New to Online Dating? Learning from Experienced Users for a Successful Match

Mo Yu*, Xiaolong Zhang*, Derek Kreager†
*College of Information Sciences and Technology, The Pennsylvania State University
{muy145, lzhang}@ist.psu.edu
†Department of Sociology and Criminology, The Pennsylvania State University
dkreager@psu.edu

***Abstract*—Online dating arises as a popular venue for finding romantic partners in recent years. Many online dating sites adopt recommender systems to help their users. However, few of current research provides solutions to cold start problem, i.e., providing recommendations to new users. In this research, we propose a new approach of providing reciprocal online dating recommendations to new users. Specifically, we detect communities from existing users, match new users to these communities, and take advantage of reciprocal activities of those community members to provide recommendations to new users. Using data from a popular U.S. online dating site, experiments show that our approach greatly outperforms existing methods.**

I. Introduction

Online dating becomes a popular venue for finding romantic partners recently. To help users make choices, many dating sites adopt recommender systems to provide suggestions.

A typical type of recommender system is user-user one, which is widely used in online social networks. Such networks are often reciprocal. A connection can only be established if two ends both agree on that arrangement. In online dating, *Bob* can send a message to *Alice*, but unless *Alice* replies to *Bob*, we cannot make an argument saying that the connection is built. We admit that a reply does not guarantee a successful date, but without a reply there will be no further opportunities. When designing recommender systems for online dating, we need to take reciprocity into consideration, and suggest those users who are likely to reply one's messages.

A main challenge for recommender system is the cold start problem. Usually, the more information a recommender system can utilize, the better recommendations it can provide. However, for a new user, since information about him/her is very limited, it is difficult for recommender systems to provide accurate suggestions. But it is usually new users who need help mostly, since they are unfamiliar with a new environment.

With reciprocity and cold start problem in mind, we propose the following research question:

- Given a new online dating user, i.e., a user who just joins online dating and has no activities, how can we recommend existing users, i.e., users who are already engaged in online dating and have activities, to the new user, such that they will establish reciprocal contacts?

For a new user, we want to recommend a list of existing users who are not only likely to be contacted by the new user, but also are likely to reply to him/her. To fulfill this task, we design a hybrid approach of combining ideas from both content-based and collaborative filtering recommender systems. We first study the activities of existing users, and detect different communities among them. Then we match new users to those communities, and take advantage of community members' activities to provide reciprocal recommendations. Using real world data from a popular U.S. online dating site, experiments show that our method greatly outperforms existing models.

We claim three contributions for our research. First, to solve cold start problem for online dating recommendation, we combine the ideas of collaborative filtering and content-based recommender systems, which overcomes drawbacks of both approaches. Second, we incorporate reciprocity in our model to provide reciprocal recommendations for new users. Third, we introduce community detection to online dating recommendation. We detect communities for existing users, and then match new users to those communities. Based on reciprocal activities from community members, we make recommendations for new members.

The rest of our paper is organized as follows. In Section II we introduce related work. Section III covers details of our approach. We evaluate our approach in Section IV. Discussion and conclusion are in Section V.

II. Related Work

Many recent works about online dating provide useful insights for designing recommender systems. Xia et al. [1] conducted research about user behaviors and preferences for a famous online dating site in China . Pizzato et al. [2][3][4] designed a series of recommender systems based on user preferences , in which the most famous one was RECON . Another stream of work in online dating recommendation is collaborative filtering. In our early work [5], we introduced a model based on user taste and attractiveness to provide reciprocal recommendations. In [6] Xia et al. compared several different models' power in providing reciprocal recommendations. There are also some works tried to tackle online dating

IEEE/ACM ASONAM 2016, August 18-21, 2016, San Francisco, CA, USA
978-1-5090-2846-7/16/$31.00

recommendation from different angles. A series of works was based on user grouping and clustering [7][8].

Our goal is to provide reciprocal recommendations for new users, and there were only a few works tried to tackle this cold start problem. The most famous model is RECON [4], and we will compare our approach with RECON as well as a derivative of RECON in experiments.

III. Methods

Our approach includes five steps. We introduce them in details.

A. Extracting Communities of Existing Users

The first step is to extract communities from existing users. For a typical community detection algorithm, input is a social network, i.e., vertexes as well as associated edges, and output are several groups of vertexes and edges within these groups.

