Predict Whom One Will Follow: Followee Recommendation in Microblogs

Hao Wu†, Vikram Sorathia, Viktor K. Prasanna
†Department of Computer Science
Ming Hsieh Dept. of Electrical Engineering
University of Southern California
Email: {†hwu732, vsorathi, prasanna}@usc.edu

***Abstract*—Microblogging services such as Twitter and Tencent Weibo have enjoyed drastic popularity in the latest few years. Recommender is essential to those microblogs as a means to find items (users or other information sources such as organizations) that might interest a user to follow. It can greatly improve user experience as well as reduce the risk of information overload might be introduced by irrelevant followees. In this paper, we examine some of the most influential factors that user might consider in selecting followees, in the hope of recommending interesting items to match each user's preferences. We investigate a large scale microblog data extracted from Tencent Weibo and conduct the evaluation of recommendations based on the guideline proposed by the challenge of Track 1 in KDD Cup 2012. Statistical analysis of the log of user actions regarding to recommendations reflect only about 7% acceptance. Experimental results show the popularity of an item is more attractive to users than other features such as the matching of item category, keywords and the influence of user actions and current followees' acceptance.**

I. Introduction

Social networking services have enjoyed phenomenal growth in latest few years, and micblogging websites such as Twitter are extremely popular ones as a medium of real-time information and news spreading [7]. People can freely share information, opinions, knowledge, insights and experience by taking advantage of the open and effective nature of microblogging system. The main activity of microblogging is tweet that is the action of a user posting a short message. For the ease of reading and disseminating, the tweet message is usually restricted to certain length, e.g., 140-characters on Twitter. Furthermore, a user can add some comments to a tweet or repost the tweet as "retweet". The visibility of tweet messages are defined by "*follower-followee*" relations that a user can follow any other users (or groups, organizations, etc) without consensus and read their tweets. A follower will receive all the microblogs from the users he follows, named *followees*. However, the "*follower-followee*" relation is directed. One can be another user's follower, but not necessarily vice versa. This is different from Facebook where friend relations are mutual and reciprocal.

Due to the large amount of users and sheer volume of real-time data in microblogging systems, it is challenging, and sometimes even frustrating for a user to find relevant users or other entities as good followees while filtering out irrelevant information that are not even worth reading. For example, as one of the largest microblogging websites in China launched in April 2010, Tencent Weibo[1] has gained its great popularity as a platform for social networking and sharing interests. Currently, there are more than 200 million registered users on Tencent Weibo, generating over 40 million tweet messages each day. This scale of information benefits the users, but it can also flood them and put them at risk of information overload. It is therefore greatly helpful if we can build a recommender with its ability to suggest users that are likely to be worth following for each user [4].

A. Related Work

Extensive studies have focused on Twitter as a pervasive public microblogging service for social networking in recent years [6], [5], [12], [7]. This line of work explored the microblogging usage and communitie [6], its role in information dissemination as news media [7] and informal communication in work places [12].

The extremely large scale of user numbers and exponentially increased microblogging activity has motivated followee recommendation strategies [1], [2]. They aim to capture information of interest while reducing information overload introduced by irrelevant followees. Followee recommendation in microblogs can be naturally modeled as link prediction problem [9], [10] from the viewpoint of social network dynamics. The structural information of the nodes such as in-degree, out-degree, common neighbors as well as topological information such as shortest path have been explored. The link prediction is usually conducted based on some proximity measure between the nodes [8]. This line of methods mainly consider each user as a node in the network while ignore the attributes the individuals can have, such as the age, gender, interests, etc. Moreover, the interaction content are valuable information for the establishment of "follower-followee" relationships. In [4], content and collaborative filtering based methods are proposed to match users with similar interests with comparison of the keywords in their and their followees/followers' profiles. Armentano et al. [2] take into account the topology of follower/followee network of Twitter and consider different factors that help can identify users worth following. But some other features that may influence a user's adoption of others

[1]http://t.qq.com/

as followees are not explored, for example, the popularity of an item is more attractive to users than other features such as the matching of item category, keywords and the influence of user actions and current followees' acceptance.

