Research on clustering method of government micro-blogging user group based on user preference

Tianhao Guo¹, Suozhu Wang*¹

¹School of Management, Capital Normal University, Beijing, 100048, China
*Corresponding author’s e-mail: wsz_wsz@163.com

Abstract. The goal of government micro-blogging is to provide the public with faster government services. Therefore, it is of great significance to study the characteristics of users and cluster analysis to provide personalized services and microblog operation management. A user preference measurement method based on the usage behavior data of users’ retweeting, commenting, and praising on the government micro-blogging platform is proposed and a user group fuzzy clustering method based on user preferences is given in this paper. An example is given to illustrate the effectiveness and feasibility of the method proposed in this paper.

1. Introduction

Government micro-blogging provides the public with new channels and windows for more convenient and quick access to government information and social dynamics. Government micro-blogging is a kind of functional microblog, which has the characteristics of public service. Serving the public is the starting point and ending point of the opening and operation of government micro-blogging, and the public is the beneficiary of the service. Users on the same government micro-blogging have different interests and behavior characteristics, which can be reflected through users' interactive behaviors such as following and commenting on the government micro-blogging platform. Through dividing user groups with similar behavior preferences, we can find the user's generic model, and the clustering results can provide decision-making basis for the public to provide personalized services and the operation and management of micro-blog. Therefore, how to use data mining technology to discover users' wishes and interests has become a hot issue in the research of government microblogging applications.

From the perspective of foreign research, there is no special term about government microblog in foreign countries. Twitter, Facebook and other social media are regarded as research objects by many studies to study the use of Weibo by government officials and the role of Weibo in government management. For example, Hong Yong Yoon[1] and others believe that the “following” and “mention” of government micro-blogging are the main ways for politicians to get in touch with other politicians, so as to gain more political support. Alhabas[2] hand others believe that marketers use indicators such as information dissemination, emotional assessment, and message review to study people's behavior on social networking sites, like Twitter. John Carlo Bertot[3] and others believe that the cooperation, participation, autonomy and timeliness of government micro-blogging can help improve the openness, transparency and anti-corruption ability of the government. In recent years, the research on government micro-blogging has also made a lot of achievements at home. J. Chen, etc.[4] have combed the research achievements at home and abroad in recent years, and believed that the domestic and foreign research mainly concentrated on the fields of “the concept, application and communication
characteristics of government micro-blogging”. There is a lack of quantitative research results based on a large amount of objective data generated by government micro-blogging activities. Research on the analysis and clustering of user behavior characteristics of government micro-blogging has not yet seen the corresponding research results. Therefore, how to use the data generated in the application of government micro-blogging, and use data mining technology and other methods to study related issues in the field of government micro-blogging, is a research field worth exploring.

Based on the above analysis, this paper based on the data generated during the interaction between users and government micro-blogging to study the measurement method of user preferences. Based on this, a fuzzy clustering method of user groups based on user preferences is proposed, and the specific steps and clustering process of this method are illustrated by an example. The purpose of this study is to provide a new idea and method for the analysis of user groups for the operation and management personnel of government micro-blogging.

Above all, a user group fuzzy clustering method based on user preferences is proposed in this paper. This method is based on the data generated during the interaction between users and government micro-blogging. An example is used to illustrate the specific steps and clustering process of the method. The results of this study are intended to provide a new way of thinking and method for the analysis and user group analysis of government microblog operations and management personnel.

2. User interaction behavior and preference

2.1. User interaction behavior

In the government micro-blogging, the interaction between users and government micro-blogging includes follow, comment, retweet and praise the content of micro-blogs. Users who follow government micro-blogging become fans, and can receive the information published by government micro-blogging at any time. In the process of browsing and reading the government micro-blogging, users will retweet the content which they are interested in, and at the same time, they can also make comments to spread and share the government information. Among the above interactive behaviors, retweet and comment are the key to the dissemination of government information. The interaction between users and government micro-blogging can be described as follows.

Define 1 user interaction behavior. Let U denote a user, and GW denote a government micro-blogging, then user U and GW have an interaction from U to GW, which is denoted as U→GW. When U and GW have one of the following behaviors.

1. U follow GW;
2. U comment on a micro-blog published by GW;
3. U retweet a micro-blog published by GW;
4. U praise a micro-blog published by GW;
5. U send a private message to GW.

It can be seen from the definition that different types of interaction may occur between a government micro-blogging user U and GW, and the frequency of interaction reflects the user's activeness on government micro-blogging. The more interactions, the more active the user is. On the same government micro-blogging platform, the frequency of interaction between different users U and GW is obviously different.

