Research on User Discovery Based on Loyalty in SNS

Yun Xue[a], Jianbin Chen[b] and Yingying Zhou[c]

Business College of Beijing Union University, Beijing, China

[a]Yun.xue@buu.edu.cn, [b]Jianbin.chen@buu.edu.cn, [c]2427252559@qq.com

Keywords: Social networking services; user discovery; loyalty

Abstract: With the rise of social networks, the diverse and informative social networking applications have become an indispensable part of the daily life of netizens. Improving user loyalty is one of the important ways for software vendors to seize and maintain market share. In this paper, we propose a user discovery method based on loyalty in social network. Firstly, we calculate the consumer's consumption value and participation value dynamically according to the double RFM model, find the standard curve, then use the similarity calculation to find out the typical loyal users and disloyal users, and then identify the potential users using the modularity-based community discovery and independent cascade propagation model. The method has obtained ideal results of user discovery when it was applied to some microblog datasets of a social network, which has certain practical significance.

1 Introduction

A social networking service (SNS)is a relational architecture that focuses on interactions and relationships among individual members. Its high viscosity characteristics promote the formation and development of virtual communities, and this high viscosity is mainly reflected in users'loyalty to the community. [1] Therefore, improving user loyalty is one of the important ways for SNS vendors to seize and maintain market share.

User loyalty, also known asuser viscosity, refers to an attitudinal and behavioral tendency to favor one specific product or service over all others, forming a "dependent" preference, and a tendency to repeat past purchases[2]. Loyalty provides a reliable reference evaluation in respect of the product sales forecast and financial growth. Therefore, "retainingcustomers" is also a focus ofSNS vendors. Early on in the analysis of loyalty, users were studied as an isolated individual, without taking into account the interaction between users and information transmission[3]. However, in the SNS, users are in the same community, a user's behavior will affect the surrounding groups, or group users will become more loyal, or group users are lost gradually. This paper will start with user discovery based on loyalty, and find out methods to improve the users' viscosity, that is, loyalty.

2 USER LOYALTY of SNS

2.1 Definition and measurement of user loyalty

Ding[5] defines the SNS user loyalty as: users feel dependent on a particular social network site, have certain emotional attachment, which is expressed in the degree of specificity, emotional attachment and good reputation; and will visit the site in a long period of time, spend some time to browse the site and interact with friends, and even recommend the site to their friends. This user loyalty discussed by the paper is based on a user's consumption value and participation value in a social networking site, and the degree of dependence on the social networking sites. In order to specify the concept of user loyalty in the practice use, we usually need to measure the user loyalty.

2.1.1 Constructing the measurement index system

The index system of user loyalty index is constructed and evaluated. The cycle is complex with high cost, and the effect is not good [7, 8, 9].

2.1.2 Formulating user loyalty program

The chain businesses or cooperative enterprises for provide a series of purchase discounts, value-added services or other incentives to customers with frequent consumption, which aims to reward loyal customers, stimulate consumption and retain core customers, however , the application of social networks is limited [10].

2.1.3 User loyalty modeling

In recent years, more and more research has been done. For example, Ni Jing constructed a complex network, using act degree to measure user loyalty [11]. Ren Jianfeng and other genetic algorithms used to influence the customer loss factor screening [12]. Xu Xiangbin et al. established the RFP model, obtaining the grading of user loyalty [13]. In this paper, the third way is used to express user loyalty.

2.2 Typical loyalty and typical disloyalty

In this paper, the online social networks and the loyalty relationship is abstracted as a loyalty graph, and a formal description is as follows: Definition 1. Loyalty is a function based on time, UL=f (L, t), wherein, L refers to the user value of a certain time, t refers to time, and t∈［m,n］.

Typical loyalty: When the curve trend of user loyalty is flat, and the average value is greater than L, the user is called typical loyalty; typical disloyalty: When the curve trend of user loyalty is not balanced, and the service time is not continuous, or the average is close to m, the user is called a typical disloyalty.

2.3 Global loyalty and local loyalty

In the calculation model of loyalty, there are two kinds of models, global loyalty and local loyalty. Global loyalty refers that each user has its own loyalty value in a social networking site, and this value is calculated from a global point of view. Local loyalty refers that a user’s local loyalty value is different from other users. The value of a user loyalty or disloyalty in a social network can affect the relevant group, resulting in increased loyalty or disloyalty of surrounding users. Global loyalty and local loyalty have their own characteristics, and are summarized in Table 1.