B. Contributions

In this paper, we take into account the various features that might be influential in attracting one to follow other users in microblogs. Our contributions lie on answering the following questions from the viewpoint of followee recommendation:

1) How does the existing social network provide cues to new "following" activity?
2) To what extent does the well-known items such as celebrities attract users and become their followees?
3) Are users more likely to follow others with similar profiles? Which domain of the profile is most significant?
4) Does the collaborative actions such as mention "@", comment, or retweet motivate new "following" activity?
5) What role does the content semantics of messages play in connecting those users that were previously not related?

II. Dataset and Task

In this paper, we use Tencent Weibo as a data subject and explore the task proposed in Track 1 of KDD Cup 2012[2]. To goal is to ***predict whether or not a user will follow an item that has been recommended to the user***.

The source dataset represents a sampled snapshot of Tencent Weibo users' preferences for various recommended ***items*** and the history of users' "following" activity. Items represent information sources in Tencent Weibo, and each can be a person, an organization, or a group. The dataset is of a larger scale compared to other publicly available datasets ever released. Also richer information in multiple domains are provided, including user profiles, social graph, item category, message keywords which may hopefully evoke deep thoughtful ideas and methodologies. The users in the dataset, numbered in millions, are provided with rich information such as demographics, profile keywords, follow history, etc. The IDs of both the users and the recommended items are anonymized for the protection of user privacy. Furthermore, their information, when in Chinese, are encoded as random strings or numbers. Timestamps for user's follow actions are also given for performing session analysis.

The basic statistics of the dataset is listed in Table I. As it shows, the numbers of recommendations and users are in millions, and there are relatively fewer items with a number in thousands.

A. Follower/followee Distributions

We first examine how much *followees* and followers each user has within the training dataset. Figure 1 shows the distributions. As we can see, the distributions, especially of the numbers of followers for each user exhibit heavy tail pattern.

[2]https://www.kddcup2012.org/c/kddcup2012-track1

TABLE I
Statistics of Tencent Weibo dataset

dataset	# of recommendations	# of users	# of items	# of edges
training	73,209,277	2,320,895	6,095	50,655,143
test	34,910,936	1,196,411	4,849	–

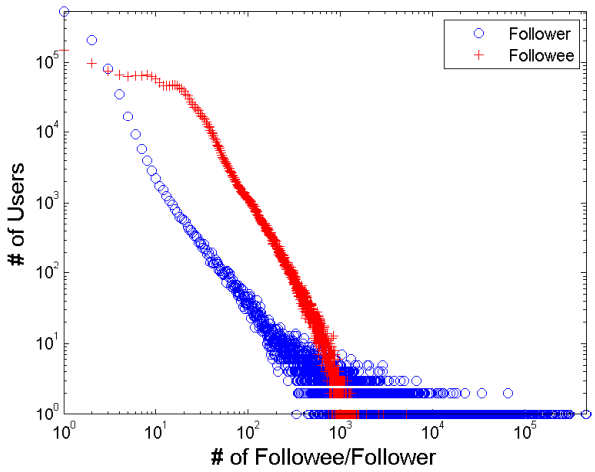

Fig. 1. Number of followees/followers

A very small proportion of users have extremely large amount of followers, e.g., tens of thousands and we can conjecture these are celebrities. However, most of users have very limited number of followers such as in tens or hundreds.

B. Recommendation Acceptance

We then examine the distributions of times of recommendations each user and each item has. Figure 2 and Figure 3 shows the distributions respectively. Meanwhile, we also illustrate the number of recommendations each user has accepted and each item has been accepted as well. A power law pattern is observed in Figure 2. Moreover, the numbers of accepted recommendations are more than 1 magnitude smaller than the overall recommendation in both figures. The statistics shows only 7.18% recommendations have been accepted by the users.

III. Evaluation

The evaluation metric provided by KDD Cup is Mean Average Precision (**MAP**). Suppose we recommend m items in a ranking list to one specific user, who may click 1 or more or none of them to follow in the period of near future for testing, the average precision[3] can be used for evaluation. The definition of Average Precision (**AP**) at n for the user is given by:

$$AP@n = \sum_{k=1}^{n} p(k) \cdot rel(k)/c(m) \quad (1)$$

where $p(k)$ is the precision at cut-off k, and $rel(k)$ is the binary function that represents whether the kth item is followed

