MATEC Web of Conferences **22**, 01002 (2015)
DOI: 10.1051/matecconf/20152201002

Classification on Web Blogger Based on Clustering

Heyong Wang, Rong Cui & Ruoyu Lei
College of E-Business, South China University of Technology, Guangzhou, Guangdong, China

ABSTRACT: In this paper, based on the clustering analysis method, the author tries to study some celebrities in web blogger groups and adopts unsupervised clustering evaluation methods, which is called silhouette coefficient, to evaluate the classification results of different clustering classification methods. It is concluded that K-means clustering is the best among the clustering methods compared with the traditional classifications. Furthermore, it is a dynamic, flexible method and can reduce restrictions of subjective consciousness using cluster analysis. As a result, K-means clustering is universal in web blogger groups' classification process.

Keywords: web blog; fame classification; cluster analysis; silhouette coefficient

1 INTRODUCTION

In recent years, the web blog has been developing fast, and it has played an increasingly important role in daily life. Compared with traditional medium, web blog has better users' viscosity. As for the transmission of information on web blog, Jiang Xin found that the key nodes always act as "opinion leaders", which make the public opinion disseminate fast on the Internet. Defining these key nodes helps to guide public opinion. Ping Liang's study also showed that "opinion leaders" who have significant effects on the transmission of information can guide the public opinion in some degree. Celebrity web blogger groups have more frequent appearance and higher attention, which makes them become "opinion leaders".

Celebrity web blogger groups have the features of a large quantity of fans, stable relationship with fans, interaction, wide spread, great personal influence and high reliability. Besides, "celebrity effect" is significant, and followers pay high attention to celebrities' acts. Great results of dissemination can be got by the way of multi-level geometric spread. So, classifying web blogger groups effectively and trying to study the celebrity groups can help both companies' marketing and governments' communication.

In this study about celebrity web blog users, Zhao Yu classified them into two categories (active and realistic celebrity and native celebrity) qualitatively or three categories (information source, opinion leader and initiator of social activities) according to the roles celebrities play. As for the way of quantitative classification, Guo Qiuyan used the reputation index to calculate the number of users' following and follower with which researchers can classify the type of users according to the artificial interval. As a foundation of classification, the former method can't classify web blog users in reality. And the latter method is limited by the formula people define, and it's difficult to deal with the formula's new feature and precise analysis.

So, in this paper, a method based on clustering to classify users comes up. To find a better clustering method, it uses K-means, Two-steps and Kohonen to do and compares the results of three methods and reputation index.

2 THEORETICAL BACKGROUND

2.1 *The reputation index*

The reputation index (RI) is used to describe the web blog users' reputation and classify the type of celebrities by collecting the number of followings and followers. The formula is shown below:

$$RI = \frac{Fol\_C}{Fri\_C} + \frac{Fol\_C}{N} \quad (1)$$

Where, Fol_C is the number of followers per user, and Fri_C is the number of followings per user, and N is the sample size. The value of Fri_C will be smaller as Fol_C becomes bigger. Meanwhile, the RI will be bigger if there is a higher proportion of Fol_C in sample, which means the user attracts people more easily in sample, in other words, he has a bigger reputation. On the contrary, the value of Fri_C will be bigger as Fol_C becomes smaller. Meanwhile, the RI will be smaller if there is a lower proportion of Fol_C in sample, which means the user attracts people more difficultly in sample, in other words, he has a smaller reputation.

Table 1. The classification of web bloggers' reputation index

RI	Type of web bloggers
	Blogger with super reputation
	Blogger with better reputation
	Blogger with good reputation
	Blogger with normal reputation
	Blogger with no reputation

By defining the range of RI we can classify the web blog users as user with super reputation or better reputation or good reputation or normal reputation or no reputation. The classification is shown in Table1.

2.2 *Cluster analysis*

Cluster analysis is a way to analyze data which are grouped and a process to divide data into sunset. Each subset is called a cluster, in which the data are similar but different from data in other clusters. A cluster is produced by cluster analysis. Different clustering algorithms may lead to different results, even if the data sets are the same. The classification goes on automatically when algorithms are adopted. In this paper, K-means, Two-steps and Kohonen are used to analyze.

