ISSN 2348 – 7968

# Mining Algorithm of Opinion Leaders Based on Analysis of User Attention Behavior

**Ying Xia, Miaomiao Cao, Xu Zhang and Hae Young Bae**

Research Center of Spatial Information System, Chongqing University of Posts
and Telecommunications, Chongqing, China

**Abstract**

Mining opinion leaders is of great significance to the dissemination of information in social network. This paper put forward a mining algorithm of opinion leaders in social network. First, the algorithm defined the concept of attention degree based on the analysis of user attention behaviors including forward, comment, point praise and other concerns. Then it adopted the ideas of voting in the PageRank algorithm to calculate user's influence. Experimental analysis based on data set from Sina Microblogging shows that the algorithm can identify the influential users more effective and accurate.

***Keywords:*** *social network, opinion leaders, attention behavior, attention degree, PageRank.*

## 1. Introduction

Opinion leaders, known as the leaders of public opinion and view communicator, refer to "active elements" that not only provide information, advice, comments, also have a certain impact on others in the interpersonal communication network [1]. Social network is an online community that provides a platform for people with the same background to share their interests and hobbies. In the process of information dissemination in social network, opinion leaders playing a more and more important role, are widely used in public opinion monitoring, information promotion, e-commerce and other fields.

With the rapid development of Internet technology and social network, users can share information anytime and anywhere on the network. Especially FaceBook, Twitter, Sina Microblogging and other large-scale online social network have become important mediums of communication, playing important roles in information communication between network users. How to find a set of users with a certain influence accurately in a large amount of data from the large-scale social network has been the research target of opinion leaders mining.

## 2. Related Work

Scholars have carried out a wide range of research on the mining of opinion leaders in social network. These studies can be broadly divided into the following three kinds of methods.

Mining methods of opinion leaders based on topology analysis of network, such as Larry Page [2] reflected the relevance and importance of web pages by using PageRank algorithm, Jon Kleinberg [3] evaluated the degree of user's authority by using HITS algorithm, Zhai [4] mined opinion leaders through classifying the data set according to the theme and adopting PageRank algorithm, Song [5] proposed InfluenceRank algorithm on the basis of considering the novelty of the blog itself and the links between blogs, Yu Xiao [6] proposed LeaderRank algorithm through increasing analysis of content and emotional of posts.

Mining methods of opinion leaders based on information interaction, mainly through the analysis on the impact of user's information and the dissemination of information to reflect the users' influence. For example, Agarwal [7] evaluated the impact of posts through the degree of novel and the length of content of posts. Li [8] found opinion leaders from the product evaluation according to the quantity and quality of blogs. Xinghua Fan [9] calculated influence value and filtered opinion leaders by the way of conveying views basing on text interaction.

Opinion leaders mining methods based on users' attributes, such as Xuefeng Ding [10] constructed attributes matrix of users and carried on the synthesis weight to sort, Shuai Zhu [11] designed X-means model, is a iterative clustering filtering model based on Bayesian information gain maximization, Yu Wu [8] put forward UI-LR, a discovery algorithm of opinion leaders based on user influence by considering the user's own influence and links between users.

Currently, due to the increasingly rich social networking platform, users' attention behaviors are also more diverse. This paper, not only taking network topology and user attributes into consideration, also adopting ideas of PageRank algorithm, analyze the behavior of users and find opinion leaders effectively.

## 3. Calculation Model of User Influence

### 3.1 Data Structure of Social Network

A user in social network is used as a node, a link between two users is used as a side, then the whole social network can constitute a graph. The network structure of users with concern relationship can be expressed by network G= (N, E, W) with weight. Among them, N is a set of nodes that represents all users. E is a collection of edges that indicates the relationship between users. W is a set of edges' weight to describe the degree of attention among users.

### 3.2 Introduction of PageRank

In social networks, If user p follows user q, then p will give his degree of attention to q, if q gets high degree of attention that means he has a large influence on others, then q has a relatively large influence on the p. This idea is similar to the basic ideas of PageRank algorithm [2].

