# Using a Critic to Promote Less Popular Candidates in a People-to-People Recommender System

**A. Krzywicki, W. Wobcke, X. Cai, M. Bain, A. Mahidadia, P. Compton, Y. S. Kim**
School of Computer Science and Engineering
University of New South Wales
Sydney NSW 2052, Australia
{alfredk,wobcke,xcai,mike,ashesh,compton,yskim}@cse.unsw.edu.au

## Abstract

This paper shows how to improve the recommendations of an interaction-based collaborative filtering (IBCF) recommender used in online dating. Previous work has shown that IBCF works well in this domain, although it tends to rank popular candidates highly, which leads to these users receiving a large number of contacts. We address this problem by using a Decision Tree model as a "critic" to re-rank the candidates generated by IBCF, effectively promoting less popular candidates. This method was first evaluated on historical data from a large online dating site and then trialled live on the same site by providing recommendations to a large number of users throughout a 9 week period. The live trial confirmed the consistency of the analysis on historical data and the ability of the method to generate suitable candidates over an extended period. Our recommendations gave higher success rates than those for a control group made with a baseline recommender.

## 1 Introduction

Recommender systems have become important tools helping users to deal with information overload and the abundance of choice. Traditionally, these systems have been used to recommend items to users. In this paper however, we report on the evaluation of a people-to-people recommender in an online dating context. The main difference between these two types of recommender systems is that recommending people to people is based on two-way interactions (Krzywicki et al. 2010; Pizzato et al. 2010). Another important difference, specific to online dating sites, is that people can only maintain contacts with a limited number of matches. On dating sites, more attractive, and hence more popular people are contacted more often, but they are less likely to respond, while less popular users tend to keep contacting more popular candidates regardless of their chances of success (Hitsch, Hortasu, and Ariely 2010). Hence it is important to restrict the number of times popular users are recommended.

We collected statistics from a large commercial dating web site that strongly support the above findings. These statistics are based on about 1.8 million user messages, each being a predefined short text, sent in March 2010. The response to each such message is also a predefined text message, either positive or negative, or no response at all. Analysing these statistics, we found that 38% of messages are sent to popular users with a positive reply rate of only 11%, while the average positive reply rate over the entire set of messages is 15%. On the other hand, users who contact "non-popular" candidates (those who receive no more than 50 contacts in the previous 28 days) have a positive reply rate of 20%. The above figures suggest that a recommender system can help users improve their success rate.

In Krzywicki et al. (2010) we showed that interaction-based collaborative filtering (IBCF) works well in the people-to-people recommendation domain, but has the problem of over-recommending candidates with high popularity. In this paper, we address this problem by proposing hybrid methods that combine recommendation methods based on interactions and profile matching. We first consider the general question of how to combine recommender systems, by normalizing and combining the ratings of candidate sets of the recommenders based on a Bayesian independence assumption. However, in our domain, this type of combination of an IBCF recommender with a Decision Tree recommender did not yield satisfactory results. This was primarily because, while the Decision Tree is able to predict negative interactions with high accuracy, accuracy for the positive interactions was much lower. Therefore, we present the idea of combining two recommenders using the second as a "critic" to modify the recommendations of the first recommender. The Decision Tree rules based on features related to the activity and popularity of users have the effect of "demoting" the ratings of candidates generated by the IBCF method. The formulation of the solution can be used as a general method of combining recommender systems where the independence assumptions hold.

The rest of the paper is structured as follows. In the next section we review related research. Then, after briefly summarizing the IBCF recommender system and the Decision Tree model, we show how to combine the ratings produced by independent recommenders using the Decision Tree as a critic. Section 4 contains a comparison of IBCF with the combined method on historical data obtained from a commercial online dating site. The next section contains the results of the live user trial showing the effectiveness of the combined recommender. Finally, we summarize the paper and potential future work.

