Friend Recommendation in Online Social Networks Combining Interest Similarity and Social Interaction

Mingqing Huang[a], Bofeng Zhang[a,*], Guobing Zou[a,*], Shulin Cheng[a], Zhuocheng Zhou[a], Furong Chang[a,b]

[a]*School of Computer Engineering and Science, Shanghai University*
Shanghai 200444, China
bfzhang@shu.edu.cn (B. Zhang), gbzou@shu.edu.cn (G. Zou)

[b]*School of Computer Science and Technology, Kashgar University*
Kashgar 844006, China
cfrkashger@shu.edu.cn (F. Chang)

***Abstract*—With the explosive popularity of online social networks (OSNs), the considerably large number of online users and their diverse activities have posed great challenges on social recommendation. However, most conventional methods for recommending friends in OSNs cannot simultaneously satisfy the requirements of accuracy and timeliness. By taking full advantage of latent Dirichlet allocation (LDA), in this study, we designed a friend recommendation approach combining interest-based features and interaction-based topologies with linear time complexity. The experimental results obtained using a real-world micro-blogging network demonstrated that the proposed hybrid scheme outperforms the other three state-of-the-art algorithms in terms of effectiveness and efficiency.**

***Keywords*—friend recommendation; online social network; interest similarity; social interaction; latent Dirichlet allocation.**

I. Introduction

Due to the pervasive popularity of social networking sites (SNSs) and their applications, people from diverse areas are increasingly depending on various SNSs for sharing their interests and interacting with one another. Better-known sites, such as Twitter and Facebook, can archive a very large number of tweets or posts updated by active individuals from all over the world every day. As a result, the sheer volume of user-generated content leads to opportunities for social computing [1] and poses a challenge in terms of the information overload.

Consistent with what people usually do in actual life, SNS users always try to enlarge their social circles to satisfy various social needs, e.g., cooperation, business, and leisure. Therefore, recommendation methods, the most notable applications of web personalization, are gradually attracting people's attention as a filtering technology facilitating individuals with the most pertinent and crucial information. In this study, we aimed to locate potential friends for target users on the basis of the knowledge extracted from the information embedded in their social links and behavioral data.

*Corresponding authors.

This work was partially sponsored by the National Key Research and Development Program of China under Grant 2017YFC0907505, the National Natural Science Foundation of China under Grant 61772128, the Shanghai Natural Science Foundation under Grant 18ZR1414400, and the Xinjiang Social Science Foundation under Grant 2015BGL100.

Most of the existing methods for recommending friends exploit and utilize the similarity of user personalities [2], the geographical vicinity [3], or the number of common friends [4]. Some other studies cast friend recommendation as a link prediction issue by identifying the most probable links among unconnected nodes [5]. Dong et al. [6] took into account heterogeneous topologies of social networks and obtained better precision. Oyama et al. [7] gathered more links among nodes in different time frames and made recommendations by accumulating this information in dynamic scenarios. By quantizing the trust of social ties, Chen et al. [8] combined collaborative filtering and social relationships to improve the recommendation accuracy for cold-start users.

However, most of the previous research focuses only on either the articulated social network structure or the diverse and dynamic interest similarity. Even for the collaborative filtering technology, which takes into account both the common interests and the social relationships, the computing complexity is prohibitively expensive for large-scale social networks. In this work, by making the most of latent Dirichlet allocation (LDA), we developed a friend recommendation approach combining interest similarities and social interactions with a relatively high success rate and low time complexity.

The major contributions and characteristics of this work can be summarized as follows:

- The most influential individuals on each latent topic are identified and sorted by comprehensively considering their own interest distributions and the number and impact of followers.
- Potential friends sharing sufficient common interests with the target user are selected by leveraging the full characteristics of LDA with linear time complexity, to scale up with an increase in the social network size.
- Interest similarity is combined with social relationships to generate the rating patterns for friend recommendation via an adaptive regularization parameter, which significantly increases the success rate.

The rest of this paper is structured as follows. The system

framework of the proposed friend recommendation scheme and the functionality of its components are described in depth in Section II. The experimental results on a real-world micro-blogging network are demonstrated and analyzed in Section III. Finally, in Section IV, we conclude the work briefly with several directions for future research.

II. Proposed Approach

A. System Overview

In this paper, Chinese micro-blogging is referred to as the study platform. Nevertheless, the proposed hybrid framework with modest adjustments can be extended to any micro-blogging site and even social media in other languages.

