A Recommendation Method for Online Dating networks based on Social Relations and Demographic Information

Lin Chen
Computer Science Discipline
Queensland University of Technology
Brisbane, Australia

Richi Nayak
Computer Science Discipline
Queensland University of Technology
Brisbane, Australia

Yue Xu
Computer Science Discipline
Queensland University of Technology
Brisbane, Australia

***Abstract*— A new relationship type of social networks - online dating - are gaining popularity. With a large member base, users of a dating network are overloaded with choices about their ideal partners. Recommendation methods can be utilized to overcome this problem. However, traditional recommendation methods do not work effectively for online dating networks where the dataset is sparse and large, and a two-way matching is required. This paper applies social networking concepts to solve the problem of developing a recommendation method for online dating networks. We propose a method by using clustering, SimRank and adapted SimRank algorithms to recommend matching candidates. Empirical results show that the proposed method can achieve nearly double the performance of the traditional collaborative filtering and common neighbor methods of recommendation.**

Keywords- online dating; clustering; SimRank

I. INTRODUCTION

Online dating networks, a community type of social networks for connecting people to people, are expanding rapidly with many people joining them. Due to a large customers' base, an online dating recommendation system has become a necessity of dating networks to suggest potential matches to its members. Different from traditional recommendation which is usually an "item to user", online dating recommendation is "user to user" and it requires two-way matching to determine that both users are interested in each other in order to start proper communication. The challenge is how to efficiently find the matches for a user considering the number of online dating network's members is usually in millions.

Content-based and collaborative-based recommendation systems are the most commonly implemented recommender systems [4], however, they have drawbacks [7][8][16]. Only a handful of work has been done related to online dating recommendation. Authors in [3] utilized the existing collaborative recommendation method using the rating information of users to the data from an online dating website. Many factors such as age, job, ethnicity, education etc. that play an important role in the match making process are not considered in this work leading to poor accuracy. More recently some preliminary works have started appearing. One piece of work proposes a system which utilizes users' past relations and user similarity [15], while another [2], proposes that users be clustered, with the, male clusters being matched with female clusters. However, the recommendation of this approach is not personalized in that all of the users in a cluster receive the same recommendation.

The work presented here will utilise various attribute information such as profile, and relations in social network for proposing a social recommendation method. The online dating network is selected because of its rich social connections and users activity. Pair to pair recommendation is time consuming; therefore, the proposed method improves the recommendation efficiency by assigning users to groups. In this paper, we propose to use the SimRank method [11] after adapting it to the social networks for finding the similar users. We also propose a variation of SimRank by taking both the user's explicit information such as profile and preference data and the implicit information such as in-link and out-link into consideration for calculating user similarity. The users' similarity information is then used in making recommendations of potential partners.

The proposed method is evaluated using the dataset collected from a live online dating website. Accuracy of the proposed method is measured as the success rate of recommendations being considered by the users. The proposed method produces higher quality recommendations in comparison to the baseline methods such as traditional collaborative filtering and Adamic/Adar common neighbor [1]. The proposed method improves the success rate from 13.9% to 36.01%.

II. THE PROPOSED FRAMEWORK

A. *Online Dating Social Networks: Basics*

Users join an online dating social network to communicate with potential partners and eventually set up the start of a good relationship. A user is usually asked to provide his/her profile and partner's preference during registration. If the registration is successful, users start communications. The forms of communication include viewing other users' profile, message, email, chat. For detailed information about the online dating can be referred to [5].

B. *Overview*

Figure 1 shows the flow chart of the proposed method. Users are divided into a female group U^F and a male group U^M initially, $U^F \cap U^M = \emptyset$ and $U^F \cup U^M = U$.

$\{u_1^M, u_2^M, ... u_n^M\} \subset U^M$ and $\{u_1^F, u_2^F, ... u_m^F\} \subset U^F$. A clustering algorithm is then applied to U^M and U^F each to divide the male and female users into smaller groups according to their explicit information i.e., profile and preference attributes. The next task is to find the similarity between each user u_i ($u_i \in U^F$ *or* $u_i \in U^M$) in a cluster with other members of the cluster. This task provides the nearest neighbors to each user in a cluster. Two similarity measures are used to find out the nearest neighbors in a cluster: the original SimRank score and the adapted SimRank score. To compute the original SimRank score between members of a cluster, a graph which carries linked node information is generated and a similarity measure is employed. To compute the adapted SimRank score, the list of users that each member of the cluster has contacted is retrieved and the similarity between users is calculated according to the contact list's profile similarity.

