Prediction Strategies: Combining Prediction Techniques to Optimize Personalization

Mark van Setten[1], Mettina Veenstra[1], Anton Nijholt[2]

[1] Telematica Instituut, P.O. Box 589, 7500 AN, Enschede, The Netherlands
{mark.vansetten, mettina.veenstra}@telin.nl
[2] University of Twente, P.O. Box 217, 7500 AE, Enschede, The Netherlands
anijholt@cs.utwente.nl

Abstract. An important step in providing personalized information, such as personalized electronic program guides, is predicting the level of interest for a specific user in information, such as TV programs. This paper describes a model that can be used to combine prediction techniques, which predict the level of interest, into prediction strategies. Our hypothesis is that prediction strategies optimize the prediction quality and that they lead to more reliable predictions, because they take into account the state of the system, the users and the information. Results of an initial experiment support this hypothesis.

1 Introduction

In this age, more and more information is made electronically available and is accessible to many people. This results in the problem of information overload. People have difficulties in getting information that is relevant and interesting for them. This problem exists in several fields, such as electronic stores, digital news provision, digital libraries, and TV systems. One of the solutions for this problem is making information systems adaptable to the user, assuring that only information interesting for the user is retrieved and presented in a way that is suitable for that user.

Our research is applied in the domain of future personalized TV systems [11] and encompasses more than research into personalization [10]. It also includes research into new user interfaces, distributed search and retrieval and efficient video browsing.

However, the focus of this paper is on personalized selection of information. We present a model that makes it possible to combine prediction techniques in order to optimize prediction quality and which results in more reliable predictions. After discussing the model, results of an initial experiment with this model are provided.

2 Personalized Information Selection

Personalized selection of information is part of several tasks, e.g. determining what programs to record, ordering and filtering electronic program guide (EPG) data and ordering and filtering search results. Personalized selection consists of two phases:

1. Predict the level of interest a user will have in a piece of information;
2. Adapt the information based on those predictions, such as re-ordering and/or filtering the set of information.

We focus on predicting the level of interest for a user. There are several types of techniques to do this. On one hand, there are content-based techniques: structured querying (e.g. SQL), information filtering [5], case-based reasoning (CBR) [6] and content-based category selection (e.g. selection based on genres). All these techniques look inside information (content and metadata) to determine how interesting it is for a user. The other techniques are social techniques, which do not look inside information, but base their predictions on (opinions of) other users: social filtering [2] [8], item-item filtering [3], social CBR (CBR that calculates similarity between users based on characteristics), Top-N (average opinion of all users) and social-based category selection (e.g. stereotypes [1]). Another technique is association rules [12], which is neither a specific content technique nor a specific social technique.

2.1 Prediction Strategies

Most of the currently available personalized information systems and research into these systems focus on the use of a single selection technique or a fixed combination of techniques [3] [9]. We believe that combining different techniques in a dynamic and intelligent way can provide better and more reliable prediction results.

By dynamic and intelligent combinations we mean that the combination of techniques should not be fixed within a system and that the combination ought to be based on knowledge about strengths and weaknesses of each technique and that the choice of techniques should be made at the moment a prediction is required, taking into account the three factors that cause the dynamics of personalization:

1. *The usage lifecycle*: When a new user starts using a personalized system, there is no knowledge available about him, making prediction techniques that rely heavily on knowledge about the user unsuitable. For new users, it is better to use techniques that rely less on knowledge about the user.
2. *The information lifecycle*: The lifecycle of information (content and metadata) also influences the suitability of techniques. On one hand, the amount and characteristics of available metadata determines the suitability of content-based techniques. On the other hand, the lack of user ratings makes social techniques unsuitable for new pieces of information.
3. *The system lifecycle*: The lifecycle of the system itself also determines the suitability of techniques. Several techniques rely on a certain amount of active users and/or an amount of available metadata that may not yet be available in new systems. In this situation, alternative techniques should be used.

Due to these dynamic factors, a model is needed that allows prediction techniques to be easily combined. Each combination should be chosen based on the most actual knowledge about the users, the information and the system at the moment of predicting. A combination of prediction techniques is called a prediction strategy. The next section describes our model for combining individual techniques into strategies.