PageRank algorithm is a classical algorithm used to measure the importance of pages in network. The structures of links between pages are used to determine the importance of pages. If page p links to page q, that is to say, p votes to q. Ranking system will assess importance of pages based on the number of votes received between pages. The expression of the algorithm for the importance of page q calculated by the importance of pages voted to q. For the PR value of page q, namely PageRank (q), as shown in formula (1):

$$\mathrm{PageRank(q)} = (1 - \mathrm{C}) + \mathrm{C}\sum_{\mathrm{p}\in \mathrm{A}} \frac{\mathrm{PageRank(p)}}{|\mathrm{p}|} \qquad (1)$$

Thus, the PR value of page q can be attained by adding all results calculated by PR value of all the pages linked into q divided by the number of every page linked out. A refers to a set of pages linked into page q, |p| refers to the number of every page linked out. In the formula, the damping factor C ($0<C\leq 1$), refers to the probability keeping to browse the pages linked out.

Thus, in the process of calculating PR value of every page in iteration, the main step is to calculate the PageRank (p) /|p|, which means, the PR value of page p average distribution to pages linked out. However, there is no consideration of the differences among pages.

### 3.3 Definition of Attention Behaviors and Attention Degree

In social network, a user can be followed by more than one user. Attention behaviors occurred in two users mainly includes one user forwards, comments, points praise and other acts on another user's information. A user's information can't get the same attention degree from others, because they don't have same friendship and hobby. So the degree of attention user p pays to user q, reflected by the attention behaviors occurred between p and q's information.

The degree of attention is reflected by attention occurred among nodes. In social networks, whether nodes can get attention or not reflected by the probability of attention behaviors occurrence. The definition of attention degree among nodes is:

$$\mathrm{W}_{(\mathrm{p,q})} = \frac{\mathrm{N}_{(\mathrm{p,q})}}{\mathrm{N}_{(\mathrm{q})}} \qquad (2)$$

In formula (2), N(p,q) indicates the number of q's information given attention by p, it can be expressed as formula (3). 1~n, which indicates that p owns the number of properties on behavior to q, such as forward, comment, point praise and collection; $\mathrm{N}_1 \sim \mathrm{N}_\mathrm{n}$ represents the number of attribute; $\alpha_1 \sim \alpha_\mathrm{n}$ represents the weight of every attribute and they add up to 1. N(q) refers to the total number of information of user q.

$$\mathrm{N}_{(\mathrm{p,q})} = \alpha_1\mathrm{N}_1 + \ldots + \alpha_\mathrm{i}\mathrm{N}_\mathrm{i} + \ldots + \alpha_\mathrm{n}\mathrm{N}_\mathrm{n} \qquad (3)$$

In the above formula, the number of attributes that are related to and the weight value of attribute can be set up according to the specific application.

### 3.4 Calculation Method of User Influence

In PageRank algorithm, the PR value of a page is uniformly transmitted to other pages link out, however it does not consider the importance of the page itself. In order to calculate user influence well, this paper proposes a mining algorithm of opinion leaders in social network (Social Network User Rank, SNURank), considering the characteristics of users' attention behaviors.

The basic idea of the algorithm is, taking the above definition of attention degree as a transfer factor of influence value distribution. A user who gets high degree of attention can get a higher influence value. On the contrary, a user who gets lower degree of attention will get a lower influence value. The problem of uniform transmission can be solved by this way and the problem of simply relying on the link among users to rank also can be overcome, so the model can reflect the actual situation more objectively. Similar to the PageRank algorithm, users in social network correspond to pages, the relationship between users is corresponding to the link between pages, thus calculation expressions of Network User Rank Social (SNURank) algorithm is as follows:

$$\mathrm{SNUR}(q) = (1 - C) + C \times \sum_{p \in E} \overline{A_{p,q\mathrm{SNUR}(p)}} \qquad (4)$$

The C, the damping coefficients are often used in the reference PageRank algorithm, is set 0.85. E represents a collection of all the users who attention about q.

In addition, as shown in the formula (5), $A_{p,q}$ refers to the proportion of p allocate attention to q, That is because user p concerned about many users, so the user q get user p's influence is relative proportion of p. F represents a set of users who are concerned by p.

$$A_{p,q} = \frac{W_{(p,q)}}{\sum_{q \in F} W_{(p,q)}} \qquad (5)$$

## 4. Experiment and Analysis

### 4.1 Experimental Data Set

Experiment selected the data set of Sina Microblogging which from Datatang platform (http://www.datatang.com/data/46758), including information belongs to 63641 users, friendship among 1391718 users, forwarding relationship in 27759 micro blogs. The data set collected are stored in the following format:
1) table of users' information: user ID, number of fans, number of attention, number of micro blogs
2) table of micro blogs' information: micro blog ID, the time published, number of forwarding, number of comments, number of points, user ID who publish
3) table of users' relationship: user ID, fans ID
4) table of micro blogs relationships: micro blog ID, original micro blog ID

