The Engine node Mining Algorithm in Microblog Information Spreading

Jing Yang
College of Computer Science and Technology,
Harbin Engineering University,
Harbin, Heilongjiang 150001, China
yangjing@hrbeu.edu.cn

Pengfei Sun
College of Computer Science and Technology,
Harbin Normal University,
Harbin, Heilongjiang 150025, China
sunpengfei@hrbeu.edu.cn

Xueyan Zhou
College of Computer Science and Technology,
Harbin Engineering University,
Harbin Engineering University Science Park
College of Engineering, Harbin University,
Harbin, Heilongjiang 150086, China
zhouxueyan@hrbeu.edu.cn

Yongshi Zhang
College of Computer Science and Technology, Harbin
Engineering University, Harbin, Heilongjiang 150001,
zhangyongshi@hrbeu.edu.cn

Jing Zhang
College of Computer Science and Technology,
Harbin Engineering University,
Harbin, Heilongjiang, China
zhangjing@hrbeu.edu.cn

***Abstract*—Retweeting is the main propagation mechanism in microblog platform. The information often spreads wider through some engine nodes. The paper proposes a new engine node mining algorithm. Firstly, the cascade (session tree) is built according to the retweeting of a microblog. Secondly the pruning strategy is used according to the timestamp. Thirdly, the cascade set (session forest) is clustered by topical relevance. Finally, the engine node with different precision can be extracted through computing the integrated diffusion capacity. We conduct an experiment to show the effective of the proposed algorithm.**

Keywords-Microblog; retweeting; engine node; generating; pruning

I. INTRODUCTION

In recent years, the microblog platform has developed rapidly; Sina microblog has more than 600 million users, among which the number of monthly active users reached 129.1 million[1][2]. The microblog platform has many characteristics such as the immediate information dissemination, the high propagation speed, the large transmission range. Since a user can share feelings and information at any time, so the information dissemination of microblog has more incomparable advantages than that of traditional media. The convenient way, however, makes the information more difficult to track and control.

Many researchers have studied microblogs, such as the microblog topic extraction and analysis, sentiment analysis, information retrieval and recommendation, the user relationship mining, information dissemination evolution and user influence analysis. The user influence analysis mainly mines the different types of nodes from different angles, such as high-impact nodes, active nodes, the nodes with high confidence, the core nodes, the source nodes and engines nodes, etc. Traditional researches of individual influence technologies include degree, closeness, betweenness, HITS, PageRank and extension methods. These studies are generally for a snapshot of the network topology, and no analysis of the specific topic. The other microblog information dissemination process is often due to the high influence of users and the rapid spread of retweeting.The basic idea is that these nodes will disseminate information more widely[3][4]. This paper proposes an engine node mining method based on topic. Engine node is the node with larger diffusion force. The mining method first builds a cascade according to the retweeting of a microblog, after which the pruning strategy is used according to the timestamp. The remainder of this paper is structured as follows. In Section II we review the related work and propose the engine node mining algorithm. In Section III, we conduct an experiment on real data. Section IV concludes the whole paper.

II. RELATED WORK

The microblog information dissemination mechanism and retweeting in microblog are introduced in this section, and then the engine node mining algorithm is proposed in detail.

A. Microblog Information Dissemination Mechanism

In recent years, lots of information diffusion algorithms on social network have been proposed. Among them, quite a number of algorithms extract information network from a group of most influential nodes. Their basic idea is that these nodes will make the information disseminate more widely, including information dissemination prediction by analyzing the blog information cascade [5][6].