2.2 Prediction Technique Model

Even though individual prediction techniques are quite different, it is possible to create a model in which all techniques can be embedded due to the basic nature of each prediction technique: each technique can calculate a predicted interest value for a piece of information for a given user (even structured querying, resulting in binary predictions). This forms the basis of our model. Naturally, each technique must normalize its predictions. We use the bipolar range from –1 to +1 (zero being neutral).

Several techniques, such as social filtering and CBR, are capable of learning from users in order to optimize their predictions. They learn from feedback provided by users. This means that feedback should also be incorporated in our model.

In order to make informed decisions within strategies, each technique exposes so-called validity indicators. These indicators are used to decide to what extent a technique should be used. E.g. a validity indicator from social filtering is the number of similar users that rated the piece of information. Social filtering can only make a good prediction if there are enough similar users who rated the item [2].

Optionally, prediction techniques can provide explanation data. Explanations provide transparency, exposing the reasoning and data behind a prediction and can increase the acceptance of prediction systems [4]. Explanations help users to decide whether to accept a prediction or not. However, we will not discuss explanations here.

The discussed aspects of prediction techniques result in our generic model shown in figure 1.

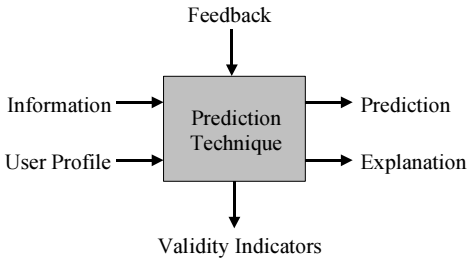

Fig. 1. Generic model of prediction techniques.

Prediction strategies consist of one or more prediction techniques and a set of rules to determine which technique(s) to use. These rules use the validity indicators of each prediction technique, which are based upon the user profile, the information for which a prediction must be made and the current state of the system. When using more than one technique, validity indicators also help to determine the weights between techniques when combining individual predictions into one prediction. However, in our initial experiment, predictions are not combined: strategies only select between techniques. Combining predictions will be investigated in future experiments. Within our generic model, a prediction strategy itself can be treated again as a prediction technique, allowing strategies to be nested.

Figure 2 shows a prediction strategy for predicting the level of interest in entertainment programs. We believe that such programs, like movies, mainly appeal to people's tastes instead of their rational interests. It uses two prediction techniques (social filtering and CBR) and one other prediction strategy (first time user strategy).

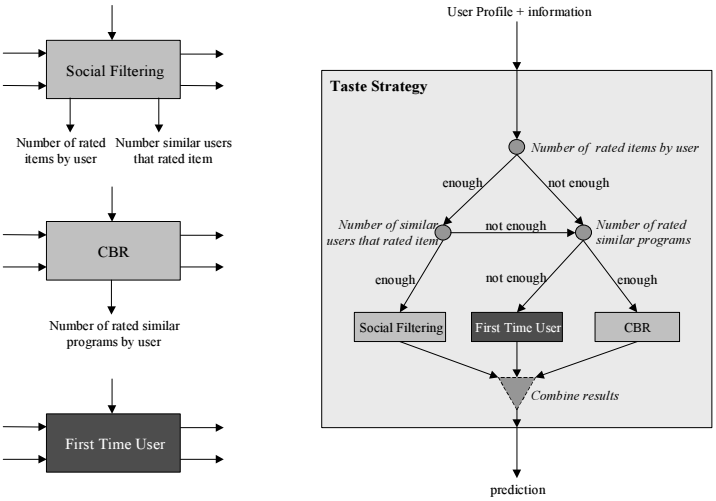

Fig. 2. Example prediction strategy for taste preferences of entertainment programs.