Finally, the system utilizes the collaborative filtering and recommends the *Top-n* potential partners to a cluster member that his/her nearest neighbors have contacted.

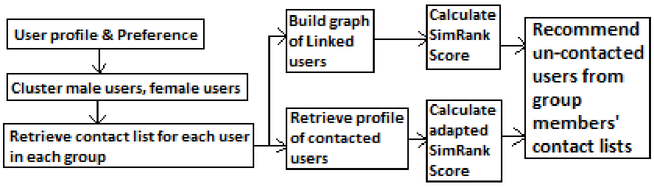
User profile & Preference
Cluster male users, female users
Retrieve contact list for each user in each group
Build graph of Linked users
Calculate SimRank Score
Retrieve profile of contacted users
Calculate adapted SimRank Score
Recommend un-contacted users from group members' contact lists

Figure 1. The Proposed Framework

C. *Clustering*

Users are clustered based on explicit information including personal profile and preference attributes. A combination of profile and preference information, or the profile information only, or the preference information only is used as an input for clustering. Work in [6] states the reasons for choosing these 3 options. Clustering input based on the preference only information is based on the assumption that people searching for similar type of partner contact similar candidates in reality. Assuming the similar people contact similar candidate, the profile only information is also considered as an input to clustering. Different similarity functions [9] including cosine similarity, Jaccard coefficient, correlation coefficient and Euclidean distance are utilized in the experiments. Repeated bisection of k-way algorithm [13] is applied in this work for clustering the users.

D. *SimRank*

The clustering process identifies the smaller but similar groups, however, the similarity between group members is yet to be found. The SimRank score is calculated to measure the similarity between each pair of members in a cluster. The basic theory of SimRank [11] is that two objects are similar if they are related to similar objects. We have applied this SimRank theory to the dating network scenario assuming that two users are similar if they contact similar users. The similarity can be defined by many means such as the number of common partners or the commonality amongst the partners' profiles. More detail, on how SimRank is applied to online dating, can be obtained in our previous work [5]

E. *Adapted SimRank*

In this paper, we have modified the SimRank function to include users' explicit information along with their network behavior. The premise is that two users in a cluster are similar if the partners they have contacted are similar based on their profile. Let $O(u_a^M)$ be the set of out-link neighbors that u_a^M has. $|O(u_a^M)|$ is the number of out-link neighbors u_a^M has. The profile information can now be used for each neighbor $O_i(u_a^M)$ to compare with other neighbors. Let $f(O_i(u_a^M))$ denotes all the profile features that $O_i(u_a^M)$ has. Let sc denote a profile feature similarity score between two neighbors $O_i(u_a^M)$ and $O_j(u_b^M)$. Cosine similarity can be employed to calculate the feature similarity sc. The adapted SimRank score can be shown in Equation 1.

$$s(u_a^M, u_b^M) = \frac{1}{|O(u_a^M)||O(u_b^M)|} \sum_{i=1}^{|O(u_a^M)|} \sum_{j=1}^{|O(u_b^M)|} sc\,(f(O_i(u_a^M), f(O_j(u_b^M))) \quad (1)$$

The adapted SimRank scores for female groups are calculated analogously. The computation time and complexity are reduced greatly with the use of adapted SimRank scores, as the original SimRank need to compare each node with all the left nodes in the graph and the number of nodes grows exponentially as the number of indirect neighbors adds in the graph [6].

F. *Recommendation*

Top- n recommendation is adopted for this work. In order to make recommendation to u_a^M, contacted female users of *n* most similar neighbor to u_a^M are recommended. Recommendations exclude u_a^M's previous contacted users. Duplications are also removed from the recommendation when neighbors have contacted the same message recipient.

