Application of Association Rule Mining Theory in Sina Weibo

Journal of Computer and Communications, **2014, 2, 19-26**
Published Online January 2014 (http://www.scirp.org/journal/jcc)
http://dx.doi.org/10.4236/jcc.2014.21004

Scientific Research

Application of Association Rule Mining Theory in Sina Weibo

Xiao Cui, Hao Shi, Xun Yi

College of Engineering and Science, Victoria University, Melbourne, Australia.
Email: xiao.cui1@live.vu.edu.au

Received November 21st, 2013; revised December 17th, 2013; accepted December 24th, 2013

ABSTRACT

A user profile contains information about a user. A substantial effort has been made so as to understand users' behavior through analyzing their profile data. Online social networks provide an enormous amount of such information for researchers. Sina Weibo, a Twitter-like microblogging platform, has achieved a great success in China although studies on it are still in an initial state. This paper aims to explore the relationships among different profile attributes in Sina Weibo. We use the techniques of association rule mining to identify the dependency among the attributes and we found that if a user's posts are welcomed, he or she is more likely to have a large number of followers. Our results demonstrate how the relationships among the profile attributes are affected by a user's verified type. We also put some efforts on data transformation and analyze the influence of the statistical properties of the data distribution on data discretization.

KEYWORDS

Association Rules; User Profiles; Sina Weibo; Social Network

1. Introduction

Online social networks such as Facebook, Twitter and Google+ have become an integral part of people's daily lives. No matter how they differentiate from one another, user profiles are a key feature. A user profile may include but not be limited to gender, age, location, occupation, social contacts, etc. The availability of the information may vary from one site to another. In spite of the fact that user profiles are less dynamic than other online behaviors, they still provide a clear signal of users' characteristics. A substantial effort has been made recently in order to obtain knowledge about users from their profile data. Lampe *et al*. [1] found that profile completion percentage on Facebook has a positive relationship with the number of friends a user has. Mislove *et al*. [2] proposed an algorithm to infer the missing part of a user profile according to other similar profiles. Quercia *et al*. [3] conducted a study on the relationship between the Big Five personality traits and user behaviors on Twitter. They introduced a novel method to predict the personality based on the number of followers, followings and tweets a user has.

As Twitter is banned in China, Sina Weibo is considered a replacement for it. Sina Weibo has reached 56 million daily active users (who spend an average of one hour per day with the service) [4]. Sina Weibo has had a significant influence on Chinese society. Unlike its predecessors, studies on Sina Weibo are still in an initial state. There are a few studies on Sina Weibo with regard to user profiles. Guo *et al*. [5] found that the connections between users are mostly one-way and the number of followers a user has changes very fast. Chen and She [6] carried out a similar study but compared verified users with unverified ones. They believed that users whose real identity has been verified are more likely to have greater influence. Wang *et al*. [7] examined the correlation between the number of followers, followings and posts. They found that the number of followers grows rapidly as the number of followings increases from 10 to 3000. They also stated that the increase in posts can lead to more followers as long as the number of posts does

not exceed 20,000.

Although considerable attention has been paid to Sina Weibo, associations among different profile attributes, such as the association between the number of reposts and comments, have not been well examined yet. Due to the fact that a large number of users on Sina Weibo have been verified according to their professional background, people on Sina Weibo are more likely to act responsibly and engage honestly with the community. It is worthwhile to explore users' characteristics on Sina Weibo especially considering they have different verified types (e.g. local authorities, news agency, and celebrity). Our research is based on a set of first-hand data collected from Sina Weibo, containing 1,192,972 users' profiles. The major contributions are summarized as follows:

- Continuous data (e.g. the number of followers) are replaced by meaningful labels (e.g. the grass roots and social star).
- The influence of the distribution of the data on data discretization is analyzed.
- Association rule mining is conducted with respect to users' verified types.
- A comparison between different types of users is made.

The rest of the paper is organized as follows. Section 2 presents the data model used in this paper. Definitions such as the number of followings a user has are given. Section 3 explains the process of data collection. The social relationships among users in Sina Weibo are illustrated. Section 4 discusses the methods for data discretization. The statistical properties of the data distribution are taken into consideration. Section 5 introduces an Apriori-based method for association rule mining and explains how we are going to conduct the association rule mining in Sina Weibo. Empirical results are given in Section 6 and conclusions are drawn in the last section.

