Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/13/20230727

Analysis of microblog user's influence based on the social network model

Xiang Yiyang

Zhejiang Gongshang University, Hangzhou, China

yanna_xiang@qq.com

Abstract. With the development of science and technology in today's world, social software and networks have become more and more advanced. Microblog is a widely used application and everyone in China has to contact it. According to calculations, thousands of microblogs can be generated every second, so the impact of microblogs is huge. At present, the research on microblogging social networks is relatively few, and the research on user influence in microblogging is even less. Even if the influence of users in microblogs is studied, the methods used are relatively complex, and it is impossible to simply and clearly select users with greater influence. Therefore, based on the social network model, this paper first uses a Web crawler to obtain microblog data. This paper obtains only one user's id, gender, age, city, number of followers, number of posts, attitudes, and comments. When this paper gets the data, this paper uses the knowledge of graph theory to simplify the specific data, so as to better study the relationship between the data and explore their influence. This paper regards every user in microblog as a node, and the relationship between the users is the edges. Then calculate each node's Degree Centrality and Eigenvector Centrality to find influential users. This paper finds that the influence gap between different users is large. Women's influence is slightly higher than men's. The youth group is significantly larger than that of the others. The influence of first tier cities and new first tier cities is slightly higher than those of second tier cities, and their influence is significantly higher than those of other cities. These results will help people suit the remedy to the case and achieve better communication results.

Keywords: microblog user, influence, the social network model, web crawler, centrality.

1. Introduction

With the development of science and technology in today's world, social software and networks have become more and more advanced. Social networks can be said to be an important symbol of the development of the Internet. People have changed from getting information through newspapers and TV to getting information through social software. After the advent of social software, the way people receive information has changed from passive to active, and they can also receive the latest news at the first time. The social network is like an invisible big network, connecting everyone together, and every user's behavior will leave traces on it.

Microblog is a widely used application and everyone in China has to contact it. Microblog is a convenient platform for expressing opinions, and every ordinary person can express their opinions here. The user is not only the transmitter of information but also the manufacturer of information.

DOI: 10.54254/2755-2721/13/20230727

According to calculations, thousands of microblogs can be generated every second, so the impact of microblogs is huge. Most people are used to finding what they want to know on microblogs through keyword searches. Many businesses also promote their products through microblogging. They tell their products to the most influential group of users and use their influence to let more people know. The government will also use microblogging to listen to the public and understand public opinions. However, since microblog is an open platform, some people use it to make negative comments. The government can let influential users speak out to clarify the facts. Therefore, it is of great significance to find out the influential users in Microblog not only for marketing but also for social stability.

At present, the research on microblogging social networks is relatively few, and the research on user influence in microblogging is even less. Analysis of influence is an important part of social network analysis. Even if the influence of users in microblogs is studied, the methods used are relatively complex, and it is impossible to simply and clearly select users with greater influence. Therefore, based on the social network model, this paper uses simple algorithms to find influential users, so as to help people identify information and use microblogging to sell goods in the future.

2. Literature review

In recent years, there are more and more methods and algorithms to analyze the influence of microblog users at home and abroad. In 2010, Weng [1] focused on the problem of identifying influential users of microblogging services and proposed the TwitterRank algorithm. But the algorithm was based on the premise of "homophily", so it had some limitations.

In 2011, Bakshy [2] tried to analyze the number of user followers, registration time, the frequency of microblogging, etc. And they used the attenuation tree model to predict user influence. This idea is feasible, but they still have no way to improve the ability to predict user influence.

In 2013, according to the relationship between users and their own interests, based on the PageRank algorithm, a hierarchical method is proposed by Su [3] to evaluate user network influence. However, when user A affects user B and user B affects user C that means A also indirectly affects user C, which they did not consider.

In 2014, Mao [4] analyzed user behavior and social network structure by predicting the number of times microblogging is forwarded and read to judge the influence of each user. This method can value the influence of every user fairly well, but it didn't take the content into consideration.

In 2016, Zhuang [5] proposed the SIRank algorithm based on the number of each user's interactions. Bartoletti [6] combined with the user's historical data, calculated the user's recent reputation using an arbitrary sorting method. Although this method improves the performance of the sorting method, especially when processing large data sets, it ignores the impact of user relationships on ranking.

