Informational friend recommendation in social media

Conference Paper · July 2013

DOI: 10.1145/2484028.2484179

Shengxian Wan
Chinese Academy of Sciences

Yanyan Lan
Chinese Academy of Sciences

Jiafeng Guo
Chinese Academy of Sciences

Chaosheng Fan
Peking University

Informational Friend Recommendation in Social Media

Shengxian Wan†, Yanyan Lan†, Jiafeng Guo†, Chaosheng Fan‡, Xueqi Cheng†
†Institute of Computing Technology, Chinese Academy of Sciences, Beijing, P.R.China
‡School of Mathematical Sciences, Peking University, Beijing, P.R.China
wanshengxian@software.ict.ac.cn, {lanyanyan, guojiafeng}@ict.ac.cn
fcs@pku.edu.cn, cxq@ict.ac.cn

ABSTRACT

It is well recognized that users rely on social media (*e.g.* Twitter or Digg) to fulfill two common needs (*i.e.* *social need* and *informational need*) that is to keep in touch with their friends in the real world and to have access to information they are interested in. Traditional friend recommendation methods in social media mainly focus on a user's social need, but seldom address their informational need (*i.e.* suggesting friends that can provide information one may be interested in but have not been able to obtain so far). In this paper, we propose to recommend friends according to the *informational utility*, which stands for the degree to which a friend satisfies the target user's unfulfilled informational need, called *informational friend recommendation*. In order to capture users' informational need, we view a post in social media as an item and utilize collaborative filtering techniques to predict the rating for each post. The candidate friends are then ranked according to their informational utility for recommendation. In addition, we also show how to further consider diversity in such recommendations. Experiments on benchmark datasets demonstrate that our approach can significantly outperform the traditional friend recommendation methods under informational evaluation measures.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering

Keywords

Friend Recommendation, Informational Utility, Diversity

1. INTRODUCTION

Recently, online social media (*e.g.* Twitter or Digg) has become one of the main platforms for users to publish, share and obtain information. For users of social media, *social need* and *informational need* are two well recognized basic needs. On one hand, users are active in social media since they can contact their friends in the real world very conveniently; on the other hand, users can access information from others more comprehensively and rapidly.

Friend recommendation is a key approach to help users discover new friends and interesting information. However, traditional friend recommendation methods in social media mainly focus on users' social need (*i.e.* suggesting potential friends one may know), but seldom address their informational need (*i.e.* suggesting friends that can provide information one may be interested in but have not been able to obtain so far). For example, some link prediction methods [8] use the network structure or users' profiles as the basis for recommendations, thus mainly considering the social need. Other methods such as [4] try to utilize user generated content and recommend friends by user-user similarity. However, these methods may only suggest similar users, but not friends that can better satisfy users' unfulfilled informational need. For example, consider user A as the target user, if the similarity between user B and A is 99%, and between user C and A is 50%, it seems that the similarity based methods will recommend B rather than C to A. However, the information from B is useless for A if most of them have already been obtained by A from other friends.

In this paper, we propose to conduct friend recommendation according to the *informational utility*, which is defined as the degree to which a friend satisfies the target user's unfulfilled informational need. We call this approach *informational friend recommendation*. To capture a user's informational need, we view typical posts in social media as items and utilize collaborative filtering (CF) techniques to predict the rating for each post. The candidate friends are then ranked according to their informational utility for recommendation. Furthermore, we discuss how to introduce diversity into the recommendation process to reduce information redundancy. Our experiments on Digg datasets show that our method can significantly outperform the traditional state-of-the-art recommendation methods with respect to informational evaluation measures.

To the best of our knowledge, this is the first paper to explore the user's unfulfilled informational need for friend recommendation, which is a novel approach for studies in social media.

The rest of the paper is organized as follows. In Section 2, we will review some related works on friend recommendation in social media. Section 3 introduces the detailed formulation of our approach. Section 4 presents our experimental results and Section 5 concludes the paper.

