Maximizing Topic Propagation Driven by Multiple User Nodes in Micro-Blogging

Chang Su, Youtian Du, Xiaohong Guan and Chenhe Wu
Ministry of Education Key Lab for Intelligent Networks and Network Security,
Xi'an Jiaotong University, Xi'an, China 710049
Email: changsu@stu.xjtu.edu.cn, {duyt,xhguan}@mail.xjtu.edu.cn, wuchenhe1987@stu.xjtu.edu.cn

***Abstract*—This work investigates the maximization of topic propagation jointly driven by multiple user nodes in micro-blogging. In this paper, we propose a new method to find a set of user nodes that jointly propagate topics approximately the most widely. First, we obtain multiple nodes with strong influence; Second, we exactly compute the breadth of information spread driven by a single node based on probabilistic models; Finally, we analyze the information propagation jointly driven by multiple nodes and derive an approximately optimal set of driving nodes. We find that the breadth of information propagation jointly driven by multiple nodes is approximately linear with both the breadth of information propagation of single driving nodes and the strength of tie among them, which indicates that selecting the optimal driving nodes needs to consider the link information among them as well as the ability of each node. Experimental results demonstrate the effectiveness of our method.**

I. Introduction

In recent years, online social networks (OSNs) play a fundamental role as a medium for the spread of information, ideas, and influence among its members. As a popular type of OSNs, micro-blogging, such as Twitter, encourages fast updating by limiting post size.

The research of OSNs mainly focuses on network structures, users' behaviors, information diffusion mechanisms and models, etc [1], [2]. Steeg and Galstyan [3] investigated the patterns of information transferring in OSN, and Yang and Counts[4] analyzed the difference of information diffusion structures between micro-bloggings and weblogs. Tang et al.[5] found that the user's ability of information propagation depends on the specific information topics. Several representative information diffusion models for OSNs have been proposed in the past research, including the linear threshold model [6], independent cascade (IC) model [7], and the SIR epidemic model [8]. Based on these models, some other extended models have also been studied. For example, Rodriguez and Leskovec[9] proposed a cascade transmission model based on IC model, and Cheng et al. [10] introduced the strength of ties among users into SIR model. These research also helps significantly improve some applications such as information recommendation[11].

In this paper, we propose a new method for finding a set of driving nodes which can approximately maximize the propagation of a certain topic in micro-blogging. This method includes three steps: First, we efficiently obtain multiple nodes with strong influence; Second, we exactly compute the breadth of information spread driven by a single node based on probabilistic models; Finally, we analyze the information propagation jointly driven by multiple nodes, in which we find that the breadth of information spread driven by multiple nodes is linear with both the breadth of information spread driven by each single node and the tie strength among the nodes. Based on the three-phase method, we finally derive an almost optimal set of driving nodes that can spread the given topics approximately the most widely.

II. Problem statement

We model a micro-blogging with graph $G = (V, E)$, where V denotes the user node, and E denotes the relation including "following" links and "retweet" behaviors between different nodes. A topic posted by a driving node will spread on the micro-blogging network via retweet. In the paper, a *driving node* is defined as a node that originally posts a topic which will consequently be spread. The breadth of the spread depends on the amount of followers, the interest to the topic, and retweet frequency. The main task, finding a set of driving nodes that can spread the given topic approximately the most widely, thus can be formulated as follows:

$$Q_P = \arg \max_{q_{p_1}, q_{p_2}, \cdots, q_{p_n}} \Psi(tp, q_{p_1}, q_{p_2}, \cdots, q_{p_n}) \qquad (1)$$

where $q_i \in V$ denotes the driving nodes, Q_P is the optimal set of n driving nodes that can spread the given topic tp the most widely, and $\Psi(\cdot)$ denotes the breadth of topic spread. It is very difficult to solve problem (1) exactly due to the large scale of data and inexplicit expression of $\Psi(\cdot)$. This paper gives a new method to obtain an almost optimal solution that can spread topics approximately the most widely on micro-bloggings.

III. Model and Methodology

A. Influence computation via extended PageRank algorithm

Inspired by [12], we introduce an extended PageRank algorithm, called *InfluentialRank* (IR), which calculates the influence of nodes based on not only following relationship of users, but also retweet behaviors and users' interests. Fig.1 shows the network on which the IR algorithm is implemented. This network consists of two types of weighted edges that represent the following relationship and retweet behavior, respectively. q_u, q_v, q_m and q_n denote user nodes, solid-line edges from q_u to q_v mean q_v is the follower of q_u, dashed-line edges from q_u to q_v denote that q_v retweets messages posted by q_u. IR algorithm assumes that the influence of each node

978-1-4799-0537-9/13/$31.00 ©2013 IEEE

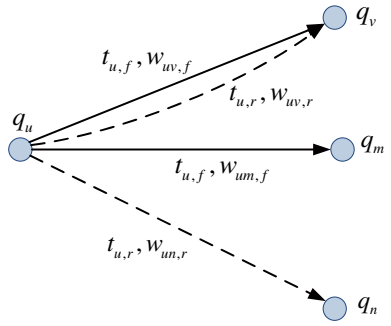


Fig. 1. Graphical network representation of micro-bloggings that consists of two types of weighted edges.