2.2. User preference and measurement

According to economics, preference refers to the degree of consumers' preference for a commodity (or a combination of commodities). Consumers rank the commodities or commodity combinations available for consumption according to their own wishes, which reflects the needs, interests and hobbies of consumers. Similarly, on the government micro-blogging platform, users will selectively praise, retweet, and comment on micro-blogging according to their own wishes and interests. Users will forward the micro-blogs they are interested in to their fans, but only browse or skip the content they are not interested in. Therefore, the concept of preference in economics can be used to measure
the interest of government micro-blogging users.

Let GW represent a government micro-blogging platform, and use \( T = \{ t_1, t_2, ..., t_j, ..., t_m \} \) to denote the collection of all micro-blog topics published by GW in a statistical period, and use \( U = \{ u_1, u_2, ..., u_i, ..., u_n \} \) to represent the collection of all users.

Define 2 user retweet preference. In a certain period, for any user \( u_i \in U \) and topic \( t_j \in T \), the number of times the user \( u_i \) retweet the topic \( t_j \) is recorded as \( f_{t,k,j} \), then the preference of user \( u_i \) for retweeting the topic \( t_j \) can be defined as follows:

\[
FP_{u_i} = f_{t,k,j} / (\sum_{k=1}^{n} f_{t,k,j}) / m \quad (i = 1, 2, ..., n; j = 1, 2, ..., m) \quad (1)
\]

Where \( f_{t,k,j} \) is the number of times users \( u_k \) retweeted the topic \( t_j \), \( m \) denotes the total number of micro-blogs.

Retweet preference represents the ratio of the number of times a user retweets a topic to the average number of retweets on the topic. It depicts the user’s preference for retweet on micro-blog on a certain topic. The larger the value, the more interested the user is in the topic.

Define 3 user comment preference. In a certain period, for any user \( u_i \in U \) and topic \( t_j \in T \), the number of times the user \( u_i \) comments the topic \( t_j \) is recorded as \( c_{t,k,j} \), then the preference of user \( u_i \) for commenting on the topic can be defined as follows:

\[
CP_{u_i} = c_{t,k,j} / (\sum_{k=1}^{n} c_{t,k,j}) / m \quad (i = 1, 2, ..., n; j = 1, 2, ..., m) \quad (2)
\]

Where \( c_{t,k,j} \) is the number of times users \( u_k \) commented the topic \( t_j \), \( m \) denotes the total number of micro-blogs.

Comment preference indicates the ratio of the number of times a user comments a topic to the average number of comments on the topic. It depicts the user's preference for comment on micro-blog on a certain topic. The larger the value, the more interested the user is in the topic.

Define 4 user praising preferences. In a certain period, for any user \( u_i \in U \) and topic \( t_j \in T \), the number of times the user \( u_i \) praises the topic \( t_j \) is recorded as \( l_{t,k,j} \), then the preference of user \( u_i \) for prasing on the topic can be defined as follows:

\[
LP_{u_i} = l_{t,k,j} / (\sum_{k=1}^{n} l_{t,k,j}) / m \quad (i = 1, 2, ..., n; j = 1, 2, ..., m) \quad (3)
\]

Where \( l_{t,k,j} \) is the number of times users \( u_k \) praised the topic \( t_j \), \( m \) denotes the total number of micro-blogs.

Praising preference indicates the ratio of the number of times a user praises a topic to the average number of praises on the topic. It depicts the user's preference for praising on micro-blog on a certain topic. The larger the value, the more interested the user is in the topic.

Define 5 user preference degree. In a certain period, for any user \( u_i \in U \) and topic \( t_j \in T \), the preference degree of user \( u_i \) for the topic \( t_j \) can be defined as follows:

\[
TP_{u_i} = f_{t,k,j} \cdot c_{t,k,j} \cdot l_{t,k,j} / (\sum_{k=1}^{n} f_{t,k,j} \sum_{k=1}^{n} c_{t,k,j} \sum_{k=1}^{n} l_{t,k,j}) / m^3 \quad (i = 1, 2, ..., n; j = 1, 2, ..., m) \quad (4)
\]

Where \( f_{t,k,j}, c_{t,k,j}, l_{t,k,j} \) have the same meanings as (1), (2), (3), \( m \) denotes the total number of micro-blogs, and \( n \) is the total number of users.

The user preference indicates the overall preference behavior of a user in the three aspects: retweet, comment, and praise. The larger the value, the more the user’s preference for a topic; otherwise, the smaller the value, the less interest the user has on the topic.
Define 6 user preference degree matrix. User preference degree matrix is defined as \((t_{ij})_{n \times m} (i=1,2,...,n; j=1,2,...,m)\), where \(t_{ij}\) is the preference of user \(u_i\) for the topic \(t_j\), it can be calculated by formula (4).

