Research on User Clustering Algorithm Based on Improved ROCK Algorithm

Rong Wang¹,* and Aimin Sun²

¹,²School of information engineering, Luohe Vocational Technology College, Luohe, China

*Corresponding author

Abstract. This paper analyzes the merits and demerits of the ROCK algorithm firstly, then indicates that the merits of Rock is easy to cluster categorical database such as Mushroom, and the demerits is that the similarity formula sim of this algorithm depends on the intuition of domain experts. The improved ROCK gets the result of clustering by calculating the similarity using Jaccard coefficient based on the reason which is that the bigger figure of the similarity is, the more alike the object is. Then we can implement personality recommendation according to the result.

1. Web Clustering

There are two kinds of clustering on Web blogs: User Clustering and Page Clustering. User Clustering tends to put similar browsing patterns users together for groups, while Page Clustering always classifies similar browsing pages into groups. When the user in the same group browses the website again, the similar pages will be recommended automatically in accordance with configuration files. This is very valuable for Internet search engines and WEB service providers [1].

The diagram 1 shows the relevance between User Clustering and Page Clustering. The longitudinal direction represents Page Clustering while horizontal direction represents User Clustering. Fig.1 includes five data sets, and U means the users sessions while F means the Web pages. The value “1” on the Diagram represents the corresponding user session is valid in this page, while “0” represents invalid. The relevance between the users and the webs could be clearly shown in the diagram, for example, {F4, F5} is the valid pages for User U1[1].

Figure 1. Relationship between user and page clustering.

As to the availability of the pages, in traditional method, value “1” represents that the page is shown in corresponding user sessions while “0” means that the page is not shown, and the factor is taken no account of what time length always results big errors when data is changed by binary values[2]. Some methods consider the amount of time and usually establish a threshold value which is defined by
specialists in this field, so undefined results will be taken. The time characters is obtained with the Law of Zipf and cluster is conducted with ROCK algorithm in this article[1].

2. The ROCK Algorithm

ROCK(Robust Clustering using links) is an algorithm which is applied to categorized character data with clustering. At first, the similar threshold value $\theta$ and the concept of common neighbor is used to calculate similar figures according to the connection numbers among the tuple. If the similar figure among a couple of tuples exceed the threshold value, then these tuples would be called neighbors[1]. Building sparse graph in specified data similarity matrix according to whether two data points are neighborhood similarity threshold $\theta$, then the hierarchical clustering algorithm is used to cluster the sparse graph. The basic definition of ROCK algorithm is introduced briefly here[3].

Definition 1. $\theta$ is the neighboring threshold value defined by the user. $\text{Sim}(p_i$ and $p_j)$ is the similarity between data points $p_i$ and $p_j$, and $[0, 1]$ is the scope for choosing values in the similarity function of $\text{Sim}(p_i, p_j) \geq \theta$. The bigger it is, the more similar they are.

Definition 2. The similarity between data points $p_i$ and $p_j$ uses the Jaccard coefficient, then

$$\text{sim}(p_i, p_j) = \frac{|p_i \cap p_j|}{|p_i \cup p_j|}$$

Definition 3. $\text{link}(p_i, p_j)$ is the crossing connection number between the group $N_i$ and $N_j$, and $N_i$ and $N_j$ are the neighboring values matrix for $p_i$ and $p_j$ among the groups.

Definition 4. The function

$$g(C_i, C_j) = \frac{\text{link}[C_i, C_j]}{(n_i + n_j)^{1/2(f(\theta))} - n_i^{1/2(f(\theta))} - n_j^{1/2(f(\theta))}}$$

$C_i$ and $C_j$ are clusters, while $n_i$ and $n_j$ are the numbers of the clusters respectively. Using $C_i$ as the referring point, there are $n_i^{f(\theta)}$ neighbors for every point; the threshold value set by the experts is $f(\theta)=(1-\theta)/(1+\theta)$.

The core concept of the ROCK algorithm is to choose n sample points randomly from the database to compose a group, Using S as the data input with the parameters “n, S, and k” (the number of the clusters formed)[1]. Every point is an isolated cluster firstly, and the connecting points should be calculated. Establishing a local heap $q[i]$ for every cluster i, and $q[i]$ has a cluster j connecting to the cluster i with at least one connecting point, and all the cluster j inside $q[i]$ are set in descending order according to the value of $g(i,j)$. A global heap Q should be set to contain the cluster j which is the largest value of the function in every q[i]. During every integration, the best cluster j of Q and the best cluster among $q[j]$ join together, then all the heaps in every area and the global heap should be calculated as well as the new formed cluster again. Repeating the step 5 in a circulating way until k clusters are left only in the local heap Q. In addition, if the number of the connecting points between every left clusters is the zero, the circulation will be stopped[3].