2. Data Model

The information in a user profile may include various attributes of a user such as geographical location, academic and professional background, interests, preferences, etc. The availability of such information may vary from one site to another. In terms of microblogging, *i.e.* Sina Weibo, the number of followers, followings and posts a user has are three indispensable parts of a user profile. Such information is always displayed at a prominent place. Besides, a verified type is added to a user profile as users on Sina Weibo may choose to verify their identity based on their professional background. In this paper, a user profile is defined as follows:

profile(*uid*) = {*username*, *province*, *gender*, *number of followers*, *number of followings*, *number of posts*, *number of reposts*, *number of comments*, *verified type*, *time since created*}

Each user has a unique identification number (*uid*). The core attributes of a profile are defined as follows:

- *NoA* refers to *number of followers*. *NoA*(*uid*) is the total number of audience who are listening to the broadcast of user *uid*. *NoA* is one of the major signs of a user's popularity.
- *NoB* refers to *number of followings*. *NoB*(*uid*) is the total number of broadcasts to which user *uid* is listening. Sina Weibo enforces that a user can listen a maximum of 2000 broadcasts.
- *NoP* refers to *number of posts*. *NoP*(*uid*) is the total number of posts that user *uid* updates. *NoP* can be a good indicator of a user's activeness.
- *NoR* refers to *number of reposts*. *NoR*(*uid*) is the total number of reposts that others forward from user *uid*. *NoR* is a sign of the capability a user has to spread out the information.
- *NoC* refers to *number of comments*. *NoC*(*uid*) is the total number of comments others leave on user *uid*. *NoC* can reveal the likelihood of a user to initiate a hot topic.
- *VT* refers to *verified type*. *VT* includes *red star* (an ordinary user whose real identity is verified), *beauty*, *e-celebrity*, *corporation*, *government*, *media*, *organization*, *campus*, *application software*, and *website* [8]. For example, user Xinhua News Agency, the official press agency of China, is classed as *media*.
- *TsC* refers to *time since created*.

3. Data Collection

Users' profiles are collected through the REST API provided by Sina Weibo. Bilateral relationships are used to expand the search of new users. Social relationships among users are defined as follows (see **Figure 1**).

Scenario 1 indicates that *uid*1 and *uid*2 have no connection between them. Scenario 2 shows that *uid*1 is a follower of *uid*2. Scenario 3 explains bilateral friendships where *uid*1 is a follower of *uid*2 and *uid*2 is also a follower of *uid*1. We assume that is two users follow each other, they are considered friends.

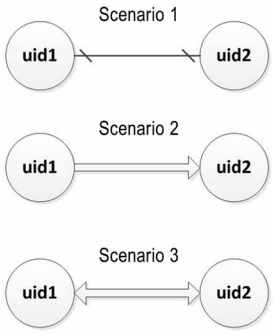

Figure 1. Social relationships.

Getting the friends of a friend is the strategy used in this paper to obtain users' ids from Sina Weibo. The REST API provides facilities to retrieve profile information according to a user's id. The implementation details are given below (see Table 1).

Unlike studies [4-6] where a user's followings are used to expand the search of new users, bilateral relationships are used in this study. Users who follow each other seem to have a closer relation between them. This method can prevent the search of new users from the spammers because no one likes to subscribe a spammer's microblog.

Finally, 1,192,972 users' profiles are retrieved. 39.58% of them are verified users. *Red star* and *e-celebrity* account for 91.08% of verified users (see Figure 2).

4. Data Discretization

Data mining process involves a preprocessing step in order to assure the data have the quality and the format required by the algorithm. Users are classified by their attributes. For example, according to *NoA*, users are classified into two groups: *the grass roots* and *social star*. Users in the latter group have much more followers than users in the former one. Other continuous data are replaced as well in a similar way (see Table 2).