To achieve the purpose of accurate prediction, the procedure for friend recommendation consists of four crucial steps of topic distillation, domain-specific influential individuals, candidates with similar inclinations, and recommendation set formation. The detailed overall architecture is displayed in Fig. 1, where the data gathering of the micro-blogging site has been plugged as a pre-existing condition.

First, text analysis was performed on the micro-blog content via the pre-procedure and the LDA model for interest quantization on specific topics. Then, the mining of top-k influential individuals on each latent topic was performed by taking into consideration both the interest degree and the impact of followers. Next, potential candidates with similar inclinations were identified by utilizing the dominant interest

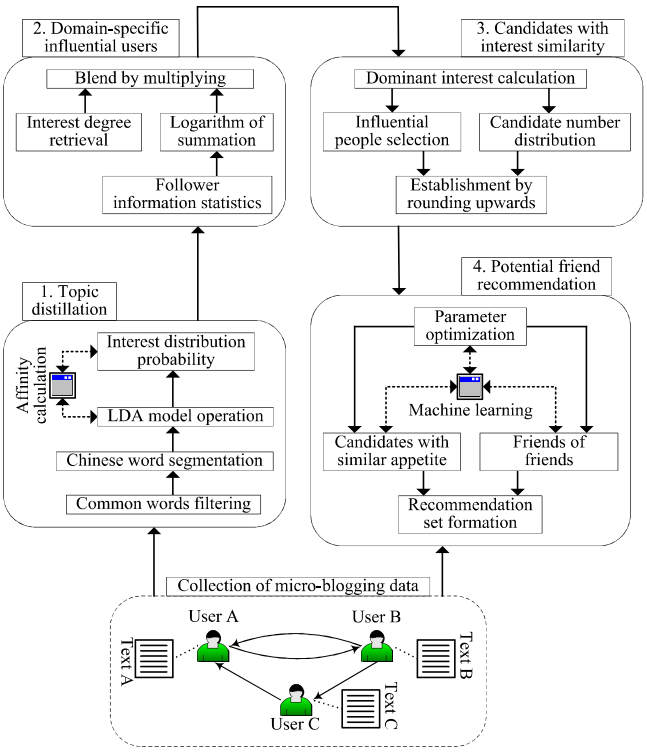


Fig. 1. System framework of the proposed approach.

distribution of the target user. Finally, the recommendation set was constructed by combining the interest similarity and the social interactions through parameter tuning. All of the components of the framework are elaborated in the following sections.

B. Interest Quantification

There are many common words in the text corpora of micro-blogs, such as modal particles, auxiliary verbs, and interjections. However, these common words demonstrate little meaningfulness in realistic applications for interest extraction. In this study, we eliminated them early to suppress interference and increase efficiency by a large margin.

We assembled the micro-blogs posted recently by the same user into an archive, as the goal was to extract the themes that each individual was interested in rather than the subject of each single micro-blog. Next, automatic Chinese word segmentation [9] was performed for each archive.

As an unsupervised machine learning scheme to identify the latent theme information from a large archive set, the LDA model [10, 11] was employed to estimate the interest degrees of a user on different topics. It considered each archive to be a bag of words, so each archive denoted a probability distribution over some themes and each theme, a probability distribution over a large number of words.

Gibbs sampling was applied to assess the model parameters obtained from the data, i.e., the document-theme associations and the theme-word associations. The result was expressed in a $D \times T$ matrix, labeled DT, where D represents the number of individuals and T denotes the number of themes. DT_{ij} counts the number of times a word in user u_i's micro-blogs has been attached to theme t_j.

Without any loss of generality, we then normalized matrix DT as DT' so that $\|DT'_{i\cdot}\|_1 = 1$ for each row $DT'_{i\cdot}$. Each row of DT' corresponds to the probability distribution of a user's interest over all of the themes; i.e., the element DT'_{ij} captures the probability that individual u_i is interested in theme t_j.

C. Domain-Dependent Influential Individuals

Obviously, the recommended friends should be those who not only share a large number of common interests with the target user but are also the influential ones in the related fields. Consequently, we only needed to keep the records of individuals with a relatively large impact in the descending order on each topic for identifying potential candidates with similar inclinations.