Two validity indicators of social filtering are used: the number of items rated by the user and the number of similar users that rated the item. These validity indicators are chosen based on research into social filtering techniques [2]. Social filtering only works well when the user has rated at least a certain amount of items and when there are enough similar users to base a prediction on. From the CBR technique, the validity indicator is the number of similar items rated by the user. No validity indicators are used from the first time user strategy, as this strategy is used as a fallback when other techniques are not suitable. The taste strategy exposes an indicator that shows how much an item appeals to the users' taste instead of their rational interests. Of course, the threshold values for these indicators, such as "enough", have to be instantiated before implementing the strategy. There are three ways to determine the thresholds: using results of existing research, determining them via experiments and/or by automatically learning them within the system.

3 Experiments

In order to test our model and validate the usefulness of prediction strategies, we performed an experiment in which individual prediction techniques were compared with a strategy. Because this was a first experiment for our model, we wanted to make sure that our prediction techniques and results were comparable with other research projects. For this reason, we decided to use the MovieLens data set of the GroupLens project[1].

In our experiment, we implemented a set of prediction techniques using our model. The prediction techniques we implemented were: a base technique that simulates a

[1] http://www.grouplens.org - the dataset consists of 100,000 ratings, 943 users and 1682 movies. The rating scale of 1 to 5 has been normalized in our experiment to –1 to +1.

system without predictions (always returns the neutral value); Top-N technique that calculates the average rate over all users; the user average technique that calculates the average rate given by the user; social filtering as used in the MovieLens system [2]; genre user average, which is a content-based category prediction technique, where for each genre the average rating by the user is calculated, which is then used to predict the interest for a movie by looking at the genres that movie belongs to; and a simple CBR technique based on genres, where movie genres are used to calculate the level of similarity between movies. The strategies implemented are the taste strategy (see figure 2) and a first time user strategy that uses the Top-N, user average and genre user average techniques.

3.1 Evaluation Measures

According to Herlocker [2], there are two good measures to evaluate prediction techniques:

1. *The mean absolute error (mae)*: this measures the average absolute deviation between a predicted rating and the user's actual rating. The lower the mae, the better the performance of a prediction technique;
2. *Coverage*: percentage of items for which a technique could generate a prediction. Some techniques cannot always provide a prediction. E.g. social filtering only generates a prediction when similar users can be found for the current user.

We believe that in some systems coverage is more important than in others. E.g. in a rental movie recommendation system, it is less important if a prediction cannot be made, as long as actual predictions are correct: it is not important if users rent a certain movie this week or in several months, as long as they will like the currently recommended movies. However, in a TV system, coverage is more important because most TV programs are only available at the moment of broadcasting. This means that we are interested in both prediction quality and coverage at the same time. For this reason, we combined both measures into a new measure: the global mean absolute error (gmae). This measure is the same as mae, but when a prediction cannot be made the neutral value is assumed (which is how users probably see a non prediction). To distinguish it from the original mae measure, we call the original measure the prediction mean absolute error (pmae). The pmae measure only calculates errors for predictions when all techniques can produce a prediction.

To compare the prediction techniques and the strategy throughout the system's life cycle, we divided the set of ratings into five groups of 20,000 ratings each. Group A consisted of the first 20,000 ratings (in time), Group B of the next 20,000 ratings, etc. When testing each group, the ratings of all previous groups were used for training.

To evaluate the prediction techniques throughout the user usage cycle, we looked in each of the five groups for users who started using the system within the time period of that group and who rated at least 150 movies in that same period. The ratings of these users where divided into two groups: the first 100 ratings and the remaining ratings. The first 100 ratings where used to evaluate techniques for users of whom little knowledge is available, the others to evaluate techniques for users of whom more knowledge is available (using the first 100 ratings as training data).

For all fifteen groups, we calculated the pmae and gmae for each technique and the taste strategy and performed paired samples T-tests (using a 95% confidence interval) to determine if differences between the results are statistically significant. In the next section, the main results are presented[2].

3.2 Results and Discussion

When comparing the results within the five groups of the system's lifecycle (left hand graph of figure 3), in every group, the taste strategy is significantly better than the individual techniques. Also for the first 100 predictions of new users (right hand graph of figure 3), the taste strategy performs significantly better or at least as good as the best individual technique. In groups B, C and E, there is statistically no significant difference between the taste strategy and the Top-N technique, but both are better than other techniques. This is acceptable, as the current implementation of the taste strategy only selects between techniques. This means that for each single prediction it can never perform better than the best individual technique. In turn, the gmae of the strategy should be statistically better or the same as the best technique in each group.