III. Experiments and Disscussion

A. *Dataset*

The underlying dating network[1] has about 2 million members. The dataset for this research contains 87,304 male users who are active during the selected 6 months period. A user is called active user if they have logged in at least once during this period. In the experiments, positive messages are used as an indicator to determine whether the recommended user is suitable. If the user sends a message to the

[1] Due to privacy reasons the details of this network are not given.

recommended user and the recommended user replies to the sender with positive message, then the recommendation is identified as being "successful". There are 1,310,551 unique messages in the selected dataset that have been sent by the 87,304 male users in this period. Among the sent messages, 182,169 are identified as being successful. This yields the baseline success rate of 13.9%.

B. *Experiment Setup*

The overall performance of the proposed recommendation approach is compared with variations of similarity measures for finding neighbors, such as SimRank, adapted SimRank, and current system success rate without applying any recommendation method. Variations of the proposed method used in experiments are shown in Table I.The proposed method with all its variations is also compared with the traditional memory-based collaborative method (CF) [10]. Another method, Adamic\Adar method [1], is also used for comparing the results as it also adopts the common neighbor principle.

TABLE I. METHOD ACRONYMS

Acronym	Method
CSAS	combined profile with preference + cosine similarity + adapted SimRank
CJAS	combined profile with preference +Jaccard similarity + adapted SimRank
CDAS	combined profile with preference +distance similarity + adapted SimRank
CRAS	combined profile with preference + correlation similarity + adapted SimRank
CSOS	combined profile with preference + cosine similarity + SimRank with out-links only
CSIOS	combined profile with preference + cosine similarity + SimRank with in-links and out-links
CSIS	combined profile with preference + cosine similarity + SimRank with in-links only
CDOS	combined profile with preference + distance similarity + SimRank with out-links only
CDIOS	combined profile with preference + distance similarity + SimRank with in-links and out-links
CDIS	combined profile with preference + distance similarity + SimRank with in-links only
EDAS	preference only + distance similarity + adapted SimRank
ODAS	profile + distance similarity + adapted SimRank
RSAS	random grouping + cosine similarity + adapted SimRank
BSR	*$BSR(U^M)$* current online dating system success rate

The Cluto software [12] is used to cluster the 87,304 male users into approximately 1,000 groups. Experiments found that 5 iterations are sufficient to stabilize the score that concur with previous SimRank works finding [11]. Once the similarity amongst all users of a cluster is calculated, we test two approaches to recommend potential partners to a user u_a^M. (1) In the first approach (labeled as Top-n all matched users), the system recommends to user u_a^M all users who were contacted by users U_{TOP}, where U_{TOP} represents the Top-*n* most similar users to u_a^M. (2) In the second approach (labeled as Top-n successful matched users), the system only recommends to user u_a^M those users who were contacted by users U_{TOP} and replied positively to U_{TOP} . If the user being considered for recommendation did not reply positively to a user in U_{TOP} then they are not recommended to user u_a^M .

C. *Evaluation Metric*

The evaluation metric for this experiment is based on deciding whether the recommended users to a given user *u* will be successful. The recommendation can be called successful if the recipient user chooses to contact the recommended people. One of the metrics to evaluate the performance is success rate (SR). *$SR(U^M)$* as defined in Equation 2 is to be compared with system success rate *$BSR(U^M)$*. *$BSR(U^M)$* is the success rate of current online dating network without using the proposed recommendation approach. Another metric is recall which is to measure the ratio of correctly identified matches from the proposed recommendation approach to the number of matches in the dataset.

$$SR(U^M) = \frac{Number\,of\ (Positive\ Partners\ \cap \text{Re}commended\ Partners)}{Number\,of\ (\text{Re}commended\ Partners)} \tag{2}$$

$$\text{Re}call(U^M) = \frac{Number\,of\,(Positive\ Partners \cap \text{Re}commended\ Partners)}{Number\,of\,(Positive\ Partners)} \tag{3}$$

D. *Results*

In terms of the success rate performance, recommending Top-n successful matched users is a better method than recommending the Top-n all matched users as shown by the results inTables II and III. Most of the time, the Top-n successful matched users success rate gives double the performance over the Top-n all matched users. From Table II we can see the CSIS method produces the best performance in Top-n all matched users experiment, followed by CDIS. In-link based SimRank is better performing than both the in-link & out-link based and out-link based SimRank for the Top-n all matched users. The reason is that in-link based SimRank retrieves the positive message information when a user receives a positive message back from the potential partner. The in & out SimRank performance is lowered by having out-link information.