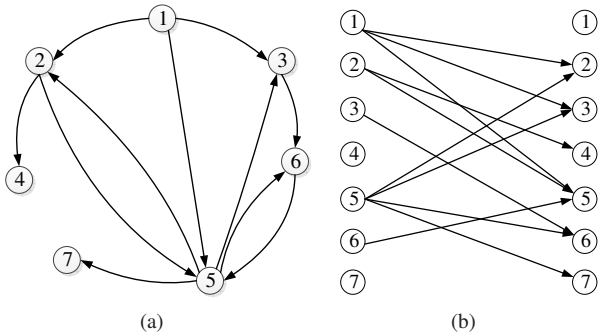


Fig. 2. The retweet network for micro-blogging. (a) a static retweet network, and (b) the dynamic Bayesian network constructed based on (a).

is evenly divided and transferred to the nodes of its outbound links as follows:

$$t_{u,f}=\frac{\alpha \cdot IR(q_u)}{OD_f(q_u)}, t_{u,r}=\frac{(1-\alpha) \cdot IR(q_u)}{OD_r(q_u)} \tag{2}$$

where $IR(q_u)$ is the influence of q_u, $OD_f(q_u)$ and $OD_r(q_u)$ denote the number of "following" edges and "retweet" edges out from q_u, respectively, α is a parameter. The weight of "following" edges, denoted by $w_{uv,f}$, is measured with the q_v's interest to the given topic, and the weight of "retweet" edges $w_{uv,r}$ is measured with the probability that q_v retweets the posts from q_u, i.e.,

$$w_{uv,f}=I_v, w_{uv,r}=\frac{M_{uv}}{M_v} \tag{3}$$

where M_v denotes the total number of posts of q_v, M_{uv} is the number of the posts that q_v retweets from q_u, and the interest I_v of q_v to certain topic are calculated by Latent Dirichlet Allocation (LDA) algorithm [13]. The influence of each node can be computed as follows:

$$IR(q_v)=\frac{d}{N}+(1-d)\left[\sum_{q_u \in B_v^f} t_{u,f} w_{uv,f} + \sum_{q_u \in B_v^r} t_{u,r} w_{uv,r}\right] \tag{4}$$

where d is jump probability and is generally set 0.15, N is the total number of nodes in OSNs, B_v^f and B_v^r are the set of nodes that q_v follows and retweets from, respectively. The top C nodes with the largest influence score comprise the candidate set Q_C.

B. The probabilistic model for topic propagation

To exactly compute the breadth of topic spread from one driving node $q \in Q_C$, we model the users' retweet behaviors with probabilistic models. Fig.2a depicts the retweet process in micro-blogging with a Directed Cyclic Graph (DCG). Each node q_i in the graph has two states: $X_i = 1$ means q_i retweets the topic from the nodes it follows, and $X_i = 0$ means not. The retweet behavior can be represented by the conditional probability $\Pr(X_i|F(X_i))$, where $F(X_i)$ denotes the state of the nodes that q_i retweets posts from.

Inspired by [14], we transform the DCG into a dynamic Bayesian network (DBN) as shown in Fig.2b, and infer the probability distribution of state X_i on the DBN. The inference at the t-th time slice of the DBN equals the t-th iterative computation on the DCG. At time t, the probability of state set $\mathrm{X}^t = \{X_1^t, \cdots, X_N^t\}$ is computed as follows:

$$\Pr(X_i^t) = \sum_{pa(X_i^t)} \Pr(pa(X_i^t))\Pr(X_i^t|pa(X_i^t)) \tag{5}$$

where $pa(X_i^t)$ is the parent nodes set of X_i^t in the DBN, and they correspond to the user nodes in Fig.2a that q_i retweets topics from. The inference result $\Pr(X_i^t)$ will converge to $\Pr(X_i)$ as t increases, and the computation will be terminated when $\max_i\{\Pr(X_i^t)-\Pr(X_i^{t-1})\} < \delta$, where δ is a predefined threshold. The size of conditional probability table (CPT) $\Pr(X_i^t|pa(X_i^t))$ increase exponentially with the number of $pa(X_i^t)$. In general, $pa(X_i^t)$ consists of a large number of elements due to one usually retweets posts from many user nodes, which makes the computation intractable.