In 2017, some studied the influence through emotional factors, some users with the highest influence are found according to the consistency between the user's emotions after forwarding the microblog and the original microblog blogger's emotions. Considering the data obsolescence and topic independence, Yuan [7] used the MapReduce framework. However, the number of user followers and user activities are factors that they did not consider.

In 2018, Wu [8] multi the evaluation of user influence based on reply content and relationship. Sun [9] analyzed the relationship network of microblog users and the actual behavior of microblog users and proposed the MBUI Rank algorithm, which introduced the weight of microblog users that makes the results more accurate. But both of these researches did not consider the initial influence of users.

In 2020, Rui [10] proposed the method of fixed neighbor scale, which is an innovative method to extract useful information from multiple neighbor levels of target objects to estimate their impact intensity. Huang [11] proposed CMIA, a solution to influence attribute networks. The operation time of this algorithm is good and it can also cover a range of areas. Nonetheless, this algorithm is based on the common network structure, without considering the uniqueness of social networks.

DOI: 10.54254/2755-2721/13/20230727

3. Methodology

3.1. Microblog data acquisition method

A web crawler is an important part of the search engine. It can automatically extract the content on the web page and download the web page from the World Wide Web for the search engine. The traditional crawler starts from the URL of one or several initial pages, obtains the URL on the initial page, and continuously extracts new URLs from the current page and puts them in the queue during the process of fetching the page until certain stop conditions of the system are met. The traditional crawler doesn't pay attention to the priority, and only climbs down all the contents of the network.

In addition to traditional crawlers, there is also a relatively complex focused crawler. In its interior, there is some code to analyze the web page, and filter out some links irrelevant to the theme based on this, leaving only useful links and putting them into the URL queue waiting to be crawled. Then, it selects the next page URL from the queue according to a certain search strategy and repeats the above process until a certain condition of the system is met. The focused crawler stored, analyzed, and indexed the crawled pages for future queries. At the same time, the results of its analysis also have an impact on the subsequent crawling.

When this paper obtain microblog data, this paper used two crawling methods comprehensively. And obtain only one user's id, gender, age, city, number of followers, number of posts, attitudes, and comments. These data not only meet the research needs but also improve the efficiency of the crawler.

Depth-first search is an algorithm that can drill down into every possible branch path to the point where it can not be deeper, and each node can only access it once. When this paper get microblog data, this paper started with one microblog and crawled the user information of all its reviewers one by one. After crawling the user information of the current microblog, this paper selected microblogs in other fields to repeat this operation. The selected fields should be broad and comprehensive, ranging from national major news, and hot search list news, to some regional minority topics, including current affairs, culture, entertainment, games, education, public health, and other aspects.

3.2. Data social network modeling

When this paper got the data, this paper used the knowledge of graph theory to simplify the specific data, so as to better study the relationship between the data and explore their influence.

Social networks include nodes, relationships, user groups, communities, and other basic concepts. Graph theory mainly includes nodes and edges. This paper regarded every user in the microblog as a node, and the relationships between the users were the edges. For example, if A is the follower of B then there is an edge between A and B.

In fact, the relationships between nodes are different. Some users interact more with each other, so their relationship is naturally closer. Therefore, different weight values should be introduced to the relationship to distinguish this kind of situation. However, when constructing the network, this paper only used the binary relationship, that was, two people were either connected or not connected, to get the simplest result.

3.3. Network centralization

This paper needs to find out the users with great influence, that is, the nodes in the center of the corresponding social network graph.

3.3.1. Degree centrality

Degree Centrality originated from social network research and is the most direct measure of node centrality in network analysis. In a graph, the greater the degree of a node, the higher the Degree Centrality of the node, and the more important the node is in the network.

The calculation formula for the Degree Centrality of a node is as follows:

$$DC=\frac{Ndegree}{n-1} \quad (1)$$

DOI: 10.54254/2755-2721/13/20230727

Ndegree: Indicates the number of edges between the existing node and the current node, it is also called the degree of the node.

n: Indicates the number of all edges connected between the current node and all the other nodes.