[3]http://en.wikipedia.org/wiki/Information_retrieval

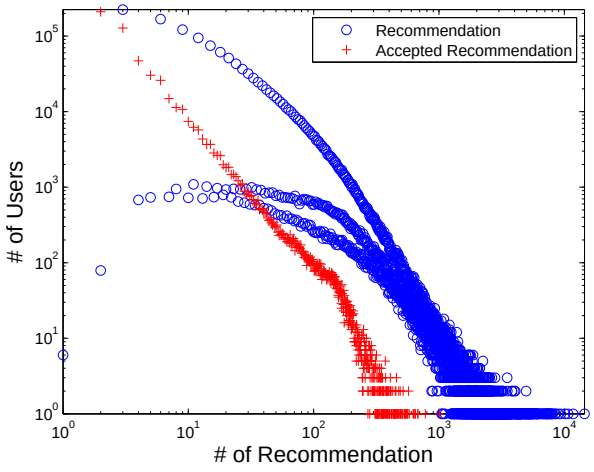

Fig. 2. Recommendations for each user

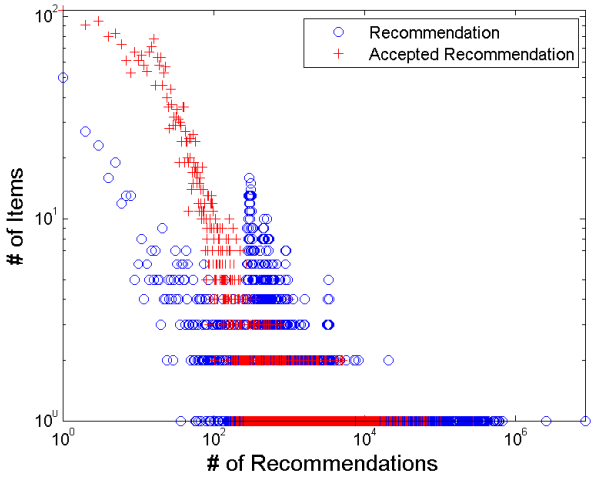

Fig. 3. Recommendations of each item

by the user. $c(m)$ denotes the number of items that the user will follow among the m items in the ranking list. Suppose there are N users, then we calculate the mean of the average precision at position n for each user j, as follows:

$$MAP@n = \sum_{j=1}^{N} AP_j@n/N \tag{2}$$

The KDD Cup organizers choose the cutoff $n = 3$, and $MAP@3$ is used for evaluation. One can submit their prediction result via leaderboard in Track 1 of KDD Cup 2012. The system will return the evaluation results, but the ground truth of user following activity regarding to the test dataset is withheld. Moreover, the leaderboard is calculated on approximately 53% of the test data with a *timestamp* $<$ 1321891200, and any data with a *timestamp* $\geq$ 1321891200 is used for the final evaluation.

In this paper, we explore the following distinct features in microblogging services and propose new methods for followee recommendation.

IV. Methodology

Notation. Let $U = \{u_1, u_2, ..., u_N\}$ be the set of users, and $T = \{t_1, t_2, ..., t_M\}$ denotes the set of items. We use $R(u_i) = \{t_1^i, t_2^i, ..., t_R^i\}$ be the set of item recommendations accepted by u_i. Let $A(u_i)$ and $A(t_i)$ be the set of item recommendations have been accepted by u_i, and the set of users that have accepted t_i.

1) **Item Category**. Items are organized in categories; each category belongs to another category, and all together they form a hierarchy. Item Category is a string "a.b.c.d", where the categories in the hierarchy are delimited by the character ".", ordered in top-down fashion (i.e., category 'a' is a parent category of 'b', and category 'b' is a parent category of 'c', and so on. For example, an item, a vip user Dr. Kaifu LEE, represented as "***science-and-technology.internet.mobile***". If a user follows Kaifu Lee, he/she may be interested in the other items of the category that Kaifu Lee belongs to, and might also be interested in the items of the parent category. Hence, we can use the categories of one's accepted item recommendations to represent his/her interests. Let $f(t_i)$ be the mapping function that return the category of t_i. And let $C = \{c_1, c_2, ..., c_P\}$ be the set of item categories. We use a vector $N_i = [n_1^i, n_2^i, ..., n_p^i]^T$ to represent how many times of each item-category occurs in the accepted recommendations within training dataset. n_k^i is defined as follows:

$$n_k^i = |\{t_j | t_j \in A(u_i) \ and \ f(t_j) = c_k\}| \tag{3}$$

For all the items in the recommendation list of a user, we rank them by the numbers of times the corresponding category pattern has been accepted by the user. The procedure can be outlined as in Algotithm 1.