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

2. RELATED WORK

Friend recommendation in social networks can be viewed as a link prediction problem [9] in complex networks. Liben-Nowell et al. [8] firstly discussed link prediction in social networks and compared lots of methods based on node proximity. These methods mainly utilize the structural information of a given social network.

Along with structural information, there are many other types of information such as the user's profile and the user generated content of the social network. For example, Hannon et al. [4] proposed different methods to profile the users by user generated content and recommend friends based on profile similarity. Recently, more and more methods utilize structural information together with other types of information for recommendation. For example, Kim et al. [5] proposed a recommendation system for both friend and tweet recommendation. They proposed a novel probabilistic model based on PLSI and combined content and structural information in the model.

Different recommendation methods may be inclined to satisfy different user needs. Chen et al. [1] compared different friend recommendation algorithms in an enterprise social network and found that the algorithms based on information about the social network can find more known users while the algorithms checking for content similarity are more suited for discovering new friends. However, all of the methods mentioned above do not explicitly explore users' unfulfilled informational need for recommendation.

3. OUR APPROACH

Differing from traditional methods, we propose to recommend friends that can better satisfy users' unfulfilled informational need, which is called informational friend recommendation. For this purpose, we need to first mine users' informational need. Here we consider users' unfulfilled informational need as the posts one may be interested in but are unable to obtained from his/her current friends so far, and mine such information with CF methods. We then introduce a quantity named informational utility, which stands for the degree to which a friend satisfies the target user's unfulfilled informational need. Finally, we rank all the friend candidates by high informational utility for recommendation.

3.1 Collaborative Filtering

As mentioned above, the first step for informational friend recommendation is to capture users' unfulfilled informational need. Here we view a post in social media as an item and consider the posts a user may be interested in but which are unable to be obtained so far as his/her unfulfilled informational need. Therefore, we need to accurately predict the rating of each unobtained post for the target user. Naturally, the rating prediction problem can be handily solved by CF techniques.

There are two popular approaches in CF: the neighbourhood-based methods and model-based methods. The basic idea of the neighbourhood-based methods is that similar users or items on observed ratings will act similarly on unobserved ratings. They identify neighbours of users or items, and conduct rating predictions based on their neighbours' ratings. The model-based methods try to uncover latent features that explain the observed ratings. They project users and items to the same latent factor space and the ratings are represented as the inner product of the latent factor vectors.

In this paper, we use several traditional state-of-the-art CF methods to conduct the rating prediction task: user-based KNN, SVD, and SVD++ [6].

Recently, recommendation is more recognized as a ranking problem; therefore ranking techniques are widely used for recommendation. In this paper, we also use the ranking approach by introducing a pairwise ranking loss into SVD and SVD++; the corresponding methods are called Rank-SVD and Rank-SVD++ [2].

With the predicted rating for each post, we sort the posts in descending order according to rating and select the top N as the post list that the target user is interested in but cannot obtain from their existing friends, denoted by list $L_{need}(u)$. We use $I_{need}(u)$ to denote the set of posts in $L_{need}(u)$. Note that in our work, we set N as 200.

3.2 Recommend with Informational Utility

We now introduce the quantity named informational utility, which stands for the degree to which a friend satisfies the target user's unfulfilled informational need. Formally, the informational utility of user u' given the target user u is defined as:

$$\mathcal{IU}_u(u') = \frac{\sum_{i \in I_{need}(u) \bigcap I_{pub}(u')} \text{weight}(i)}{|I_{pub}(u')| + C}, \qquad (1)$$

where $I_{pub}(u')$ is the set of posts from user u' (by sharing, publishing or other activities), weight(i) is the weight function of post i which decreases with the rank position in the ranking list $L_{need}(u)$, and C is a smoothing factor. Note that the numerator denotes the weighted information gain the target user u can obtain from user u', while the denominator denotes the cost for user u when browsing the posts from user u'. The smoothing factor C is used to eliminate inactive users with few posts and to prevent over-fitting.