### 4.2 Evaluation Factors

Influence coverage rate of nodes put forward in [12] will be used as evaluation factor in this paper, which means the proportion of the number of nodes affected by influential nodes directly or indirectly to the number of total nodes. The formula is shown as (6):

$$p(k) = \frac{\sum_{k=1}^{N} L(k)}{N} \qquad （6）$$

As shown in the above formula, p(k) indicates the influence coverage rate of the former k users, N represents the number of users to be studied, L (k) indicates that the number of users is affected by user number k.

In general, the greater impact coverage rate of the previous users, the greater influence users are.

### 4.3 Influence Coverage Analysis

In order to evaluate the effectiveness of SNURank algorithm (SNUR), compared with mining algorithm of opinion leaders based on PageRank algorithm (PR), considering the influence coverage rate of top 20 users in two algorithms.

In experiment, we main comprehensive the number of forwards, comments and points as attributes to calculate degree of attention of users. Taking into account user p transmits user q's information can help the q's information be further spread. That is, it may help q get a higher degree of attention. So the weight of this three attributes can be set to 0.4, 0.3 and 0.3. The results of experiment are shown in figure 1. It can be seen that the influence coverage rate of SNURank algorithm is significantly higher than the PageRank algorithm's.

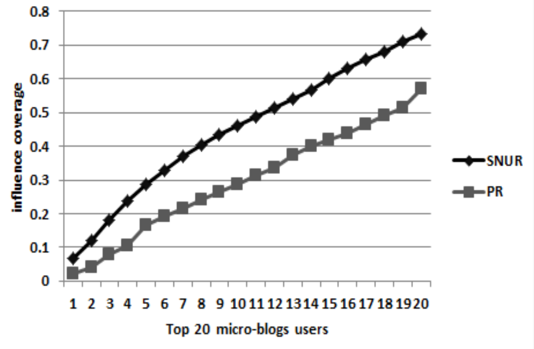

Fig. 1 Influence Coverage Rate for comparison

### 4.4 Case Analysis

Compared with PageRank algorithm and SNURank algorithm proposed in this paper, calculate the influence of sina Weibo users. Finally get the ranking results of users' influence. Results of two algorithms obtained as shown in Table 1 and Table 2.

Table 1: Influence Ranking of PageRank

| *Rank* | *userid* | *fans* | *follows* | *micro-blogs* | *PR* |
|---|---|---|---|---|---|
| 1 | 1618051664 | 33787450 | 88798 | 88798 | 0.002152728 |
| 2 | 1191258123 | 37607791 | 951 | 204 | 0.002105834 |
| 3 | 2656274875 | 18249317 | 172 | 31901 | 0.001809606 |
| 4 | 1496852380 | 4534645 | 184 | 483 | 0.001779502 |
| 5 | 2803301701 | 19977043 | 272 | 28677 | 0.001429315 |
| 6 | 1197161814 | 51287110 | 517 | 13828 | 0.001398102 |
| 7 | 1761179351 | 7834504 | 178 | 1247 | 0.001167495 |
| 8 | 1182389073 | 20287367 | 161 | 69177 | 0.001156822 |
| 9 | 1705586121 | 11640156 | 295 | 3097 | 0.001038854 |

Table 2: Influence Ranking of SNURank

| *Rank* | *userid* | *fans* | *follows* | *micro-blogs* | *SNUR* |
|---|---|---|---|---|---|
| 1 | 1216766752 | 1250609 | 1640 | 3397 | 0.003442212 |
| 2 | 1974808274 | 1541472 | 321 | 13826 | 0.003056597 |
| 3 | 2834256503 | 2521620 | 198 | 7409 | 0.003039269 |
| 4 | 1904769205 | 2023180 | 1261 | 13564 | 0.002968050 |
| 5 | 1496852380 | 4534645 | 184 | 483 | 0.002776082 |
| 6 | 1850988623 | 1898923 | 1962 | 23000 | 0.002755934 |
| 7 | 1235457821 | 2126566 | 164 | 731 | 0.002280324 |
| 8 | 1649159940 | 2813741 | 614 | 73914 | 0.002049567 |
| 9 | 1197362373 | 3589483 | 719 | 817 | 0.001414612 |