Table 1. Pseudocode for data collection.

```
enqueue i in q
while q is not empty do
    get friends_uids according to i
    for each j in friends_uids do
        if j does not exist in q then
            insert j into q
        end
    end
    dequeue k from q
    get profile according to k
    set i = k
end
```

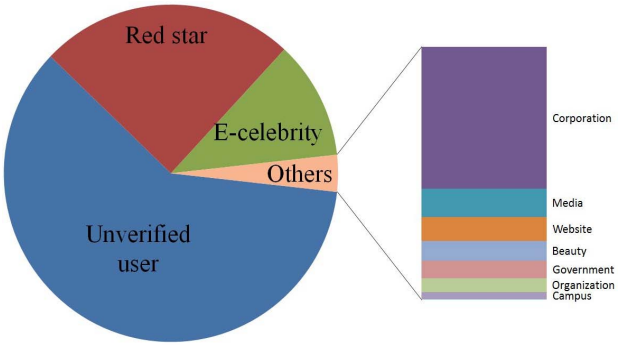

Figure 2. The proportion of users.

4.1. The *K*-Means Method and the Pareto Principle

This paper experimented with two methods: the *k*-means clustering algorithm and the Pareto principle. The purpose of clustering is to search for similar examples and group them into clusters such that the distance between examples within cluster is as small as possible and the distance between clusters is as large as possible [9]. Let $P=\{p_1,p_2,\cdots,p_n\}$ be a set of data points to be clustered and k is the number of clusters (Here, $k=2$). Randomly select k data points from P as the initial centroids of the clusters, $C=\{c_1,c_2\}$. Then, following steps are repeatedly performed until the convergence is obtained: 1) Assign each data point $p_i \in P, i=\{1,2,\cdots,n\}$ to the closest centroid either c_1 or c_2. 2) Re-compute the centroids of the clusters $C=\{c_1,c_2\}$. Centroid is the mean of the points in cluster.

The Pareto principle (also known as the 80 - 20 rule) [10] originally referred to the observation that 80% of Italy's wealth belonged to only 20% of the population. Here, we assume that, for example, a user whose followers are more than 80% of the other users is classed as *social star*. The quantile function used to calculate the cut points between the groups (e.g. *the grass roots* and *social star*), is defined as follows [11]:

$$F^{-1}(p)=\min\{X\in\mathbb{R}:F(X)\geq p\},\ p=0.8 \qquad (1)$$

Here, X may refer to one of the variables in Table 2. The distribution function of X is given by $F(X)=P(X\leq x)$ where $F(X)$ represents the probability that X is less than or equal to x. Equation (1) determines the place where 80% of the data lies below it, e.g. 80% of *NoA* is less than or equal to 1140% and 80% of *NoR* is less than or equal to 294.

4.2. Discretization Index

In this paper, a discretization index (di) is proposed to measure the quality of the discretization produced from above methods. Let $X=\{x_1,x_1,\cdots,x_n\}$ be a set of data points to be split. Suppose X is partitioned into two groups $G=\{g_1,g_2\}$. A di is defined as follows:

$$di=\max_{i\in\{1,2\}}(\delta_i)\sum_{j=1}^{2}\sum_{x_k\in g_j}(x_k-\mu_j) \qquad (2)$$

where δ_i denotes the proportion of g_i in X and μ_j denotes the mean of data points in g_j. The method with the smallest di is considered the best method based on the following criteria: 1) Minimize the distances within the clusters and maximize the distances between the clusters. 2) Split as equally as possible. The reason why both criteria are needed is that using the first criterion (*i.e.*, clusters that are coherent internally but clearly different from each other) alone to split the data may cause an extremely uneven partition (see Section 4.3). As asso-

Table 2. Data discretization.

Continuous Data	Statistical Properties		Partition			*di* (10^9)	
	Standard deviation	*Skewness*	*Class*	*Quantity*[a]	*Interval*[b]	*K-means*	*Pareto*
NoA	219583.30	116.64	The grass roots	1,192,865	[1.1140)	15.31	12.67
			Social star	107	[1140.63717128]		
NoB	485.80	2.00	Self-centered	1,029,534	[1.894)	0.19	8.06
			Scout	163,438	[894.2000]		
NoP	3032.53	12.25	Lurker	1,152,976	[1.2034)	12.36	0.88
			Blog zealot	39,996	[2034.413549]		
NoR	62489.64	131.02	Valuelss	1,192,897	[0.294)	4.10	3.44
			Propagator	75	[294.18645439]		
NoC	56012.94	353.34	Uninterested	1,192,939	[0.206)	3.05	2.22
			Topic inititator	33	[1206.39344085]		

[a]Calculation was based on the partitions generated from the k-means method; [b]Calculation was based on the partitions generated from the Pareto principle.

ciation rules are generated from frequent itemsets (see Section 5), data in the minority, for example, 200 *social star* users in 1,192,972 users, are very likely to be overlooked. More explanations for why partitioning as evenly as possible is important to this study are given in Section 5. We propose *di* aiming to build a balance between the criteria.