Information dissemination speed on the microblog platform is much faster than that on the blog platform, and their propagate model are also different. Dabeer [7], analyzed the factors affecting the microblog information dissemination, including information characteristics and the activity, response and out-degree of fans nodes. He proposed a decision-making framework based on Markov to measure the effectiveness of information dissemination. Lehmann et al. [8] tracked the HashTag diffusion process in Twitter network, and discovered that the epidemic spread model played an important role. Yang et al. [9] predicted the microblog information dissemination speed, size and scope. Tsur et al. [10] employed a linear regression model to predict the diffusion of information within a given time according to the content and network topology. Wang [11] provided a network-traffic based web traffic computing technology for hot topic discovery, incident detection, real-time tracking and other applications. The “fission” engine node mining algorithms are rare, and these nodes which make the information “Second outbreak” or even “N times outbreak” have a great potential for application. On the one hand, grasping the engine node is helpful for the information control. On the other hand, seeking out the engine nodes is significant in the field of advertising and other commercial applications.)

B. Retweeting in microblog

Retweeting in microblog is the main object of data mining, because it is the base of information dissemination, influence analysis, sentiment analysis, topic discovery and evolution etc. Therefore, the study of microblog retweeting contributes to the understanding of the information diffusion mechanism. Macskassy et al. [12] showed that the majority of users do not necessarily retweet their familiar topic. Pal and Counts [13] assessed and sorted the user’s authority using the number of the original tweeting, participated in the session and retweeting as primary index. They used a Gaussian mixture model to calculate the user’s influence. Since the computation complexity of the model is very high, it is not suitable for traceability research. In addition, the user influence assessment based on information needs to deal with many insoluble problems, such as languages, dialects, pictures and videos. Therefore, we must use a simple way to get information as accurate as possible, such as only the cascade and the topology are used to study traceability research without semantic mining, and only the positive and negative emotions to assess the influence while ignoring the impact of information in different formats.

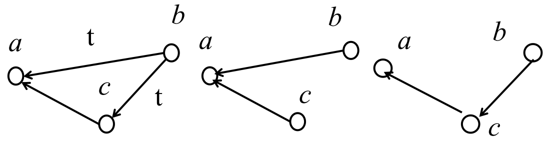
The microblog information propagation model can be described as G= (V, E), where V is a user and E is the retweeting or comment relationships. An information cascade is also known as a session tree, and the out-degree of its initiator is 0, the others link to the initiator or participators to form the information cascade through retweeting, sharing or commenting etc. So the influence is opposite to the direct edge. The information cascade may contain loops, and there are following relationships in microblog, so strictly speaking it is not a tree structure. The timestamps have been concerned to pruning and generating the cascade tree, and then the engine nodes can be digged out through attribute measuring.

C. The Engine node Mining Algorithm

Engine node mining algorithms are designed to extract nodes with high information diffusion ability. Firstly, the information is built according to the retweeting. Secondly the pruning strategy is used according to the timestamp. Thirdly, the cascade set is clustered by topical relevance. Finally, the engine node with different precision can be extracted through computing the integrated diffusion capacity. Algorithm 1 shows the main steps.

```
Algorithm 1: EngineNodes
Input: G (V, E)
Output: Mining Engine node ENT for each topic T
Begin
 C←ExtractCascade (G);
 C←PruningStrategy(C);
IF C is context-aware then
        ←ExtractSubgraph(C);
   else
        ={C}
   for each GT∈  do
    if GT is not a tree then
      {GT ←GeneratingStrategy(GT);
       GT ←PruningStrategy(GT);}
       ENT← EngineNodes(GT);
End
```

The GeneratingStrategy and PruningStrategy are used to amend the cascades to the tree structure. The PruningStrategy will delete certain retweeting edges according to the timestamp to remove the circle structure, and the GeneratingStrategy will add certain retweeting edges according to the timestamp and following factor to form a larger cascade tree. The PruningStrategy is shown in Figure 1.