Table1: Comparison of global and local loyalty

Content	Global loyalty	Local loyalty
Formalized definition	$C: L \rightarrow [0,1]$	$C: L \times L \rightarrow [0,1]$
Computing cost	Less	Great
Computational accuracy	Higher	Lower
Computational complexity	Higher	Lower

3 DYNAMIC USER LOYALTY MODEL

3.1 Building of user loyalty model

RFM model [17] is a widely used customer value analysis model in various industries. The study found that the smaller R is (or the greater of F or M value), customers are more likely to reach a new deal with an enterprise, commonly used in data mining customer segmentation. The specific meaning of the RFM model is as follows:

R(Recency): The number of days, weeks, months lasted from recent purchasing date to statistics day.

F(Frequency): The frequency of consumption in the most recent period.

M(Monetary): Monetary in the most recent period.

Initial researches on loyalty focus on repeated purchases or sustained use behavior, and later studies gradually tale the emotional factors into account, stating loyalty from a more comprehensive perspective [18], the attitudinal (emotional attachment) and behavioral perspective (purchase repetition). Thus, the loyalty of user i in a social network is not only reflected in the value contribution to a community, but also in the behavioral contribution of a community.

Definition 2: User i’s loyalty in the moment of t:

$$\mathrm{UL}(\mathrm{i},\mathrm{t}) = \varepsilon_1 \mathrm{RFM}_{\mathrm{Behave}} + (1 - \varepsilon_1)\mathrm{RFM}_{\mathrm{Business}} \qquad (1)$$

$\mathrm{RFM}_{\mathrm{Behave}}$ represents the users’ participation value; and $\mathrm{RFM}_{\mathrm{Business}}$ represents the users’ consumption value. $\varepsilon_1$component value method is a difficulty of the value model, there is no weight vector suitable for all industry background.

3.1.1 User participation value

User participation value is embodied in the willingness of reuse. The intention refers that a user will visit the site again for a long period of time in the future.

Definition 3: User participation value

$$\mathrm{RFM}_{\mathrm{Behave}} = \alpha_1 \mathrm{R} + \beta_1 \mathrm{F} + \gamma_1 \mathrm{M} \qquad (2)$$

The meaning of each parameter is as follows: R: The interval of visiting the site the last time; F: the frequency of visiting in the most recent period; M: the number of collection/label/comment in the most recent period; α_1, β_1, γ_1: value of each component based on the characteristics of the industry background.

3.1.2 User consumption value

The users' consumption value is reflected in the degree of repeat consumption (free or paid), which is the amount of time that the user spent on the social networking site.

Definition 4: User consumption value:

$$\mathrm{RFM}_{\mathrm{Businesss}} = \alpha_2 \mathrm{R} + \beta_2 \mathrm{F} + \gamma_2 \mathrm{M} \qquad (3)$$

In which, R: interval of last reading / buying / ... frequency; F: Reading / purchasing / ... frequency in the most recent period; M: Number of reading / purchasing / ... in the most recent period; According to expert,$\varepsilon_1 = 0.5$, and α_1, β_1, γ_1 and α_2, β_2, $\gamma_2$is 0.2,0.5.0.3.

3.2 Discovery of typical users

The discovery of typical users refers to the discovery of typical loyal and typical disloyal users. For an SNS site, analysis of loyalty is in favor of analyzing the user experience, for example, user counts of a SNS site showed continuous decline for 13 consecutive months in Fig. 1.It needs to study which kind of users have been losing and which are typical loyal users and whether more users will become typical loyal users of SNS.

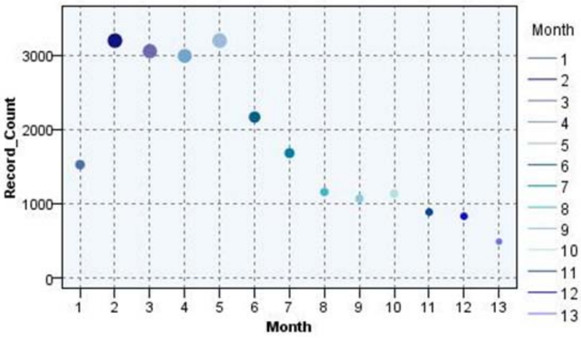

Fig. 1. User counts of a SNS site for 13 consecutive months

According to the formula (1)(2)(3), the loyalty value of a given user can be calculated. The user loyalty trends can be expressed by the evolution of user loyalty values on the time axis. On this basis, a clustering analysis is conducted according to the mean and standard deviation of user loyalty within a period of time and number of visits to find out partial users who continue to use with high loyalty and users who cannot continue to use with low loyalty. The users similarity measurement is used to find more typical loyaland typical disloyal users.

