Research on Clustering Method for Government Micro-blogging User Segments Based on User Interaction Behavior

Suozhu Wang[1, a], Jun Wang[2, b]

[1,2] School of Management, Capital Normal University, Beijing, 100089, China

[a] wsz_wsz@163.com, [b] wangjuncnu@163.com

Keywords: Government micro-blogging, Interaction behavior, User segments clustering

Abstract. One purpose of government micro-blogging is to provide services for the public. Therefore, it is of great significance to study the users' characteristics and clustering user group for providing personalization services and micro-blogging's operation and management. In this paper, a measuring interaction strength method of user and micro-blogging is presented by mean of user behavior features such as retweet, comment and so on in the same government micro-blogging, and a fuzzy clustering method for user segments is put forwarded based on user interaction behavior strength. An example demonstrates effectiveness and feasibility of the method.

Introduction

Government micro-blogging refers to micro-blogging that it is opened by the party and government organizations or party and government officials through real name authentication. Its purpose is to release government information, promote the disclosure of government information, strengthen the communication between officials and the people, shape new government, and strengthen public services. With the rapid development of government micro-blogging, it has become a significant way for government providing public services, and a new access to social dynamics and government affair information for the public. Government micro-blogging, one of the functional micro-blogging, is user services oriented. Government micro-blogging's starting point and destination is serving for the public who will benefit a lot from such service. Users have differentiated interests and behavioral characteristics in the same government micro-blogging, which can be reflected by the user's retweet, comment and other interactive behavior on micro-blogging [1]. The user's generic pattern can be discovered by dividing the user groups with similar behavior preferences. Clustering results can provide decision support for offering personalization services to the public as well as micro-blog's operation and management.

There is no term for government micro-blogging in abroad, a large number of researches focus on the issues, such as government officers using micro-blog, its influence on government management, etc., based on the study about Twitter, Facebook and other social media. For example, Ho Young Yoon, etc. [2] hold that following and mention on micro-blogging are main approaches for politicians linking each other in order to gaining more political support. Alhabash, etc. [3] hold that marketers use index, such as information dissemination, sentiment assessments, information reviews and etc., to study people's behavior on social network websites. John Carlo Bertot, etc.[4] consider that the collaboration, participation, autonomy and timeliness of the government micro-blogging will help to enhance governmental openness, transparency and anti-corruption capability. In recent years, study about government micro-blogging has obtained a great quantity of achievements at home. J. Chen, etc.[5] summarized research results at home and abroad recent years, claim that domestic and foreign researches mainly concentrate on the issues, such as government micro-blogging concept, application, propagation, etc., while quantitative study on a large number of objective data produced by government micro-blog activities is not quite enough. By far, research findings in government micro-blogging user behavior characteristics analysis and clustering have not been seen yet. Therefore, it is worth discovering that how to use the data produced by the government micro-blogging activity and adopt data mining methods to study related issues about government micro-blog.

Above all, this paper will propose a user interaction behavior characteristics' measure method based on the data generated by user interaction with government micro-blog, and user group fuzzy clustering method depend on user interaction behavior according to the method is put forwarded in this paper. Specific steps and clustering process are illustrated by experiment. The research result is aimed at providing a new idea and method for government micro-blogging operator and manager, when they analyze user groups.

User Interaction Behavior and Strength

User interaction behavior

Spreading accurate information issued by the government as fast as possible is one of the intentions of government micro-blogging operating. The followers of government micro-blogging will find out what they are interested in and retweet and comment simultaneously, which make official information shared and distributed, when they are browsing and reading those micro-blogs. Through the direct analysis on government micro-blogging, interaction behaviors between users and government micro-blog mainly includes: follow, retweet, comment, mention (@) and etc. All of these inter-behaviors are essential for information dissemination. Therefore, user interaction behavior can be defined as follows.

Definition 1 *Interaction behavior*. Let *U* be one user of the government micro-blogging, *GW* stands for one of the government micro-blogs. An interaction behavior from a user to government micro-blog can be expressed as $U \to GW$, when one of following activity occurs

(1) *U* comment on a micro-blog posted by *GW*.

(2) *U* retweet a micro-blog posted by *GW*.

(3) *U* post a micro-blog mentioned (@) *GW*.

Through the definition above, a variety of interaction behaviors can take place between *U* and *GW*, and the number of interaction behaviors shows user's activity on micro-blog, the more numbers of interaction behaviors the user more activity. Obviously, the interaction strength will go to differ due to different users and *GW* on the same government micro-blogging. To measure a user's interaction strength, this paper intends to present user interaction strength by defining user activity vector. Activity means the users' influence on a micro-blogging. The more user activity is, the more likely to engage in the spread of hot topics and attract other users' attention.