Based on Eq. (1), the top-k friend recommendation task can be accomplished by ranking the candidate users according to their informational utility for the target user.

3.3 Diversity

Beyond recommendation with respect to each user's informational utility independently, diversity is another important factor which is usually considered in information retrieval (IR) and recommender systems. In this context, diversity means that the users in a recommendation list should cover as many different aspects of informational need of the target user as possible. In IR, one popular strategy is to consider documents as composed of information nuggets and diversity is formalized as a max-COVER problem solved by a greedy algorithm [3]. Here we borrow this idea to achieve diversity in informational friend recommendation.

While it is non-trivial to identify nuggets in documents, our approach can naturally use the posts as the nuggets. To achieve diversity, we introduce a weight decay factor α into Eq. (1). Then in the greedy process, as long as a post is covered by the previously selected friend, the weight of the post will be punished by α to prevent information redundancy. The modified informational utility for diversity is defined as:

$$\mathcal{IU}_u(u') = \frac{\sum_{i \in I_{need}(u) \bigcap I_{pub}(u')} \text{weight}(i)\alpha^{ct(i)}}{|I_{pub}(u')| + C}, \qquad (2)$$

where $ct(i)$ is the time of post i being covered. Essentially, if α is equal to 1, it turns back to the original informational utility.

4. EXPERIMENTS

In this section, we conduct experiments on real social media datasets and compare different CF methods and friend recommendation methods.

4.1 Experimental Settings

In this work, we use data from Digg [1] collected by Lerman et al. [7] in June 2009 through Digg API for experiments. Digg is a social news aggregator where social need and informational need are both important for the users. Besides, users of Digg can access posts from both their friends and other sources such as the front page. As a result, we can construct the ground-truth of any given user's informational need by the posts digged from the front page but not from their friends. The data contains the information of social network and user digging behaviours. Overall, there are 3,553 stories, 3,018,197 votes, 139,409 users and 1,731,658 asymmetric links between these users.

We sort the story posts by the time they were published and separate them into training, validation and testing set, with sizes 2553, 500, 500, respectively. For the training set, we separate it into two subsets for each user, one is the set of posts from other friends and the other is the set of posts from other sources, denoted as train1 and train2 respectively. Train1 is used for the training of CF and train2 is used as the test data to evaluate the performance of CF methods. After that, the ranking strategies are conducted over the whole training set, and the validation set is used for tuning the smoothing factor C. Finally, measurements on test data are used to evaluate the performance of the recommendation methods.

4.2 Experiments on Collaborative Filtering

Firstly, we compare the results of different CF methods on our dataset, including KNN, SVD, SVD++, rank-SVD and rank-SVD++. In the experiments, we define the positive examples (*i.e.* a user diggs a post) as rating 1. We can see that this is a one-class CF problem, so we sample equal amount of negative samples defined as rating 0 in each round of iteration in the optimization process of these methods. For SVD++, we use all the positive examples as user feedback since we have no implicit user feedback data. In this section, we use precision@n as the evaluation measure, since the output of the first step is a ranking list.

From the experimental results in Fig.1, we can see that: (1) All the model-based CF methods (SVD, SVD++, Rank-SVD, Rank-SVD++) work better than neighbourhood-based CF methods (KNN); (2) The methods based on ranking loss (Rank-SVD, Rank-SVD++) are better than the corresponding methods based on regression loss (SVD, SVD++). These results are in accordance with previous studies.

In the following friend recommendation experiments, we choose Rank-SVD as the CF method due to its good performance and simpler implementation over Rank-SVD++.

4.3 Experiments on Friend Recommendation

Now we compare our method (denoted as Inf-Rec) with some state-of-the-art friend recommendation methods.