3.2.1 Selection of scale loyalty curve

Choosing standard curve mainly considers distribution characteristics of the curve.The mean value, standard deviation, skewness and kurtosis can be considered as an indicator of the standard curve. A user loyalty mean value over a period of time is calculated as follows:

$$\overline{\mathrm{UL}(\imath, \mathrm{t})} = \frac{1}{\mathrm{N}} \sum_{\mathrm{t=m}}^{\mathrm{t=n}} \mathrm{UL}(\mathrm{i}, \mathrm{t}) \qquad (4)$$

The standard deviation of the user loyaltycurve over a period of time is calculated to express its smoothness, and to illustrate the degree of deviation of the possible value of time curve to its average value. It is a measure of the fluctuation of time curve in the mean value. The larger the value is, the more distant from the past the average value and more unstable; on the contrary, the smaller the value is, representing a more stable state. Suppose that UL (i, t) is the loyalty degree of user i at time t, its smoothness is calculated as follows:

$$\sigma(i) = \sqrt{\frac{1}{N}\sum_{t=m}^{t=n}(UL(i,t) - \overline{UL(\imath,t)})^2} \qquad (5)$$

3.2.2 Searching of similar loyalty curve

The similar loyalty curve search is mainly used to determine the correlation between the value change and time series of typicalloyal users and typical disloyal users, and take it as the basic criterion for discovering users. In this paper, the user loyalty curve is based on the time series of some periods of time. Searching the similarity of time series involves the main issues including: similarity measure, and the search index strategy, etc. Suppose $UL(i,t) = \{UL(i,t_1), UL(i,t_2), \dots UL(i,t_N)\}$ is the time series of user loyalty, T=$\{t_1, t_2, \dots t_N\}$is the value on the time axis, ULDB=$\{i \in U, t \in T | UL(i,t)\}$is the user loyalty curve, and U represents all the users. Thus, the loyalty similarity searching is presented as follows:

Given a scale loyalty set D, a time series database ULDB, similarity measure function sim () and similarity search strategy Find (), the similarity search is to find out the sequence set R that is similar to scale loyalty D in the database of ULDB, namely:

$$R = \{x \in ULDB | Find(sim(D,x), ULDB\} \qquad (6)$$

Similarity measure is the basis or criterion of similarity between two sequences. It is the basic of similarity search. The variation range of time series studied in this paper is small, so we use Euclidean distance [19]. Suppose U (i, t) and UL (j, t) are two time series of user loyalty, t ∈ [m, n] whose length is L, then the Euclidean distance between them is defined as:

$$D(UL(i,t), UL(j,t)) = \sqrt{\sum_{t=m}^{t=n}(UL(i,t) - UL(j,t))^2} (7)$$

Time series contains a large amount of data, in order to improve the efficiency of search, indexing is needed for the time series. Because the similarity measure is based on the Euclidean distance, its indexing method adopts the spatial index structure, such as R-tree.

3.3 Discovery of potential users

In the social network, users exist in a complex network. They can communicate with other users, and can form a social network. Each node represents a user, and node size indicates the users' loyalty value; different nodes build relationships through common acts (such as co-consumption, forwarding, comments, etc.) and to characterize the intensity of the relationship based on the cumulative co-consumption or behaviors. In the social network, typical loyal users and typical disloyal users are used as the initial user set to study the impact of the initial user set, so as to effectively determine the user classification.

3.3.1 Division of user community

Therefore, using typical loyal users and typical disloyal users as the initial user set has become the key object for us to tap potential typical loyal and disloyal users. But not all users in the community are likely to be affected, for example, if the community size is too large or internal relations are not close within the community, the impact of individual loyal and disloyal users cannot be spread to the entire community.

Definition 7: For the network graph G (V, E, W), V is the set of nodes in the graph, i.e., the set of users in the social network, E is the set of edges, W is the weight between the nodes, and the weight between the nodes records the behavior of the two nodes.