In Table III, it is shown that CDAS performs the best and achieves a success rate of 36.01% for Top-1 successful matched users. CSIOS is the second best method. The in & out-links based method outperforms in-link based only and out-link based only methods. Positive message information is known in this experiment when the potential partners who have returned a positive message are recommended. Therefore, methods containing in-link information only do not benefit.

Top-n all matched users approaches offer more potential partners for recommendation than Top-n successful matched users approaches. In terms of recall for Top-n matched users (Table IV) and recall for Top-n successful matched users (Table V), SimRank methods offer more recommendations to users than adapted SimRank methods.

TABLE II. SUCCESS RATE OF TOP-N ALL MATCHED USERS

	Top-1	Top-3	Top-5	Top-10
CSAS	14.8%	12.48%	12.73%	12.38%
CJAS	12.85%	12.71%	11.81%	11.28%
CDAS	15.35%	13.58%	12.92%	11.88%
CRAS	10.71%	12.12%	11.98%	11.45%
CSOS	16.15%	14.4%	13.58%	12.56%
CSIOS	16.24%	13.97%	13.0%	13.0%
CSIS	22.06%	18.62%	17.27%	16.01%
CDOS	15.11%	13.36%	12.7%	11.87%
CDIOS	15.02%	12.87%	12.31%	11.73%
CDIS	19.89%	17.21%	16.16%	15.19%
EDAS	10.01%	10.96%	11.22%	11.02%
ODAS	12.71%	11.84%	12.38%	11.78%
RSAS	11.7%	11.56%	10.82%	10.66%
BSR	13.9%			

TABLE III. SUCCESS RATE OF TOP-N SUCCESFUL MATCHED USERS

	Top-1	Top-3	Top-5	Top-10
CSAS	30.9%	26.43%	26.9%	25.68%
CJAS	25.77%	25.35%	24.56%	23.54%
CDAS	36.01%	32.9%	29.26%	25.9%
CRAS	23.85%	25.88%	25.37%	23.15%
CSOS	23.58%	24.07%	24.16%	23.9%
CSIOS	32.16%	27.04%	25.27%	24.14%
CSIS	31.37%	28.18%	25.87%	25.10%
CDOS	23.45%	23.85%	23.94%	23.74%
CDIOS	30.08%	25.9%	24.89%	24.03%
CDIS	29.58%	26.5%	25.4%	24.84%
EDAS	23.24%	25.37%	24.95%	24.19%
ODAS	32.84%	29.55%	28.13%	26.22%
RSAS	18.98%	21.4%	19.59%	18.98%
BSR	13.9%			

TABLE IV. RECALL OF TOP-N ALL MATCHED USERS

	Top-1	Top-3	Top-5	Top-10
CSAS	0.08%	0.39%	0.90%	2.44%
CJAS	0.09%	0.43%	0.92%	2.49%
CDAS	0.13%	0.50%	0.95%	2.04%
CRAS	0.07%	0.32%	0.73%	2.04%
CSOS	3.08%	7.04%	9.23%	11.47%
CSIOS	3.89%	7.90%	9.75%	11.46%
CSIS	2.82%	5.88%	7.22%	8.53%
CDOS	1.92%	4.03%	4.98%	5.81%
CDIOS	2.35%	4.43%	5.23%	5.91%
CDIS	1.74%	3.28%	3.90%	5.26%
EDAS	0.08%	0.41%	0.79%	1.55%
ODAS	0.11%	0.47%	0.98%	2.44%
RSAS	0.51%	2.35%	5.39%	13.9%

In most cases, the success rate decreases as *n* increases in Top-n (all/successful) matched users. But in some cases, the success rate increases as *n* increases. For example, for CSOS in Table III the success rate increases initially. The reason for this is that Top-1 recommendation is recommending the most similar user's contacted partners to the user. The number of contacted partners varies. The Top-1 most similar user may have a huge number of contacted partners and in this case the chance of achieving a high success rate is less than that achieved if similar users who have a smaller number of contacted partners were used. When using Top-3 users, the success rate of the 3 users could be averaged out if one of the Top-3 user's success rate is not high. As expected, recall increases as *n* increases in Top-n (all/successful) matched users.