The impact ranking counts a user's fascination to establish friendships in a social network. That is, the higher the ranking is, the easier it is for the user to be endorsed as a friend. The domain-specific impact depends on both the probability distribution of interest over topics and the topological connection of social relationships, which includes the following three aspects: 1) how much interest the potential candidate

devotes to the relevant content; 2) how many following edges sponsored by other users exist; and 3) how much influence (measured by the number of followers) the user's followers gain.

On the basis of the above discussion, the impact ranking $r_j(i)$ of user u_i on topic t_j was derived mathematically as follows:

$$r_j(i) = DT'_{ij} \times \sum_{u_v \in \Gamma_i} \left(log\left(\lambda \times k_v^{(in)} + 1\right) + 1 \right). \quad (1)$$

Where Γ_i denotes the set of followers of u_i; $k_v^{(in)}$ is the in-degree of u_v representing the number of followers; the first term "+1" accounts for the elimination of the minus values of the logarithm; the second term "+1" results from the contribution of u_v itself; and λ stands for a tunable parameter, which can be set equal to the base of the logarithm.

D. Potential Candidates with Similar Inclinations

In this section, the most important interests of the target user, i.e., the topics with larger probability values, were first defined as *dominant interests*. According to the probability distribution of the dominant interests, we selected the appropriate proportions of potential candidates from the most influential individuals on the corresponding topics.

1) *Dominant Interests*

The interests of higher significance, i.e., a certain number of topics with a higher level of attention, were defined as the dominant interests for a specified user.

Definition 1 (Dominant Interests). *The set of dominant interests of user u_i, marked D_i, is a subset of topics addressing the following demands:*

(1) *The total probability value of the set is equal to or larger than η, which is a pre-defined threshold.*

(2) *The probability distribution of any topic in the set is equal to or larger than that of any topic not in the set.*

(3) *The set should keep the minimum number of topics.*

Labeling the probability vector of topics for user u_i as $P_i = [p(z_1|u_i), p(z_2|u_i), \ldots, p(z_{|Z|}|u_i)]$, we sorted the probability distribution in the descending order to gain the updated vector $\hat{P}_i = [p(\hat{z}_{i1}|u_i), p(\hat{z}_{i2}|u_i), \ldots, p(\hat{z}_{i|Z|}|u_i)]$, where $p(\hat{z}_{im}|u_i) \geq p(\hat{z}_{in}|u_i)$ if $m < n$. The size of the dominant interest set was calculated as follows:

$$q_i = \arg \min_q \left(\sum_{j=1}^{q} p(\hat{z}_{ij}|u_i) \geq \eta \right). \quad (2)$$

Finally, we derived the dominant interests as set $D_i = \{\hat{z}_{i1}, \hat{z}_{i2}, \ldots, \hat{z}_{iq_i}\}$.

For the sake of simplicity, we set $\eta = 0.5$ in this study. A more sophisticated and precise parameter value can be acquired through machine learning in a realistic application, which we intend to pursue in our future work.

2) *Interest Candidate Set*

Given the domain-dependent influential individuals on each latent topic and the probability distribution of the dominant interests of the target user, the establishment procedure for potential candidates with the interest similarity for friend recommendation is formally represented in Algorithm 1.

Algorithm 1: Congenial candidate selection.

Input: target user u_t, existing friend set FS, k_i; /*k_i denotes the number of potential friends with similar inclinations.*/

Output: interest candidate set CS_i;

$\hat{p}_t = \sum_{\hat{z}_{tr} \in D_t} p(\hat{z}_{tr}|u_t)$; /*$D_t$ expresses the dominant interest set of u_t.*/

$CS_i = \varnothing$;

for $r \leftarrow 1$ **to** q_t /*q_t *captures the number of topics in* D_t.*/ **do**

  $N_r(t) = \left\lceil k_i \times \frac{p(\hat{z}_{tr}|u_t)}{\hat{p}_t} \right\rceil$;

  if $k_i - |CS_i| \geq N_r(t)$ **then**

    select $N_r(t)$ number of individuals with the maximum domain-specific impact on topic $\hat{z}_{tr}$ that are not in $FS \bigcup CS_i$, and add them to CS_i;

  else if $0 < k_i - |CS_i| < N_r(t)$ **then**

    pick $(k_i - |CS_i|)$ number of persons with the maximum domain-dependent influence on topic $\hat{z}_{tr}$ that do not belong to $FS \bigcup CS_i$, and append them to CS_i;

  else

    break;

  end

end

return CS_i;

E. Friend Recommendation

Characterizing the users' existing social relationships as 1-hop social ties, we located the potential friends with graph approximation by exploiting 2-hop social chains, i.e., friends-of-friends. Then, we constructed the final recommendation set consisting of potential candidates with interest similarities or social intimacy.