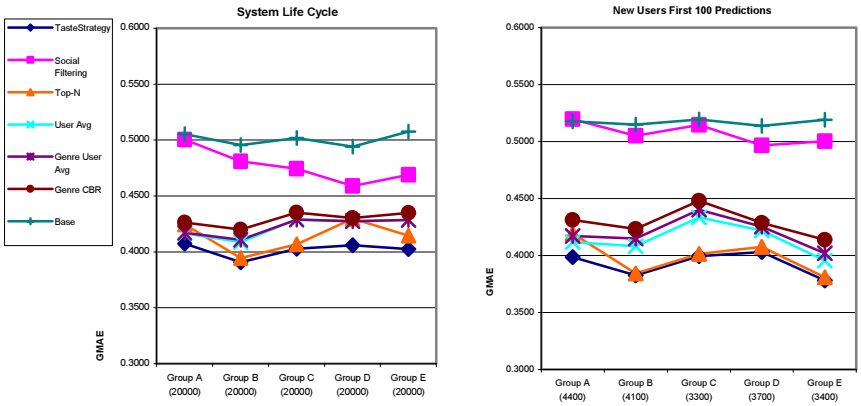

Fig. 3. Global Mean Absolute Error for the system life cycle test and the first 100 predictions of new users. The numbers below each group indicate the sample size (number of predictions).

The same argument applies to the test for the predictions for new users after the first 100 predictions (left hand graph of figure 4). In the first three groups, the taste strategy is statistically the same as the best techniques (the genre CBR technique in group A and the Top-N technique in groups B and C). In the last two groups the taste strategy is statistically better than any of the individual prediction techniques.

We also looked at the prediction accuracy only (pmae) in the three tests. The taste strategy performs almost always statistically better or at least as good as the best prediction technique. There are two exceptions. In two situations, the taste strategy does not predict as accurately as an individual technique. In the first stage of the

[2] Detailed data results can be found at: https://doc.telin.nl/dscgi/ds.py/View/Collection-4586

system lifecycle, genre CBR performs better. Social filtering also predicts statistically better in the second and third group for new users after the first 100 predictions. We believe this is not a reason to dismiss the idea of using strategies, but a reason to optimize the strategy. This can be done, because the current rules that select between techniques are not all based upon research into individual techniques, some are educated guesses using knowledge of how an individual technique works.

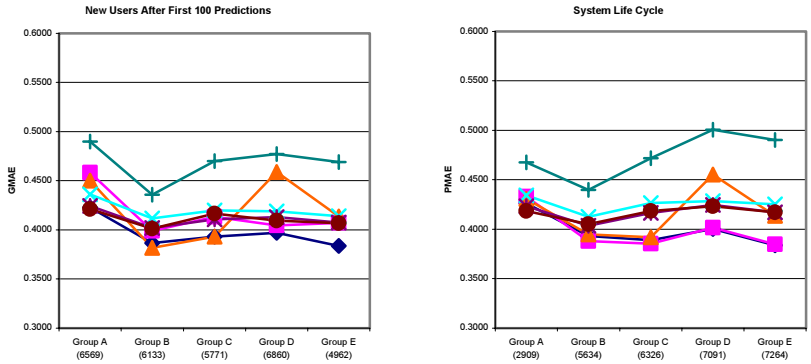

Fig. 4. Global Mean Absolute Error for the predictions after the first 100 for new users and the Prediction Mean Absolute Error throughout the system lifecycle.

Due to the very low coverage of social filtering for new users, the results of the pmae on the first 100 predictions of new users (left hand side of figure 5) is statistically less valid, as the sample size (number of predictions) in those groups was small. In the results of that test, there is almost no statistical difference between any of the techniques and the strategy.