2.2.1 *K-means*

K-mean clustering is also called fast clustering, and it's an algorithm about numerical division. The principle of division is used for clustering and as for the result every sample point belongs to the only one cluster. The process of K-means is shown below:

1) Define the number of clusters. In K-mean clustering, K needs to be defined first.

2) Define K initial clustering centers. After defining K, choose k points randomly from data and think them as initial clustering centers.

3) Cluster according to distance. Calculate the Euclidean distance between each point and the initial clustering centers, and cluster the points to the nearest clusters according to the Euclidean distance, and then form K clusters. The Euclidean distance between points x and y is the length of the line segment connecting them and it is given by:

$$EUCLID(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2} \tag{2}$$

Where, x_i is the i-th variable of point x, and y_i is the i-th variable of point y.

4) Define K clustering centers again. Calculate and define centers of K clusters. The new centers are the mean points of each cluster.

5) If conditions of termination are met, then end it. If not, go back step three and repeat the procedure again and again until the conditions are met. The two conditions are the current numbers of iterations equal to the specified ones and the maximum offset of new clustering centers is less than specified one.

2.2.2 *Two-steps*

Two step clustering is an improved algorithm proposed by Chiu. It can deal with both numerical variable and categorical variable. This method can define clustering number according to some rules and cluster in two steps.

The process of two-steps is shown below:

1) Pre-clustering. In this step, classify the data previously.

First, construct CF by BIRCH and then compress data into subsets which are easy to analyze. The pointer can show the hierarchical relationship of nodes in the tree. Leaf nodes are subclasses, and a class formed by some subclasses which have the same father node is called intermediate node and these classes merge with each other to from a higher-level node until the root node which represents all data belong to one class.

Second, CF tree is a data processing method of compression and storage. Each node in the tree just store summary statistics required in distance calculation during clustering.

2) Clustering. Do re-clustering and define the final clustering method, and then take two steps to make sure the number of clustering classification.

First step: take Bayes information criterion (BIC) as standard. If we set the number of clustering as J, then

$$\begin{aligned} BIC(J) &= -2\sum_{j=1}^{J}\xi_j + m_J \log(N) \\ m_J &= J(2K^A + \sum_{k=1}^{K^B}(L_k - 1) \end{aligned} \tag{3}$$

The former shows the sum of J logarithmic likelihood. It is a total measure about inter-class difference. The latter is a multiplication formula of model complexity. With the given sample, the value of latter formula will become bigger when J becomes bigger. A good clustering will produce high quality clusters with reasonable cluster number and high intra-class similarity. Defining the cluster number is to find the J which makes BIC minimum.

In this paper, based on Clementine, $dBIC$ and $R_1(J)$ are used to define cluster number.

$$\begin{aligned} dBIC(J) &= BIC(\mathrm{J}) - \mathrm{BIC}(\mathrm{J}+1) \\ \mathrm{R}_1(J) &= \frac{dBIC(J)}{dBIC(1)} \end{aligned} \tag{4}$$

In the beginning, if $dBIC$ is less than 0, then the cluster number is 1, and the next algorithm is given up. If not, find the minimum of $R_1(J)$, which means to find a J that makes the decrease rate of BIC minimum, and then evaluate the cluster number roughly. Second step: correct the rough value J referred above. The method is

$$R_2(J) = \frac{d_{\min}(C_J)}{d_{\min}(C_J + 1)} \tag{5}$$

Where, $d_{\min}(C_J)$ is the minimum log-likelihood distance between two clusters, when the cluster number is J. $R_2(J)$ is the relative change of inter-class difference minimum in process of merging clusters. The larger the value is, the more inappropriate the merger between $J+1$ and J is. Calculate the value of $R_2(J-1), \mathrm{R}_2(\mathrm{J}-2), \ldots, \mathrm{R}_2(2)$ one by one, and find the maximum and the second largest value. In Clementine, if the maximum is more than 1.5 times as large as the second largest value, the J corresponding to the max-

imum is the final cluster number. If not, choose the larger one between cluster numbers corresponding to the maximum and the second largest value.