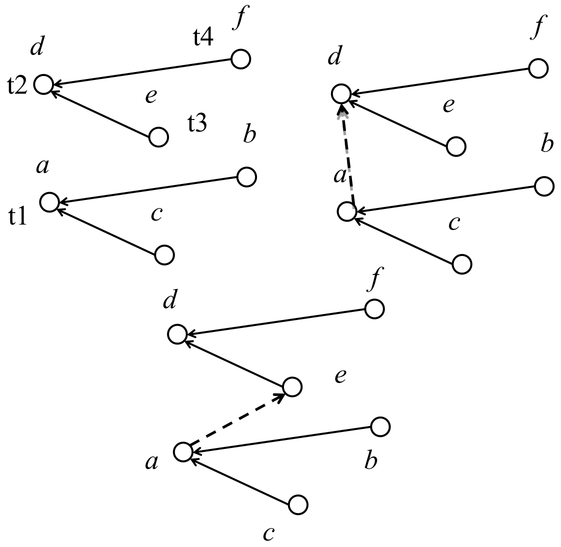


(a) Retweeting with circle structure (b) PruningStrategy while t1<t2 (c) GeneratingStrategy while t1>t2

Figure 1. PruningStrategy Schematic Diagram

PruningStrategy eliminates loops based on the first retweeting. User b forwards the same message from a and c to form a loop, then the PruningStrategy is used according to the time stamp, such as the edge b to c will be cut off while t1<t2, and vice versa.

At the same time, GeneratingStrategy adds virtual edges combine scattered information to cascade into larger information cascade tree. The GeneratingStrategy is shown in Figure 2.



(a)Two independent cascade tree(b)while d,e,f follow a and t2<t1(c)while e,f follow a,t3<t4 and t3<t1

Figure 2. GeneratingStrategy Schematic Diagram

The basic principle of GeneratingStrategy regards the user who post the same message later than ones' friends, and then adds virtual edges. Specific steps are as following: (1) for the root node, look for its friends with the earliest timestamp within the cascade; (2) regarding the root node as children of the friend with earliest timestamp by adding virtual edges, which will increase the size of the cascade; (3) if the GeneratingStrategy brings in some loops, the PruningStrategy should be used.

Diffusion coefficient α in formula (1) is used to measure the information diffusion force of the node. For example, each node of information cascade tree in Figure 3. has a different diffusion coefficient.

$$\alpha = \frac{O_i}{O_i - N_{i1} + \sum_{k=1}^{N_{i1}} O_k} \qquad (1)$$

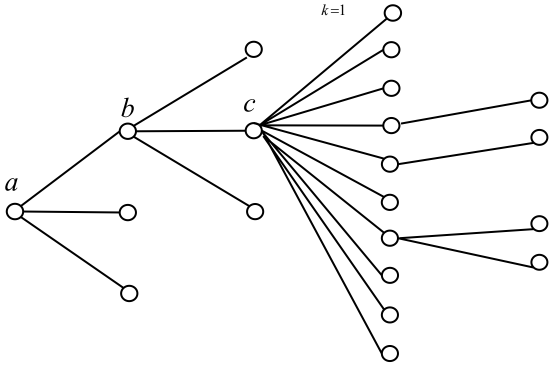


Figure 3. The diffusion coefficient of an information cascade tree

Where, diffusion coefficient α denotes the information diffusion force of node i, Oi is the outdegree of node i, and Ni1 is the number of the one-hop non-leaf of node i. The denominator of the diffusion coefficient is equal to the number of the leaf nodes. For example, the diffusion coefficient of node a, b and c respectively are 3/15, 3/13 and 10/11, and the node c is the maximum one. So the diffusion coefficient can be applied to mine the engine node.

III. Experiments

The data set is taken from a well-known China microblog site - Sina microblog, which was opened to the public in Oct. 2009. It has nearly 600 million registered users, among which nearly 100 million are daily active users. Microblog information has strong timeliness, and most topics will fade in a short time. The experiment data take partial data in Oct. 2014(including 75,526,147 posts) for analysis. The proposed algorithm is mainly to extract the engine node in the information propagation.

