Opinion Leader Mining of Social Network Combined with Hierarchical Sentiment Analysis

Hang Ye and Junping Du

1 Introduction

Nowadays, with the booming of internet, social networks are more and more popular in daily communications around the world. Everyone can post ideas and be the source of information at any time in anywhere on the social network. It is of great value to do research about social networks with mass data. However, because of the unrestraint and anonymity, the authenticity of the information on social networks can't be easily judged compared with the traditional official media. A huge challenge related to personal safety and social stability appears in public. Hence, it is meaningful to discuss about mining opinion leaders related with national security in social network.

Currently the methods of opinion leaders mining are divided into three patterns: the analysis of user's attributes, the analysis of information exchange and the analysis of network structure.

Considering bloggers' attributes, Zhang et al. [1] use the Markov networks to analyze the relevance of the intrinsic attributes of each user. The methods concerned with information interaction excavate opinion leaders by analyzing the propagation properties of the microblogs. Agarwal et al. [2] consider forwarding number, comment number and other attributes. To finding influential users, Li et al. [3] assess the quality of the bloggers in calculation. However, such methods have shortcomings in different aspects.

The last strategies become mainstream gradually because of exploiting graph models to represent the data structure and the better results can be obtained. There are two directions among these methods. One is complex network and the other is

H. Ye · J. Du (✉)
Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China
e-mail: junpingdu@126.com

Z. Deng (ed.), *Proceedings of 2017 Chinese Intelligent Automation Conference*, Lecture Notes in Electrical Engineering 458,
https://doi.org/10.1007/978-981-10-6445-6_70

topology analysis. Cho [4] measures the user's influence by intimacy, sociality and centrality in the network. Microblog-Rank algorithm [5] pays attention on comments from the microblogs. Twitter-Rank algorithm [6] focus on the social relationship between the user and their theme similarity.

This paper mainly utilizes the method that combines network structure and microblog's attributes. Besides, we adopt the emotional analysis method of hierarchical structure to identify the malicious users who are possible threat to national security from the set of opinion leaders.

2 Expression of Mining Opinion Leaders of Social Network

Providing a microblog's dataset with specific topic, how can we find the set of leading users, how can we discover the dubious ones among the set? Our main target is to answering these questions. The solution is based on the algorithm provided by Subbian et al. [7] associated with Yu et al.'s [8]. The process is composed of two parts, the first step is to calculate each user's influence score, the second is to pick up the ones who are probably harmful to social stability.

2.1 Construction of Information-Flow Tree

The construction can be divided into two process which are building a tree and calculating. With communication networks, we need an accessible and computable pattern to represent the connections. Under this circumstance, we use the information-flow tree to save interactions and attributes of texts and users without taking textual information into consideration. Imitating from the creation of frequent pattern tree, the information-flow tree is constructed based on the behaviors like forwarding and commenting, and each node stands for a user.

2.2 Attributes of Nodes and Calculation of Users' Influence

As we discuss in Sect. 3.1, the nodes' weight is determined by social and users' properties. The users' attribute is composed of the quantities of microblogs, followers as well as followings. As for the social attribute, we take account for the number of posting and the number of comments for the formulas. The calculations are given by:

$$wi = \beta_1 N_1 + \beta_2 N_2 + \beta_3 N_3 \tag{1}$$

$$p_i = \lambda N_4 + (1 - \lambda) N_5 \tag{2}$$

Besides, it is noticeable that the overall path weight from root to a particular node is equal to the node's weight itself.

After completing the construction of information-flow tree, it comes to the calculation of each node's influence score. The influence between two nodes is obtained by calculating the sum of the information flows in the two pairs of nodes. The information flows between paired nodes is defined as the sum of the flows of all paths between two nodes for a certain keyword for some time.

The flows between paired nodes are given by following:

$$V(a_j, a_k, K_i, t_c) = \sum_{P \in S_{jk}} A(P, K_i, t_c) = \sum_{P \in S_{jk}} \sum_{a_i \in P} w(a_i, K_i, t_c) \tag{3}$$

w is determined by weight of each node.