We generate two social networks for male and female existing users separately. Here we use male user network as an example. Let $G = \{V, E\}$ be the male user network. We have a set of vertexes V as the set of all male users. We define $E = \{e_{x,y}, x \in V, y \in V, x \neq y\}$ as the set of edges. For any $x \in V$, let C_x be the set of female users who have contacts with x. In C_x we not only include female users who have been contacted by x, but also we include those who contacted x and got $x's$ reply. We define $e_{x,y}$'s weight as:

$$w_{e_{x,y}} = \frac{|C_x \cap C_y|}{|C_x \cup C_y|}$$

$w_{e_{x,y}}$ is the Jaccard coefficient between C_x and C_y, which measures the overlap between $x's$ and $y's$ contacts. We find weights for all edges in E. For simplicity, we remove those edges which have 0 as weight.

After getting G, we detect communities using the algorithm introduced in [9], for its speed and overall performance. The output is a set of communities T. For each community $t \in T$, $t = \{V_t, E_t\}$, where $V_t \subset V$ is the set of users who belong to t, and $E_t \subset E$ is the set of edges among all users in V_t.

B. Finding Reciprocal Contacts for Communities

For each community $t \in T$, we have V_t as the set of users who belong to t. For a male user x, we define CR_x as the set of female users who have been contacted by and also replied to x. Then we create a distribution R_t for t, which is defined as follows:

$$R_t = \{(u, n) : u \in \bigcup_{x \in V_t} CR_x\}$$

where n is the number of users in V_t who have contacted u and also got u's replies. We normalize n with regards to the total of all ns in R_t. We normalize n by the sum of all ns in R_t. Also, we collect R_t for all $t \in T$ and denote this set as R, $R = \{R_t, t \in T\}$.

C. Generating Community Profile

To generate community profile, we follow the same fashion to find user preferences in [4]. In user profiles, there are several attributes, such as race, height, and body type. We define the list of attributes as A. We also define the user profile of a user x as

$$U_x = \{val_a : for\ all\ attributes\ a \in A\}$$

where val_a is a value of attribute a.

For community t, attribute a in its profile is represented as $p_{t,a}$, which is defined as follows:

$$p_{t,a} = \{(val, n) : for\ all\ unique\ discrete\ values\ val\ of\ a\}$$

where n is the number of times val occurred in V_t

The profile P_t for community t can then be represented as:

$$P_t = \{p_{t,a} : for\ all\ a \in A\}$$

In general, community profile is represented as a set of distributions, and each distribution shows the number of times each discrete value of an attribute has occurred. We collect profiles for all communities, and denote the set as P, $P = \{P_t, t \in T\}$.

D. Calculating Similarities Between New Users and Communities

To find a new user's similarities to communities, we take the similar fashion of fitting a user to another's preferences introduced in [4], but here we are fitting a new user x to each community t's profile. We denote the similarity score between new user x and community t as $Sim(x, t)$. We normalize $Sim(x, t)$ with regards to the total of all $Sim(x, t)$s for $t \in T$.

The similarity score between a new user and a community measures to what extent the new user belongs to that community. This idea is similar to that behind content-based recommender systems. Note that we calculate similarity scores between a new user and all communities. These scores will be used for the final step of recommendation.

E. Recommendation

To recommend an existing user x to new user y, we find a recommending score $RS(x, y)$ between them. The algorithm is described in Algorithm 1.

We evaluate recommending scores between new user x and all existing users, and generate a recommended list in descending order of recommending scores for x.

IV. Evaluation

A. Dataset

We use data from a popular U.S. online dating site for experiments. The dataset covers all users from a U.S. city for a time span of 196 days. It includes both user profiles and activity logs. A profile contains a user's demographic information and habits. Each record in activity logs represents a contact between a pair of users. It contains sender ID,

Algorithm 1 Calculate $RS(x, y)$

Input:
Existing User x, New User y, Community Profile P, Reciprocal Contact Distribution R
Output: Recommending score $RS(x, t)$
$RS(x, y) = 0$
for each $t \in T$ **do**
Calculate $s = Sim(y, P_t)$
Get (x, n) if $(x, n) \in R_t$
$RS(x, y) \leftarrow RS(x, y) + s * n$
end for
return $RS(x, y)$

TABLE I
DESCRIPTION OF USER PROFILE

Attribute	Value
Race	White, Black, Multiple, Other
Education Level	High School, University, Postgrad
Education Status	Dropped, Graduated, Working On
Body Type	Overweight, Average, Fit, Thin
Smoking	No, Sometimes, When drinking, Yes, Trying to quit
Drinking	Never, Rarely, Socially, Often, Very often, Desparately
Quickmatch Rating	Discreted, from 1 to 5
Height	Discreted, from 0 to 9
Age	Discreted, from 3 to 23

receiver ID, time stamp of first message, as well as whether receiver replied or not.