3. Improved ROCK Algorithm

Step 1: To set the value 1 and 0, “1” in this Diagram represents that the corresponding user session in this page is valid, while “0” represents invalid.

Step 2: User clustering
① To use the ROCK algorithm, calculate the similarity of object domain with the Jaccard goodness, make the Two-dimensional table.

② Based on the Two-dimensional table of the similarity, to cluster by the rule that the value of similarity is larger means the object is more similar.

Part of the algorithm of Step2:

```
procedure cluster(S, k)
Begin
1. \( d(i,j) := |P_i \cap P_j| / |P_i \cup P_j| \)
2. for each \( s \) in \( S \) do
3. \( q[i] := \text{build_local_heap}(d, s) \)
4. \( Q := \text{build_global_heap}(S, q) \)
5. while |\( Q \)| > \( k \) do {
6. \( u := \text{extract_max}(Q) \)
7. \( v := \text{max}(q[u]) \)
8. delete (\( Q, v \))
9. \( w = \text{merge}(u, v) \)
10. deallocate(q[u]); deallocate(q[v])
11. }
```

The concept of improved ROCK algorithm is to choose \( n \) sample points randomly from the database and to compose a group \( S \) as the data input with the parameters “\( S \)” and “\( k \)” (the number of the clusters formed). Using the Jaccard parameter, a local heap \( q[i] \) is established for every cluster \( i \), and every \( q[i] \) has a cluster \( j \) which connects the cluster \( i \) with at least one connecting point \( r \) according to the value of \( g(i,j) \), all the cluster \( j \) inside \( q[i] \) are set in descending order. A global heap \( Q \) should be set to contain the cluster \( j \) which is the largest value of the function in every \( q[i] \). During every integration, the best cluster \( j \) of \( Q \) and the best cluster among \( q[j] \) combine together. Repeating the step 5 in a circulating way until \( k \) clusters are left only in the local heap \( Q^{[i]} \).

For example: Table 1 is a two-dimensional binary variable table (data derived from [4]), ROCK algorithm and improved ROCK algorithm respectively clustering. Experiments can compare the clustering effect. To use the ROCK algorithm, calculate the similarity of object domain with the Jaccard goodness, make the Two-dimensional table. See Tab.2.

Firstly, ROCK algorithm is applied to the data in Tab.2, if the neighbor’s threshold is 0.2, the neighbors are \{\( (2,3), (2,4), (3,4) \}\), because a point is the neighbor by itself, so the other neighbors are \{\( (1,1), (2,2), (3,3), (4,4) \}\}. See Tab.3, there have Common neighbors of two points. So ROCK algorithm forms two clusters of \{\{U1\}, \{U2, U3, U4\}\}, see Fig.3. Then, improved ROCK algorithm is applied to the data in Tab.2, the largest similarity is 0.6, so there is a cluster of \{U2, U3\}; the similarity is 0.2, see Tab.4 (the Tab.4 is a child table of Tab.2). In Tab.4, there is a cluster of \{U2, U3, U4\}, see Fig.4. It could be seen from the example that the effect is the same by ROCK algorithm and improved ROCK algorithm respectively clustering, see Fig.3 and Fig.4. Experiments can prove that the work for clustering using improved ROCK Algorithm is well.
Table 1. The table of binary-valued variables.

|   | F1 | F2 | F3 | F4 | F5 | F6 |
|---|----|----|----|----|----|----|
| U1| 1  | 0  | 0  | 0  | 0  | 0  |
| U2| 0  | 1  | 1  | 1  | 1  | 0  |
| U3| 0  | 1  | 1  | 0  | 1  | 1  |
| U4| 0  | 0  | 0  | 1  | 0  | 1  |

Table 2. The calculated similarity based on Jaccard coefficient

|   | U1 | U2 | U3 | U4 |
|---|----|----|----|----|
| U1| 1  | 0  | 0  | 0  |
| U2| 0  | 1  | 0.6| 0.2|
| U3| 0  | 0.6| 1  | 0.2|
| U4| 0  | 0.2| 0.2| 1  |