A. Cascade extraction

Cascade can be divided into two categories, namely chain and star. The chain has only one node in each layer, focusing on depth propagation, and the star has multiple nodes in a layer, focusing on breadth propagation. In the database, the frequency of the star is higher than that of the chain under the same cascade size. According to the definition of engine node, namely the "nuclear fission" central node, these nodes are more likely to exist in the star cascade. In fact, most of the cascades are between the two cases, for the complex shape of chain and star. In addition, due to the special nature of microblog retweeting, it will not appear in the case of multiple initiators, namely there is only one root in the topology. And a user can forward the message many times, so there may be a loop in a cascade. Figure 4 shows six basic high frequency cascade topology structures, and it can be seen that most cascades are too short to mine the engine node, so the GeneratingStrategy and PruningStrategy are used to amend the cascades to the tree structure.

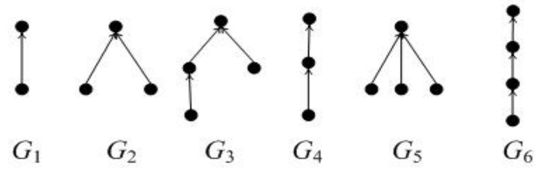


Figure 4. The basic high frequency cascade topology structure

B. Engine node mining

To calculate the diffusion coefficient of all nodes in all topics is not necessary, and the computation complexity is high, so the algorithm analyzes the diffusion capacity of nodes in a certain topic. The keyword matching algorithm is used to extract the sub graph based on topic, and the core idea is to put the cascades with same keywords into a collection. The topic T has the keyword {k1, k2, k3}, and GT is the cascade set with same keywords of topic T. The GeneratingStrategy and PruningStrategy are used to amend the cascades to cascade sets based on topics, and then the largest connected component is the main data for analysis. The largest connected component may contain more engine nodes. Assuming that the branch contains n nodes, n-1 retweeting edges and m leaf nodes, according to the definition of diffusion coefficient, we can get α= 0 on the leaf nodes and α= 1 on the nodes whose one-hop nodes are leaf nodes. Diffusion capacity of these two nodes is limited. But the diffusion coefficient of other nodes in cascade tree is between 0 and 1. Taking the "Lanxiang" event with duration of 11 days and "persistent haze" event with duration of 12 days as an example, in the topic-based cascade, the diffusion capacity of three largest branches are analyzed respectively. TABLE I shows the nodes analysis of each cascade branch. A high engine node percent means topology structure is closer to the star topology.

TABLE I. THE NODES ANALYSIS OF EACH CASCADE BRANCH

Event	**Lanxiang**			**Haze**		
Branch No.	1	2	3	4	5	6
nodes	8065	7619	5210	7324	4301	3847
edges	8064	7618	5209	7323	4300	3846
Leaf nodes	2516	2734	1837	3519	1042	1566
Leaf-to-be	1027	1804	1094	1249	841	931
Engine node	4522	3081	2279	2556	2418	1350
Percent	56.1	40.4	43.7	34.9	56.2	35.1

The node diffusion coefficient in each branch is calculated, and the nodes with α> 0.9 should be labeled as engine node. The distribution is shown in Figure 5.The red section represents the nodes size with α> 0.9, whose proportion is small in each branch. The highest proportion is 16% in No.4, and the lowest proportion is 10% in No.2. It is obvious that the proportion of high coefficient diffusion nodes is low. The result set is beneficial for the network public opinion analysis, and the engine nodes are very important for the advertising and other applications.