2.2.3 *Kohonen*

Used in clustering analysis, Kohonen is a self-organizing feature map (SOM) belonging to neural network, and is also an unsupervised learning algorithm in data mining. The process is shown below:

1) Preprocessing of data. The degree of "closeness" is based on Euclidean distance, so preprocess the data first. Get p clustering variables $x^i(i=1,2,,...,p)$ ranging 0 to 1, and consider N sample data as points in p -dimensional space.

2) Define the initial clustering center.

3) At time t, calculate the Euclidean distance $d(t)$ between $X(t)$ chosen from sample data randomly and K clustering centers. Find the closet center and output $W_c(t)$. $W_c(t)$ is the "winner" and the best match for the t-th sample now.

4) Adjust the location of $W_c(t)$ and its adjacent nodes. Set the weight of $W_c(t)$ as:

$$W_c(t+1) = W_c(t) + \eta(t)[\mathrm{X(t)} - W_c(t)] \tag{6}$$

Where, $\eta(t)$ is the rate of learning at time t.

Nodes in the circle which have $W_c(t)$ as circle center and distance from $W_c(t)$ within a given value as radius are all adjacent nodes. Set the weight of adjacent nodes as:

$$W_j(t+1) = W_j(t) + \eta(t)h_{jc}(t)[X(t) - W_j(t)] \tag{7}$$

Where, $h_{jc}(t)$ is kernel function showing the distance between adjacent node $W_j(t)$ and the "winner" $W_c(t)$ at time t. Adjusted Chebyshev distance is adopted in Clementine.

$$h_{jc}(t) = \max(\left|w_{ij}(t) - w_{ic}(t)\right|) \ (i = 1, 2, ..., p) \tag{8}$$

The formula above takes maximum distance of single dimension as the measure of distance.

5) Judge if the iterative ending condition is met. If not, return step three. Repeat the process, until the condition is met, which means the weights are basically stable or the specified number of iterations are reached.

2.3 *Silhouette coefficient*

Silhouette coefficient is an intrinsic method about evaluating clustering quality when no data set standard is available. Using similarity measure among objects in data set, separation and compactness of clusters are used to evaluate.

For data set D having objects, suppose that D is divided into K clusters $C_1,...,C_k$. For each object $o \in D$, calculate the average distance $a(o)$ among o and other objects in the same cluster. Similarly, $b(o)$ is the minimum average distance between o and other clusters not including o . Suppose $o \in C_i$ $(1 \le i \le k)$, and then

$$a(o) = \frac{\sum_{o' \in C_i, o \ne o'} dist(o,o')}{|C_i| - 1} \tag{9}$$

And

$$b(o) = \min_{C_j : 1 \le j \le k, j \ne i} \left\{ \frac{\sum_{o' \in C_j} dist(o,o')}{|C_j|} \right\} \tag{10}$$

The silhouette coefficient of object o is defined as

$$S(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}} \tag{11}$$

The value of $S(o)$ is between -1 and 1. The value of $a(o)$ shows the compactness of the cluster included means the smaller $a(o)$ is, the more compact the cluster is. The value of $b(o)$ shows the separation between o and other clusters means the larger $b(o)$ is, the more separate o and other clusters are. So, when the value of $S(o)$ is close to 1, the cluster including o is compact and far from other clusters. On the contrary, when the value of $S(o)$ is less than 0 ($b(o) < a(o)$), o is closer to objects in other clusters than in the same cluster.

3 EMPIRICAL ANALYSIS

3.1 *Source of data*

In this paper, research Sina web blog and use crawlers to crawl and collect information including users' ID, nickname, followers, followings and web blog number. Taking "College entrance examination" as search keyword, 643 users' information is available, including 233862296 followers, 378929 followings and 5741151 web blogs. The data is shown in Table 2.

3.2 *Experimental results*

3.2.1 *Results of classification*

Using the reputation index, K-means, Two-steps and Kohonen separately, show each result and make a summary. In this paper, "mean of cluster followers" show the average value of all sample users' followers in a cluster, and "mean of cluster followings" show the average value of all sample users' followings in a cluster, and "mean of cluster blogs" show the average value of all sample users' web blogs in a cluster, and "cluster sample" show the number of users in a cluster. These are all for classification results. The results tables list the number referred above to find the difference.

