Online Social Network User Behavior Analysis — With RenRen Case

Wenqian Wang(✉) and Yinghong Ma(✉)

College of Management Science and Engineering, Shandong Normal University, Jinan, China
{wenqiansuk,yinghongma71}@163.com

Abstract. This paper investigated typical user behaviors in RenRen and used a clustering algorithm that assigns users to groups through a distance measure that is computed based on the values of user feature vector. The user feature vector consists of four attributes and we got six user groups from the clustering process. By analyzing the six different user behavior patterns, we considered some strategies for providers to improve their service quality.

Keywords: RenRen · Clustering · User behavior · Strategy

1 Introduction

Exploring user habit and psychology hidden behind the user behavior is very important in the research of online social networks. For example, some users are enthusiastic to express themselves by updating status and uploading as many logs or photos as they can, whereas there are users that act like free-riders [1] and just want to enjoy the contents that made publicly available. When tapping into user behavior, [2] analyzed a large online community system and investigated the relationship and engagement between users and groups of users. In [3] it showed that the user behavior can improve the accuracy of a web search ranking algorithm. User behavior models have also been extensively studied from a community of users with a single behavior to multiple classes of users [2, 3].

Our work differs fundamentally from the aforementioned references. We investigated user behaviors of a concrete online social network and based on users personal and social attributes to assigns them to different groups. In Sect. 2, we compared several clustering algorithms and expounded the clustering algorithm. In Sect. 3 we defined user feature vector based on personal attributes and social attributes and then applied K-means algorithm to get the final results. In Sect. 4, we analyzed characteristics of user behaviors and think strategies to improve the service quality of RenRen. Finally we give the conclusion and future directions.

2 Analyze Clustering Algorithm

We compare different clustering algorithms: the density-based approach such as DBSCAN; partition clustering algorithm; and agglomerative hierarchical clustering algorithm etc. [4]. For DBSCAN, when the amount of data increases, it requires large

Q. Zu and B. Hu (Eds.): HCC 2016, LNCS 9567, pp. 925–929, 2016.
DOI: 10.1007/978-3-319-31854-7_101

memory support and the I/O consumption is great. For agglomerative hierarchical clustering algorithm, it has high time and space complexity O(n2). And it cannot be separated by the way of the split to the previous state [5]. Considered the characteristics of user behavior attributes and the distinction of algorithms, we choose K-means as the clustering algorithm and the Euclidean distance as the distance measure.

For a given D dimensional space R^d, an evaluation function $E:\{p:p \in C\} \rightarrow R^+$ is defined in R^d. Give each cluster a quantitative evaluation and then input object set C in R^d and one integer k. Output one division of $C:C_1, C_2, \ldots C_k$ which makes the evaluation function E minimize. We used the mean square error as the evaluation function and it is defined as $E = \sum_{i=1}^{k} \sum_{p \in c_i} |p - m_i|^2$.

E is the sum of squared error of all the objects and p is point in R^d which represents a given data, m_i is the mean value of cluster C_i (p and m_i are multidimensional), $|p - m_i|^2$ represents the distance between the data object and the centroid.

Algorithm flow:

(1) Select K initial centroids: for example, $c[0] = data[0], \ldots, c[k-1] = data[k-1]$;
(2) For $data[0]$, …, $data[n]$, compare with $c[0], \ldots, c[k-1]$ respectively and if the difference between it and $c[i]$ is the least, it is marked as i;
(3) For all the points marked as i, recalculate $c[i] = \{$the sum of all the $data[j]$ which marked as i the number of points which marked as i..
(4) Repeat 2 and 3 until all the change values of $c[i]$ less than the defined threshold. (The threshold in this paper is defined as 10^{-5}).

3 Users Behavior in RenRen

3.1 Users' Behavior Data Preprocessing

Define the user feature vector as a one-dimensional vector of length four, where each position contains information about the referred user: $user_i = [f_1,f_2,f_3,f_4]$. The four features are detailed:

(1) Number of uploads (f_1): f_1 represents the number of uploaded status, logs and photos by user. It indicates the potential of a user as a content producer.
(2) Number of watches (f_2): f_2 represents how many videos, logs, albums or other Internet information have been watched by the user. It could indicate the potential of the user as a content consumer.
(3) Age (f_3): We consider user age as the time elapsed between its join date and last login.
(4) Number of comments (f_4): f_4 represents the number of comments which other people give for their status and logs. It indicates the popularity of the user and his willingness to interact with others.

Calculate the correlation coefficient between each two attributes (Table 1). All the correlation coefficients are less than 0.5 which indicates their correlations are low enough to put these four attributes as the representative features of user behavior vector. These four features are of different units and magnitudes. To ensure the distance is computed with features of equal weight, we normalized the data by the maximum value of each feature so that every feature ranges from 0 to 1.

Table 1. The correlation coefficients

Correlation coefficient	f_1	f_2	f_3	f_4
f_1	1.000	0.385	0.207	0.459
f_2	0.385	1.000	0.167	0.306
f_3	0.207	0.167	1.000	0.200
f_4	0.459	0.306	0.200	1.000

3.2 Cluster User Behavior According to Feature Vector

We used an error value J to measure the effect of clustering under different number of clusters (k) in which J is represented by the sum distances from every point to their centroids in each cluster.

$$J = \sum_{j=1}^{n} \alpha\left(data[j] - c[i]\right)^2.$$

n is the total number of points which marked as i and α is a parameter ($0 < \alpha < 1$). The smaller the error value, the better the clustering effect. Through the experiment of iterating different k values, the J values are obtained and presented in Fig. 1. When k reaches 6, the cluster error value is obviously decreased, and it changes a little when k value increased. So we choose 6 as the number of cluster. Then we work on 2390 user behavior vectors and get result as Table 2.