Community division is based on modular optimization [20]. The algorithm can be divided into two stages, namely, division and folding, and repeated iterations. Suppose the social network has N nodes, the specific process is as follows:

Stage I: Community division

Initial state: Each node is assigned a community label, the network has N communities. For each user node i, its adjacency node j should be considered; suppose the community of i becomes community of j, and calculate how this action affects the value of the modularity degree. If this change is positive, we will accept the change, otherwise, the original distribution should be remained.

$$\Delta Q = \left[\frac{\sum_{in} + 2k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \quad Q \in [-1,1] \qquad \text{Formula (8)}$$

$\sum_{in}$ represents sum of the connection weights of the community; $\sum_{tot}$ indicates the sum of weights of all edges connected to the community; $k_i$is the sum of weights of all edges connected to node i; $k_{i,in}$ denotes the sum of weights of all the nodes in the community from i to other nodes; M is the total weights of the entire network connection; this division will maximize Q, namely, the higher the degree of modularity, the better. When the whole process cannot be extended, it will stop.

Stage II: Folding in the same community

Folding the same community and a new network is formed after folding, in which: the connection weight between the communities is the sum of weight connecting nodes of the two communities; the connection within the community forms a ring, and the weight is the sum of internal connection of the community.

3.3.2 Discovery of potential affected users

Discovery of potential affected users refers to users who are affected by typical loyal and typical disloyal users. We use the independent cascade model algorithm [21,22] to discover the potential users. In order to reflect the relationship between user value and time, we introduce the time factor as the weight on the basis of the independent cascade model to dynamically observe the changes of potential users. The weight of the time factor and the probability of propagation in the independent cascade model are combined to form a weighted independent cascade model. Assuming typical loyal users and typical disloyal users are the most influential nodes, as the initial set of users, the neighbor node that can be activated by each of the typical loyal user nodes and the typical disloyal user nodes is calculated by the weighted independent cascade model.

Definition 8: The ability of nodes to influence other nodes is called influence. An influential node can activate other nodes; and non-influential nodes cannot activate other nodes.

Definition 9: The time factort_{ij}: represents the number of times that node $v_i$interacts withnodev_j in the period T, as follows:

$$t_{ij=} \frac{\sum_{i=m}^{n} \text{times}}{|T|} \quad T \in [m, n] \qquad (9)$$

Definition10: Influence propagation probability$\lambda_{i,j}$: represents the propagation probability from the node v_i to nodev_j, as follows:

$$\lambda_{ij} = \varepsilon * t_{ij} * \frac{w_{ij}}{w_{imax} + w_{imin}} \qquad (10)$$

The influence propagation probability can be defined according to different industry backgrounds. The initial value of ε is 1 and the next is the result value of the previous node operation. wij represents the weight of the node Vi to the node Vj, Wimax represents the maximum weight in the nodes from node i and Wimin represents the minimum weight in the nodes from node i. During the propagation of the independent cascade model, at time t, whether node V succeeds in activating its neighbor node or not, at a later time, V itself is still active, but it no longer has the influence. A node that is activated at time t is still active at time t + 1 after trying to activate its neighbor node. However, it can’t activate any other nodes. This type of node is called non-influential active node. When there is no influential active node in the network, the propagation process ends [22].

4 EXPERIMENTAL COMPARISON and ANALYSIS

4.1 Experimental framework

Aiming at the user discovery algorithm based on user loyalty, the experimental framework is shown in Fig.2.

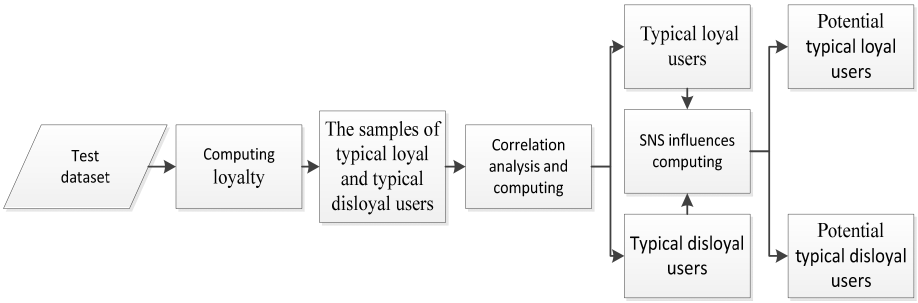
Test dataset
Computing loyalty
The samples of typical loyal and typical disloyal users
Correlation analysis and computing
Typical loyal users
SNS influences computing
Typical disloyal users
Potential typical loyal users
Potential typical disloyal users

Fig.2. Experimental Framework of User Discovery Algorithm based on User loyalty

This paper tests on a real online social network data set. The data set is the data of a social network platform from July 2010 to July 2011, a total of 12 months of data, which contains about 23314 users, 2050627 piece of microblog information, 184400 pieces of reply messages. The loyalty network is built according to users' posts and replies.