3. User group clustering method

3.1. User behavior data collection

In the process of interaction between users and government micro-blogging, the micro-blogging platform records the behavior of different users, such as retweet and comment, which can be collected by network crawler tools or self-edited web crawlers. According to the definition of user preferences, it is necessary to collect the relevant data of the user's retweets, comments and praises for each micro-blog topic. The specific data format is “user identity, the micro-blog ID, number of retweets, number of comments and praises”. Once the data collection is complete, the data will be preprocessed into user-topic retweet behavior data table, user-topic comment behavior data table and user-topic praise behavior data table. These three data tables are the basis for subsequent cluster analysis.

3.2. Clustering method

Clustering is one of the common methods of data mining analysis. A cluster is a group of physical or abstract objects that are divided into classes based on how similar they are, where similar objects form a class\(^5\). A class is a set of objects that are similar to each other, and objects in different classes are different. The description of similarity between objects is based on the value of the object's properties, which is usually represented by the distance (between objects).

There are many commonly used clustering algorithms, including division-based clustering methods, hierarchical-based clustering methods, density-based clustering methods, model-based clustering methods and grid-based clustering methods. However, these methods have many shortcomings in dealing with large-scale, high-dimensional, fuzzy and dynamic data, such as complex calculation process, low time efficiency and so on. Through comparative analysis, this paper adopts the method of fuzzy clustering analysis, which can eliminate the repeated calculation process of multiple iterations, greatly reduce the calculation amount, and greatly improve the time efficiency. At the same time, according to the relevant data in the database to construct fuzzy similar matrix, directly process the similar matrix without repeatedly scanning the database, which greatly improves the calculation efficiency. In addition, this method has good scalability and makes it easy to find isolated points.

In this article, preference is used to measure users' interest in the micro-blog topic. Firstly, according to the user-topic retweet behavior data table, user-topic comment behavior data table and user-topic praise behavior data table, the user preference degree matrix in definition 6 is calculated on the basis of the above formula 4. Then use the fuzzy clustering method to cluster. There are many methods to calculate the statistics of the similarity degree among the classified objects in the fuzzy clustering method, such as Euclidean distance method, quantity product method, geometric average minimum method, arithmetic average minimum method, angle cosine method etc. In calculating the degree of similarity between objects, the angle cosine method are used in this paper, which is

\[
 r_{ij} = \frac{\sum_{k=1}^{m} a_{ik} a_{jk}}{\sum_{k=1}^{m} a_{ik}^2 \sum_{k=1}^{m} a_{jk}^2} (i=1,2,...,n; j=1,2,...,m)
\]  

(5)

Assume that \(U\) is a collection of users to be classified, the specific steps of the clustering method are as follows.

Step 1: On the government micro-blogging GW, the user's interactive behavior data is collected using the crawler tool, including the number of users' topic retweets, comments and praises, and the data is pre-processed to form a user-topic retweet behavior data table, user-topic comment behavior data table, user-topic likes data table;

Step 2: According to the above three data sheets, use the formula (1) (2), (3) and (4) calculate the value of the preference degree of each user \(u_i\), to form the user preference degree matrix in definition 6;
Step 3: Based on the user preference degree matrix, adopt the angle cosine similarity measurement method, that is, use formula (5) to calculate the similarity $r_{ij}$ between $u_i$ and $u_j$, so as to construct the fuzzy similarity matrix $R^F = [r_{ij}]_{m \times m} (i = 1, 2, \ldots, n; j = 1, 2, \ldots, m)$;

Step 4: If the fuzzy similar matrix $R^F$ is a fuzzy classification relationship, cluster analysis can be carried out directly, otherwise, go to the next step;

Step 5: Transform the fuzzy similarity matrix $R^F$ into a fuzzy classification relation by cyclic self-multiplication of the fuzzy similarity matrix, that is: $R \circ R = R^2$, $R^2 \circ R^2 = R^4$, until it is met $R^{2k} = R^k$, then $R^k$ is a fuzzy classification relation;

Step 6: According to the fuzzy classification relationship $R^k$, select $\lambda \in [0, 1]$, the fuzzy cluster analysis can be carried out by $\lambda$ cut-matrix. If the bigger the $\lambda$, the more accurate the classification, otherwise the more inaccurate the classification.

Step 7: Output clustering mode.