For baselines, we use three commonly used methods which only use the network structure, including: (1)*friend-of-friend* (FdFd): this method simply recommends the friends of friends to the target user; (2)*follower-of-friend*(FlFd): this method recommends users who have similar friend sets as the target user; (3)*friend-of-follower-of-friend*(FdFlFd): this method goes further than FlFd by recommending the friends of FlFd to the target user. In addition, we also propose two other baselines which focus on users' informational need for recommendation[2], including: (1)*sharing similarity*(Share-Sim): this method ranks users by Jaccard similarity of digging behaviours with the target user; (2)*sharing efficiency*(Share-Effi): this method ranks users by the ratio of common posts with the target user to the total posts they digged.

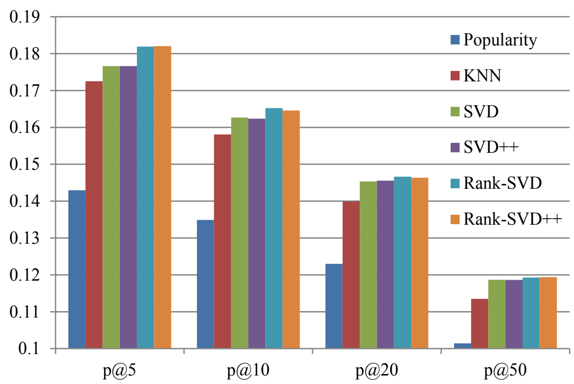

Figure 1: Results of different CF methods

In this paper, we mainly focus on satisfying users' informational need in friend recommendation. Therefore, we use an informational evaluation measure for comparison, which is defined as the efficiency of the potential information gain that one can obtain from the recommendation:

$$EIG@k = \frac{1}{k}\sum_{i=1}^{k}\frac{|I_{need}(u)\bigcap I_{pub}(u_i)|}{|I_{pub}(u_i)|}, \quad (3)$$

where k stands for the number of friends recommended.

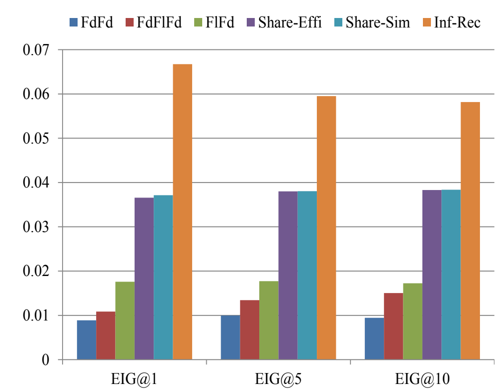

Figure 2: Evaluation results of recommendation lists

From the experimental results in Fig.2, we can see that: (1) Our method can significantly outperform the baselines. For example, the relative improvement of Inf-Rec over the

[1] *http://digg.com/*. Digg's home page has changed several times in recent years and now it is very different from the old version when the dataset was crawled.

[2] Since in the Digg datasets, there is no content information, we could not implement the traditional methods on content information.

worst baseline is 497%, and the relative improvement of InfRec over the best baseline is 56.4% on $EIG@5$; (2) The methods based on network structure (FdFd, FlFd, FdFlFd) perform the worst amongst all the methods. For example, the average $EIG@1$ is 0.0124, which is much smaller than that of the two designed baselines based on user digging behaviours (0.0369). Therefore, we can conclude that both methods based on network structure and user digging behaviours cannot properly address the user's unfulfilled informational need. Our method fully explores a user's informational need by the defined informational utility, and thus achieves better performance.

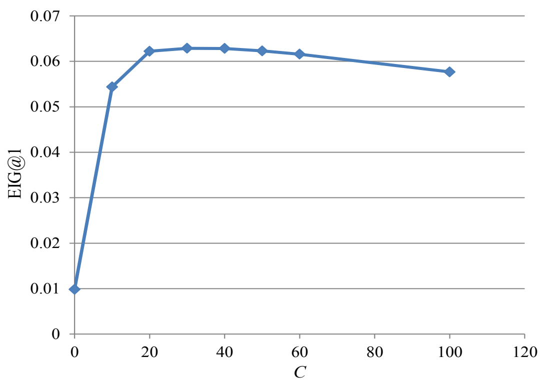

Figure 3: Results with different smoothing factor

4.4 Discussions

In this section, we discuss the effect of the smoothing factor C and diversity mentioned in Section 3.3.