4.2Analysis of results

4.2.1 Discovery of typical users

The dynamic user loyalty model was used to calculate the user value in each segment, taking the length of one year as the time period, or 12 months. A total of 4723 users were analyzed removing invalid data. The scale typical loyal users are obtained using formula(1)~(5). Finally, we get 94 typical loyal users, and 85 typical disloyal users.

4.2.2 Discovery of potential typical users

The social network is a directed network composed of 5667 user nodes and 33818 edges. The average size is 5.968 and the average path length is 4.11. Then, it is divided into 41 communities using the community-partitioning algorithm, as shown in Fig. 3. From partial results of node propagation of UserID=3330, the potential UserID=9368 is selected in the primary transmission, the potential UserID=1327,UserID=6111 is selected in the secondary transmission according to the threshold decision, and these three users are not in the similar curve, as shown in Table 2.

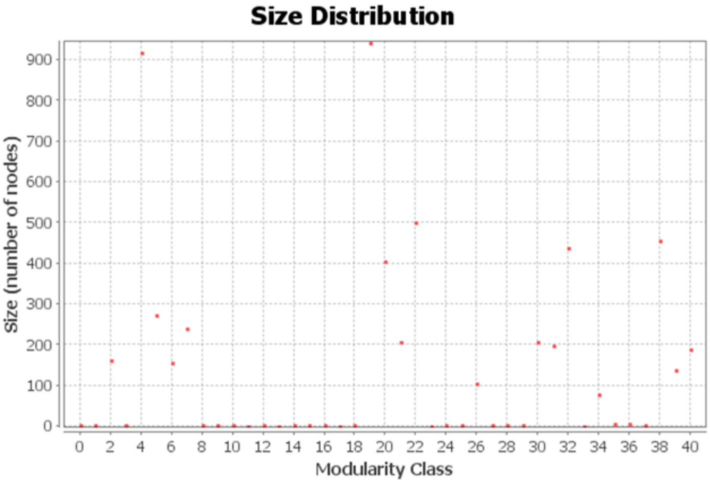

Fig.3. Community Division Result of a SNS

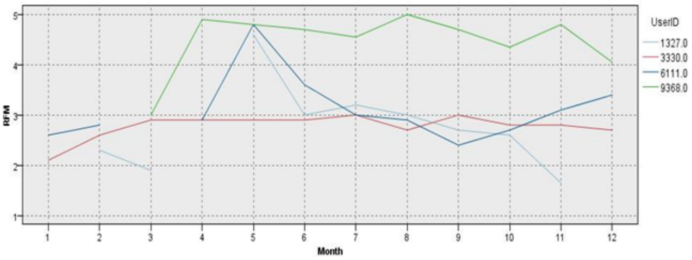

Fig.4. Potential Users of UserID=3330

Table2: Node transmission result of user ID=3330

UserID	***Link UserID***	ε	$\frac{w_{ij}}{w_{imax}+w_{imin}}$	t_{ij}	λ_{ij}	***UserID***	***Link UserID***	ε	$\frac{w_{ij}}{w_{imax}+w_{imin}}$	t_{ij}
3330	9368	1	0.37777778	0.916667	0.346296	9368	3399	0.346296	0.0075	0.833333
	5589	1	0.71111111	0.25	0.177778		3416	0.346296	0.01995012	0.333333
	3262	1	0.28888889	0.333333	0.096296		4597	0.346296	0.01496259	0.75
9368	298	0.346296	0.99492386	0.166667	0.057423		4618	0.346296	0.01995012	0.916667
	299	0.346296	0.535	0.083333	0.015439		9562	0.346296	0.00997506	0.916667
	3253	0.346296	0.01	0.916667	0.003174		1327	0.346296	0.98756219	0.75
	3276	0.346296	0.015	0.916667	0.004762		6111	0.346296	0.5	0.916667
	3284	0.346296	0.11	0.666667	0.025395					
	3310	0.346296	0.27	0.166667	0.015583					
	3335	0.346296	0.01	0.916667	0.003174					
	3350	0.346296	0.015	0.75	0.003896					
	3398	0.346296	0.0125	0.583333	0.002525					

Experimental results show that based on the standard curve, the typical loyal users and the typical disloyal users cannot be fully covered. Then explore potential users according to the relationships among users in the social network, as shown in Figure 4. Potential loyal users are discovered through UserID=3330 in the social network. We can effectively discover and classify users with methods introduced in this paper.