3.3.2. Eigenvector centrality

Degree Centrality thinks each neighbor has the same importance. However, in the real microblog network, each node has different importance, and their impact on the central node is also different, so this paper also used Eigenvector Centrality to reflect the centrality of a node more accurately.

First, this paper converted the graph to its corresponding Adjacency Matrix A. If two nodes do not have directly connected edges, it is recorded as 0, otherwise, it is recorded as 1. This paper put the Degree Centrality of each node into a column matrix M. Then this paper multiplies matrix M by matrix A and takes the obtained matrix as the new matrix A. Iterate this process until the proportion between the data is close to a fixed value. This vector A is what this paper is looking for. The data in vector A represent the importance of each node.

$$M \cdot A = \lambda \cdot A \tag{2}$$

λ:Is a constant

4. Results

When obtaining user data, this paper uses python technology to crawl 2350 users and generate 3578 relationships. This paper classifies all the data according to the three labels of gender, age, and province. The subjects were divided into Juvenile, Youth, Middle age, Old age these four groups according to their different age. And according to the prosperity of their cities, the subjects are also divided into four groups. The classification of the age group and the city group as well as the proportion of each group obtained are shown in detail in Table 1 and Table 2.

Table 1. Age group.

Age range	Age Group	Proportion
0-18	Juvenile	29%
19-40	Youth	63%
40-55	Middle age	6%
56 years old and above	Old age	2%

Table 2. City group.

City	City Group	Proportion
Shanghai, Beijing, Shenzhen, Guangzhou	First-tier cities	37%
Chengdu,Chongqing,Hangzhou, Wuhan,Xi'an,Zhengzhou,Qingdao, Changsha,Tianjin,Suzhou,Nanjing, Dongguan,Shenyang,Hefei,Foshan	New first tier cities	33%
Kunming,Fuzhou,Wuxi, Xiamen,Harbin,Changchun,Nanchang,Jinan, Ningbo,Dalian,Guiyang, Wenzhou,Shijiazhuang,Quanzhou,Nanning, Jinhua,Changzhou,Zhuhai,Huizhou, Jiaxing,Nantong,Zhongshan,Baoding,Lanzhou, Taizhou, Xuzhou,Taiyuan, Shaoxing,Yantai, Langfang	Second tier cities	23%
Others	Others	7%

DOI: 10.54254/2755-2721/13/20230727

Table 3. Collected data (only some are listed).

Number	Eigenvector Centrality	Gender	Age Group	City Group
6	0.08845	female	Youth	Second tier cities
25	0.10134	female	Youth	New first tier cities
137	0.07793	female	Middle age	First-tier cities
487	0.00684	male	Juvenile	Second tier cities
526	0.11146	female	Juvenile	First-tier cities
672	0.24692	female	Youth	First-tier cities
896	0.2524	female	Youth	New first tier cities
987	0.11065	female	Youth	New first tier cities
998	0.03866	male	Juvenile	First-tier cities
1072	0.26966	female	Youth	New first tier cities
1098	0.19322	male	Middle age	New first tier cities
1243	0.02875	female	Middle age	Others
1289	0.12129	male	Juvenile	First-tier cities
1762	0.05858	male	Juvenile	Others
1972	0.10742	male	Youth	First-tier cities
2054	0.00929	female	Juvenile	Second tier cities
2178	0.11994	male	Juvenile	New first tier cities
2263	0.29106	female	Youth	First-tier cities
2298	0.2152	male	Juvenile	First-tier cities
1350	0.00755	female	Old age	Second tier cities