-The reputation index

Calculate each user's reputation index by using

Table 2. Sample of bloggers' information.

ID	Nickname	Followers	Followings	Blogs
1225314032	Sina education	2849784	1747	15654
1651428902	Economic reports in 21th century	2268741	462	27144
2185450707	Guide for art examination	85722	1536	9351
3290121547	SHINHWA	149467	2	217
2709577332	Happy Zhang jiang	2542976	2282	6839
……	……	……	……	……
2841812674	Wen jia Yu da ye	230	144	39
3844238381	Fantasy num.21	349	370	217
2717265785	GY-be sunshine for myself	335	244	2608
3854327817	User3854327817	520	605	248
3814671665	Gossip	18	47	33
Total	643	233862296	378929	5741151

Table 3. The result of RI.

Type	Mean of followers	Mean of followings	Mean of blogs	Cluster sample	Proportion
Blogger with super reputation	4123995	552	22010	27	4.20%
Blogger with better reputation	977457	685	18912	111	17.26%
Blogger with good reputation	88114	715	9645	139	21.62%
Blogger with normal reputation	11074	673	7403	145	22.55%
Blogger with no reputation	738	412	2866	221	34.37%

Note: The type of blogger is defined by mean of followers mainly.

Table 4. The result of K-means.

Type	Mean of followers	Mean of followings	Mean of blogs	Cluster sample	Proportion
Blogger with super reputation	5753223	1457	44235	7	1.09%
Blogger with better reputation	4932916	616	18677	15	2.33%
Blogger with good reputation	894924	728	35438	59	9.18%
Blogger with normal reputation	278436	1737	11311	102	15.86%
Blogger with no reputation	83467	303	4145	460	71.54%

Note: The type of blogger is defined by mean of followers mainly.

Table 5. The result of Two-steps.

Type	Mean of followers	Mean of followings	Mean of blogs	Cluster sample	Proportion
Blogger with super reputation	5164834	957	30199	22	3.42%
Blogger with better reputation	2035792	795	15184	25	3.89%
Blogger with good reputation	510186	863	35595	64	9.95%
Blogger with normal reputation	122629	1597	7899	102	15.86%
Blogger with no reputation	56235	279	3752	430	66.87%

formula (1). Referring to classification table, classify the user to one of user with super or better or good or normal or no reputation according to the RI.

We can conclude from Table3 that there is a big difference among different types in cluster average followers and both average web blogs and average followers have a gradient transformation, but there is no big change in cluster average followings.

-K-means

The result is shown in Table 4 by K-means clustering algorithm. There are two great clusters: user with super reputation and user with better reputation. The change of cluster average blog is not as same as average followers'.

The type of user with better reputation has large cluster average web blogs, and the average followings of each cluster differ. Compared to the result of RI, the difference is apparent in the result of K-means.

-Two-steps

The result is shown in Table 5 by Two-steps clustering algorithm. The difference of each cluster's average followers is smaller than the one getting from RI, but larger than the one getting from K-means. Similar to the result of K-means, the change of cluster average web blogs is not as same as average followers'.

The type of blogger with better reputation has large cluster average web blogs, and the average followings of each cluster differ. Compared to the results of RI and K-means, the result of Two-steps is closer to K-means' result.

-Kohonen

The result is shown in Table 6 by Kononen clustering algorithm. The difference of each cluster's average followers is small, and both average web blogs and average followers have a gradient transformation, which is just like the result of RI. But the average followings of each cluster differ. Compared to the results of RI, K-means and Two-steps, the result of Kohonen is the closest to RI's result but it has a balanced sample number.

-Summary

In this part, standardized Euclidean distance is used to analyze the feature of each cluster under four methods. Find the feature by analyzing "cluster average followers", "cluster average followings", "cluster average web blogs" and "proportion", and then explore the similarity among four methods and their own feature according to the results.