TABLE V. RECALL OF SUCCESSFUL MATCHED USERS

	Top-1	Top-3	Top-5	Top-10
CSAS	0.013%	0.051%	0.12%	0.31%
CJAS	0.012%	0.050%	0.11%	0.29%
CDAS	0.019%	0.070%	0.12%	0.24%
CRAS	0.009%	0.041%	0.091%	0.25%
CSOS	0.71%	1.11%	1.30%	1.44%
CSIOS	0.66%	1.14%	1.31%	1.42%
CSIS	0.64%	1.11%	1.27%	1.38%
CDOS	0.34%	0.56%	0.64%	0.69%
CDIOS	0.36%	0.59%	0.65%	0.70%
CDIS	0.35%	0.57%	0.64%	0.67%
EDAS	0.010%	0.057%	0.11%	0.21%
ODAS	0.015%	0.060%	0.12%	0.29%
RSAS	0.010%	0.041%	0.088%	0.22%

E. Profile or Preference or Combined

Intuitively using user's preferences (what they want in their partner) as the input data for clustering should generate better recommendations than using user's profiles only and a combination of user's profiles and preferences. However, in our experiments, profile and preference combined input for clustering results is the highest performance in terms of recommendation. In Tables II, III, IV, &V, CDAS performs better than ODAS and ODAS performs better than EDAS. Results ascertain that users who have more in common in both their profile and preference are likely to choose similar people as their ideal partners. In other words, consideration of "what a user wants their ideal partner to be" or "what a user is like" alone plays an inferior role when deciding who they contact as a potential ideal partner.

F. Clustering or no clustering

Comparing performance of recommendation utilizing a clustering method against the performance of recommendation without a clustering method (RSAS – random grouping), the proposed idea of recommendation with clustering achieves higher performance except in the case of EDAS, which is worse than RSAS in a few cases. Therefore, clustering does contribute to better recommendation performance in general.

G. SimRank or Adapted SimRank

SimRank is the best performing method in Top-n all matched users, with CSIS; which is a SimRank variation; with combined profile and preference, cosine similarity and in-link information; giving the highest success rate, as shown in Table II. CSIOS gives the highest recall, as shown in Table IV. In the Top-n successful matched user experiment, an adapted SimRank method – CDAS achieves a higher success rate score than the success rates of all SimRank methods. However, SimRank methods achieve a higher recall score. The reason why SimRank has higher recall is that it favors similar users who have lots of network activities (initiate or receive lots of messages) as the Top-n matched users. Those similar users who only have a few message activities are less likely share a common neighbor with the user and therefore the SimRank score is low. Adapted SimRank compares the user's contacted partners' profiles with the similar user's contacted partners' profiles, instead of comparing links between two users. Thus neighbours can be discovered even if the two users have no common links. In most cases, similar users only have a handful of contacted partners and thus the number of recommendations from adapted SimRank is less than the number of recommendations from SimRank, however the quality of recommendations is better.

H. SimRank vs. Baseline methods

Due to space limitation, results of the CF and Adamic are compared with "all" recommendations suggested by the proposed method rather than each of the top-1, 3, 5 & 10. Table VI shows that the best performing CDAS and the worst performing RSAS from Top-n (all) successful match methods outperform the memory-based collaborative (CF) and Adamic/Adar [1] methods in terms of success rate. However, CDAS performed worse than Adamic/Adar in terms of recall. The reason is that the number of neighbors from the clustered SimRank method is limited. Adamic/Adar method searches the whole training dataset for the neighbors, but the higher recall sacrifices the success rate.