Based on the factorization introduced in [15], we simplify the CPT $\Pr(X_i^t|pa(X_i^t))$ as follows:

$$\Pr(X_i^t|pa(X_i^t)) = \sum_{X_j^{t-1} \in pa(X_i^t)} \theta_i^j \cdot \Pr(X_i^t|X_j^{t-1}) \tag{6}$$

where θ_i^j is the parameter that needs learning. For simplicity, we set all θ_i^j for q_i to $1/|pa(X_i^t)|$, where $|pa(X_i^t)|$ denotes the size of $pa(X_i^t)$. $\Pr(X_i^t|X_j^{t-1})$ is constant for different t, and just equals $\Pr(X_i|X_j)$, where $X_j \in F(X_i)$. Based on the achieved data, we can estimate $\Pr(X_i|X_j)$ as follows:

$$\Pr(X_i|X_j) = \frac{M_{ji}}{M_i} \tag{7}$$

where M_{ji} is the number of posts retweeted by q_i from q_j, M_i is the total number of posts of q_i.

Once q_i retweets a post (i.e., $X_i = 1$), its followers will receive it. In other words, the probability of the follower q_k receiving the message equals $\Pr(X_i = 1)$. When q_k is the common follower of a set of multiple nodes Q_F, the probability that it receives the message is defined as

$$p_k = \max_{i:q_i \in Q_F} \{\Pr(X_i = 1)\} \tag{8}$$

C. Maximizing topic propagation driven by multiple nodes

In general, the breadth of information spread jointly driven by multiple users doesn't equal the summation of that driven by each single node due to the overlap among the topic spread.

Inspired by [10], we define the strength of tie between two driving nodes q_u and q_v as:

$$w(q_u, q_v) = \frac{o_{uv}}{k_u - 1 + k_v - 1 - o_{uv}} \tag{9}$$

where k_u and k_v are the out-degrees (i.e., the number of followers) of q_u and q_v, respectively, o_{uv} is the number of the common followers of driving nodes q_u and q_v. The strength of tie between a set of driving nodes Q and a single driving node q_v can be simply extended as follows:

$$w(Q, q_v) = \frac{1}{|Q|} \sum_{q_u \in Q} w(q_u, q_v) \tag{10}$$

where $|Q|$ is the size of Q.

A node is considered to be *activated* when it receives the objective topic, and the probability that it receives the topic is called activation probability. To measure the driving ability of a single driving node q_i, we introduce the concept of *activation expectation* (AE), defined as the expectation of activation probability on the micro-blogging:

$$E_a(q_v) = \frac{1}{N} \sum_{k=1}^{N} p_k^v \tag{11}$$

where p_k^v denotes the activation probability of q_k caused by driving node q_v which can be computed based on Eq.8, and N is number of nodes in the network. Similarly, we can get the *joint activation expectation* (JAE) of a set of driving nodes Q: $E_a(Q) = \sum_{k=1}^{N} p_k^Q / N$, where p_k^Q is the activation probability of q_k given the set Q.

We find, in general, the joint driving ability of a given set of driving nodes and another single driving node linearly depends on both the driving ability of this single node and the strength of tie between them, and can be formulated as:

$$E_a(Q) = Aw(Q', q_v) + BE_a(q_v) + b + \varepsilon \tag{12}$$

where Q' is the given set of driving nodes, $Q = \{Q', q_v\}$, A, B and b are parameters, and ε is a random noise.

Based on the above analysis, we give the following algorithm to derive an almost optimal set of driving nodes Q_P that spread topics approximately the most widely. Initially we let $Q_P = \emptyset$.

1) Calculate the driving ability of each node in Q_C, choose $q_p \in Q_C$ with the largest driving ability and move it into Q_P, i.e., $Q_P \leftarrow Q_P \cup q_p$, $Q_C \leftarrow Q_C \backslash\ q_p$.
2) Calculate the strength of ties between Q_P and each node in Q_C based on Eq.10, choose $q_p \in Q_C$ that maximizes the Eq.12 and move it into Q_P, i.e., $Q_P \leftarrow Q_P \cup q_p$, $Q_C \leftarrow Q_C \backslash\ q_p$.
3) Repeat step 2 until the size of Q_P equals n.