4.3. Comparison between the Methods

We found that the use of discretization methods depends on the statistical properties of the data distribution. The spread of the data (*i.e.* standard deviation) and the symmetry of the data (*i.e.* skewness) may have significant influence on the performance of the discretization. Higher standard deviation implies greater spread of data. Positive values for the skewness indicate that the distribution is skewed right. Higher skewness implies longer tail in the right side. A normal distribution has a skewness of 0. We found that the *k*-means method is very good at creating clusters coherent internally but different from each other. However, the *k*-means method tends to partition data in an extremely uneven way when the distribution is skewed (see **Table 2**). On the other hand, partition based on the Pareto principle produces a lower *di* in most cases (see **Table 2**). Data are partitioned in a 80-20 way without impairing the internal coherence and the external difference of the clusters.

We use examples to illustrate how the statistical properties of the data distribution can have influence on the data discretization. As shown in **Figure 3**, the distribution of *NoB* is much closer to a normal distribution with a skewness of 2, at the same time it has the lowest standard deviation compared with other variables. In this case, the *k*-means method produces a lower *di* than the Pareto principle. In comparison, data points in *NoA* are spread out over an extremely large range of values 1 to 63,717,128. A skewness of 116.64 indicates that the distribution of *NoA* has a very long tail at the right side (see **Figure 3**). As a consequence, the majority of data points in *NoA* fall within a very small range of values and very few of data points fall within an extremely wide range of values. Actually, 80% of the data points in *NoA* fall within the interval [1, 1140) and the rest falls within the interval [1140, 63,717,128]. In this case, the *k*-means method tends to group almost all data points into one cluster and put the rest into another one. Actually, only 0.01% of users were classed as *social star* in the *k*-means method (see **Table 2**). Partition based on the Pareto principle is applied in this study because it makes a trade-off between the criteria.

5. Mining Association Rules in Sina Weibo

5.1. Association Rule Mining

The association rule mining can be conceptualized as follows [9]: Let $f=\{I_1, I_2, \cdots, I_n\}$ be the set of all items. Let DB be a set of database transactions where each transaction T is a set of items such that $T \subseteq f$. Let A be a set of items. A transaction is said to contain A if and only if $A \subseteq f$. An association rule is an implication of the form $A \Rightarrow B[s,c,l]$, where $A \subseteq f$, $B \subseteq f$, $A \cap B = \varnothing$. The support s, confidence c and lift l of the rule $A \Rightarrow B$ are defined as:

$$s = P(A \cup B) = F(A \cup B)/|DB| \quad (2)$$

$$c = P(B|A) = F(A \cup B)/F(A) \quad (3)$$

$$l = P(A \cup B)/P(A)P(B) \quad (4)$$

where $F(A)$ stands for the number of transactions containing the set X in DB and $|DB|$ denotes the total number of transactions in DB. Rules with the support more than a minimum support threshold $s_{\min}$ and the confidence more than a minimum confidence threshold $c_{\min}$ are called strong. A set of items is refe-