Standardized Euclidean distance is an improved method aiming at the shortage of simple Euclidean distance. The idea is: since the distribution of components in each dimension is different, then we should "standardize" each component to have equal mean and variance. In this paper, standardized distance is used to measure "followers", "followings" and "web blogs". The formulas are shown below:

$$x1 = \frac{\text{(followers-mean of cluster followers)}}{\text{variance of cluster followers}} \tag{12}$$

$$x2 = \frac{\text{(followings-mean of cluster followings)}}{\text{variance of cluster followings}} \tag{13}$$

$$x3 = \frac{\text{(blogs-mean of cluster blogs)}}{\text{variance of cluster blogs}} \tag{14}$$

$$d = \sqrt{x_1^2 + x_2^2 + x_3^2} \tag{15}$$

Where, x_1 is the standardized value of "followers", and x_2 is the standardized value of "followings", and x_3 is the standardized value of "web blogs", and d is the standardized Euclidean distance for single sample. Figure1 shows each cluster's overall standardized Euclidean distance is the mean of sample standardized Euclidean distance in a cluster under four methods.

From Figure1, we can see that the standardized Euclidean distance of each cluster using three clustering algorithms are shorter than the one using RI. Overall, the result is the same and even apparent. The shorter the distance is, the more compact the cluster is. It means there is a higher intra-class similarity by using these clustering algorithms than RI.

Using the mean of cluster followers in Table 2, 3, 4 and 5, Figure 2 is formed. In this part, the difference of each cluster is small by using Kohonen, and each value of clusters is the largest one by using K-means in four methods. Large magnitude of change reflects big differences among clusters and low inter-class similarity. And the method with that feature is a better clustering algorithm.

Using the mean of cluster followings in Table 2, 3, 4 and 5, Figure 3 is formed. In this part, the difference of each cluster is small by using RI, but the differences of each cluster are great by using other algorithms and the characteristics are obvious. A large magnitude of change reflects big differences among clusters and low inter-class similarity. And the method with that feature is a better clustering algorithm.

Using the mean of cluster blogs in Table2, 3, 4 and 5, Figure 4 is formed. In this part, the result of K-means is similar to the result of Two-steps, and the result of RI is similar to the result of Kohonen. The changes of each cluster using the first two methods are bigger than ones using the last two methods. Large magnitude of change reflects big differences among clusters and low inter-class similarity. And the method with that feature is a better clustering algorithm.

Using the proportion in Table2, 3, 4 and 5, Figure 5 is formed. In this part, the result of K-means is similar to the result of Two-steps, and the result of RI is similar to the result of Kohonen. The dotted line is a boundary line of 50% in Figure 5. Only K-means and Two-steps have a class with more than 50% proportion: user with no reputation. Compared to the results of RI and Kohonen, results of K-Means and Two-steps are more centralized.

In summary, combined with the analysis of figure 1, 2, 3, 4 and 5, clustering analysis is better than RI. Furthermore, clustering analysis can accept new index and is dynamic and flexible. In these four methods, K-Means is similar to Two-steps and RI is similar to Kohonen.