For user profile, we drop those attributes that are left blank by most users. We then select users who completed all remaining attributes. We also combine similar values under some attributes, and discrititize continuous values. A description of user profile is in Table I. In the online dating site, when a user registers an account, he/she will be randomly assigned five user profiles, and this user can rate these profiles in the scale of 1 to 5. The mean value of all collected ratings will be assigned as *quickmatch* rating for a user.

To detect communities, we use data from a span of 100 days, and we call such period as training phase. If a user has any activities during training phase, then we identify him/her as an existing user. Community detection is based on existing users' activities during training phase. We use another 30 days

TABLE II
DATASET DESCRIPTION

Item	Experiment				
	1	2	3	4	5
Existing user	6686	6863	7052	7210	7355
New users	1060	1129	1094	1037	1009
Contacts sent by new users	5047	5614	5164	5066	4965
Contacts from new to existing users	79%	76%	77%	78%	79%
Reciprocal contacts sent by new users	1204	1383	1373	1367	1263
Reciprocal contacts from new to exisitng users	76%	73%	74%	76%	79%

TABLE III
EXPERIMENT RESULTS(10^{-3})

	TopK	RECON	rRECON	CBR
Precision	10	1.12	1.63	**1.70(+4.5%)**
	20	1.58	1.58	**1.86(+17.7%)**
	30	1.28	1.58	**1.80(+13.9%)**
	40	1.42	1.46	**1.71(+17.1%)**
	50	1.40	1.46	**1.72(+17.8%)**
Recall	10	6.21	**9.32**	7.96(-14.6%)
	20	14.64	16.75	**17.03(+1.2%)**
	30	18.30	**25.67**	24.90(-3.0%)
	40	27.73	31.81	**32.05(+0.8%)**
	50	35.27	38.49	**40.73(+5.8%)**
NDCG	10	5.01	7.77	**8.51(+9.5%)**
	20	10.07	11.60	**13.65(+17.7%)**
	30	11.52	15.12	**16.86(+11.5%)**
	40	14.91	17.23	**19.46(+12.9%)**
	50	17.30	19.86	**21.81(+9.8%)**
MAP	10	1.25	1.91	**2.36(+23.6%)**
	20	1.75	2.41	**3.03(+37.7%)**
	30	1.90	2.77	**3.36(+21.3%)**
	40	2.17	2.94	**3.56(+21.1%)**
	50	2.34	3.08	**3.77(+22.4%)**

after training phase to identify new users, and we call this period as selecting phase. For users who have activities in selecting but not training phase, we identify them as new users. Note that new users who join early in selecting phase will become existing users for those who join late, and there are interactions among these two groups. To mitigate such affects, we choose a relatively short selecting phase. Also, our analyses confirmed that most new user activities are with existing users. We remove those "new-to-new" interactions during our experiment.

The last step is to collect testing data. We find that most activities of new users happen within a short period after they join online dating sites, and such finding was also pointed out by previous research [6]. Thus for a new user x, we look at his/her activities for 10 days after first appearance, and generate CR_x for him/her. Note that we removed "new-to-new" interactions, so here CR_x is a set of existing users only.

A brief description of our data is shown in Table II.

B. Experiments

To evaluate our approach, we first calculate precision and recall, which were commonly used in previous works. We also include normalized discounted cumulative gain (NDCG) and mean average precision (MAP), which both take result positions into consideration. We compare our model with two baselines. The first one is *RECON* [4]. To make recommendations to new users, *RECON* finds existing users' preferences based on their sending activities, and recommends those existing users who will be interested in the new users to them. Another baseline is a derivative of *RECON*, which only utilizes existing users' reciprocal contacts to learn their preferences, and we call this model $rRECON$. Both baselines share the same training phase with our proposed model.

We follow the idea of cross-validation to conduct a fair comparison of different models. We run five rounds of experiments based on different training and associated testing phases, and

take the mean value of all metrics for each model to report. Table III show performances based on different top-K levels.

In general our proposed model achieves the best results for precision, NDCG and MAP. For recall, $rRECON$ achieves best results for top-10 and top-30, and its preformance is close to our model for other top-Ks. The original $RECON$ model in general is not comparable to its derivative model $rRECON$. Overall our model achieves the best performance.