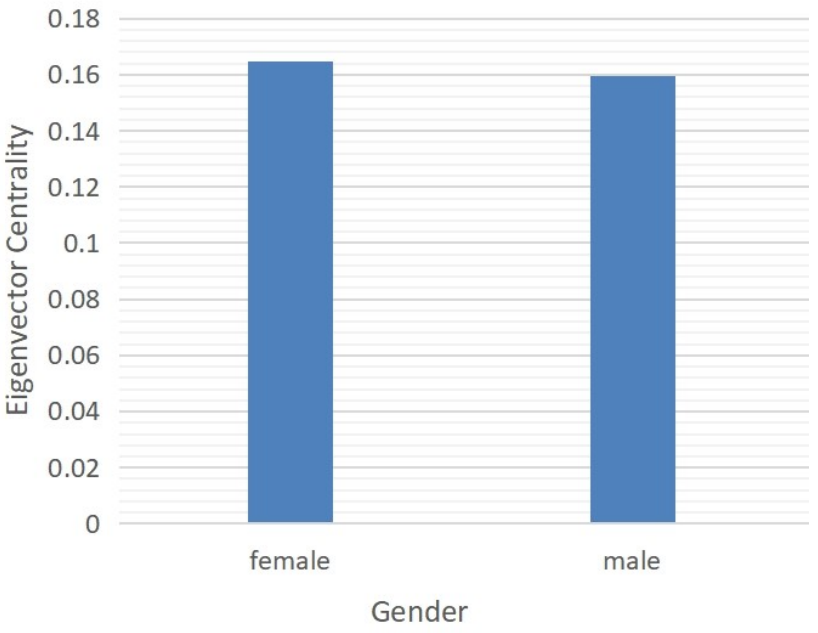

Figure 1. The influence of gender on Eigenvector Centrality.

DOI: 10.54254/2755-2721/13/20230727

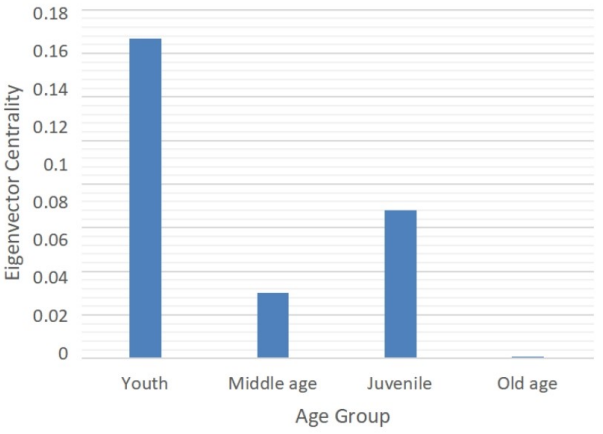

Figure 2. The influence of age on Eigenvector Centrality.

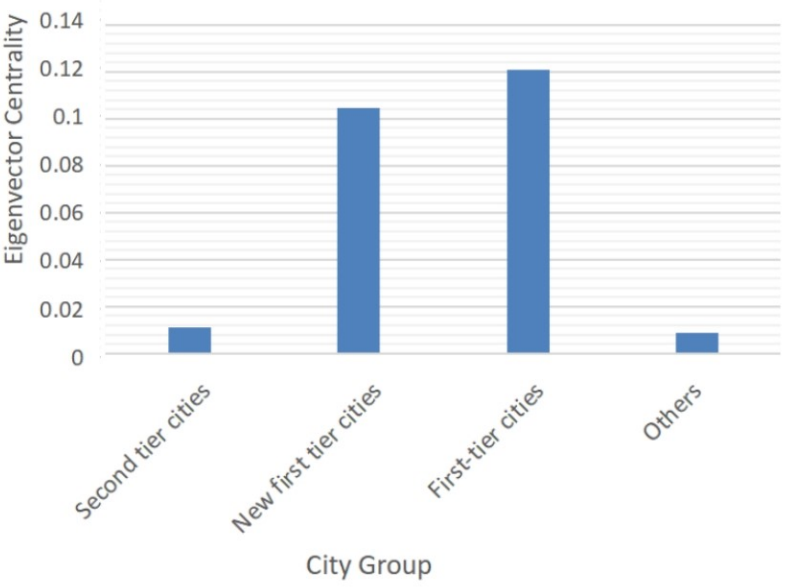

Figure 3. The influence of city on Eigenvector Centrality.

According to Figure 1, this paper finds that gender has a relatively small impact on influence, in which women's influence is slightly higher than men's. However, there is a large gap in the impact of age on influence, and the data of the youth group is significantly larger than that of the other three groups, according to Figure 2. Cities also have a certain impact on the influence. The data of first tier cities and new first tier cities are slightly larger than those of second tier cities and other cities, according to Figure 3.