1) *Sociable Candidate Set*

In fact, the types of interactions that take place routinely in a social network play an important role in the assessment of the relationship strength [12]. There are six prominent kinds of interactions in micro-blogging, which include private messages, chatting, wall posts, comments, forwarding, and profile visits. However, the modeling of these ties is inherently difficult as it could be a combination of various activities. Note that these interactions are inadequate unless the two important features of each interaction, i.e., frequency and recency are taken into consideration.

In this paper, only the "following" relationships among the entities in micro-blogging are leveraged to compute the

interaction intensities because of space and time limitations. The exact strength values of social ties that counted for all of the transactional events are deferred to the full version of the paper.

The identification process of the potential candidates with social interactions for friend recommendation involves the following crucial steps:

(1) Retrieve all of the friends-of-friends of the target user, who are not contained in CS_i (see Algorithm 1), and add them as keys to an empty mapping collection MC.

(2) For each existing friend u_i of the target user, if $u_j \in MC$ is endorsed as a friend by u_i, then increase the mapping value of u_j by one.

(3) Return k_s number of keys with the maximum mapping values from MC as the sociable candidate set CS_s.

2) *Final Recommendation Set*

Adopting a regularization parameter α, whose value is determined through machine learning in a realistic application, we selected appropriate proportions of recommended friends from the potential candidates with similar inclinations and graph approximation, respectively. By leveraging α, we calculated the sizes of interest and sociable candidate sets, denoted k_i and k_s separately, as follows:

$$\begin{cases} (1-\alpha)\, k = k_i \\ k_i + k_s = k \end{cases}. \tag{3}$$

Taking the interest similarity and the social interactions into account, we have described the pseudo-code of the comprehensive mechanism for friend recommendation in Algorithm 2.

Algorithm 2: Potential friend recommendation.

Input: target user u_t, existing friend set FS, k, regularization parameter α; /*k represents the number of recommended friends.*/

Output: final recommendation set RS;

```
k_i = (1 − α) k;
k_s = αk;
invoke Algorithm 1 with parameters u_t, FS, and k_i, and obtain the interest candidate set CS_i;
gain the sociable candidate set CS_s by taking the identification procedure with parameters u_t and k_s (see Section II-E.1);
RS = CS_i ⋃ CS_s;
return RS;
```

F. Computation Complexity

At first blush, it appears that the Gibbs sampling procedure operates very slowly. But in fact, after the training process, the forecasting process for topic distillation has a linear correlation with the number of users.

For each latent topic, only a certain number of most influential individuals were retrieved and preserved, which avoided the sorting in the ascending or descending order of all of the users. Therefore, the time taken to locate potential candidates with similar inclinations was $O(n)$, here n indicates the number of entities in the social network.

Identifying friends-of-friends of the target users and estimating the social intimacy only involved counting the "following" relationships, which led to the time complexity of $O(m)$, where m stands for the number of social ties.

In general, as $m > n$ in the connectivity network, the overall computational complexity of the proposed scheme was $O(m)$. In a sparse social network, the time complexity was further reduced to $O(n)$ due to $m \propto n$.

III. Experimental Evaluation

A. Gathering of Micro-blogging Data

Sponsored by Tencent Technology Co. Ltd., Tencent micro-blogging is an online social networking service similar to Twitter. By the end of 2012, the number of registrants reached up to 540 million. In particular, the number of daily active members has surpassed 100 million. Moreover, Tencent micro-blogging provides an open and free application program interface (API) [13] for developers to collect the global user data. Given these conspicuous advantages, we took Tencent micro-blogging as the testing platform for the performance assessment.

Statistics indicate that every member has dozens of friends on average. Hence, randomly selecting some participants to construct a "following" relationship sub-network did not accurately demonstrate the social properties of these persons. Therefore, we started from a specified individual who was treated as the initial entity in the social sub-network, and all of the followers of this user were then recruited to the sub-network. In a similar fashion, all of the followers of these newcomers were added again. This procedure was repeated till the number of participants met the experimental requirements. During this period, the persons who had more than 1000 friends or followers were eliminated, as we considered them to be stars or friend abusers.