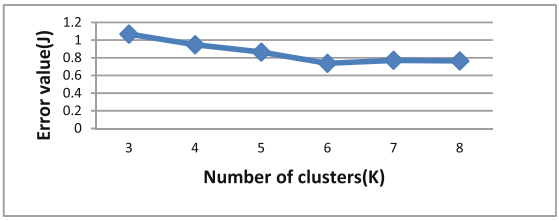

Fig. 1. Corresponding error value J calculated under different k

3.3 Clustering Results and Analysis

Present above results as Fig. 2: the horizontal coordinates are created clusters and the percentage of users assigned to each cluster; the vertical coordinate is the intergroup relative feature value. We can analysis the behavior of different groups:

Table 2. Results of clustering

Attribute values and proportion / Groups	center	f_1	f_2	f_3	f_4	proportion	features
Group 1	U_0	0.819	0.476	0.722	0.559	26.47%	Upload lots, watch and communicate less,
Group 2	U_1	1.399	0.883	0.749	2.008	15.55%	Upload and communicate lots but watch less,
Group 3	U_2	0.153	0.143	0.593	0.073	34.03%	All values are less,
Group 4	U_3	1.367	1.613	0.715	5.736	5.88%	Upload and watch less and communicate lots,
Group 5	U_4	6.972	2.440	0.725	3.773	4.20%	All values are large,
Group 6	U_5	1.737	2.639	0.746	0.891	13.87%	Upload and watch lots but communicate less.

Group 1: They like show their daily life by uploading status, logs and photos but they are not much interested in Internet resources and they get less comment from their friends. They do not usually communicate with others. They prefer to be a kind of pure content producer and they are 26.47 % of all users.

Group 2: They upload a lot but browse Internet resources less. The contents uploaded by them can be commended by more people. It reflects they are more popular and they are more enthusiastic in communication. They have 15.55 % of all and we can call them content producer & energetic communicator.

Group 3: All the attribute values of this group are quiet low. Most of them abandoned their accounts after they registered. It represents the largest fraction of users (34.03 %). This group of users is inactive user.

Group 4: The users are unwilling to upload status or photos as well as pay attention to web information. They are more enthusiastic to communicate with their friends. They are pure energetic communicators. The proportion of this group of users is relatively low, is about 5.88 %.

Group 5: Uploads, watches and communication number are all large in this group. They not only produce and consume contents, but also frequently communicate with friends. They are active users but only with a low proportion (4.2 %).

Group 6: Compared to communicate with others, they are more willing to show themselves and browse web information. They have both characteristics of content producers and consumers, so we name them producer & consumer (13.87 %).

The user age is roughly equal distributed among clusters. It does not affect the clustering result of user behavior so that we can delete this attribute to reduce the occupancy of the data storage space and improve the clustering efficiency.

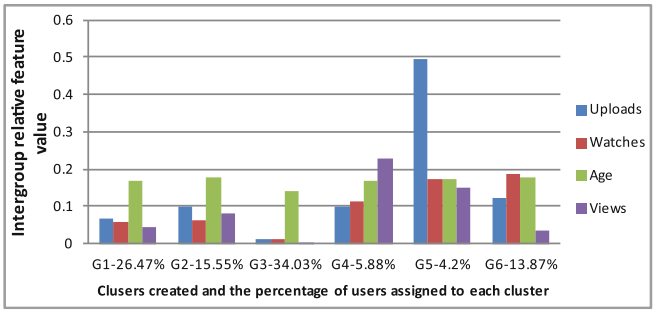

Fig. 2. Relative feature values for the six clusters created

4 Consider Implementation Strategies

Identifying different user behaviors is helpful to improve business and resource management in social networks. For example, for pure content producers, they show their lives but get no more attention. So in order to improve their initiative in interacting with others, something can be done to search and introduce them some friends with same interests through the key words of the contents they have uploaded. For content producer & consumer, considering they are less willing to communicate with others, we can support them expansive display platforms as well as more attractive network resources. For pure energetic communicator, they are interested in neither display platform nor web resources, so more energetic should be put in improving the quality of interaction services.

In this article, we identified six distinct behaviors of RenRen users. Generally, one can use the methodology we presented to assign users to groups with similar behavior. For future directions we could investigate more online social network user behavior and explore their difference. More work should be done to study the influence of user behaviors to the structure of network.

References

1. Feldman, M., Papadimitriou, C., Chuang, J.: Free-riding and whitewashing in peer-to-peer systems. IEEE J. Sel. Areas Commun. **24**, 1010–1019 (2006)
2. Backstrom, L., Kumar, R., Marlow, C., Novak, J.: Preferential behavior in online groups. In: Proceedings of the ACM Web Search and Data Mining, Stanford, CA, USA, February 2008
3. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., Moon, S.: I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: Proceedings of the ACM Internet Measurement Conference (IMC), San Diego, CA, USA, October 2007
4. Gong, X., Ning, Q., Zhou, A.: Clustering in very large databases based on distance and density. J. Comput. Sci. Technol. **18**(1), 67–76 (2013)
5. Zhou, Y., Peng, F., Zhou, J.: Complex surface fitting based on interpolation and approximation. J. Eng. Graph. **4**, 47–54 (2014)