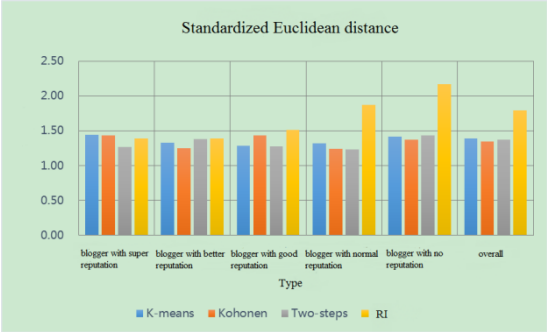

Fig.1 Comparison of standardized Euclidean distance

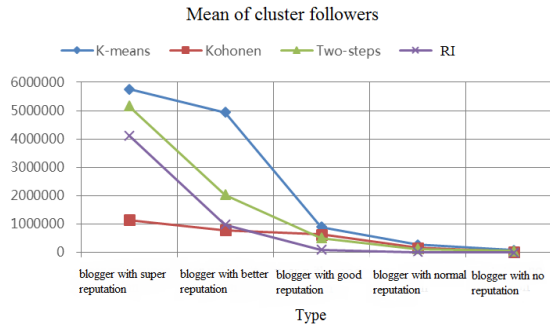

Figure.2 Comparison of mean of cluster followers

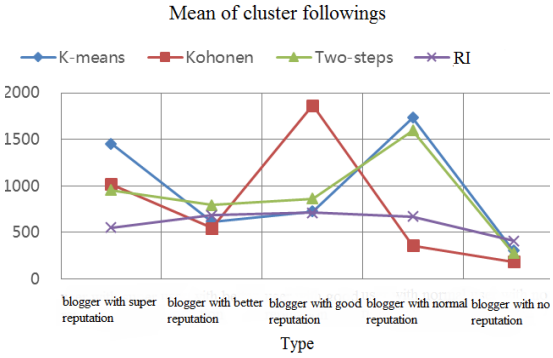

Figure 3. Comparison of mean of cluster followings

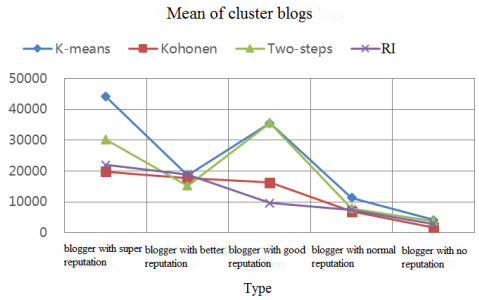

Figure 4. Comparison of mean of cluster blogs

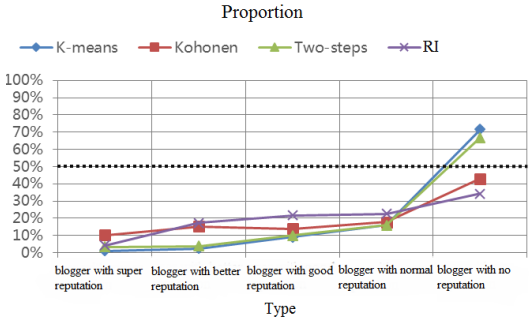

Figure.5 Comparison of proportion

3.2.2 *Silhouette coefficient*

To emphasize the advantages of three clustering algorithms better, silhouette coefficient is used to analyze in this paper. Under each method, data set is classified into five clusters $C_1,...,C_5$: user with super reputation, user with better reputation, user with good reputation, user with normal reputation and user with no reputation. And calculate the silhouette coefficient by formula 11 defined before. Fitting of clustering and quality of clustering are calculated by formula 16 and 17 separately, and the results are shown in Table 7.

In Table 7, fitting of clustering (FOC) is the mean of silhouette coefficient in a cluster:

$$\mathrm{FOC}=\frac{\sum_{i=1}^{n_{C_k}} S(o)_i}{n_{C_k}} \quad (16)$$

Where, $S(o)$ is silhouette coefficient, and $o\in C_k$ $(1\le k\le 5)$, and n_{C_k} is the sample number of C_k.

$$\mathrm{QOC}=\frac{\sum_{i=1}^{N} S(\mathrm{o})_{\mathrm{i}}}{N} \quad (17)$$

Where, $S(o)$ is silhouette coefficient, and $o\in C_k$ $(1\le k\le 5)$, and N is the total sample number.

We can find that K-means is best followed by Two-steps, RI and Kohonen successively. There is a cluster with best fitting of clustering and largest number of sample: user with no reputation. And this cluster has great effect on quality of clustering. In K-means and Two-steps, the number of sample is larger and more concentrated than in RI and Kohonen, so the quality of clustering of the first two methods is better. Kohonen is the only method having negative value in QOC. Furthermore, the number of sample is balanced in Kohonen, and only the FOT of user with no reputation is positive which have limited effect on total QOC. Therefore, QOC is influenced by sample number and FOC together. If a cluster has larger sample number and better FOC, then the method is better.

4 CONCLUSIONS

In this paper, based on clustering analysis, compared with traditional artificial formula RI, three clustering algorithms are used to classify web blog users more flexibly. The results show that: (1) the classes have more characteristics and are more concentrated based on clustering analysis; (2) the QOC of K-means is best based on silhouette coefficient, which means clustering analysis is more effective to classify web blog users and then find web blog celebrities. The fame classification of Sina web blog users is feasible, and furthermore, when accurate classification about certain users is needed, clustering analysis which can accept new index and is universal in the web blog users classification process can help.