The performance testing was executed with two batches of micro-blogging data relevant to the same 12,746 members, gathered in June and August 2015 for the friend recommendation and the performance evaluation separately. The total number of social ties among these persons reached 83,809 in June.

B. Evaluation Metrics

To assess the success rate of the proposed scheme, the three most popular measurement criteria in the case of binary recommendation, i.e., *Precision*, *Recall*, and *F-Score*, were calculated as follows:

$$Precision = \frac{|E_r \bigcap E_a|}{|E_r|}. \tag{4}$$

$$Recall = \frac{|E_r \bigcap E_a|}{|E_a|}. \tag{5}$$

$$F\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \quad (6)$$

Where E_a stands for the set of the actually formed "following" relationships and E_r represents the set of the recommended social ties.

C. Compared Methods

In this work, three other existing state-of-the-art algorithms were adopted to compare with the proposed scheme, which are the typical representatives of interest-based, sociability-based, and collaborative-filtering-based methods, respectively.

1) *Interest-based approach*

Xie [14] designed a general recommendation framework for exploring new friends with common interests. The cosine similarity measure was employed to calculate the correlation between two vectors of the users' favorite items as follows:

$$Similarity\,(u_i, u_j) = \frac{\sum_{k=1}^{n} I_{i,k} * I_{j,k}}{\sqrt{\sum_{k=1}^{n} (I_{i,k})^2 \sum_{k=1}^{n} (I_{j,k})^2}}. \quad (7)$$

Where $I_{i,k}$ represents the interest degree of individual u_i on the kth item and n indicates the number of items.

The computational complexity was $O(n^2)$, here n denotes the number of individuals in the network.

2) *Sociability-based method*

We tested the conventional algorithm for friend recommendation, i.e., *friends-of-friends*, which recommended potential friends by using the number of 2-hop social relationships. For each candidate edge l (with the arrow pointing from follower i to friend j), we leveraged the number of 2-hop social ties between the two endpoints to compute its recommendation score as follows:

$$Score\,(l) = |\{x\,|(i,x) \in E \wedge (x,j) \in E\}|\,. \quad (8)$$

The time complexity was $O\,(m)$, where m indicates the number of edges in the network.

3) *Collaborative-filtering-based scheme*

Agarwal et al. [15] incorporated collaborative filtering into the exploitation of new friends with similar inclinations for online users. They computed a weighted average $P\,(r_{i,j})$ to predict the rating score that user u_i would assign to user u_j as follows:

$$P\,(r_{i,j}) = \frac{1}{\sum\limits_{u_k \in N(u_i)} sim\,(u_i, u_k)} \sum_{u_k \in N(u_i)} sim\,(u_i, u_k) \times r_{k,j}. \quad (9)$$

Where $N\,(u_i)$ indicates the set of users similar to u_i; $sim\,(u_i, u_k)$ expresses the attribute similarity between users u_i and u_k, which is calculated by using Pearson's correlation coefficient; and $r_{k,j}$ indicates the rating score of u_k to u_j.

The computing complexity was $O(n^2)$; here, n indicates the number of participants in the network.

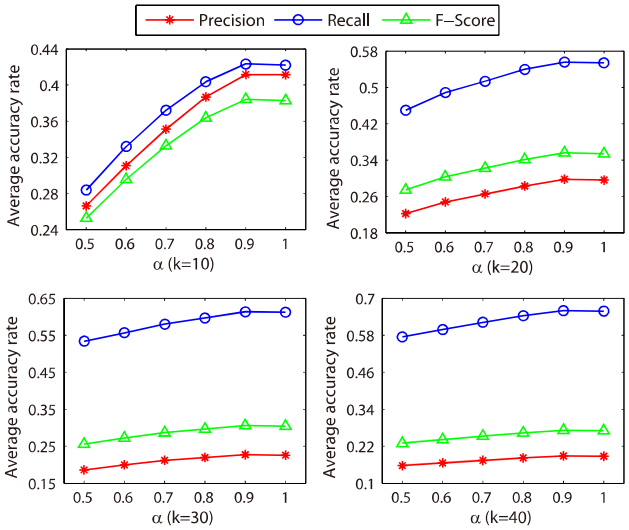


Fig. 2. Relationship between the accuracy rate and the regularization parameter with different numbers of recommended friends.