5. Discussion

It can be seen from the results of the study that the influence gap between different users is large. The data on women is slightly larger than that of men, indicating that women have a little more influence. Perhaps the information transmitted on microblogs is more symbolic of women's taste.

The data of young users is obviously large, which is also consistent with reality. Because microblogging has the characteristics of convenient information, cross media, and fragmented communication, which is consistent with the needs of young people for speed, innovation, lively, diverse forms, and strong interaction, it is welcomed by young people. In reality, the people who use microblogging most are also young people.

DOI: 10.54254/2755-2721/13/20230727

The regional differences are also obvious, and each place has its own characteristics. Among them, the first tier cities and new first tier cities have faster economic development than other cities, rich commercial resources, strong urban population activity, and diverse lifestyles, so there are relatively more news and people's demand for social networking on the Internet will be higher. In some inland areas, the traffic is relatively blocked, the population is small, or the city is civilized by natural landscape users' influence will be slightly less.

The classification of users plays an important role in studying the influence of users. Select different types of users when spreading messages, so that you can suit the remedy to the case and achieve better communication results.

6. Conclusions

The research aims to analyze the influence of microblog users. Based on a quantitative analysis of Eigenvector Centrality corresponding to every user. It can be concluded that the influence gap between different users is large. Gender, age, and city are all important factors when analyzing the influence.

Although this research method can easily analyze the influence of each user and screen out the users with greater influence, the selected data has a greater impact on the results. If the selected data are limited to such minor aspects as entertainment or current affairs news, the results will be biased. Although this paper selects a variety of data when obtaining data, there are still limitations. In the research, researchers should also consider the difference between the user's city information on the microblog and the actual city information, or the difference between the age information on the microblog and the real information, so the researchers should collect as much data as possible.

To better understand the implications of these results, future studies should be grouped the users in a more optimized way, and negative impacts among users should also be considered at the same time to make the results more accurate. This paper uses a simple way to analyze the influential groups in microblogging, and draws a conclusion, which is conducive to further research on this basis.

References

[1] Weng J, Lim E P, Jiang J, et al. Twitterrank: finding topic-sensitive influential Twitterers[C]//Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010: 261-270.

[2] Bakshy E, Hofman J M, Mason W A, et al. Everyone's an influencer: quantifying influence on Twitter[C]//Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011: 65-74.

[3] Su C, Du Y, Guan X, et al. Maximizing topic propagation driven by multiple user nodes in micro Blogging[C]//38th Annual IEEE Conference on Local Computer Networks. IEEE, 2013: 751-754.

[4] Mao J, Liu Y, Zhang M, et al. Social Influence Analysis for Micro-Blog User Based on User Behavior [J]. Chinese Journal of Computers, 2014,37(4): 791-800.

[5] Zhuang K, Shen H, Zhang H.User spread influence measurement in microblog[J] Multimedia Tools and Applications.2016: 1-17.

[6] Bartoletti M, Lande S, Massa A. Faderank: an incremental algorithm for ranking Twitter users[C]//International Conference on Web Information Systems Engineering. Springer, Cham, 2016: 55-69

[7] Yuan Y, Li C, Tian L. Refinement and Application of Microblog-oriented PageRank Algorithm [J]. Computer Applications and Software, 2017, 4(3):31- 37.

[8] Wu L, Yang B, Jian M, et al. MPPR: Multi Perspective Page Rank for User Influence Estimation[C]//2018 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE, 2018: 25-29.

[9] Sun H, Zuo T. Research on PageRank-based Algorithm of Micro-Blog User Influence [J]. Application Research of Computers, 2018, 35(4):1028-1032

DOI: 10.54254/2755-2721/13/20230727

[10] RUI X，YANG X，FAN J，et a1．A neighborhood scale fixed app1．oach for influence maximization in social networks[J].Computing,2020,l02(2):427—449
[11] HUANG H M, SHEN H, MENG Z Q.Community—based influence maximization in attributed network’s[J].Applied Intelligence,2020,50(2):354—364