5 CONCLUSION

The main contribution of this paper is to propose a user discovery method based on user loyalty in social networks, which dynamically quantifies user loyalty through the double RFM model, and classify users through clustering analysis and influencing propagation model in social networks. First, the loyalty standard curve is obtained by clustering analysis and standard deviation analysis according to the user loyalty degree by using the double RFM model;then, using the similarity degree of time series to find out more typical users having similar standard curve;finally, calculate the potential users based on the initial set of typical loyal and typical disloyal users that have been found, and using influencing propagation model of SNS. In the future research, we will continue to explore the quantitative research of user loyalty in SNS, for example, we will further consider the impact of users' emotional value on SNS website loyalty, and how to collect the data of users' participation in SNS more conveniently.

References

[1] Ning Lianju, Zhang Yuhong, An Empirical Study of the Impact of Sense of Virtual Communities on User loyalty [J]. Technical Economy, 2014,11:7-15+35.

[2] http://baike.baidu.com/view/341649.htm

[3] Huang Wanqiu, A Potential Losing Customer Detection Method Based on Social Network Services [J]. Journal of Beijing Jiaotong University,2014,03:123-127.

[4] Zhao Jian, The Influence of Consumption Value of Virtual Communities on Brand Image and Community Loyalty [J]. Commercial Times,2014,32:60-62.

[5] Ding Yiqiong, Zhang Song, Research Review of User loyalty of Social Network Service (SNS) [J]. Journal of Information,2013,03:106-112+100.

[6] Zhang Liang, Zhang Di, User Discovery with SNS Influence based on Web2.0 [J]. Journal of Information,2015,06:158-162+173.

[7] Deng Aimin, Ma Yingying, An Empirical Study on the Influencing Factors of User loyalty in Online Shopping [J]. Chinese Journal of Management Science,2014,06:94-102.

[8] Yan Huijuan, Zhang Xingzhou, Liu Zirui, Yu Qi, Personal User loyalty Evaluation Model Based on Transaction Behavior [J]. Modernization of Management, 2015,06:70-72.

[9] Zhou Yun, Zhu Mingxia, Research on the Measurement of Brand Loyalty [J]. Economic Problems,2015,10:92-98.

[10] Zhu Zhiwen, Zhang Li, Impact of Customer Recommendation Program on Existing User loyalty [J]. Business Economics and Management, 2016,01:53-61.)

[11] Ni Jing, Yan Guangle, Ye Lin, Zhong Liangwei, Research on E-commerce Clustering Consumption Propagation Model Based on Complex Networks [J]. Journal of Computer Applications,2011,03:1003-1006.

[12] Ren Jianfeng, Zhang Xinxiang, Research on Modeling and Forecasting of E-commerce Customer Loss [J]. Computer Simulation, 2012,05:363-366.

[13] Xu Xiangbin, Wang Jiaqiang, Tu Huan, Mu Ming, Customer Segmentation Based on Improved RFM Model for E-commerce [J]. Journal of Computer Applications,2012,05:1439-1442.)

[14] Guo Chong, Modeling and Simulation of Online Shopping User loyalty Based on Big Data Analysis [J]. Computer Simulation, 2015,10:239-242+304.)

[15] Gu Bin, Xu Jing, User loyaltyMining based on Knowledge Sharing in Professional Virtual Communities [J]. Information Science,2015,01:105-110.)

[16] Tang Huxin, Simulation Research on E-commerce User loyalty Model [J]. Computer Simulation, 2016,01:413-415+424.

[21] Ma Yin, Research on the Maximization Algorithm of Social Network Influence and Its Propagation Model [D]. Lanzhou University, 2012.

[22] Li Lei, Research on Social Network Influence Model and Its Algorithm [D]. Beijing Jiaotong University, 2010.