D. Experimental Results

The LDA model depends on three arguments, i.e., theme number T, and document-topic and topic-word hyper-parameters α and β. In this work, we set them as $T = 100$, $\alpha = 50/T$, and $\beta = 0.1$. With different values of these parameters, the results of the interest quantification were adjusted overall. However, this was not investigated because the focus of this work was on the enhancement of the accuracy of friend recommendation in social networks. In fact, the results and revelation of this work were not limited by the very particular assignment of these three arguments.

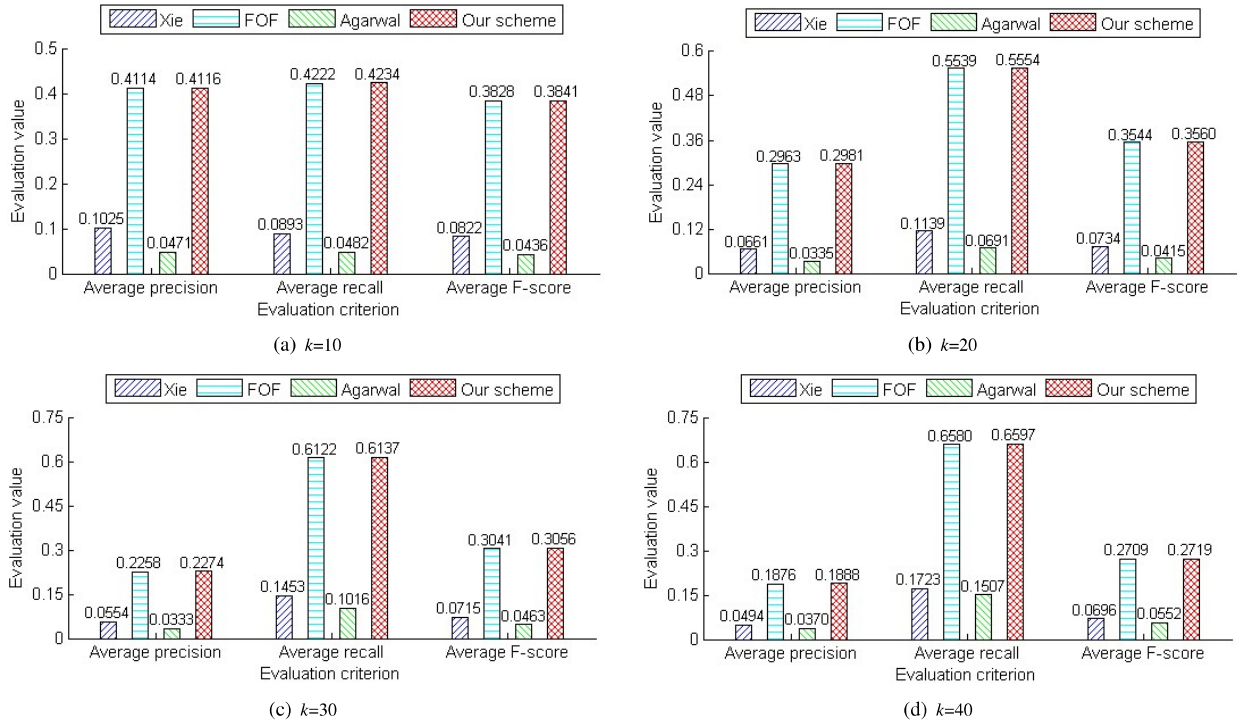
Out of the 12,746 persons, we selected 121 active ones as the target users for the testing of the friend recommendation. Different from the remaining persons who did not update their social ties frequently, the selected users almost established or terminated at least 5 "following" relationships during the considered two-month interval.

1) *Regularization Parameter Tuning*

The theoretical analysis and estimation results revealed that the variation in the regularization parameter $\alpha \in [0, 1]$ of our scheme severely affected the recommendation list. As illustrated in Fig. 2, the success rate of the friend recommendation varied for different values of α. At point $\alpha = 0.9$, the algorithm achieved the optimal effect. Thus, this parameter value was involved as the default configuration in the following comparative experiments.