IV. Experimental Results and Analysis

The experimental data set contains 29514 user nodes, 421140 following edges, and 3248734 posts including 776641 retweets crawled from the Sina micro-blogging. The average degree in the data set is 29, and its properties of small-world and scale-free are verified. In the experiment, we choose the "fashion" topic as the target topic.

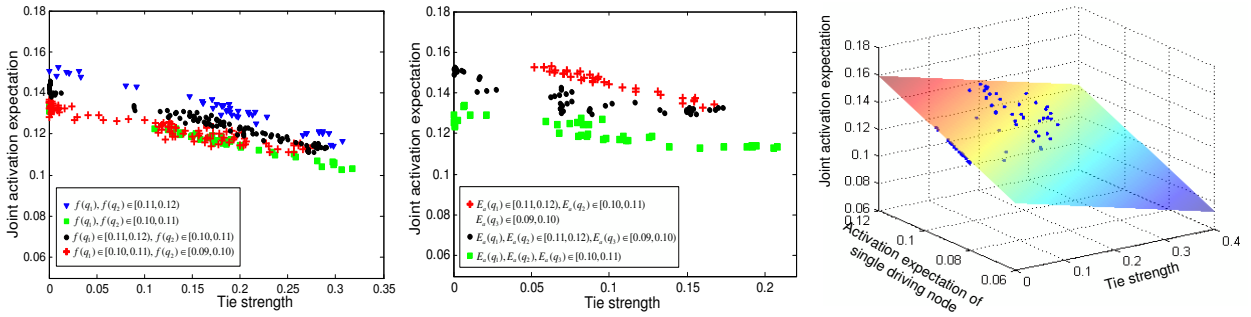
TABLE I
Comparison of three methods in measuring the influence and information spread ability of nodes.

PageRank	IR $\alpha = 0.4$	IR $\alpha = 0.6$	IR $\alpha = 0.8$	Activation Expectation
1664971205	10473	10473	10473	3004005
10473	79660	79660	79660	10473
10484	1015414785	1015414785	26063195	1664971205
10452	26063195	26063195	10514	104360
21110	10514	10514	31928289	79660
10392	31928289	31928289	10413	10413
104016	10413	10413	35364455	10514
10470	35364455	35364455	10484	10484
104629	10484	10484	104629	104629
104360	104629	104629	10392	34103

Table 1 compares the users' influence computed with InfluencialRank and traditional PageRank. We set parameter α in Eq.2 with three values: 0.4, 0.6 and 0.8. With the results, we find that InfluencialRank algorithm encodes the users' interests and retweeting behaviors besides the following relationship that considered in traditional PageRank algorithm, and is more suitable in the analysis of micro-blogging. For example, the user No.79660 has a relatively fewer followers, but its posts are mostly related to the topic "fashion", and thus it has a high rank according to IR algorithm, while node No.1664971205 is opposite.

In the last column of Table 1, we also list the top 10 nodes based on the activation expectation of a single driving node. From the table, we find that the rank based on activation expectation has some difference with which based on the IR algorithm. It is because that activation expectation accurately measures the breath of topic spread on micro-bloggings, which leads to a better solution. We also find that $\alpha = 0.6$ leads to a more reasonable results, and the following experiments will use this setting.

As introduced in Eq.12, the joint activation expectation linearly depends on both the activation expectation of the single driving node and tie strength among the driving nodes. Fig.3 illustrates this linear relationship. In Figs.3a and 3b, the horizontal axis denotes the tie strength between two driving nodes q_1 and q_2 and that between q_3 and the set Q' containing q_1 and q_2, respectively, and the vertical axis denotes the joint activation expectation. Fig.3a shows that the joint activation expectation of two driving nodes are approximately linear with the strength of tie between them, when the activation expectation of each single driving node is constant (in the experiment, we limit each activation expectation into a narrow interval, e.g., $[0.11, 0.12)$). Similarly, as shown in Fig.3b, the joint activation expectation of the set $Q' = \{q_1, q_2\}$ and the third node q_3 approximately follows the linear relationship with the tie strength. Fig.3c shows the dependance of the joint activation expectation of $Q' = \{q_1, q_2\}$ and the third node q_3



(a) JAE of two nodes *vs*. tie strength (b) JAE of two nodes and the third one *vs*. tie strength (c) JAE of two nodes and the third one *vs*. AE of the third node and tie strength

Fig. 3. The approximate linearity among the JAE of multiple driving nodes, the AE of a single node and the tie strength.

