were randomly placed into one of three groups: a group that received a system-controlled rating interface, one that received a user-controlled rating interface, and one that received a mixed-initiative rating interface.

The system-controlled interface is based on the log(popularity) * entropy movie selection metric presented in [10]. The system presented ten titles at a time, randomly chosen from the top 250 titles as scored by the metric. This interface is similar to the current signup method used in MovieLens, and is represented in the top half of Figure 1.

The user-controlled interface presented the user with five empty text fields per page. The user is asked to enter the titles of movies to rate. Upon submission, the user confirmed her entries by selecting her choice from a list of possible title matches as generated by a nearest edit distance algorithm. The page gave no hints or suggestions as to what movies should be rated and is similar to the bottom half of Figure 1.

The mixed-initiative interface is a hybrid of the system and user-controlled interfaces and is shown in Figure 1. Five system-selected movies and three text fields are shown on each page. The user can choose to rate the presented movies, enter titles of her own choice, or both. Since this interface both presented movies to rate and provided open text boxes, the presented movies could have acted as hints for items to rate via the textbox interface.

4.1 Experimental Method

Users were asked to rate movies using the given interface until they had rated 12 movies. They were then given the opportunity to continue rating movies using their assigned interface or to end the signup process and start receiving recommendations. Once the user finished rating items, they were asked to complete a brief survey before entering MovieLens. The survey consisted of the following five questions, each asked on a 5-point scale.

1. How well did you understand this signup process? Answers ranged from "Did Not Understand" to "Completely Understood".

2. How hard was it for you to find/enter movies that you have seen? Answers ranged from "Easy" to "Difficult".

3. How much time did it take you to complete this signup process? Answers ranged from "Short Period" to "Long Period".

4. How would you describe this interface for entering movie ratings? Answers ranged from "Simple" to "Complex".

5. How well do the movies you rated represent your movie tastes? Answers ranged from "Not At All" to "Perfectly".

4.2 Results

A total of 225 new users participated in the experiment. 192 users successfully completed the signup process and 163 completed the survey.

For each group, Table 1 shows the percentage of users who completed the signup process, the average number of ratings provided, and the average time required for each user to provide 12 ratings. Mixed-initiative users had the highest completion rate while system-controlled users rated the most items on average. Users receiving the user-controlled interface took far more time to complete the process and did not provide as many ratings as users in the mixed-initiative or system-controlled groups.

Table 1. Statistics on the Completion of the Signup Process

	User	Mixed	System
Percentage of Users Finishing Signup	80%	92%	88%
Mean Number of Ratings	14	30	36
Mean Time to 12 Ratings	565 sec	382 sec	286 sec

We used Mean Absolute Error (MAE), a common measure of recommendation accuracy [7], to measure the quality of the user models for each interface. MAE is the average absolute difference in actual rating for an item compared to the prediction the system would have generated for that item. Thus, a high quality user model would have low MAE.

The "30 based on 12" and "30 based on N" MAE calculations were run by taking either the first 12 or all N ratings from the signup process as the user's model and then generating predictions for 30 randomly selected ratings the user entered into the system after finishing signup. For the "All But One" calculation [2], one of the first 12 user ratings was removed and a prediction was generated for it using the other 11 ratings as the user model. This was repeated for each of the first 12 ratings. In all cases, we computed the MAE for each individual user, and averaged the per-user MAE's to compute an overall measure of quality.

Table 2 shows the MAE results. At odds with our hypotheses, the user-controlled interface had the best MAE, followed closely by the system-controlled interface. The mixed-initiative interface generated the worst user models. Using either the first 12 ratings or all ratings users provided during the startup process made no difference. The "All But One" condition has higher MAEs across the board and shows a difference between the user-controlled interface and the other interfaces.

Table 2. Mean Absolute Error (MAE) calculations for the three signup interfaces, based on a 5-star rating scale. Lower is better

	User	Mixed	System
30 based on original 12	0.79	0.88	0.81
30 based on all of signup (N)	0.78	0.88	0.81
"All But One" on first 12	0.90	0.99	1.03

Table 3 shows the survey results. Users understood the signup process and felt the interface was not complex. Even though the user-controlled group took twice as long to finish, most reported that the signup process took a short amount of time to complete. In agreement with our hypotheses, the user-controlled group also felt that the items they entered more accurately represented their tastes than those entered by the other two groups. Finally, the system-controlled interface was voted to be the hardest to locate and rate items that the user had already seen.