2) *Comparison with Other Algorithms*

The results of the comparison tests between the proposed approach and the other three methods are depicted in Fig. 3. In particular, note that the success of *friends-of-friends* could be attributed to the fact that Tencent adopts a topology-based algorithm for link recommendation. However, the proposed



(a) k=10 (b) k=20

(c) k=30 (d) k=40

Fig. 3. Effect comparison of four algorithms on a micro-blogging sub-network for friend recommendation. Here, *friends-of-friends* is abbreviated as FOF for ease of presentation.

scheme obtained the highest performance with linear computational complexity.

IV. Conclusions and Future Work

The analysis and mining of online social networks is an inherently interdisciplinary academic activity. By sufficiently leveraging the characteristics of the LDA model, in this paper, we presented a novel friend recommendation scheme taking both interest similarities and social interactions among users into consideration. Theoretical analysis and experimental results verified that the proposed approach was superior to the other three state-of-the-art algorithms with linear runtime complexity; it could satisfy the practical application requirements of large-scale social media.

The current design of the proposed method takes into account only the semantic similarities and the "following" relationships among users in a social network, leaving several aspects to be improved in the future. First, we intend to integrate the contextual information into the attribute similarity for higher accuracy, including location and time. Another important direction is to improve the strength calculation of a social relationship by incorporating more interpersonal interactions, such as forwarding and comments. Finally, we plan to further verify the performance of our scheme in large-scale social network datasets.

Acknowledgments

We would like to express our sincere appreciation to all of the anonymous reviewers for their sharp comments and constructive suggestions that have obviously upgraded the presentation of this paper.

References

[1] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, and et al, "Life in the network: the coming age of computational social science," *Science (New York, NY)*, vol. 323, no. 5915, 2009, article. 721.

[2] Z. Wang, J. Liao, Q. Cao, H. Qi, and Z. Wang, "Friendbook: a semantic-based friend recommendation system for social networks," *IEEE transactions on mobile computing*, vol. 14, no. 3, pp. 538–551, 2015.

[3] X. Yu, A. Pan, L. A. Tang, Z. Li, and J. Han, "Geo-friends recommendation in gps-based cyber-physical social network," in *Proceedings of the IEEE International Conference on Advances in Social Networks Analysis and Mining*, 2011, pp. 361–368.

[4] Z. Yu, C. Wang, J. Bu, X. Wang, Y. Wu, and et al, "Friend recommendation with content spread enhancement in social networks," *Information Sciences*, vol. 309, pp. 102–118, 2015.

[5] R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer, "Folks in folksonomies: social link prediction from shared metadata," in *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, 2010, pp. 271–280.

[6] Y. Dong, J. Tang, S. Wu, J. Tian, N. V. Chawla, and et al, "Link prediction and recommendation across heterogeneous social networks," in *Proceedings of the 12th*

IEEE International Conference on Data mining, 2012, pp. 181–190.

[7] S. Oyama, K. Hayashi, and H. Kashima, "Cross-temporal link prediction," in *Proceedings of the 11th IEEE International Conference on Data mining*, 2011, pp. 1188–1193.

[8] L. Chen, C. Shao, P. Zhu, and H. Zhu, "Using trust of social ties for recommendation," *IEICE Transactions on Information and Systems*, vol. 99, no. 2, pp. 397–405, 2016.

[9] "Stanford word segmenter software package," *<https://nlp.stanford.edu/software/segmenter.shtml>*.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[11] M. Huang, G. Zou, B. Zhang, Y. Gan, S. Jiang, and et al, "Identifying influential individuals in microblogging networks using graph partitioning," *Expert Systems with Applications*, vol. 102, pp. 70–82, 2018.

[12] M. K. Sohrabi and S. Akbari, "A comprehensive study on the effects of using data mining techniques to predict tie strength," *Computers in Human behavior*, vol. 60, pp. 534–541, 2016.

[13] "Tencent micro-blogging open platform," *<http://dev.t.qq.com>*.

[14] X. Xie, "Potential friend recommendation in online social network," in *Proceedings of the IEEE/ACM International Conference on Cyber, Physical and Social Computing*, 2010, pp. 831–835.

[15] V. Agarwal and K. K. Bharadwaj, "A collaborative filtering framework for friends recommendation in social networks based on interaction intensity and adaptive user similarity," *Social Network Analysis and Mining*, vol. 3, no. 3, pp. 359–379, 2013.