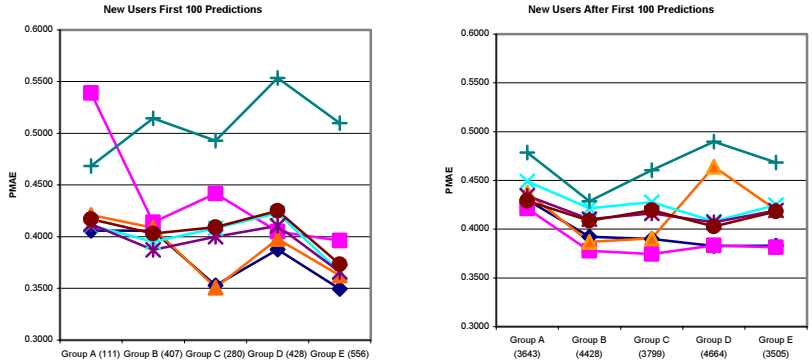

Fig. 5. Prediction Mean Absolute Error for the first 100 predictions of new users and for the predictions of new users after the first 100.

A good indication that prediction strategies are more reliable can be seen in the results of group D in most tests. In most groups, Top-N is almost as good as the strategy, but not in group D. The taste strategy is capable of handling this reduced performance of the Top-N technique, by relying on other techniques.

4 Conclusions and Future Research

The results of our initial experiment indicate that prediction strategies indeed improve the prediction quality and provide more reliable predictions by always using the best (combination of) techniques in a particular situation. With our model, it is possible to quickly create, use and test different prediction strategies using several prediction techniques. We will apply the model and strategies in a personalized TV system, providing recommendations for personalized EPGs and offering personalized search facilities. We will also experiment with actually combining predictions of individual techniques, instead of only selecting between techniques. Because our model is generically applicable, it will also be applied and tested in other domains.

Acknowledgements

This research is part of the PhD project Duine, for which most of the work is done in the GigaPort project (http://www.gigaport.nl) at the Telematica Instituut (http://www.telin.nl). The authors like to thank Betsy van Dijk, Andrew Tokmakoff and Harry van Vliet for their helpful comments. Also thanks to the researchers at the university of Minnesota for making the MovieLens dataset publicly available.

References

1. Gena, C.: Designing TV Viewer Stereotypes for an Electronic Program Guide. In: Bauer, M., Gmytrasiewicz, P.J. and Vassileva, J.: User Modeling 2001. Springer, Sonthofen, Germany (2001) 274-276
2. Herlocker, J.: Understanding and Improving Automated Collaborative Filtering Systems. University of Minnesota (2000)
3. Herlocker, J. and Konstan, J.A.: Content-Independent Task-Focused Recommendation. IEEE Internet Computing, Vol. 5 (2001) 40-47
4. Herlocker, J., Konstan, J.A. and Riedl, J.: Explaining Collaborative Filtering Recommendations. Proceedings of CSCW'2000. ACM, Philadelphia, PA (2000)
5. Houseman, E. M. and Kaskela, D. E.: State of the art of selective dissemination of information. IEEE Trans Eng Writing Speech III (1970) 78-83
6. Jackson, P.: Introduction to Expert Systems. Addison-Wesley, Reading, MA (1990)
7. Shardanand, U. and Maes, P.: Social information filtering: algorithms for automated "Word of Mouth". Proceedings of Human factors in computing systems 1995. ACM, New York (1995) 210-217
8. Smyth, B. and Cotter, P.: A personalised TV listings service for the digital TV age. Knowledge-Based Systems, Vol. 13 (2000) 53-59
9. Tokmakoff, A. and van Vliet, H.: Home Media Server Content Management. In: Smith, J.R., Panchanathan, S., Jay Kuo, C.-C. and Le, C. Internet Multimedia Management Systems II, Volume 4519 (2001) 168-179
10. van Setten, M., Tokmakoff. A. and van Vliet, H.: Designing Personalized Information Systems - A Personal Media Center. Proceedings of workshop Personalisation in Future TV, Sonthofen, Germany (2001). http://www.di.unito.it/~liliana/UM01/vanSetten.pdf
11. Witten, I. H. and Frank, E.: Data mining: practical machine learning tools and techniques with Java implementations. Morgan Kaufmann Publishers, San Diego (2000)