Next, we can acquire each node's influence score by Influence Function which is shown as below:

$$\alpha(a_j, Q, t) = \frac{I(a_j, *, Q, t)}{I(*, *, Q, t)} = \frac{\sum_{K_i \in Q} V(a_j, *, K_i, t)}{\sum_{K_i \in Q} V(*, *, K_i, t)} \tag{4}$$

The nodes are ranked by the influence scores. If one user get a high influence score, he is more likely to become an opinion leader of relevant topics.

2.3 Sentiment Analysis of Users

Assumed that we obtain the ordered set of conceivable leading users, it's time to take sentiment analysis into account, where we use the method called hierarchical emotional analysis. The microblog is separated into three parts which are sentence, sub-sentence and phrase. According to the bottom-up rule, each phrase sentiment score is calculated and then to the level of sentences.

When we concentrate on the sentiment of one word, the PMI has been used to calculate semantic similarity in accordance with sentiment dictionary.

Next, the emotional vector of the phrase focus on the structure and dependencies between these different words. In the sub-sentence level, the whole text is divided by punctuations like "、", ",", ";", and relations among this short sentences become a significant feature to calculate. Finally, we reach the top level which consists of long sentences. With effect of different punctuations, we can make a result of a sentence.

$$Sen(a_i) = \frac{1}{N}\sum_{a_i \in V} Sen(tweet_i) = \frac{1}{N}\sum \frac{1}{N_i}\sum_{j=1}^{N_i} \beta_{ij} \cdot \left(\frac{\overrightarrow{p_{v_{ij}}} \cdot (\overrightarrow{\varphi_{v_{ij}}} \circ \overrightarrow{\eta_{v_{ij}}})}{n_{v_{ij}}} + \frac{\overrightarrow{p_{a_{ij}}} \cdot (\overrightarrow{\varphi_{a_{ij}}} \circ \overrightarrow{\eta_{a_{ij}}})}{n_{a_{ij}}} \right), \tag{5}$$

The formula (5) presents the integration of all three levels. From it, we can arrive at the sentiment score of a microblog.

2.4 Integration of Users' Influence Score and Emotion Score

In this section, each user's degree of threatened in a certain period can be elicited on basis of influence score and sentiment score. The user list facilitates further tracking of the relevant users. The combination process is shown as below:

$$D(a_i) = \delta \times \alpha(a_i, Q, t_i) + (1 - \delta) \times Sen(a_i) \tag{6}$$

Due to the dimension between the results of α and Sen, it is essential to normalize data before formula (7).

3 Experiments on Real Data Set

This section is an application to our procedure and empirical valuation. We set the same weights of the social attributes and the combination factor as following in all experiments:

$$\beta_1 = \beta_2 = \beta_3 = 0.33, \quad \alpha = 0.6 \tag{7}$$

3.1 Data Set of Sina Weibo

We gather two datasets from Sina Weibo for assessing several aspects of our method. The total number of microblogs is 2.25 million. The period is from 2011 to 2014. Among them, we concentrate on the pieces related to national security. Therefore, some keys are used to filter to get truly meaningful contents. The keys contains The"Xiao Yueyue" Event, Bullet Train Rear-End Collision, Ya'an Earthquake, MH370 Accident etc. After pretreatment, we finally get nearly 30 thousand microblogs with 7 thousand users.

3.2 Results and Evaluation

To illustrate the advantages of our methods (Information-Flow Tree with Attributes), we introduce some influence evaluation techniques as the comparisons. They are TopicLeaderRank (LR) [9], ProfileRank (PRF) [11], Repuser (SD, SS3) [10] and PageRank (PR). The former two methods are supplements and improvements of PageRank. The Repuser algorithm is to maximize objective function by stratified sampling and diversity sampling. All this comparative methods are set with default parameters. The standard list adopted is the sequence of users that only takes the number of forwarding and comment as the sorting index.

Figure 1 shows the core rate of four methods. The core rate measures the ability of interaction of an opinion leader. The core rate of a user is determined by the number of posting and the number of comments associated with the user's microblogs. Figure 2 shows the assessments of single coverage rate. From the point view of network topology, coverage rate measures the dominant power by calculating the number of affected users. Single-step coverage only consider the direct touched neighbors of each user. As expected, our IFwA method achieve better than others in terms of the two metrics. It is a obvious fact that the IF method performs outstandingly in top-150 users.

Table 1 shows more details for this experiment with Tables 2 and 3.