3.3. Example and analysis

In order to verify the feasibility of the above clustering analysis method, taking a local government microblogging as an example to illustrate the specific cluster analysis process in this section. In a certain period of time, the government microblogging published 10 themes, a total of 200 microblogs, collected relevant data of 20 users by using the crawler tool, and compiled into user-topic retweet behavior data table (Table 1), user-topic comment behavior data table (Table 2), and user-topic praise behavior data table (Table 3) after preprocessing.

| Table 1. User-topic retweet behavior data table. |
|------------------------------------------------|
| $U_1$ | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ | $T_6$ | $T_7$ | $T_8$ | $T_9$ | $T_{10}$ |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|------|
|       | 1     | 2     | 2     | 4     | 4     | 4     | 10    | 10    | 10    | 10   |
| $U_2$ | 7     | 4     | 7     | 4     | 9     | 1     | 4     | 6     | 10    | 0    |
| $U_3$ | 5     | 8     | 10    | 9     | 4     | 0     | 2     | 2     | 9     | 3    |
| $U_4$ | 6     | 4     | 7     | 7     | 5     | 2     | 8     | 0     | 2     | 1    |
| $U_5$ | 10    | 4     | 8     | 0     | 6     | 0     | 2     | 5     | 2     | 16   |
| $U_6$ | 1     | 7     | 7     | 1     | 2     | 7     | 6     | 6     | 7     | 1    |
| $U_7$ | 0     | 8     | 3     | 8     | 10    | 3     | 5     | 10    | 9     | 10   |
| $U_8$ | 4     | 0     | 9     | 9     | 8     | 8     | 1     | 6     | 12    | 0    |
| $U_9$ | 0     | 0     | 3     | 7     | 3     | 5     | 10    | 4     | 1     | 0    |
| $U_{10}$ | 1    | 2     | 10    | 1     | 8     | 8     | 1     | 9     | 9     | 2    |
| $U_{11}$ | 8   | 9     | 0     | 7     | 4     | 6     | 9     | 1     | 6     | 7    |
| $U_{12}$ | 8   | 5     | 7     | 9     | 3     | 2     | 10    | 2     | 1     | 10   |
| $U_{13}$ | 8   | 6     | 8     | 10    | 5     | 10    | 10    | 9     | 2     | 3    |
| $U_{14}$ | 7   | 9     | 10    | 0     | 6     | 10    | 8     | 9     | 3     | 6    |
| $U_{15}$ | 2   | 0     | 1     | 4     | 5     | 7     | 0     | 1     | 1     | 8    |
| $U_{16}$ | 8   | 3     | 9     | 0     | 4     | 6     | 3     | 8     | 8     | 7    |
| $U_{17}$ | 5   | 11    | 10    | 3     | 7     | 2     | 7     | 8     | 5     | 7    |
| $U_{18}$ | 5   | 8     | 3     | 10    | 6     | 8     | 10    | 3     | 5     | 9    |
| $U_{19}$ | 0   | 7     | 5     | 10    | 0     | 3     | 2     | 9     | 4     | 8    |
| $U_{20}$ | 7   | 7     | 18    | 9     | 10    | 2     | 2     | 7     | 8     | 9    |

| Table 2. User-topic comment behavior data table. |
|------------------------------------------------|
| $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ | $T_6$ | $T_7$ | $T_8$ | $T_9$ | $T_{10}$ |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|------|
| $U_1$ | 3     | 11    | 2     | 4     | 15    | 1     | 7     | 14    | 3    |
| $U_2$ | 5     | 8     | 6     | 15    | 8     | 4     | 1     | 6     | 6    |
|       |       |       |       |       |       |       |       |       | 12   |
According to the user interaction behavior data in Table 1, Table 2, and Table 3, the user preference degree matrix can be calculated according to step 2 above. The results are as follows.