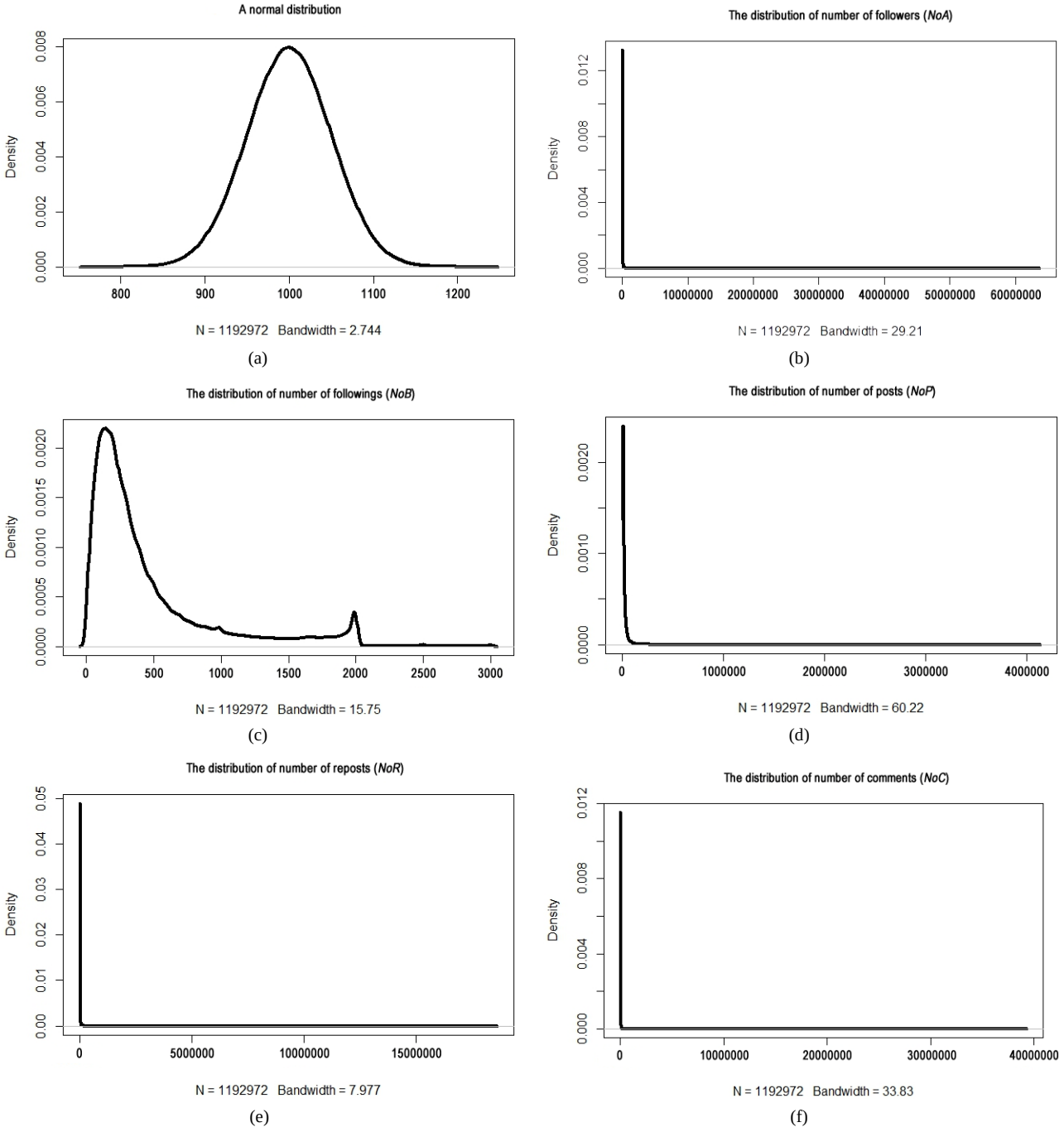

Figure 3. Data distribution.

reed as an itemset. An itemset that contains (k) items is a k-itemset. The support count of an itemset is the number of transactions containing the itemset. The minimum support count is defined as $s_{\min} \cdot |DB|$. An itemset is frequent if its support count is not less than the minimum support count.

5.2. Apriori Algorithm

Apriori is an influential algorithm for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses the Apriori property, *i.e.*, all nonempty subsets of a frequent itemset much also be frequent. Let L_i be the set of frequent i-itemsets. Given L_{k-1}, Apriori algorithm finds L_k using join and prune actions as follows: 1) Join: To find L_k, a set of candidate k-itemsets, denoted as l_k is generated by joining L_{k-1} with itself. Any two (k − 1)-itemsets A and B are joinable if they contain (k −