## 2 Related Work

Over-recommending popular items has often been signalled as an issue for recommender systems (Garcin et al. 2009; Wang and Tan 2011; Park and Tuzhilin 2008), however we are not aware of any research addressing this problem for people-to-people recommendation. The closest solution in the literature seems to be work addressing the "long tail" problem (Park and Tuzhilin 2008), where, in the context of item-to-people recommendation, there are many items with very few ratings provided by users. For people-to-people recommendation, the analogous issue is that many users on social networking or online dating sites receive very little attention (receive a small number of contacts). The manifestation of the "long tail" in this context, however, is very different. While recommending popular *items* to users generally increases the accuracy of the recommender (Jambor and Wang 2010), in the people-to-people context where accuracy is expressed as the likelihood of receiving a positive reply from the suggestion, recommending popular *users* actually decreases the accuracy. This is simply because users are not able to maintain too many contacts at the same time.

Park and Tuzhilin (2008) consider a number of cutting points to separate the long tail from the short head, combined with a number of clusters for each cut. The error rate is calculated for each such combination. These methods, however, are arbitrary and may not generalize well across different datasets. Our method based on Decision Tree learning combined with collaborative filtering does not require defining any arbitrary popularity limit.

Jambor and Wang (2010) introduce a framework to parameterize a recommender system to meet multiple objectives, reducing the "long tail" being one of them. This is done by assigning a positive weight to each user-item predicted rating and calculating weights in such a way as to recommend popular items to users who may really be interested in them. The weights are calculated using the mean and variance of the item ratings. Our method is different from this approach in that each user-candidate pair is weighted by a value learned from the tree model. Another difference is that we consider not only the taste of the initiating user, but also the likelihood of success with the candidate.

Other relevant research includes that of Ishikawa et al. (2008), who describe a method to recommend web pages from the "long tail" that can be relevant to users. The method is based on information diffusion theory and an observation that the popularity of an item may increase rapidly once noticed by interested users. This observation, although not directly used in our paper, may be applicable to people-to-people recommenders. Less popular candidates, once they receive more contacts and respond positively, may then receive higher ranks.

Finally, Burke (2002) provides an exhaustive survey of hybrid recommendation techniques and applications available at the time, including a variety of ways to combine recommendations given by two systems (weighting, switching, cascading). Our use of the critic is similar in spirit to cascading, however rather than breaking ties, the critic is used to re-rank candidates to avoid over-recommending popular users.

## 3 Two Stage Recommender Using a Critic

In this section, we show how to combine people-to-people recommender systems to improve user success rates by "demoting" highly popular users. This is done by combining the ratings of two recommenders (IBCF and Decision Tree based) using a Bayesian independence assumption. This is a two stage recommendation process where the Decision Tree model is used as a "critic" to improve the recommendations of IBCF, which has the effect of promoting less popular users. We begin by summarizing the two methods.

### 3.1 Interaction-Based Collaborative Filtering

In Krzywicki et al. (2010) we introduced an application of interaction-based collaborative filtering (IBCF) to people-to-people recommendation, and defined a number of IBCF methods that are variants of collaborative filtering. An *interaction* consists of a user sending a message to a receiver, which may have a positive, negative or null reply. The notion of user similarity is based on these interactions. In particular, two users are *similar senders* to the extent they have contacted other users in common, and are *similar recipients* to the extent they have been contacted by other users in common. By considering the links in the network of interactions, we defined various collaborative filtering methods based on the two notions of similarity.

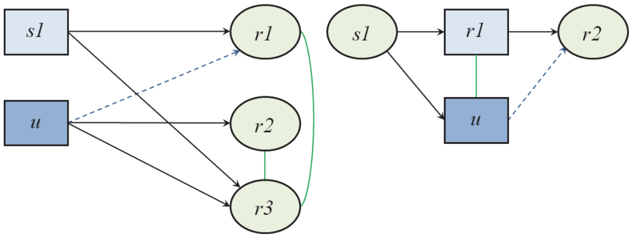

Figure 1: Basic CF+ and Inverted CF+ Recipient

It was found that these IBCF methods provide much better recommendations if they are based only on positive interactions, and those methods were denoted 'CF+' instead of 'CF'. Testing also showed that the best results were obtained by combining the best two of these methods, Basic CF+ and Inverted CF+ Recipient. Figure 1 illustrates the best two methods. In this figure, an arrow from $s$ to $r$ indicates a message sent from $s$ to $r$ that receives a positive reply, lines represent similarity, and the convention is that users of the same gender are in nodes of the same type (boxes or ovals), ignoring same-sex interactions for simplicity. It was shown that the two methods complement each other very well, notably with Inverted CF+ Recipient able to recommend candidates to users who have received no positive replies (but who must have replied positively to some messages), while Basic CF+ can recommend candidates to users who have received positive replies.