User interaction strength

Suppose that *GW* is a government micro-blogging, let $W = \{w_1, w_2, ..., w_j, ..., w_n\}$ be a set of all *GW* issued micro-blogs within a period time, $U = \{u_1, u_2, ..., w_i, ..., u_m\}$ be a set of all users.

Definition 2 *Retweet activity degree*. Let $W_{u_i} (W_{u_i} \subseteq W)$ be the set of micro-blogs retweeted by the user u_i within a period time, retweet activity degree is defined as

$$FA_{u_i} = \sum_{w_j \in W_{u_i}} ft_j / ((\sum_{t=1}^{n} ft_t) / n) \qquad (i = 1, 2, ..., m) \tag{1}$$

Where ft_t denotes the number of micro-blogs w_t retweeted, n denote the total number of micro-blogs.

Retweet activity degree indicates the ratio of the number of micro-blog retweeted by a certain user to the average number of all retweeted micro-blogs, which describes the active degree of a user retweet behavior on a government micro-blog, the bigger the value, the user is more active.

Definition 3 *Comment activity degree.* In a certain period of time, let $W_{u_i} (W_{u_i} \subseteq W)$ represents a set of micro-blogs commented by the user u_i , then comment activity degree is defined as

$$CA_{u_i} = \sum_{w_j \in W_{u_i}} ct_j / ((\sum_{t=1}^{n} ct_t) / n) \qquad (i = 1, 2, ..., m) \tag{2}$$

Where ct_t denotes the number of the micro-blogs *t* commented, n denotes the total number of micro-blogs.

Comment activity degree indicates the ratio of the number of micro-blogs commented by a certain user to the average number of all commented micro-blogs, which describes the active degree of a user comment behavior on a government micro-blogging, the bigger the value, the user is more active.

Definition 4 *Mention (@) activity degree*. In a certain period of time, let n_{u_i} represents the number of *GW* mentioned (@) by the user u_i, then mention (@) activity degree is defined as

$$A_{u_i} = n_{u_i} / (N / m) \qquad (i = 1, 2, \ldots, m) \tag{3}$$

where *N* denotes the number of *GW* mentioned (@) by all users, *m* denotes the total number of users.

Mention(@) activity degree indicates the ratio of the number of micro-blogs mentioned(@) by a certain user to the average number of mentioned(@) by all users, which describes the active degree of a user mention(@) behavior on a government micro-blogging, the bigger the value, the user is more active.

Definition 5 *Interaction behavior strength*. In a certain period of time, interaction behavior strength between user u_i and *GW* is represented by a vector $IB_{u_i \to GW} = (FA_{u_i}, CA_{u_i}, A_{u_i})$, the norm of vector is $| IB_{u_i \to GW} |$, which represents the interaction strength between u_i and *GW*, where $FA_{u_i}, CA_{u_i}, A_{u_i}$ calculated by Eq.(1), Eq.(2) and Eq.(3) separately.

Interaction strength vector denotes the overall behaviors of a certain user in retweet, comment, and mention (@). The bigger the norm of vector, the greater the interaction strength between user and government micro-blogging. The user interaction behavior strength matrix can be given by using the vector as the row of matrix.

Definition 6 *Interaction behavior strength matrix*. After data standardization processing for $FA_{u_i}, CA_{u_i}, A_{u_i}$, $UIB = [FA_{u_i}, CA_{u_i}, A_{u_i}]_{m \times 3}$ *(i=1, 2,..., m)* is said to be user interaction behavior strength matrix.

User group clustering method

Clustering method

Clustering is that a group of physical or abstract objects are divided into several classes by the similarity degree between them, and similar objects constitute one class [6]. A class is a set consisted of objects that are similar to each other, while objects are heterogeneous among different classes. Similarity degree can be measured by object attribute data.

There are many commonly clustering algorithms in field of data mining. They are partition method, hierarchical clustering method, method based on density, and so on. However, these methods have some disadvantages, especially in dealing with large-scale, high dimension, fuzzy, dynamic data.

In practice, the words (high, low, general, etc.) that describe the degree of user activity are ambiguous. Therefore, in this paper, cluster user groups will use fuzzy clustering method. Fuzzy clustering analysis is generally refers to construct fuzzy matrix based on the properties of the research subject, and on this basis, according to the degree of membership to determine the cluster relationship. Fuzzy mathematics method was used to determine the fuzzy quantitative relation between samples, objectively and accurately for cluster

Assume that $W = \{wid_1, wid_2, \ldots, wid_j, \ldots, wid_n\}$ (wid_i is micro-blog's ID) be a set of a government micro-blogging issued all micro-blogs, $U = \{u_1, u_2, \ldots, w_i, \ldots, u_m\}$ be all users of *GW*, and represents objects to be classified. Specific steps of clustering analysis method are as follows:

Step 1 On a government micro-blogging, collect users' interaction behavior data, including retweet, comment, mention (@), by using scraping tool, and the data is processed by statistics and preprocessing to form user interaction behavior data table.