Compared with the results of user ranking obtained by two algorithm can be found that the ranking of top 9 users' influence is different. Because PageRank algorithm only considers link between two nodes, that is to say influence value is mainly determined by the sum of out degree and in degree. Here, the number of fans and attention a user owned refers to the node's out degree and in degree. For example, the user whose ID is “1496852380” owns the number of fans is far greater than the top 2 users' in table 2. The reason for this is because of many of his fans not pay attention to him. So the algorithm proposed in this paper to calculate the influence value, the ranking of this user is lower than the ranking of PageRank algorithm attained.

Known from analysis, SNURank algorithm considers attention behaviors like forwards, comments, point praise and other concerns between two users. The influence value of user can be calculated by the degree of attention. It is more accurate to identify the opinion leaders in social networks.

## 5. Conclusions

Mining opinion leaders for information dissemination in social network is of great significance. This paper proposes SNURank algorithm, a mining algorithm of opinion leaders in social network. It can be used to calculate user's influence and select opinion leaders by defining degree of attention based on analysis of user's attention behaviors and the ideas of PageRank algorithm. Through experimental analysis of data set from Sina Microblogging, the algorithm can identify the influential users more effective and accurate.

### Acknowledgments

This work is funded by Chongqing Natural Science Foundation (cstc2014kjrc-qnrc40002), Scientific and Technological Research Program of Chongqing Municipal Education Commission (KJ1500431), and WenFeng Creative Foundation of CQUPT (WF2014-05).

## References

[1] Paul F.Lazarsfield. The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign[M] New York: Duell,Sloan & Pierce, 1944

[2] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 1998, 30(1): 107-117

[3] Kleinberg J M. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999, 46(5): 604-632

[4] Zhongwu Zhai, Hua Xu, Peifa Jia. Identifying Opinion Leaders in BBS[C]. 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent AgentTechnology, Sydeny, NSW, Australia, 2008:398-401

[5] Xiaodan Song, Yun Chi, Koji Hino, et al. Identifying opinion leaders in the blogsphere[C]. Proceedings of the 16th ACM Conference on Information and Knowledge Management, Lisboa, Portugal, 2007: 971-974

[6] Yu Xiao, Wei Xu, Lin Xia. Networking Groups Opinion Leader Identification Algorithms Based on Sentiment Analysis. Computer Science, 2012, 39(2): 34-37+46

[7] Agarwal N, Liu H, Tang L, et al. Identifying the influential bloggers in a community[C].Proc of International Conference

on Web Search and Web Data Mining. New York: ACM Press,2008: 207_218

[8] Li Feng, Du T C. Who is talking? An ontology-based opinion leader identification framework for word-of-mouth marketing in online social blogs[J]. Decision Support Systems, 2011, 51(1): 190-197

[9] Xinghua Fan, Jing Zhao, Binxing Fang, Yuxiao Li. Influence Diffusion Probability Model and Utilizing It to Identify Network Opinion Leader. Chinese Journal of Computers, 2013,36(2)：360-367

[10] Xuefeng Ding，Feng Hu，Wen Zhao. Research on the characteristics of network opinion leaders. Journal of Sichuan University: Engineering Science Edition, 2010,42(2): 145-149

[11] Shuai Zhu，Xiaolin Zheng，Deren Chen. Research of algorithm for automatic opinion leader Detection in BBS. System Engineering-theory & Practice, 2011, 31(S2): 7-12

[12] Yu Wu, Lulu Ma, Mao Lin, HongTao liu. Discovery Algorithm of Opinion leaders Based on User Influence. Journal of Chinese Computer Systems,2015,36(3)：561-565

**Ying Xia** is a professor of Chongqing University of Posts and Telecommunications. Her research area mainly includes database and data mining, spatial information processing, etc.

**Miaomiao Cao** is a graduate student at the College of Computer Science and Technology, Chongqing University of Posts and Telecommunications. Her research area mainly is social network data mining.

**Xu Zhang** is an associate professor of Chongqing University of Posts and Telecommunications. His research area mainly includes cloud computing and big data processing.

**Hae Young Bae** is tenured full professor of Inha University of Korea and honorary professor of the Chongqing University of Posts and Telecommunications of China. His research area mainly includes database and spatial information processing.