The ranking of a candidate for each method is given by the number of "votes" of similar users for that candidate. For the combined method Best CF+, the ranking is obtained by adding together the votes of Basic CF+ and Inverted CF+ Recipient.

## 3.2 Decision Tree Model

A Decision Tree for interaction data is constructed in a standard way using See5 Release 2.06 provided by Rulequest Research, which is a commercial implementation of the popular C4.5 Decision Tree software. We also experimented with numerous WEKA algorithms. However, these algorithms (e.g. Ripper) are not scalable to the size of problem being addressed in this paper, with millions of examples. See5 was chosen because of its efficiency and suitability for commercial applications.

The feature sets used in the construction are of two types: user profile features and temporal features derived from historical interactions. The user profile features include basic attributes such as age, location, education, family status, body type, smoking and drinking habits, etc. Derived attributes are also used, such as differences in the values of basic attributes between pairs of users. Temporal features are measures of activity (number of messages sent) and popularity (number of messages received) in the previous 7 and 28 days. The input to the learner is a set of attribute pairs corresponding to the senders and receivers of messages in a training set along with a Boolean value indicating either a successful (positive reply received) or unsuccessful (negative or no reply) interaction. The output is a Decision Tree classifying whether an interaction between an arbitrary pair of users is successful or unsuccessful.

Weights were used when building the Decision Tree using the cost parameter of See5 that penalize positive and negative misclassifications differently. Decision Trees are trained separately for male-female (M–F, 1.4 million examples, cost=0.5) and female-male (F–M, 400,000 examples, cost=0.34) interactions and then converted into decision rules. Each decision rule corresponds to a branch of the Decision Tree and implicitly defines two subgroups of users, senders satisfying the sender conditions and receivers satisfying the receiver conditions, such that both sender and receiver satisfy the conditions on derived attributes.

## 3.3 Combining Independent Recommenders

Each recommender provides a list of user-candidate pairs $(u, v)$ each with a numerical rating. The general problem is to define a new list of pairs combining the two given lists, along with a combined rating for pairs included in both lists.

We treat each rating as determining a probability, here denoted $SR$ (success rate), that an interaction between $u$ and $v$ is successful. The quality of a recommendation from $v$ to $u$ is represented by a quantity we call $SRI$ (success rate improvement), the probability that an interaction between $u$ and $v$ is successful given that $v$ has been recommended to $u$, divided by the prior probability that an interaction between $u$ and $v$ is successful. As long as their success probabilities are conditionally independent, the combined SRI for two recommenders is simply the product of the SRIs for the recommenders, which can be justified using Bayesian reasoning.

In order to apply this method, the ratings of two (or more) recommenders need to be converted into probabilities (SR), numbers between 0 and 1. Here, the independence assumption is reasonable, since the IBCF method is based on interactions, while the Decision Tree model is based only on user profile and temporal features. We now summarize how these probabilities are computed.

The IBCF method provides ratings of candidates as a number of "votes" derived from related successful interactions, as mentioned above. For each number of such votes $r$, we determine an average SRI from the data as a ratio $SR_r/BSR_r$, where $SR_r$ is calculated as the number of successful interactions in the training set for all pairs $u_r, v_r$ generated by IBCF with $r$ votes divided by number of all interactions for these pairs. $BSR_r$ (Baseline SR) is a similar ratio calculated for users $u_r$ contacting anyone in the training set.

For the Decision Tree rules, as discussed above, each rule defines a subgroup pair of users, and each user pair $(u, v)$ is contained in exactly one such subgroup pair since the rules are mutually exclusive. Thus the SRI for the pair $(u, v)$ can be estimated as the SRI of this rule and calculated as above based on interactions for these pairs in the training set. The rating of the combined recommender is obtained by multiplying the two SRIs for the pair.