Table 3. The neighbor matrix of ROCK algorithm.

|   | U1 | U2 | U3 | U4 |
|---|----|----|----|----|
| U1| 1  | 0  | 0  | 0  |
| U2| 0  | 3  | 3  | 3  |
|   | 3  | 0  | 3  | 3  |
| U4| 0  | 3  | 3  | 3  |

Table 4. The calculated similarity based on Jaccard coefficient (child table).

|   | U2 | U3 | U4 |
|---|----|----|----|
| U2| 1  | 0.6| 0.2|
| U3| 0.6| 1  | 0.2|
| U4| 0.2| 0.2| 1  |

Figure 3. The effect of clustering based on ROCK algorithm.
4. Conclusion

Now, the target of cluster analysis on Web logs is to obtain user’s browsing pattern, and provide customized personalization service for various user groups. Improved ROCK algorithm has the distinct advantage: fast calculation speed and high efficiency. The experimental results prove that it can obtain a better cluster result with this method.

References

[1] Wang, Rong, and Bo Ping Zhang. “The Research of Web Clustering Algorithm Based on Binarized Zipf’s Law.” Applied Mechanics and Materials, 2013.

[2] Yiling Yang. Analysis on Intelligent Web Sites Based on Web Blogs Excavating Technologies. [Thesis For Doctor’s Degree]. Shanghai Jiaotong University. 2002:53-73.

[3] S. Guha, R. Rastogi, and K. Shim. Rock: A robust clustering algorithm for categorical attributes. Proceedings of the 15th International Conference on Data Engineering. Sydney, Australia, 1999. Los Alamitos (CA, USA): IEEE Computer Society Press, 1999: 512-521.

[4] Guo Chonghui. Courses of Data Excavation. Tsinghua University Press. 2006, 1: 107-138.

[5] Lyu Xueqiang, wangteng, lixuewei, dongzhian. Research on movie recommendation algorithm based on content and interest drift model [J]. Application Research of Computers, 2018, 35(3): 717-720, 802. doi:10.3969/j.issn.1001-3695.2018.03.016.

[6] Lei Shi, Xin Ding, Yong-cai Tao, Lin Wei. Meta Path Recommendation Model with Interest Drift Detecting. Journal of Chinese Computer Systems. 2019, 40(3): 612-617.

[7] Lei Zou, Jing Zhu, Xiao-hui Nie, Ya Su, Dan Pei, Yu Sun. Detecting Hotspot in Multi-dimensional Data Through Clustering. Journal of Chinese Computer Systems. 2019, 40(3): 465-471.

[8] Chuan Hu, Xiang-Wu Meng, Yu-Jie Zhang, Yu-Lu Du. Enhanced Group Recommendation Method Based on Preference Aggregation. Journal of Software. 2018, 29(10): 3164-3183.

[9] CASTRO, JORGE, QUESADA, FRANCISCO J., PALOMARES, IVAN, et al. A Consensus-Driven Group Recommender System [J]. International Journal of Intelligent Systems, 2015, 8(8): 887-906.

[10] Wei Dezhi, Chen Fuji, Lin Lina. Microblog hotspot detection method based on MFIHC and TOPSIS. Application Research of Computers. 2018, 35(4): 1014-1017, 1041.

[11] Luo Da-xiong, YE Jun-min, Liao Zhi-xin, WANG Zhi-feng, Chen Shu. Problem Solver Recommendation Algorithm Based on Implicit Behavior. Journal of Chinese Computer Systems. 2019, 40(3): 606-611.

[12] Min Zhang, Biyuan Ding, MA Weizhi, Tan Yunzhi, Liu Yiqun, MA Shaoping. Hybrid recommendation approach enhanced by deep learning. Journal of Tsinghua University (Science and Technology). 2017, 57(10): 1014-1021.

Figure 4. The effect of clustering based on modified ROCK algorithm
[13] Zhong Zhao-Man, Guan Yan, Hu Yun, Li Cun-Hua. Mining User Interests on Microblog Based on Profile and Content. Journal of Software. 2017, 28(2): 278-291.

[14] Cui Yi-hui, Song Wei, Wang Zhan-Bing, Shi Cheng-Liang, Cheng Fang-Quan. Privacy Preserving Cluster Mining Method Based on Lattice. Journal of Software. 2017, 28(9): 2293-2308.