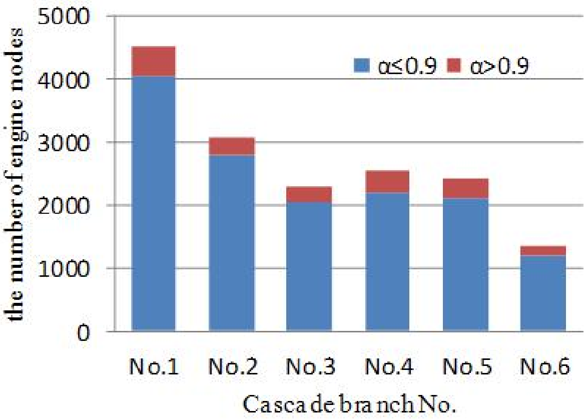


Figure 5. The nodes with α > 0.9 in the diffusion nodes

Determining the engine node in different network topology according to the nodes (α > 0.9) is unreasonable. In order to control the number of engine nodes, our experiment usually chooses a certain percentage. As there are 8065 nodes in No.1 cascade branch, if 2% nodes are required to be advertised, the engine node is 161 with α> 0.937, which is far less than the red section of 477 nodes in Figure 5. Therefore, the engine node set is a variable data field based on actual demands. By manual verification, the resulting engine node has many fans and retweeting. They are always celebrities or grassroots heroes.

IV. CONCLUSIONS

The paper proposes an algorithm to extract the engine node with high information diffusion capacity on the microblog data. The algorithm combines knowledge of information cascade, the time factor, the topic factors and graph theory, and the engine node set is variably based on the actual demands. On one hand, these nodes have reasonable size to be used for public opinion analysis and early warning. On the other hand, the advertising has broad prospects. Therefore the research of marketing strategy based on engine nodes computation will be our future work.

ACKNOWLEDGMENT

This paper is supported by (1) the National Natural Science Foundation of China under Grant nos. 61672179, 61370083 and 61402126, (2) the Research Fund for the Doctoral Program of Higher Education of China under Grant nos. 20122304110012, (3) the Natural Science Foundation of Heilongjiang Province of China under Grant no. F201320.(4) 2017-2020 The National Natural Science Fund (Project No.51679058). (5) 2013-2016 China Higher Specialized Research Fund (PhD supervisor category) (20132304110018). (6) Heilongjiang social science research Fund (16XWB01, 16TQD03). (7) Harbin

University Young Doctor start-up fund scientific research projects.

REFERENCES

[1] Ding Zhaoyun, Jia Yan, Zhou Bin.Survey of Data Mining for Microblog. Journal of Computer Research and Development. 51(4): pp.691-706.2014.

[2] Hu Q, GAO Y, Ma P, et al. A new approach to identify influential spreaders in complex networks. In: Web-Age Information Management 2013. New York: Springer, pp.99-104.2013.

[3] Chen D B, Lü L, Shang M S, et al. Identifying influential nodes in complex networks. Physica A, 391: pp.1777-1787.2012.

[4] Yang J, Counts S. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter. ICWSM,10:pp.355-358P.2010.

[5] Leskovec, Jurel. “Cascading Behavior in Large Blog Graphs.” Trials15. pp.1-11, 2007.

[6] Li, Hui. “Affinity-driven blog cascade analysis and prediction.” Data Mining & Knowledge Discovery 28.2. pp.442-474,2014.

[7] Dabeer, Onkar. “Timing Tweets to Increase Effectiveness of Information Campaigns.” International Conference on Weblogs & Social Media 2011.

[8] Janette Lehmann, et al. “Dynamical classes of collective attention in twitter”. Computer Science . pp.251-260.2011.

[9] Yang, Jiang, S. Counts. “Predicting the Speed, Scale, and Range of Information Diffusion in Twitter”. International Conference on Weblogs & Social Media 2010.

[10] Tsur, Oren, A. Rappoport. “What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities”. International Conference on Web Search & Web Data Mining . pp..643-652. 2015

[11] Wang, Bai Ling, et al. “Research on Network-Traffic Based Web Traffic Computing Technology.” Acta Electronica Sinica 41.4.pp.751-756. 2013.

[12] Macskassy, Sofus A., M. Michelson. “Why do People Retweet? Anti-Homophily Wins the Day!.” International Conference on Weblogs & Social Media 2015.

[13] Pal, Aditya, and S. Counts. “Identifying topical authorities in microblogs.” ACM International Conference on Web Search & Data Mining. pp.:45-54.2011.