Initial experiments showed that the best results are obtained by using the Decision Tree as a "critic" where rules with SRI $< 1$ are used to demote the candidates generated by IBCF. This combined method is called IBCF+DT in this paper.

Out of 76 female-male rules obtained from the tree model, 18 rules had SRI $< 1$. Similarly, 88 male-female rules were derived from the tree model, 11 of which had SRI $< 1$. Of these 29 rules, 22 contain a condition on the popularity of the candidate. These rules were tested on several datasets from different time periods and produced consistent results, therefore there was no need to re-learn the Decision Tree model every time recommendations are generated.

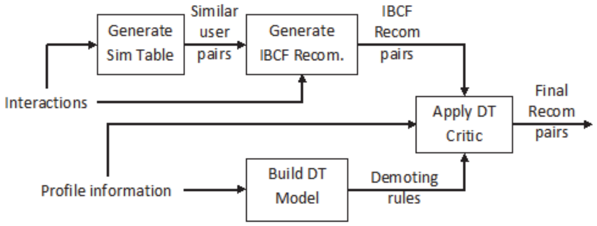

Figure 2: Generating IBCF+DT Recommendations

Figure 2 shows the two stage recommendation process, where the upper branch represents the IBCF recommender and the lower branch the Decision Tree "critic". Similarity pairs (e.g. (*r1*,*r3*) and (*r2*,*r3*) in Figure 1) are created based on user interactions and are used to generate IBCF recommendations. Rules with SRI $< 1$ produced by the Decision Tree are supplied to the "Apply DT Critic" module, where IBCF SRIs are multiplied by rule SRIs. Finally, candidates are re-ranked based on this new rating. Oracle SQL is used for most of the processing except the DT rules, where See5 is used.

# 4 Evaluation on Historical Data

In this section, we discuss how the IBCF+DT method, combining interaction-based collaborative filtering (IBCF) with a Decision Tree critic, improves success rates compared to IBCF, with special emphasis on how less popular candidate rankings are improved using the combined method.

For training and evaluation, we used a historical dataset from a commercial online dating site, which records both profile information about each user and interactions between users, as described above. Each interaction is recorded with a date/time stamp, the type of message and the response message type, which is pre-determined as being either positive or negative, or may be null if no reply has been received. Null replies are significant in this domain, with around a third of all messages going without a reply. In our evaluations, null replies are counted as negative interactions, since they would correspond to an unhelpful recommendation.

Table 1 shows basic information about the training and test sets. The training set contains around 1.8 million interactions recorded from around 133,000 users. The test set consists of around 87,000 users with around 638,000 interactions, of which roughly 15% are positive.

Table 1: Summary of Training and Test Datasets

| | Training | Test |
|---|---|---|
| #all interactions | 1,800,000 | 638,000 |
| #all unique users (senders and recipients) | 133,000 | 87,000 |
| #interactions among non-popular users | 1,000,000 | 396,000 |
| #unique non-popular users (senders and recipients) | 126,000 | 81,000 |

## 4.1 Experimental Setup and Metrics

Two recommenders are evaluated and compared: interaction-based collaborative filtering (IBCF), (Krzywicki et al. 2010)), and the combined system IBCF+DT. For evaluation, the IBCF recommender was trained on two interaction datasets (Table 1), one containing all interactions and one containing only interactions for which both sender and receivers are "non-popular", where a user is defined as "popular" if they receive more than 50 contacts in the 28 days prior to the start of the test period. This reduced dataset, denoted by "non-pop to non-pop", is used in the live trial (Section 5) for computational reasons. In Table 1, it can be seen that the 5% of popular users account for over 40% of interactions (either as senders or receivers).

The candidate list for each dataset is constructed using only interactions prior to the start of the test period. For computational efficiency, the number of candidates for each target user was limited to the top 200. The candidate list constructed from all interactions contains about 12 million user-candidate pairs and the list constructed from interactions among non-popular users contains about 7 million pairs.