ACKNOWLEDGEMENT

This research was supported by Project of National Social Sciences Foundation, (GN: 13BTJ005), the Fundamental Research Funds for the Central Universities, (GN: 2013XZD01).

REFERENCES

[1] Wu Minqi. 2013. An empirical study of web bloggers' everyday life information seeking behaviors and the influencing factors. *Information Science*, 1: 86-90

[2] Jiang xin. 2012. Research on "little word" phenomenon of information dissemination in web blog: Take tencent

Table 7. Silhouette coefficient under four methods

Clustering algorithms	Type	Sample	FOC	QOC	Ranking
Silhouette coefficient(RI)	Blogger with super reputation	27	0.03	0.12	3
	Blogger with better reputation	111	-0.20		
	Blogger with good reputation	139	-0.17		
	Blogger with normal reputation	145	-0.22		
	Blogger with no reputation	221	0.71		
Silhouette coefficient (K-means)	Blogger with super reputation	7	-0.18	0.33	1
	Blogger with better reputation	15	0.10		
	Blogger with good reputation	59	-0.23		
	Blogger with normal reputation	102	-0.51		
	Blogger with no reputation	460	0.60		
Silhouette coefficient (Two-steps)	Blogger with super reputation	22	0.27	0.23	2
	Blogger with better reputation	25	0.37		
	Blogger with good reputation	64	-0.39		
	Blogger with normal reputation	102	-0.32		
	Blogger with no reputation	430	0.43		
Silhouette coefficient (Kohonen)	Blogger with super reputation	66	-0.58	-0.22	4
	Blogger with better reputation	98	-0.61		
	Blogger with good reputation	89	-0.72		
	Blogger with normal reputation	114	-0.65		
	Blogger with no reputation	276	0.80		

web blog as example. *Information Science*, 30(8): 1139-1142.

[3] Ping Liang. & Zong Yongli. 2010. Research on microblog information dissemination based on SNA centrality analysis: A case study with Sina microblog. Document, *Information & Knowledge*, (6): 92-97.

[4] Guo Qiuyan. & He Yue. 2013. Study on the celebrity users' characteristics mining and the effects of Sina Web blog. *Journal of Intelligence*, 32(2): 112-117.

[5] Zhao Yu. 2012. Analysis on "celebrity effect" in web blog: Take Sina web blog as example. *News World*, 9: 149-151.

[6] Xue Wei. & Chen Huange. 2012. *Data Mining Based on Clementine*. Beijing: China Renmin University Press.

[7] Jiawei Han, Micheline Kamber. & Jian Pei. 2013. Data Ming: *Concepts and Techniques (Third Edition)*. Beijing: China Machinery Press.

[8] Wang Xiaoguang. 2010. Empirical analysis on behavior characteristics and relation characteristics of web blog users: Take "Sina Web blog" for example. *Library and Information Service*, 54(14): 66-70.

[9] He Li, He Yue. & Huo Yeqing. 2011. Analysis on web blog users' characteristics and data mining on core users. *Information Systems*, 34(11): 121-125.

[10] Yang Chengming. 2011. Empirical analysis of microblog users' behavioral characteristics. *LIS*, 55(12): 21-25.

[11] Yang Kai. & Zhang Ning. 2013. Structure and cluster analysis on web blog users' relationship networks. *Complex Systems and Complexity Science*, 10(2): 37-43.

[12] He Jing, Guo Jinli. & Xu Xuejuan. 2013. Analysis on statistical characteristic and dynamics for user behavior in microblog communities. *Analysis and Research on Information*, 7/8: 94-100.

[13] Sheng Qizhi. & Gao Senyu. 2013. Opinion leaders of web blog in China: Characteristics, type and trend of development. *Journal of Northeastern University (Social Sciences Edition)*, 15(4): 381-385.

[14] Wang Lu. 2011. Analysis on new social media web blog base on communication. *Journal of Zhengzhou University (Philosophy and Social Sciences Edition)*, 44(4): 142-144.

[15] Zhao Ling. & Zhang Jing. 2013. Multidimensional analysis on web blog users' behavior. *Information and Documentation Services*, 5: 65-70.