2) common items. For example, $A=\{x_1,\cdots,x_{(k-2)},x_{(k-1)}\}$ and $B=\{x_1,\cdots,x_{(k-2)},x_k\}$ are joinable. The resulting candidate *k*-itemset is $\{x_1,\cdots,x_{(k-2)},x_{(k-1)},x_k\}$. 2) Prune: l_k can be huge. To reduce the size of l_k, the Apriori property is used as follows. If any $(k-1)$-subset of an candidate *k*-itemset is not in L_{k-1}, the candidate cannot be frequent either and so can be removed from l_k. The set of remaining candidates in l_k is a superset of L_i, that is, its elements may or may not be frequent, but all of the frequent *k*-itemsets must be included in l_k. A scan of the database to determine the count of each candidate in l_k would result in the determination of L_k, *i.e.*, all candidate having a count no less than the minimum support count are frequent and therefore belong to L_k. By Apriori algorithm, all frequent itemsets along with their support counts can be found efficiently.

5.3. Experimental Design

Suppose dataset *U* contains all the data collected from Sina Weibo. Association rules are mined from both *U* and its subsets.

Considering the property of association rule mining described above, rare types of users are very likely to be pruned due to their relatively low support counts. Splitting *U* into disjoint subsets based on *VT* and mining association rules from them separately is necessary so as to avoid overlooking some interesting patterns that are hidden in the rare types of users. The dataset *U* is divided into 2 subsets: *verified_accounts* and *unverified_accounts*. A comparison in terms of association rules, between *verified_accounts* and *unverified_accounts*, is made to identify the difference between verified users and unverified users. If necessary, the dataset *verified_accounts* can be further divided according to *VT*. In this paper, a comparison between *red star* and *e-celebrity* is conducted for two reasons: 1) *red star* and *e-celebrity* together account for 91.08% of verified users. 2) *red star* refers to the masses, opposite to *e-celebrity* who are public figures and professionals and well known in local communities.

Considering the fact that users who have large number of followers (followings, posts, reposts, and comments) only account for a very small part, we have to give a relatively small $s_{\min}$ so as to assure the rules for *social star* (*scout*, *blog zealot*, *propagator*, and *topic initiator*) can be elicited. Association rules are sorted by lift values. A lift equals to 1 means the occurrence of A is independent of the occurrence of B if an association rule is in the form of $A\Rightarrow B[s,c,l]$. A lift is greater than 1 indicates that the occurrence of A has a positive effect on the occurrence of B. We are interested in the profile attributes which are dependent on each other.

6. Empirical Results

We found that both *NoR* and *NoC* play important roles in a user's popularity (see **Figure 4**). If a user's posts are welcomed, either the posts are forwarded by many times or many people leave comments about the posts, the owner of the posts is much more likely to be tagged as a *social star*. Another finding was that *NoC* is positively correlated with *NoR*. 3 of 4 rules for "*NoR* = Propagator" were attributed to "*NoC* = Topic initiator" (*i.e.* Rule #3, 5, and 6). Also, an *e-celebrity* user is always accompanied by a large number of followers (*social star*). Thus, *social star* is a good indicator of *e-celebrity*. We found that a 5-year *social star* user is an *e-celebrity* with a confidence of 54.16%.

A comparison between association rules derived from *unverified_accounts* and *verified_accounts* was made (see **Figures 5** and **6**).

A positive correlation between *NoC* and *NoR* exists in both of them; however, rules in *unverified_accounts* have higher lift than that in *verified_accounts*. In other words, *NoC* and *NoR* are more dependent on one another in *unverified_accounts*. According to above findings, we could state that, for a verified user, an increase in *NoC* may not enhance the probability of an increase in *NoR*. On the other hand, for an ordinary user who has not been verified yet, saying something controversial to receive more comments (*NoC*) is a good way to increase the rate of