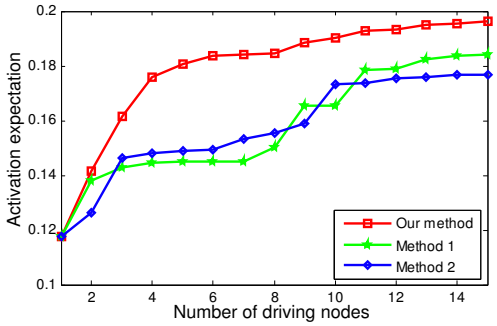


Fig. 4. The comparison of the breadth of topic spread with three methods.

on the tie strength and activation expectation of q_3. In the figure, the points are distributed near the plane. The estimated parameters of Eq.12 are $A = -0.0973$ and $B = 0.8875$, which means that the joint activation expectation will increase as the tie strength decreases and the activation expectation of the single driving node increases.

Fig.4 compares the topic spread from the driving nodes based on our method and two other related methods. Method 1 denotes selecting driving nodes randomly from Q_C, and method 2 denotes selecting the driving node q_p with $E_a(q_p) = \max_{Q_C} E_a(q_i)$. In the experiment, we first choose a candidates set Q_C of top 50 nodes with the largest influence based on InfluentialRank algorithm, and derive the optimal driving nodes from Q_C. The choosing of optimal driving nodes from Q_C begins with the node No.3004005 that has the largest activation expectation. As shown in Fig.4, every method can increase the breath of topic spread as the number of driving nodes increases. Compared to method 1 and method 2, our proposed method achieves significantly higher joint activation expectation, which means that our method spreads topics on micro-blogging more widely.

V. Conclusion

In this paper, we propose a new method to find a set of user nodes that can jointly propagate topics approximately the most widely. In the method, we first choose a initial set of nodes that have large influence based on InfluentialRank algorithm, and compute the breadth of topic spread jointly driven by the driving nodes in the initial set for a given topic. We employ activation expectation to measure the breadth of topic spread, and find activation expectation in the case of multiple driving nodes is linear with the activation expectation driven by each single driving node and the strength of ties among them. Experimental results demonstrate that our method can derive a set of driving nodes that can spread topics significantly more widely than two compared methods.

Acknowledgment

This work is supported in part by the National Natural Science Foundation (60905018, 61221063) and 111 International Collaboration Program, of China.

References

[1] K. Musiał, P. Kazienko. *Social networks on the Internet*. World Wide Web Journal, 2012, DOI: 10.1007/s11280-011-0155-z.
[2] J.C. Zhao, J.J. Wu, K. Xu. *Weak ties: Subtle role of information diffusion in online social networks*. Physical Review E, 2010, 82(1).
[3] G.V. Steeg, A. Galstyan. *Information transfer in social media* . In International Conference on World Wide Web, 2012, pp.509-518.
[4] J. Yang, S. Counts. *Comparing information diffusion structure in weblogs and microblogs*. In AAAI Conference on Weblogs and Social Media. 2010.
[5] J. Tang, J.M. Sun, C. Wang, Z. Yang. *Social influence analysis in large-scale networks*. In ACM SIGKDD, 2009, pp.807-816.
[6] M. Granovetter. *Threshold models of collective behavior*. American Journal of Sociology, 1978: 1420-1443.
[7] J. Goldenberg, B. Libai, E. Muller. *Talk of the Network: A complex systems look at the underlying process of word-of-mouth*. Marketing Letters, 2001, 12(3): 211-223.
[8] M.E.J.Newman. The structure and function of complex networks. SIAM review, 2003, 45(2): 167-256.
[9] M.G. Rodriguez, J. Leskovec, A. Krause. *Inferring networks of diffusion and influence*. In ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2010, pp.1019-1028.
[10] J.J. Cheng, Y. Liu, B. Shen, W.G. Yuan. *An epidemic model of rumor diffusion in online social networks*. The European Physical Journal B, 2013, 86(1): 1-7.
[11] V. Chaoji, S. Ranu, R. Rastogi, R. Bhatt. *Recommendations to boost content spread in social networks*. In International Conference on World Wide Web, 2012, pp.529-538.
[12] Y. Yamaguchi, T. Takahashi, et al. *Turank: Twitter user ranking based on user-tweet graph analysis*. In Web Information Systems Engineering-WISE 2010 (pp. 240-253). Springer Berlin Heidelberg.
[13] D.M. Blei, A.Y. Ng, M.I. Jordan. *Latent Dirichlet allocation*. The Journal of Machine Learning Research, 2003, pp.993-1022.
[14] S. Kim, S. Imoto, S. Miyano. *Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data*. Biosystems, 2004, 75(1-3): 57-65.
[15] L.K. Saul, M.I. Jordan. *Mixed memory markov models: decomposing complex stochastic processes as mixtures of simpler ones*. Machine Learning, 1999, pp.75-87.