TABLE VI. TOP-N (ALL) PERFORMANCE OF BASELINE METHODS & SIMRANK

Method	SR	Recall
CDAS	23.6%	0.72%
RSAS	18.1%	0.49%
CF	12.8%	0.46%
Adamic/Adar	16.8%	3.4%

IV. CONCLUSION

This paper has applied social recommendation concepts to online dating networks by considering explicit information and social network connections. The proposed method clusters users into groups to reduce the computation time and complexity. The similarity based SimRank has been adapted in this paper. Two interpretations of SimRank methods are developed. In the first version, SimRank scores, the similarities of users depend on how similar the people they have contacted are. The similarities scores purely depend on their social network connections in this version. In the second version, adapted SimRank scores, the similarities of users in the cluster depend on the similarity of their contacted users' explicit information (users' attributes). In this version, both explicit information and user connection relations are taken into consideration.

The proposed methods have been evaluated on an online dating network dataset. The best performing method has improved the success rate from 13.9% to 36.01%. To generate a better recommendation, a combination of user's profile and preference information should be fed in as input for clustering. The result also proves that a clustering method works better than randomized grouping. In future, improving the recall is essential. The proposed method will be extended to include capability for handling new users.

ACKNOWLEDGEMENT

The content presented in this paper is part of an ongoing cooperative study between Queensland University of Technology (QUT) and an Industry Partner, with sponsorship from the Cooperative Research Centre for Smart Services (CRC-SS). The authors wish to acknowledge CRC-SS for partly funding this work and the Industry Partner for providing data for analysis. The views presented in this paper are of the authors and not necessarily the views of the organizations.

REFERENCES

[1] L. Admic and E. Adar. Friends and Neighbors on the Web, vol. 25, pp.211-230, 2001.

[2] S. Alsaleh, et al. Improving Matching Process in Social Network Using Implicit&Explicit Information. APWeb 2011. Pp313-320.

[3] L. Brozovsk and V. Petricek. (2007, December 10th). *Recommender System for Online Dating Service.* Available: www.occamslab.com/ petricek/papers/dating/brozovsky07recommender.pdf

[4] R. Burke, "Hybrid Recommender Systems: Survey and Experiments," *User Modeling and User-Adapted Interaction*, vol. 12, pp. 331-370, 2002.

[5] L. Chen, R. Nayak, &Y. Xu, Improving matching process in social network. Presented at ICDM Workshops, pp.305-311

[6] L. Chen, R. Nayak, & Y. Xu. How people really behave in Online Dating Networks? Interesting Findings with Social Network Analysis.

[7] W. Chu and S. T. Park, "Personalized Recommendation on Dynamic Content Using Predictive Bilinear Models," presented at the WWW2009, Madrid, pp. 691-700, 2009.

[8] I. Guy, *et al.*, "Personalized Recommendation of Social Software Items Based on Social Relations," presented at the RecSys'09, New York, pp. 53-60, 2009.

[9] J. W. Han and M. Kamber, Data Mining Concepts and Techniques: Elsevier Inc.,2006. ISBN: 1558609016.

[10] J. L. Herlocker et.al., An Algorithmic Framework for Performing Collaborative Filtering. Presented at SIGIR'99, pp. 230-237, 1999.

[11] G. Jeh and J. Widom, "SimRank: A Measure of Structural-Context Simiarity," presented at the KDD'02, pp. 538-543, 2002.

[12] G. Karypis, "Cluto: A Clutering Toolkit," ed, 2003. Available: glaros.dtc.umn.edu/gkhome/fetch/sw/**cluto**/manual.pdf

[13] G. Karypis and V. Kumar, "Multilevel k-way Hypergraph Partitioning", 36th Design Automation Conference, pp343-348, 1999.

[14] P. Kazienko and K. Musial, "Recommendation FrameWork for Online Social Networks," 4th Atlantic Web Intelligence Conference (AWIC'06), Washington D.C., pp. 111-120, 2006.

[15] R. Nayak, "Utilizing Past Relations and User Similarities in a Social Matching System" PAKDD 2011,

[16] X. Xin, *et al.*, "A Social Recommendation Framework Based on Multi-Scale Continuous Conditional Random Fields," CIKM'09, HongKong, pp. 1247-1256, 2009.