Step 2 Calculate the strength vector of each user's interaction behavior by Eq.(1), Eq.(2) and Eq.(3), according to user interaction behavior data table, and then standardize these data to form the user interaction behavior strength matrix *UIB* defined in definition 6.

Step 3 Calculate the similarity S_{ij} between u_i and u_j by using Eq.(4) and Max-min similarity measurement based on the user interaction behavior strength matrix *UIB* , and form fuzzy similar matrix $R^F = [S_{ij}]_{m\times m}$:

$$S_{ij} = \frac{(FA_{u_i} \wedge FA_{u_j}) + (CA_{u_i} \wedge CA_{u_j}) + (A_{u_i} \wedge A_{u_j})}{(FA_{u_i} \vee FA_{u_j}) + (CA_{u_i} \vee CA_{u_j}) + (A_{u_i} \vee A_{u_j})} \qquad (i, j = 1, 2, ..., m) \qquad (4)$$

Step 4 If fuzzy similar matrix R^F is fuzzy classification relation, then cluster analysis directly, otherwise, go to the next step.

Step 5 Transform from fuzzy similar matrix R^F to fuzzy classification relation, by the means of fuzzy similar matrix squares matrix-self circularly, $R \circ R = R^2, R^2 \circ R^2 = R^4, ...$ until $R^{2k} = R^k$, then R^k is fuzzy classification relation.

Step 6 Select $\lambda \in [0, 1]$ and adopt λ cut-matrix to fuzzy cluster analysis according to fuzzy classification relation R^k, If the bigger the λ, the more accurate the classification, otherwise the more inaccurate the classification.

Step 7 Output clustering pattern.

Example and analysis

To testify the feasibility of the clustering method discoursed above, take a local government micro-blogging as an example to demonstrate the specific clustering analysis process. The government micro-blogging issued 100 micro-blogs in a period of time, 7 users behavior data collected by scraping tool has been preprocessed shown in table 1.

Table 1 User interaction behavior data

User Id	Retweet	Comment	@	Micro-blog Retweeted	Micro-blog Commented
U$_1$	3	2	3	{wid$_1$, wid$_5$,wid$_{10}$}	{wid$_1$, wid$_5$ }
U$_2$	5	3	2	{wid$_1$,wid$_2$,wid$_5$,wid$_7$,wid$_{15}$}	{wid$_1$, wid$_3$,wid$_6$ }
U$_3$	6	1	2	{wid$_1$,wid$_2$,wid$_5$,wid$_9$,wid$_{10}$ wid$_{16}$}	{wid$_1$ }
U$_4$	2	2	1	{wid$_{16}$,wid$_{100}$ }	{wid$_4$, wid$_6$ }
U$_5$	4	2	3	{wid$_3$,wid$_4$,wid$_{20}$,wid$_{26}$ }	{wid$_{21}$, wid$_{19}$ }
U$_6$	2	1	0	{wid$_5$,wid$_8$ }	{wid$_2$ }
U$_7$	3	4	4	{wid$_{17}$,wid$_{25}$,wid$_{30}$ }	{wid$_{17}$,wid$_4$,wid$_{25}$,wid$_{26}$ }

According to Table 1 user interaction behavior data, calculate the user interaction behavior strength matrix by step 2 given above, results as follows:

$$\mathrm{UIB} = \begin{bmatrix} 0.5 & 0.5 & 0.75 \\ 0.833 & 0.75 & 0.5 \\ 1 & 0.25 & 0.5 \\ 0.333 & 0.5 & 0.25 \\ 0.667 & 0.5 & 0.75 \\ 0.333 & 0.25 & 0 \\ 0.5 & 1 & 1 \end{bmatrix}$$

According to the user interaction behavior strength matrix *UIB*, calculate user fuzzy similar matrix by step 3 given above, results as follows:

$$R^F = \begin{bmatrix} 1 & 0.643 & 0.556 & 0.619 & 0.913 & 0.333 & 0.7 \\ & 1 & 0.704 & 0.52 & 0.714 & 0.28 & 0.618 \\ & & 1 & 0.417 & 0.63 & 0.333 & 0.417 \\ & & & 1 & 0.565 & 0.538 & 0.433 \\ & & & & 1 & 0.304 & 0.656 \\ & & & & & 1 & 0.233 \\ & & & & & & 1 \end{bmatrix}$$

As a result of $\mathrm{R^F}$ is not fuzzy classification relation, it is necessary to squares matrix-self circularly by step 5 given above, the fuzzy classification relation as follows：

$$R^K = \begin{bmatrix} 1 & 0.714 & 0.704 & 0.619 & 0.913 & 0.538 & 0.7 \\ & 1 & 0.704 & 0.619 & 0.714 & 0.538 & 0.7 \\ & & 1 & 0.619 & 0.704 & 0.538 & 0.7 \\ & & & 1 & 0.619 & 0.538 & 0.619 \\ & & & & 1 & 0.538 & 0.7 \\ & & & & & 1 & 0.538 \\ & & & & & & 1 \end{bmatrix}$$

According to the fuzzy classification relation R^K, select different confidence level $\lambda, \lambda \in [0,1]$ and cluster analyze.