Table 1 is the mean average precision of all algorithms under different chosen ranges. It is clear to observe that IFwA achieves a better result than other baseline methods.

Table 2 is the results of accuracy, which means we concern about involvement rather than sequence within measurements. Moreover, the IFwA algorithm has more advantage of NDCG in Table 3.

It is probably owing to the IF tree not only rely on the communications among users, but also consider user attributes and interaction attributes, rather than only take the number of forwarding for the construction of IF tree. This procedure is learnt from the principle of LeaderRank algorithm to complement the original influence calculation. Compared with the sampling strategy in the SD and SS3

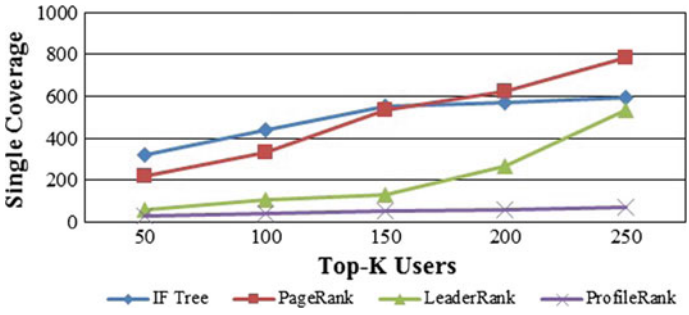

Fig. 1 Single coverage of the PageRank, LeaderRank, IF tree with attributes and ProfileRank

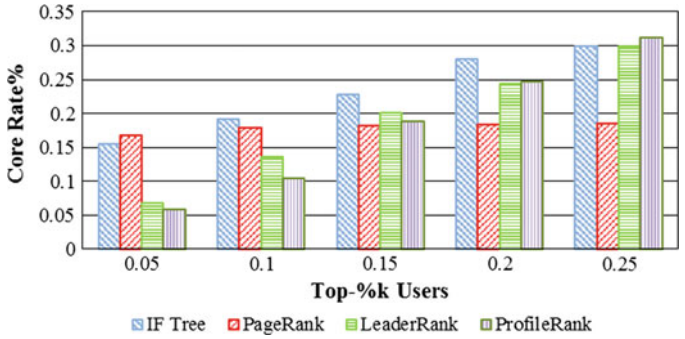

Fig. 2 Core rate of the PageRank, LeaderRank, IF tree with attributes and ProfileRank

Table 1 Evaluation of the PageRank, LeaderRank, IF tree, ProfileRank in terms of MAP

MAP	PR	LR	IFwA	PFR
@50	0.50	0.01	0.71	0.32
@100	0.51	0.02	0.72	0.33
@200	0.56	0.15	0.60	0.34
@300	0.59	0.23	0.55	0.20
@400	0.61	0.26	0.53	0.30
Avg.	0.54	0.17	0.57	0.31

Table 2 Evaluation of the PageRank, LeaderRank, IF tree, ProfileRank, SS3, SD in terms of accuracy (without sequence)

Accuracy	PR	LR	IFwA	SSD	S3	PRF
@50	0.48	0.0	0.86	0.16	0.2	0.02
@100	0.62	0.0	0.83	0.19	0.22	0.01
@150	0.72	0.006	0.83	0.17	0.17	0.01
@200	0.72	0.04	0.67	0.15	0.15	0.005
@400	0.73	0.235	0.38	0.15	0.19	0.02

Table 3 Evaluation of the methods of NDCG (Top-100)

NDCG	PR	LR	IFwA	SSD	S3	PRF
@50	0.557	0.0	0.640	0.193	0.379	0.002
@100	0.554	0.0	0.644	0.192	0.377	0.002
@200	0.601	0.0001	0.650	0.193	0.380	0.0004
@300	0.600	0.105	0.649	0.167	0.379	0.003
@400	0.604	0.096	0.648	0.153	0.379	0.003

algorithm, the IFwA method pays more attention to the specific behavior of the users. Hence, for the chosen baseline as the evaluation, the IFwA improves superiorly.