The main metric used is the success rate improvement (SRI), calculated as defined in Krzywicki et al. (2010) and discussed in Section 3.3. This metric has been specifically designed to measure how the method can increase user success and thus improve user experience and retention on the dating web site. It measures how much more likely users are to receive a positive reply from a recommended candidate compared to their baseline success rate. SRI is measured for the top $N$ recommendations, where $N$ = 10, 20, $\cdots$, 100.

We also examine the popularity and activity of candidates, where *popularity* is the number of messages received in the 28 days before the recommendations were generated, and *activity* is the number of messages sent in the same period.

## 4.2 Discussion of Results

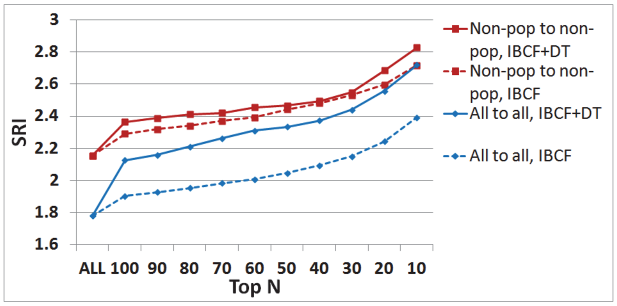

Figure 3: Comparison of SRI for IBCF and IBCF+DT

Figure 3 compares the SRI for the top N recommendations for both IBCF and IBCF+DT with the "all to all" and "non-pop to non-pop" datasets. The SRI on the "non-pop to non-pop" dataset is higher, which can be explained by the fact that removing popular users' interactions also removes many unsuccessful interactions from less to more popular users. In fact, these account for about 75% of removed interactions, while there are only 11% of removed interactions from more to less popular users. The SRI for the IBCF+DT method is higher than that for IBCF in both datasets, although this difference is much smaller for "non-pop to non-pop". We also noticed that the SRI for candidates with very low popularity, those who received fewer than 5 contacts in the previous 28 days, is also higher for "non-pop to non-pop", and IBCF+DT has higher SRI than IBCF in both datasets.

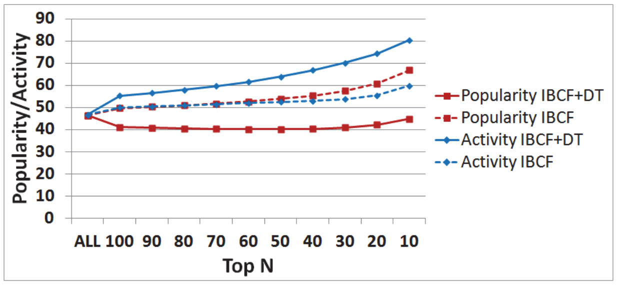

Figure 4: Activity and Popularity for "all to all" Dataset

Figures 4 and 5 show the average activity and popularity for the top N recommendations for each user. IBCF+DT recommends less popular but more active candidates and this is

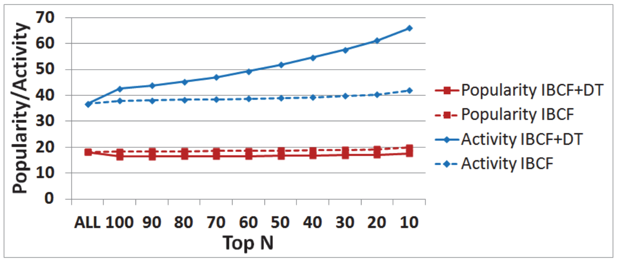

Figure 5: Activity and Popularity for “non-pop to non-pop” Dataset

quite consistent across the whole range of rankings. It is interesting to note, however, that for “non-pop to non-pop” the difference is much smaller for popularity than it is for activity. In fact, the activity rises much faster with rank for the IBCF+DT method. This is because many tree rules used in this method decrease the rating for less active users while less popular users are generally more active.

## 5 User Trial

### 5.1 Method

The IBCF+DT method combining interaction-based collaborative filtering (IBCF) and the Decision Tree critic was evaluated in a live user trial on the commercial dating web site which previously made the historical data available for analysis.