	LHS		RHS	Support	Confidence	Lift
1	{NA=Social star, TsC=5th}	=>	{VT=E-celebrity}	0.01295423	0.5416375	4.764781
2	{NB=Scouts, NC=Topic initiator}	=>	{NA=Social star}	0.02841147	0.8927697	4.462610
3	{NB=Scouts, NC=Topic initiator}	=>	{NR=Propagator}	0.02830920	0.8895562	4.445802
4	{NB=Scouts, NR=Propagator}	=>	{NA=Social star}	0.04854192	0.8880250	4.438894
5	{NA=Social star, NC=Topic initiator}	=>	{NR=Propagator}	0.08272386	0.8833265	4.414667
6	{NC=Topic initiator, VT=E-celebrity}	=>	{NR=Propagator}	0.04252835	0.8821024	4.408550
7	{NB=Scouts, VT=E-celebrity}	=>	{NA=Social star}	0.02135931	0.8794740	4.396150
8	{NP=Blog zealot, VT=E-celebrity}	=>	{NA=Social star}	0.03280890	0.8650680	4.324140
9	{NC=Topic initiator, VT=E-celebrity}	=>	{NA=Social star}	0.04168843	0.8646811	4.322206
10	{NA=Social star, VT=Corporation}	=>	{NR=Propagator}	0.01131463	0.8647021	4.321587

$s_{min}=0.01, c_{min}=0.5$

Figure 4. Top 10 rules derived from *U*.

diffusion of his or her posts (*NoR*). Actually, it has already happened in many online social networks where people initiate some controversial topics in order to become famous [12].

Although *red_star* is a disjoint subset of *verified_accounts*, the dependence, in terms of lift, between *NoC* and *NoR* in *red_star* is much stronger than that in *verified_accounts* itself (see **Figure 7**). On the other hand, rules derived from *e-celebrity* are less interesting in terms of lift (see **Figure 8**).

We found that if an *e-celebrity* user's posts are welcomed, then he or she is a *blog zealot* with a confidence greater than 65%. Actually, it happens in many kinds of user types. Unlike *red star* users, users such as *corporation*, *media*, and *application software*, have a strong motive for promoting themselves or something else. As a consequence, they are likely to send as many messages as possible. At the same time, due to their high reputation,

	LHS		RHS	Support	Confidence	Lift
1	{NB=Scouts, NC=Topic initiator}	=>	{NA=Social star}	0.01065137	0.8342025	9.460253
2	{NB=Scouts, NR=Propagator}	=>	{NA=Social star}	0.02242365	0.8255363	9.361974
3	{NA=Social star, NC=Topic initiator}	=>	{NR=Propagator}	0.02723050	0.8816871	8.728310
4	{NB=Scouts, NC=Topic initiator}	=>	{NR=Propagator}	0.01108281	0.8679922	8.592736
5	{NB= Scouts, NP=Blog zealot}	=>	{NA=Social star}	0.01829101	0.6623298	7.511135
6	{NP=Blog zealot, NC=Topic initiator}	=>	{NR=Propagator}	0.02702519	0.6867003	6.798027
7	{NA=Social star, NP=Blog zealot}	=>	{NR=Propagator}	0.02162043	0.6540076	6.474384
8	{NR=Propagator, TsC=4th}	=>	{NC=Topic initiator}	0.02992595	0.7003896	6.410191
9	{Gender=m, NC=Topic initiator}	=>	{NR=Propagator}	0.02477089	0.6412872	6.348458
10	{NP=Blog zealot, NR=Propagator}	=>	{NC=Topic initiator}	0.02702519	0.6728723	6.158344

$s_{min} = 0.01, c_{min} = 0.5$

Figure 5. Top 10 rules derived from *unverified_accounts*.

	LHS		RHS	Support	Confidence	Lift
1	{NB=Scouts, NC=Topic initiator}	=>	{NR=Propagator}	0.05461077	0.8964570	2.551409
2	{NA =Social star, NC=Topic initiator}	=>	{NR=Propagator}	0.16745213	0.8837344	2.515200
3	{NB=Scouts, NR=Propagator }	=>	{NA=Social star}	0.08841976	0.9148386	2.466740
4	{NB=Scouts, NC=Topic initiator}	=>	{NA=Social star}	0.05552790	0.9115121	2.457771
5	{NC=Topic initiator, TsC=5th}	=>	{NR=Propagator}	0.03513491	0.8187158	2.330150
6	{NC=Topic initiator, TsC=2nd}	=>	{NR=Propagator}	0.01130214	0.7978469	2.270755
7	{NB=Scouts, TsC=5th}	=>	{NA=Social star}	0.01784493	0.8397289	2.264217
8	{NA=Social star, TsC=5th}	=>	{NR=Propagator}	0.03774228	0.7831151	2.228826
9	{Gender=m, NC=Topic initiator}	=>	{NR=Propagator}	0.12442097	0.7800648	2.220145
10	{NB=Scouts, TsC=5th }	=>	{NP=Blog zealot}	0.01562093	0.7350743	2.211185

$s_{min} = 0.01, c_{min} = 0.5$

Figure 6. Top 10 rules derived from *verified_accounts*.