Next, we evaluate the complete proposed procedure using 2500 randomly chosen data because of the limitation of runtime. The comparisons are the combination of hierarchical sentiment analysis with PR as well as with LR. The baseline

Table 4 MAP of performance by integration of influence score and emotional score

MAP	PR_Sentiment	LR_Sentiment	IFwA_Sentiment
@50	0.06	0.03	0.27
@100	0.11	0.37	0.34
@150	0.36	0.33	0.38
@200	0.39	0.41	0.41
@500	0.40	0.38	0.39

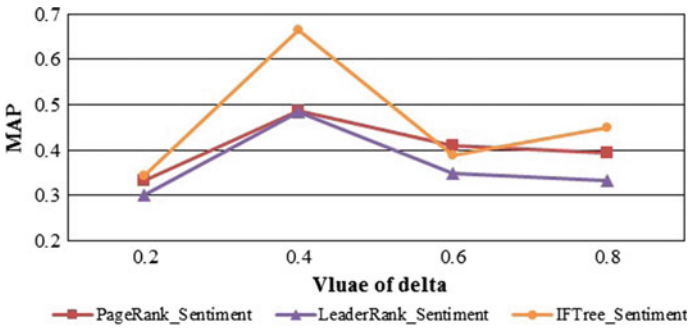

Fig. 3 Evaluation of MAP about the integration with different values of delta

of the sentiment analysis is the array only considering the number of sentiment words. δ in formula (10) is identified as 0.5. The evaluation displays in Table 4.

We can find that the Information-Flow tree with Attributes with hierarchical emotional method is slightly better than the other two integration.

Figure 3 presents the impact of the value of δ on MAP. The greater the δ value is, the more dominant the emotional factor is. Therefore, we can figure it out that the integration of our procedure perform outstandingly under most conditions. And when the δ reaches nearly 0.5, this three methods all have greater effect.

4 Conclusions and Future Work

We propose to tackle the problem of mining opinion leaders related to national security by quantizing the leading degree and threat degree of each user. The two-step procedure is used to realize our goal: an Information-Flow tree is first constructed for the calculation of influence score. Furthermore, we apply a Hierarchic Emotional analysis to discover the radicalness from microblogs. The experiment results show our procedure meet the requirement in real dataset. Quantitative analysis on synthetic data demonstrates our method performs more excellent than other baseline algorithms.

In the future, we could extend our procedure in source trustworthiness analysis besides improvement in influence calculation. Moreover, it would be possible to produce a more fine-grained word-level analysis in sentiment evaluation.

Acknowledgements This work was supported by the National Natural Science Foundation of China (No. 61532006, No. 61320106006, No. 61502042).

References

1. Zhang WZ, Li XQ, He H et al (2014) Identifying network public opinion leaders based on Markov logic networks. Sci World J 435–444
2. Agarwal N, Liu H, Tang L et al (2010) Identifying the influential bloggers in a community. In: International conference on web search and web data mining, Palo Alto, California, pp 207–218
3. Li F, Du TC (2011) Who is talking? An ontology-based opinion leader identification frame00work for word-of-mouth marketing in online social blogs. Decis Support Syst 51 (1):190–197
4. Cho Y, Hwang J, Lee D (2012) Identification of effective opinion leaders in the diffusion of technological innovation: a social network approach. Technol Forecast Soc Chang 79(1):97–106
5. Lin Y, Li HX, Liu XQ et al (2013) Hot topic propagation model and opinion leader identifying model in microblog network. Abstr Appl Anal 6:1–13
6. Weng J, Lim EP, Jiang J et al (2010) Twitterrank: finding topic-sensitive influential twitters. In: Proceedings of the third ACM international conference on Web search and data mining, New York City, pp 261–270
7. Subbian K, Aggarwal CC, Srivastava J (2016) Querying and tracking influencers in social streams. In: ACM international conference on web search and data mining, San Francisco
8. Yu ZW, Wang ZT, Chen LM et al (2016) Featuring, detecting, and visualizing human sentiment in Chinese micro-Blog. ACM Trans Knowl Discov Data 10(4):48
9. Wu XH, Zhang H, Zhao X (2015) Mining algorithm of microblogging opinion leaders based on user-behavior network. Appl Res Comput 9:2678–2683 (in Chinese)
10. Tang J, Chenhui Z, Keke C et al (2015) Sampling representative users from large social networks. In: 2015, Association for the Advancement of Artificial Intelligence, Austin
11. Silva A, Meira W, Zaki M (2013) ProfileRank: finding relevant content and influential users based on information diffusion. The Workshop on social network mining and analysis, pp 1–9