The effective trial was conducted for over 9 weeks, from February to April 2011. Results reported in this paper were recorded for two groups of users: the IBCF+DT group, which used our method recommendations, and the Control group, which used a recommender developed by the dating site company based on matching user preferences. Recommendations were generated three times a week. Each time recommendations were generated, the same number of target users for both the IBCF+DT and Control groups were randomly selected from the new users joining the site since the previous run of the recommenders, and added to the user groups previously allocated to the two groups. Thus the composition and size of the two groups increased over the course of the trial, and once a user was assigned to a group, they remained assigned to that group. Recommendations were delivered to users via e-mail. This method of delivery was decided by the dating site company. The IBCF+DT recommendations were generated on the day before the e-mail was sent using interactions occurring in the previous 28 days and included only the top 10 recommendations per user. Each run of the IBCF+DT recommender took around half an hour. For computational efficiency, the most popular users (more than 50 contacts received in the 28 days before recommendations were generated) were removed. In other words, we did not provide recommendations for very popular users and we did not recommend them to others. This corresponds to the “non-pop to non-pop” dataset above.

### 5.2 Results and Discussion

Since the IBCF method requires positive interactions between users prior to making recommendations, not all users in the IBCF+DT group received recommendations. For this reason, as shown in first section of Table 2, the number of target and candidate users in the IBCF+DT group is much smaller than in the Control group. We found, however, that the performance of the Control group recommender is similar for active and non-active users.

Table 2: Trial Results

| | IBCF+DT | Control |
|---|---|---|
| #all users | 66429 | 158202 |
| #target users | 5442 | 11352 |
| #candidate users | 63221 | 151949 |
| #recommendations sent | 371940 | 1870586 |
| #resulting interactions | 4282 | 10484 |
| SRI | 1.8 | 0.77 |
| SRI low popularity candidates | 2.14 | 0.99 |
| AVG popularity of contact recipients | 21 | 33 |

There were around 372,000 recommendations sent to IBCF+DT users and over 1.87 million to the Control group users. Contacts from target to candidate users were recorded after recommendations were sent. Unfortunately, we were not able to directly track click-through actions from the recommendation e-mails, therefore these counts include all contacts from target to candidate users, some of which could be initiated by users without looking at the recommendations. It is noticeable that the number of these contacts is very small compared to the number of recommendations, as expected. Nevertheless, the number of interactions per recommendation from IBCF+DT (4282 divided by 371,940) is double those in the Control group, which is an encouraging result. Likewise, the success rate improvement (SRI) is 2.33 times higher than that of the Control group recommender. We also recorded very good results for very low popularity candidates, those who in the 28 days prior to recommendation generation received fewer than 5 contacts from other users. The SRI from recommendations for this group of candidates is 2.14 for IBCF+DT compared with 0.99 for the Control group (also 0.99 for active Control group users). This result is consistent with the historical data evaluation. The average popularity of candidates for IBCF+DT is also substantially lower (21) than that in the Control group (33). The above results indicate that the IBCF+DT method’s recommendations are much more successful in general and for non-popular candidates in particular.

It can be noticed that the trial SRI values are lower than those in the evaluation based on historical data. This difference is due to a number of differences between the two evaluation settings, such as the timing of recommendations, the result collection and the fact that trial runs had to generate new candidates each time.

Figures 6 and 7 show the weekly SRI and average candidate popularity over the trial period. The SRI is consistently higher and the candidate popularity consistently lower for

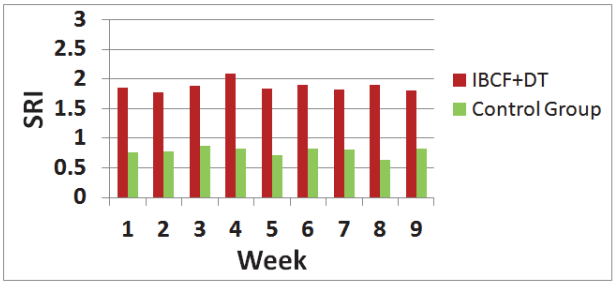

Figure 6: Trial SRI by Week

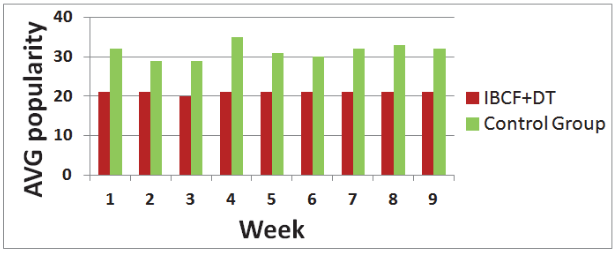

Figure 7: Trial Popularity of Candidates by Week

the IBCF+DT method over the entire period. We do not show weekly SRI results for low popularity candidates, as these numbers are small, therefore for these users we give only the overall SRI in Table 2.