When $0.913 < l \le 1$, then users are divided into 7 classes: $\{U_1\}$, $\{U_2\}$, $\{U_3\}$, $\{U_4\}$, $\{U_5\}$, $\{U_6\}$, $\{U_7\}$.

When $0.714 < l \le 0.913$, then users are divided into 6 classes: $\{U_2\}$, $\{U_1, U_5\}$, $\{U_3\}$, $\{U_4\}$, $\{U_6\}$.

When $0.704 < l \le 0.714$, then users are divided into 5 classes: $\{U_1, U_2, U_5\}$, $\{U_3\}$, $\{U_4\}$, $\{U_6\}$, $\{U_7\}$.

When $0.7 < l \le 0.704$, then users are divided into 5 classes: $\{U_1, U_2, U_3\}$, $\{U_4\}$, $\{U_5\}$, $\{U_6\}$, $\{U_7\}$.

When $0.619 < l \le 0.7$, then users are divided into 3 classes: $\{U_1, U_2, U_3, U_5, U_7\}$, $\{U_4\}$, $\{U_6\}$.

When $0.538 < l \le 0.619$, then users are divided into 2 classes: $\{U_1, U_2, U_3, U_4, U_5, U_7\}, \{U_6\}$.

When $0 < l \le 0.538$, then users are divided into 1 classes: $\{U_1, U_2, U_3, U_4, U_5, U_6, U_7\}$.

According to the results above, users can be clustered in different levels by government micro-blogging management and operation officers based on the actual demands. Depend on the clustering result, calculate the average interaction behavior strength of each user group based on the matrix *UIB*, then the most active user groups can be found by ranking. In the light of Table 1, after discovering intersection of all users retweeted or commented micro-blogs, all users retweeted or jointly commented micro-blogs are calculable. Furthermore, hot topic, the public concerned, can be discovered.

Conclusions and Future Work

It is an urgent problem for the applications of government micro-blogging how to develop the public behavior patterns accurately, which fertilizes to optimize government micro-blogging and improves government services. Clustering government micro-blog users will help the government micro-blogging operation and management officers to offer more personalized information services to the public, and provide guidance for the public services decisions of government agencies. Based on the interaction behavior data between users and government micro-blogging, this paper put forward the user interaction behavior strength measurement model and the clustering method of user subdivision. The method of this paper can provide a new idea for the analysis of government micro-blogging user groups. In this paper, only take 3 indexes, including retweet, comment and mention(@), into account, while like, collection and other behavior factors, which reflects users preference and attention degree on government micro-blogging as well, are not considered. So our future work will add more factors to the measurement of user interaction behavior to calculate user Interaction behavior strength more accurately. With the development of the data mining technology and other network information mining technologies, this will provide a lot of valuable tools for government micro-blogging managers. And we will develop more efficient clustering algorithms related to government micro-blogging users.

Acknowledgements

This research was supported by National Natural Science Foundation of China (grant No. 61371177) and by the Science and Technology Plan of the Beijing Education Committee (grant No. KM201510028019).

References

[1] N. Zhang, H. Huang, M. Duarte, J. Zhang, Risk analysis for rumor propagation in metropolises based on improved 8-state ICSAR model and dynamic personal activity trajectories. Physica A: Statistical Mechanics and its Applications, Vol. 451(2016), p. 403-419.

[2] Ried, Christoph, K.bler, Felix, Goswami, Suparna, Krcmar, Helmut. Tweeting to Feel Connected: A Model for Social Connectedness in Online Social Network. International Journal of Human － Computer Interaction, Vol. 10-29 (2013), p. 670-687.

[3] Ho Young Yoon, Han Woo Park. Strategies affecting Twitter －based networking pattern of South Korean politicians: social network analysis and exponential random graph model. Quality ＆ Quantity, Vol. 1-48 (2014), p. 409－423.

[4] John Carlo Bertot, Paul T. Jaeger, Derek Hansen. The impact of polices on government social media usage: Issues, challenges, and recommendations. Government Information Quarterly, Vol. 29 (2012), p. 30-40.

[5] Jing Chen, Qinjian Yuan. A Review of Administrative Microblog Study in China and Abroad. Information Science, Vol. 32-6 (2014), p. 156-161.

[6] Zhangcheng Qiu, Hong Shen, User clustering in a dynamic social network topic model for short text streams. Information Sciences, Vol. 414(2017), p. 102-116.