Table 3. Percentage of users giving positive responses to survey questions

	User	Mixed	System
Understood signup process	95%	96%	93%
Easy to find items to rate	59%	67%	40%
Short amount of time to complete	70%	75%	69%
User interface was not complex	77%	88%	88%
Preferences reflect tastes	63%	30%	19%

Although Table 2 showed that the user-controlled group had the lowest percentage of users finishing the signup process, the longevity statistics in Table 4 show that these users had the highest number of logins, highest percentage of active users (defined as users who gave at least 10 ratings after signup), and the highest number of total ratings provided in the system after signup was completed. In contrast, the mixed-initiative group had the most users finish the signup process, but tied for the lowest percentage continue to use MovieLens and provided fewer ratings than the user-controlled group.

Table 4. Statistics about longevity of users from the different interfaces. Longevity was measured as active use in the system after 25 days

	User	Mixed	System
Percentage of active users	35%	21%	21%
Mean ratings entered post signup	225	159	125
Mean number of user logins, post signup	5.1	3.8	3.2

5 Discussion

Even though the user-controlled group did not feel they were spending a long time in the signup process, compared to users from the other two groups, they in fact spent nearly twice as long (9.4 vs. 5.6 minutes on average). Because they received no prompting or hints for movie titles, it took time to recall items to rate. We believe that users in the user-controlled group did not notice the effort because asking them to recall titles created focus and engagement with the system. These users are the most likely to keep returning to the system and provide more ratings. They also thought the system best understood their tastes. In some sense of the word, we could say these users were more "loyal" to the system.

The fact that the user-controlled group had the best MAE results is surprising and very intriguing. To confirm our results, we re-ran the MAE calculations against another CF-based recommender engine. This second engine produced uniformly higher MAEs but still maintained the relative ordering of the interfaces for all three experiments. These results confirm that users who directly choose items can do as good of a job defining their user models as a system probing users for the information. Users know what is good for them.

The difference in MAE for the system-controlled interface between the "All But One" test and the "Predict 30" tests also requires investigation. The items rated in the system-controlled interface were taken from the log(popularity) * entropy metric. Further analysis shows that high entropy movies are harder to make accurate predictions on; these items have a higher MAE for all users. That is, even though high entropy items lead to good user models, even the best user model will have trouble making predictions for these items. Because of this property, we expect the "All But One" metric to have higher MAEs for both the system-controlled and mixed-initiative interfaces, and this is exactly what we see in Table 2.

While this may explain the high MAE for the mixed-initiative interface in the "All But One" test, it does not explain its poor performance for the other two conditions. Just as the mixed-initiative interface can be split into two halves, the ratings gathered from users in this group can be split into two distributions: a system-controlled part and a user-controlled part. Each distribution looks similar to the distributions found for its respective parent interface. More importantly, each half has different MAE scores: for example, in "30 Based on N", the system-controlled interface had an MAE of 0.90 and the user-controlled interface had an MAE of 0.64.

We theorize the two interfaces interacted with each other; by explictly listing titles, the system-controlled half provided movie hints that could have been used by users to generate titles for the user-controlled half. Seeing this, users internally renormalized their ratings, rating items in the user-controlled half higher and items in the system-controlled half of the interface lower than the users from the respective parent interfaces. This interaction is important as it altered the makeup and lowered the quality of the user model the interface generated. As system designers, we need to be aware of how subtle interactions in interface elements can profoundly affect the user models our systems generate.

Although the mixed-initiative interface produced the worst models, its users voted the interface high in terms of ease of rating items they had seen. Moreover, users rated this interface as taking the shortest amount of time to complete the signup process, even though it took longer for users to complete than the system-controlled interface. Again this can be traced to higher user engagement with the system. Mixed-initiative users, while not as loyal as the user-controlled, provided more ratings than either of the other groups, and felt the system understood their tastes better than the system-controlled group. They did not, however, come back as often.