| User | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | T10 |
|------|----|----|----|----|----|----|----|----|----|-----|
| U1   | 8  | 1  | 2  | 9  | 8  | 5  | 0  | 2  | 16 | 16  |
| U2   | 3  | 0  | 9  | 20 | 17 | 3  | 8  | 8  | 1  | 14  |
| U3   | 17 | 19 | 4  | 15 | 12 | 6  | 2  | 0  | 8  | 3   |
| U4   | 4  | 8  | 11 | 17 | 9  | 19 | 16 | 8  | 3  | 7   |
| U5   | 12 | 20 | 5  | 16 | 4  | 18 | 15 | 19 | 13 | 3   |
| U6   | 3  | 11 | 16 | 14 | 7  | 5  | 12 | 14 | 9  | 2   |
| U7   | 18 | 5  | 6  | 6  | 18 | 19 | 17 | 16 | 5  | 2   |
| U8   | 7  | 20 | 8  | 15 | 20 | 14 | 10 | 20 | 13 | 0   |
| U9   | 10 | 1  | 17 | 1  | 10 | 19 | 6  | 10 | 20 | 13  |
| U10  | 1  | 8  | 12 | 5  | 9  | 12 | 11 | 4  | 1  | 7   |
| U11  | 19 | 1  | 16 | 4  | 4  | 13 | 19 | 20 | 9  | 20  |
| U12  | 20 | 1  | 7  | 9  | 3  | 5  | 5  | 3  | 6  | 3   |
| U13  | 19 | 11 | 0  | 19 | 13 | 10 | 13 | 11 | 7  | 9   |
| U14  | 2  | 2  | 2  | 6  | 13 | 0  | 7  | 14 | 13 | 6   |
| U15  | 7  | 15 | 8  | 6  | 10 | 1  | 15 | 17 | 19 | 18  |
| U16  | 0  | 9  | 2  | 10 | 0  | 12 | 18 | 15 | 11 | 0   |
| U17  | 5  | 2  | 15 | 14 | 12 | 6  | 19 | 11 | 10 | 15  |
| U18  | 17 | 20 | 5  | 15 | 16 | 13 | 13 | 2  | 0  | 16  |
| U19  | 16 | 0  | 6  | 6  | 20 | 20 | 5  | 16 | 1  | 11  |
| U20  | 0  | 8  | 15 | 0  | 8  | 15 | 10 | 18 | 1  | 12  |
After circular self-multiplication, the fuzzy classification relationship can be obtained as follows.
microblog user groups. Repost, comment and praise are the three indexes of the model. A new method for analyzing user groups for the operation and management of government microblogging has also increased, so it is more important to understand the preferences of government microblogging users. Based on the data of the interaction behavior between government microblogging and users, a user preference measurement model and its clustering method are proposed in this paper. This method provides a new method for analyzing user groups for the operation and management of government microblogs. Repost, comment and praise are the three indexes of the model. The angle cosine similarity method is used to perform cluster analysis on user preferences, and divide government microblog user groups. However, user behaviors such as "Focus" and "Private Messaging" have not been taken into consideration in this paper, so adding more user behaviors to improve the accuracy of

According to the fuzzy classification relation $R^k$, we can select different confidence levels $\lambda$ ($\lambda \in [0,1]$), according to the actual needs, and carry out cluster analysis. As shown in Figure 1.

![Figure 1. User preference dynamic clustering graph.](image)

4. Conclusions and future work

With the expansion of micro-blog user groups, the influence of government micro-blogging has also increased, so it is more important to understand the preferences of government micro-blogging users. Based on the data of the interaction behavior between government micro-blogging and users, a user preference measurement model and its clustering method are proposed in this paper. This method provides a new method for analyzing user groups for the operation and management of government microblogs. Repost, comment and praise are the three indexes of the model. The angle cosine similarity method is used to perform cluster analysis on user preferences, and divide government microblog user groups. However, user behaviors such as "Focus" and "Private Messaging" have not been taken into consideration in this paper, so adding more user behaviors to improve the accuracy of
user preference metrics will become future work. With the development of data mining technology and Internet information extraction technology, it will provide more methods and tools for government micro-blogging operation and management personnel to analyze user groups, and we will also build a new clustering model to analyze the user group of government micro-blogging.

Acknowledgements
This research was supported by National Natural Science Foundation of China (grant No. 61371177), and the Scientific Research Foundation of Beijing Municipal Education Commission (grant No.KM201810028021).

References
[1] Yoon, H. Y ., Park, H. W.. (2014). Strategies affecting twitter-based networking pattern of south korean politicians: social network analysis and exponential random graph model. J. quality & quantity, 48(1), 409-423.
[2] Riedl, C. , Koebler, F. , Goswami, S. , Krcmar, H. . (2013). Tweeting to feel connected: a model for social connectedness in online social networks. J. International Journal of Human Computer Interaction, 10(29), 670-687.
[3] Bertot, J. C. , Jaeger, P. T. , Hansen, D. (2012) The impact of polices on government social media usage: Issues, challenges, and recommendations. J. Government Information Quarterly., (29):30-40.
[4] Chen, J. , Yuan, Q.J. (2014) A Review of Administrative Microblogging Study in China and Abroad. J. Information Science., 32(6):156-161.
[5] Qiu, Z.C. , Shen, H. (2017) User clustering in a dynamic social network topic model for short text streams. J. Information Sciences., 414: 102-116.