(1) Fig.3 shows the varying curve of EIG with C on the validation set. We can see that the algorithm will be severely over-fitting without the smoothing factor, and the result will stay stable when C becomes large enough. In our experiments, we set $C = 30$.

(2) Fig.4 shows the SET_COST/SET_GAIN curve on the test set to compare the diversity of recommendation lists, where SET_COST is the sum of users' costs in the recommendation list and SET_GAIN is defined as

$$SET\_GAIN@k = \left| \left(\bigcup_{i=1}^{k} I_{pub}(u_i) \right) \bigcap I_{need}(u) \right|, \qquad (4)$$

where k stands for the number of friends recommended. A recommendation with higher gain and lower cost at set level is better in the sense of diversity. We can see that by further introducing the decay factor $\alpha = 0.3$ in our method, it can achieve better performance in terms of diversity.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed to conduct friend recommendation according to users' informational need, which is called informational friend recommendation. Firstly, we utilized collaborative filtering techniques to capture users' unfulfilled informational need. Secondly, two ranking strategies were conducted according to informational utility to obtain the recommendation. Our experimental results on Digg datasets show that our method significantly outperforms the current state-of-the-art friend recommendation methods.

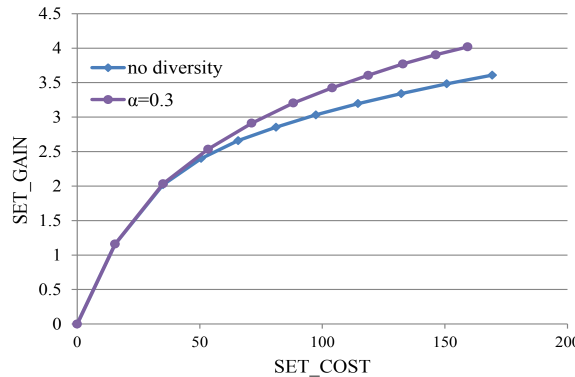

Figure 4: Diversity result of recommendation lists, where the points on the curve corresponding to different ks.

As for future work, we plan to investigate how to conduct informational friend recommendation with a unified approach other than the two-step method in this paper.

6. ACKNOWLEDGEMENTS

This research work was funded by the National Natural Science Foundation of China under Grants No. 61232010, No. 61203298, No. 61003166, 863 Program of China under Grants No. 2012AA011003, and National Key Technology R&D Program under Grants No. 2011BAH11B02, No. 2012BAH39B02, No. 2012BAH39B04.

7. REFERENCES

[1] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy. Make new friends, but keep the old: recommending people on social networking sites. CHI '09, pages 201–210, 2009.

[2] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, and Y. Yu. Collaborative personalized tweet recommendation. SIGIR '12, pages 661–670, 2012.

[3] C. Clarke, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. SIGIR '08, pages 659–666, 2008.

[4] J. Hannon, M. Bennett, and B. Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *ACM Recsys*, pages 199–206, 2010.

[5] Y. Kim and K. Shim. TWITOBI: A Recommendation System for Twitter Using Probabilistic Modeling. *ICDM '11*, pages 340–349, 2011.

[6] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. KDD '08, pages 426–434, 2008.

[7] K. Lerman and R. Ghosh. Information Contagion: an Empirical Study of Spread of News on Digg and Twitter Social Networks. In *ICWSM*, 2010.

[8] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. CIKM '03, pages 556–559, 2003.

[9] L Lü and T Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, (6):1150–1170, 2011.