## 6 Conclusion and Future Work

In this paper, we addressed in the context of people-to-people recommendation, the problem that collaborative filtering over-recommends popular items. The problem is serious in this domain since, in contrast to product recommendation, recommending popular users results in a decrease in user success rate. After introducing a general method for combining recommenders, we proposed a new two stage recommendation process, where one method (a Decision Tree) is used as a critic to re-rate the candidates provided by an initial recommendation method (interaction-based collaborative filtering). Evaluation on historical data shows that the combined recommender promotes less popular candidates and improves user success rates. Using an additional rating of candidates also improves the overall ranking by helping to break ties for lower rated candidates.

We conducted a live user trial on a commercial online dating web site, where our recommendations were sent via e-mail over a 9 week period. The trial results were broadly consistent with the evaluation on historical data: users who used our recommendations had higher success rates. In addition, the combined recommender outperformed a proprietary recommender used for a Control group based on two-way preference matching. The relative number of contacts resulting from recommendations was twice as high for our method compared to the Control group. The trial also confirmed the feasibility of the method and its ability to generate suitable candidates over an extended period as the user group changes. This further strengthens the argument that interaction-based collaborative filtering is an effective method for people-to-people recommendation, and that the critic-based approach addresses problems with basic collaborative filtering in this domain.

The method of combining recommenders using the critic technique described in this paper has led to the development of other hybrid approaches to people-to-people recommendation, particularly to provide recommendations to all users, rather than only those with positive interactions. In addition, we plan to conduct a further trial where the recommendations will be delivered online.

## Acknowledgements

This work was funded by Smart Services Cooperative Research Centre. We would also like to thank our industry partners for providing the datasets and conducting the trial.

## References

Burke, R. 2002. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction* 12:331–370.

Garcin, F.; Faltings, B.; Jurca, R.; and Joswig, N. 2009. Rating aggregation in collaborative filtering systems. In *Proceedings of the Third ACM Conference on Recommender Systems (RecSys'09)*, 349–352.

Hitsch, G.; Hortasu, A.; and Ariely, D. 2010. What makes you click? Mate preferences in online dating. *Quantitative Marketing and Economics* 8:393–427.

Ishikawa, M.; Geczy, P.; Izumi, N.; and Yamaguchi, T. 2008. Long tail recommender utilizing information diffusion theory. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 785–788.

Jambor, T., and Wang, J. 2010. Optimizing multiple objectives in collaborative filtering. In *Proceedings of the Fourth ACM Conference on Recommender Systems (Recsys'10)*, 55–62.

Krzywicki, A.; Wobcke, W.; Cai, X.; Mahidadia, A.; Bain, M.; Compton, P.; and Kim, Y. S. 2010. Interaction-based collaborative filtering methods for recommendation in online dating. In Chen, L.; Triantafillou, P.; and Suel, T., eds., *Web Information Systems Engineering – WISE 2010*. Berlin: Springer-Verlag. 342–356.

Park, Y.-J., and Tuzhilin, A. 2008. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM Conference on Recommender Systems (Recsys'08)*, 11–18.

Pizzato, L.; Rej, T.; Chung, T.; Koprinska, I.; and Kay, J. 2010. RECON: A reciprocal recommender for online dating. In *Proceedings of the Fourth ACM conference on Recommender systems (RecSys'10)*, 207–214.

Wang, K., and Tan, Y. 2011. A new collaborative filtering recommendation approach based on Naive Bayesian method. In Tan, Y.; Shi, Y.; Chai, Y.; and Wang, G., eds., *Advances in Swarm Intelligence*. Berlin: Springer-Verlag. 218–227.