V. Discussion and Conclusion

$RECON$ has been proven as an effective method for reciprocal online dating recommendation, but its effectiveness is mainly for existing users, since it needs to infer their intestes based on their activities, and see how a pair of users fit into each other's interests. Under cold start scenario, however, since new users have no activities, it is impossible to get their preferences. Thus, $RECON$ only takes one end of a possible relationship into consideration. On the other hand, some other research showed that $RECON$ is not comparable to some collaborative filtering based algorithms [6]. The power of collaborative filtering comes mostly from the ability of utilizing similar users' activities. In online dating recommendation, instead of just looking at one's own activities, it is also useful to look at similar users' activities and learn from them. We believe it is the lack of reciprocity as well as the absence of learning from similar users' activities that lead to *RECON*'s failure of providing reciprocal recommendations to new users.

Instead of using a user's sending contacts to learn his/her preferences, in $rRECON$ we use reciprocal contacts that are initiated by that user. This modification mitigates the problem of lacking of reciprocity in $RECON$, and also provides considerable improvement over the original model. However, a potential threat to $rRECON$ is that most of sending contacts in online dating get no responses. Similar to previous research [6], in our dataset there are only limited sending contacts that finally turn into reciprocal ones. The sparsity of reciprocal contacts limits $rRECON$'s learning ability. On individual basis, even sending activities are limited, neither to say reciprocal ones. Also, like $RECON$, $rRECON$ is lack of the ability to utilize similar users' activities.

The success of our proposed approach is based on multiple factors. First, we find existing users who are similar to new users to help with reciprocal recommendations. Instead of calculating similarities between a new user and all existing users, we conduct community detection for existing users, and match new users to these communities. This step avoids finding customized similarity functions, and thus eliminates drawbacks of content-based recommender systems. To conduct community detection, we borrow ideas from collaborative filtering, and build two networks which are based on similarities of the sending and replying patterns of existing male and female users separately. The modularity scores for male and female communities were around 0.3 and 0.4 respectively, which indicates strong community structures within networks. Second, with communities in hand, we borrow the idea from $RECON$, and create a community profile based on attributes of all community members. We then match each new user to all communities with different similarity scores. This step enables new users to take advantage of all existing users' activities. For each community, we retrieve reciprocal contacts initiated by its community members. Reciprocity is introduced in this step. In the final step of recommendation, for each pair of new and existing users, we fit the new user to every community, and check whether the existing users were contacted by any of this community's members, and then generate a matching score between the new and existing users. With the summation of matching scores generated from all communities, all user similarities and reciprocity are taken into consideration to generate the final reciprocal recommendation list. Experiments validate effectiveness of our approach.

There are spaces for further improvement. First, we define networks based on existing users' sending and reciprocating activities for community detection. There are other ways of defining networks for existing users. Comparing communities detected from different networks will be an interesting work. Second, since our method only considers reciprocal contacts from community members, it may be better if we also include sending contacts. Third, our experiments are conducted with data from a particular U.S. online dating site. It will be great if we can test the power of our approach on other datasets.

References

[1] P. Xia, B. Ribeiro, C. Chen, B. Liu, and D. Towsley, "A study of user behavior on an online dating site," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 2013, pp. 243–247.

[2] J. Akehurst, I. Koprinska, K. Yacef, L. Pizzato, J. Kay, and T. Rej, "Explicit and implicit user preferences in online dating," in *The Behavior Informatics 2011 (BI2011) Workshop - The 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2011)*. Springer, 2011, pp. 15–27.

[3] L. Pizzato, T. Rej, T. Chung, I. Koprinska, K. Yacef, and J. Kay, *Learning User Preferences in Online Dating*, 2010.

[4] L. Pizzato, T. Rej, T. Chung, I. Koprinska, and J. Kay, "RECON: a reciprocal recommender for online dating," in *Proceedings of the fourth ACM conference on Recommender systems*, no. TBA. ACM, 2010, pp. 207–214. [Online]. Available: http://portal.acm.org/citation.cfm?id=1864708.1864747

[5] M. Yu, K. Zhao, J. Yen, and D. Kreager, "Recommendation in reciprocal and bipartite social networks–a case study of online dating," in *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer Berlin Heidelberg, 2013, pp. 231–239.

[6] P. Xia, B. Liu, Y. Sun, and C. Chen, "Reciprocal recommendation system for online dating," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 2015, pp. 234–241.

[7] L. Chen, "A Recommendation Approach Dealing with Multiple Market Segments," in *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, vol. 1. IEEE, 2013, pp. 89–94.

[8] L. Chen, R. Nayak, and Y. Xu, "A Recommendation Method for Online Dating Networks Based on Social Relations and Demographic Information," in *2011 International Conference on Advances in Social Networks Analysis and Mining*. IEEE, 2011, pp. 407–411.

[9] M. E. J. Clauset, Aaron and and C. Moore, "Finding community structure in very large networks," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 70, no. 6 2, pp. 1–6, 2004.