Algorithm 1 Ranking Using Item Category

```
Training:
for all u_i ∈ U do
    compute each n_k^i according to eq. 3
end for
Test:
for all u_i ∈ U do
    for all t_j ∈ R(u_i) do
        if f(t_j) = c_k then set s_j = n_k^i
        end if
    end for
    rank t_j according to s_j in decreasing order.
end for
```

2) **Item Popularity**. The celebrities have prominent attractiveness compared to ordinary people. Intuitively, the more popular a item is, the more likely it will be accepted by the users. To measure the popularity of a item p_j, we use the number of times of the item has been accepted by the users in training dataset, as follows:

$$p_j = |A(t_j)| \tag{4}$$

We rank the items in the recommendations of a user according to their popularity, as shown in Algorithm 2.

Algorithm 2 Ranking Using Item Popularity

```
Training:
for all t_i ∈ T do
    compute p_j = |A(t_j)| according to eq. 4
end for
Test:
for all u_i ∈ U do
    for all t_j ∈ R(u_i) do
        set s_j = p_j
    end for
    rank t_j according to s_j in decreasing order.
end for
```

3) **Followees' Acceptance** We then look into how a user's acceptance of a item is influenced by his/her followees. According to the information theory, the more users in one's neighborhood that adopt a item, the more likely one will accept the item. We can consider this as a process of "diffusion of innovation" [11]. Intuitively, the more followees one has that have accepted the item, the more likely the user will accept the item to follow. We use $F(u_i)$ to denote the set of followees of u_i, the number of followees of u_i that have accepted t_j can be computed as

$$a_{ij} = |\{u_j | u_j \in F(u_i) \ and \ u_j \in A(t_j)\}| \quad (5)$$

The ranking is performed in Algorithm 3.

Algorithm 3 Ranking Using Followees' Acceptance

```
Test:
for all u_i ∈ U do
    for all t_j ∈ R(u_i) do
        set s_j = a_ij as computed in eq. 5.
    end for
    rank t_j according to s_j in decreasing order.
end for
```

4) **Semantic Keywords**. The data contains the keywords extracted from the tweet/retweet/comment by each user. Keywords are in the form *"kw1:weight1;kw2:weight2; ...; kw3:weight3"*. The greater the weight, the more interested the user is with regards to the keyword. Every keyword is encoded as a unique integer, and the keywords of the users are from the same vocabulary as the Item-Keyword. Specially, Item-Keyword contains the keywords extracted from the corresponding Weibo profile of the person, organization, or group. The format is a string *"id1;id2; ...;idN"*, where each unique keyword is encoded as an unique integer such that no real term is revealed. We then try to match the semantic keywords extracted from a user and a item for recommendation.
Let $W(k_m^i)$ denote the weights of the keyword $k_m^i \in K(u_i) = \{k_1^i, k_2^i, ..., k_W^i\}$ that are extracted from u_i, and let $K(t_j)$ be the set keywords extracted from u_i and t_j respectively. Then the ranking algorithm is proceeded as in Algorithm 4.

Algorithm 4 Ranking Using Semantic Keywords

```
Test:
for all u_i ∈ U do
    for all t_j ∈ R(u_i) do
        set s_j = max {W(k_m^i) | k_m^i ∈ K(t_j)}.
    end for
    rank t_j according to s_j in decreasing order.
end for
```

5) **Action Influence**. The data contains the records of user actions. For example, user A has retweeted user B 5 times, has "at" (@) mentioned B 3 times, and has commented user B 6 times, then there is one line "A B 3 5 6" in user action data. We can make fully use of it since those actions make one's followees exposed to one's followers, which may potentially motivate new following activity. We can construct a new affinity graph based on the existing bipartite graph of users and items, where edge weight is defined as:

$$w_{ij} = \alpha r_{ij} + \beta m_{ij} + \gamma c_{ij} \quad (6)$$

where r_{ij}, m_{ij} and c_{ij} are the times of retweets, mentions and comments from u_i to t_j, and $\Theta = [\alpha, \beta, \gamma]^T$ is the vector of weights, which can be learned by optimization using the training data. For simplicity, here we explore several settings where $\Theta = [1, 0, 0]^T$, $\Theta = [0, 1, 0]^T$, $\Theta = [0, 0, 1]^T$ and $\Theta = [1, 1, 1]^T$.
The ranking procedure for recommendations is described in Algorithm 5.