6 Conclusion

Both system- and user-controlled interfaces generate high quality user models. New users tended to take a longer period of time to complete the signup process when using a user-controlled interface when compared to a system-controlled interface. Moreover, these users did not notice the extra time required to complete the signup process and also felt that the system better understood their tastes. This comes at a price, however. Only 80 percent of users in the user-controlled group finished the signup process, compared to 92 percent for mixed-initiative and 88 percent for system-controlled. Of those that finished the signup process, there was a 15 percent higher user retention rate from the user-controlled group than the other two groups.

We theorize these users were more enagaged and focused when using the system because of the higher burden and extra control associated with the user-controlled interface. The economic theory of reciprocity suggests in part that users who put forth more effort into a system will expect more from the system in return [4]. Thus, by creating a signup process with higher burden and extra control, those users who complete the process are more likely to be loyal users.

A hybrid mixed-initiative interface does not provide a sensible alternative to the two other interfaces. Users of this interface were the least likely to return and had the worst user models. We also believe that the components of a hybrid interface may interact with each other in unexpected ways, affecting the quality of the user model the system generates for users of the interface.

It appears that there will often be a tradeoff between giving users control and increasing their workload. Depending on the goals of the application being built, system designers need to decide whether low user burden or high user loyalty is more important and choose an interface to elicit user preferences and generate user models appropriately.

7 Acknowledgements

We would like to thank Dan Cosley and Istvan Albert for their suggestions and comments on early drafts of this paper. This work is supported by grants from the NSF (DGE 95-54517, IIS 96-13960, IIS 97-34442, IIS 99-78717, and IIS 01-02229), and by Net Perceptions, Inc.

References

1. Allen, J. (1999) Mixed Initiative Interaction. Proceedings IEEE Intelligent Systems. 14 (6), pp. 18-23.
2. Breese, J., Heckerman, D. and Kadie, C. (1998) Empirical Analysis of Predictive Algorithms for Collaborative Filtering. Proceedings of UAI 1998. Pp. 43-52.

3. Brusilovsky, P. (1996) Methods and techniques of adaptive hypermedia. User Modeling and User-Adapted Interaction. 6 (2-3), pp. 87-129.
4. Fehr, E. and Gächter, S. (2000) Fairness and Retaliation: The Economics of Reciprocity. Journal of Economic Perspectives. 14(3), pp. 159-181.
5. Flemming, M. and Cohen, R. (1999) User Modeling in the Design of Interactive Interface Agents. Proceedings of User Modeling 1999. Pp. 67-76.
6. Fu, M. C., Hayes, C. C., and East, E. W. (1997) SEDAR: Expert Critiquing System for Flat and Low Slope Roof Design and Review. Journal of Computing in Civil Engineering. 11(1), pp. 60-69.
7. Herlocker, J., Konstan J, A., Borchers, A., and Riedl, J. (1999) An Algorithmic Framework for Performing Collaborative Filtering. Proceedings of SIGIR 1999. Pp. 230-237.
8. Horvitz, E. (1999) Principles of Mixed-Initiative User Interfaces. Proceedings of CHI 1999. Pp. 159-166.
9. Pennock, D., and Horvitz, E. (2000) Collaborative Filtering by Personality Diagnosis: A Hybrid Memory- and Model-based Approach. Proceedings of UAI 2000. Pp. 473-480.
10. Rashid, A. M., Albert, I., Cosley, D., Lam, S. K., McNee, S. M., Konstan, J. A., and Riedl, J. (2002) Getting to Know You: Learning New User Preferences in Recommender Systems. Proceedings of IUI 2002. Pp. 127-134.
11. Resnick, P., Iacovou, N., Sushak, M., Bergstrom, P., and Riedl, J. (1994) GroupLens: An Open Architecture for Collaborative Filtering of Netnews. Proceedings of CSCW 94. Pp. 175-186.
12. Shneiderman, B. (1998) Designing the User Interface, third edition. Reading, MA: Addison Wesley.
13. Stolze, M. and Strobel, M. (2001) Utility-Based Decision Tree Optimization: A Framework for Adaptive Interviewing. Proceedings of User Modeling 2001. Pp. 105-116.
14. Terveen, L. G. (1993) Intelligent Systems as Cooperative Systems. International Journal of Intelligent Systems. 3, 2-4, pp. 217-250.