Algorithm 5 Ranking Using Action Influence

```
Training:
for all u_i ∈ U do
    for all t_j ∈ T do
        compute w_ij according to eq. 6
    end for
end for
Test:
for all u_i ∈ U do
    for all t_j ∈ R(u_i) do
        set s_j = Σ_k w_kj, where u_k ∈ F(u_i).
    end for
    rank t_j according to s_j in decreasing order.
end for
```

A. Results and Discussion

We train our methods using training dataset and use the testing dataset for evaluation. The evaluation is performed by the leaderboard system of KDD Cup 2012. Table II shows the evaluation results for all the five methods. For action influence, we use different parameter settings, and when $\Theta = [1, 1, 1]^T$, it achieves best performance. While when $\Theta = [1, 0, 0]^T$, $\Theta = [0, 1, 0]^T$ and $\Theta = [0, 0, 1]^T$, the performances are somewhat identical, i.e., about 0.21697, 0.21850, 0.20095 respectively. This indicates all the three actions, retweet,

TABLE II
RESULTS OF USING THE FIVE METHODS

Methods	Item Category	Item Popularity	Followees' Acceptance	Semantic Keywords	Action Influence
MAP@3	0.20481	0.29101	0.20189	0.19600	0.22187

mention, comment have positive influence on the following activity of one's followers to one's followees. However, the comment action is not as influential as retweet and mention. Among all the five methods, we find item popularity based ranking achieves best performance, which suggest the celebrity effect is one of the most influential factor that attract users. The keywords based algorithm doesn't perform well, we conjecture users' following activity is not very semantics related.

V. CONCLUSION AND FUTURE WORK

In this paper, we have examined the task of predicting whether a user will follow an item that has been recommended in microblogging system. The statistical analysis shows the numbers of followees/followers of each user exhibit power law distributions, and only 7.18% recommendations have been accepted by the users within the training dataset. Several influential factors including item popularity, item category, followees' acceptance, extracted keywords and actions (e.g., retweet, mention and comment) are considered in prediction. The experimental results show that item popularity is the most prominent feature in users' adoption of recommendations. This reflects the fact that the effect of celebrity is prevalent in Tencent Weibo. The result also indicate all the three actions, retweet, mention, comment have positive influence on the following activity of one's followers to one's followees.

Due to the consideration of algorithm efficiency to deal with the extremely large-scale data, we only consider very intuitive ranking algorithms in this research. For future work, we can use learning based algorithm such as logistic regression by taking advantage of all the above-mentioned features. Moreover, it is also interesting to use latent sematic analysis tools such as LDA [3] to map the keywords into *semantic topic* space and capture the semantic similarity between each user-item pair.

VI. ACKNOWLEDGEMENT

This work was supported in part by the US National Science Foundation under grant CCF-1048311. We would like to thank Kristina Lerman for her insightful suggestions.

REFERENCES

[1] M.G. Armentano, D. Godoy, and A. Amandi. Towards a followee recommender system for information seeking users in twitter. In *Proceedings of the Workshop on Semantic Adaptive Social Web (SASWeb 2011). CEUR Workshop Proceedings*, volume 730, pages 27–38.

[2] M.G. Armentano, D. Godoy, and A. Amandi. Topology-based recommendation of users in micro-blogging communities. *Journal of Computer Science and Technology*, 27(3):624–634, 2012.

[3] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] J. Hannon, M. Bennett, and B. Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 199–206. ACM, 2010.

[5] B. Huberman, D. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. 2008.

[6] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[7] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[8] K. Lerman, S. Intagorn, J.H. Kang, and R. Ghosh. Using proximity to predict activity in social networks. *Arxiv preprint arXiv:1112.2755*, 2011.

[9] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[10] L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.

[11] E.M. Rogers. *Diffusion of innovations*. Free Pr, 1995.

[12] D. Zhao and M.B. Rosson. How and why people twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 243–252. ACM, 2009.