	LHS		RHS	Support	Confidence	Lift
1	{NA=Social star, TsC=2nd}	=>	{NB=Scouts}	0.01168051	0.6491129	4.746265
2	{NB=Scouts, NR=Propagator}	=>	{NA=Social star}	0.04027213	0.8317200	4.246981
3	{NB=Scouts, NC=Topic initiator}	=>	{NA=Social star}	0.02562309	0.7727925	3.946081
4	{NB=Scouts, NC=Topic initiator}	=>	{NR=Propagator}	0.02481812	0.7485146	3.556166
5	{NA=Social star, NC=Topic initiator}	=>	{NR=Propagator}	0.06402035	0.7417362	3.523962

$s_{min} = 0.01, c_{min} = 0.5$

Figure 7. Top 5 rules derived from *red_star*.

	LHS		RHS	Support	Confidence	Lift
1	{NB=Scouts, TsC=5th}	=>	{NP=Blog zealot}	0.03181895	0.7480929	2.242229
2	{NB=Scouts, NC=Topic initiator}	=>	{ NP=Blog zealot }	0.06849739	0.6910944	2.071389
3	{NB=Scouts, NR= Propagator }	=>	{ NP=Blog zealot }	0.10027210	0.6526205	1.956073
4	{NR=Valueless, TsC=2nd }	=>	{NA=The grass roots}	0.02864812	0.6668383	1.938533
5	{NA=The grass roots, NC=Uninterested}	=>	{NR=Valueless}	0.24381503	0.8507179	1.917633

$s_{min} = 0.01, c_{min} = 0.5$

Figure 8. Top 5 rules derived from *e-celebrity*.

other users prefer to forward their posts or have discussion with them. Posts are welcomed is independent of having a large number of posts. For this reason, lift values in **Figure 8** are very close to 1.

7. Conclusion

In this study, we explored the relationships among different profile attributes through the techniques of association rule mining. We found that a user is more likely to have a large number of followers (*NoA*) if his or her posts are forwarded by many times (*NoR*) or many people get involved in the discussion he or she initiated (*NoC*). Our results indicate that *NoR* and *NoC* are strongly dependent on each other with respect to ordinary users (both unverified users and *red star* users). Profile attributes for verified users are relatively independent on one another. We also examined both the *k*-means method and the Pareto principle as a method for data discretization. We found that the statistical properties of data distribution can have significant influence on data discretization. Due to the fact that data used in this study are skewed heavily, we suggested using the Pareto principle to partition data.

REFERENCES

[1] C. A. Lampe, N. Ellison and C. Steinfield, "A Familiar Face (Book): Profile Elements as Signals in an Online Social Network," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2007, pp. 435-444.

[2] A. Mislove, B. Viswanath, K. P. Gummadi and P. Druschel, "You Are Who You Know: Inferring User Profiles in Online Social Networks," *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 2010, pp. 251-260.

[3] D. Quercia, M. Kosinski, D. Stillwell and J. Crowcroft, "Our Twitter Profiles, Our Selves: Predicting Personality with Twitter," 2011 *IEEE Third International Conference on Privacy, Security, Risk and Trust* (*Passat*) *and* 2011 *IEEE Third International Conference on Social Computing* (*Socialcom*), 2011, pp. 180-185.

[4] D. Clark, R. Crandall and Y. Mei, "4th Annual China 2.0 Conference Underscores Business Innovation, Social Impact and U.S.-China Links," 2013. http://sprie.gsb.stanford.edu/news/4th_annual_china_20_confe-rence_underscores_business_innovation_social_impact_and_uschina_links_20131022/

[5] Z. Guo, Z. Li, H. Tu and L. Li, "Characterizing User Behavior in Weibo," 2012 *Third FTRA International Conference on Mobile, Ubiquitous, and Intelligent Computing* (*MUSIC*), 2012, pp. 60-65.

[6] J. Chen and J She, "An Analysis of Verifications in Microblogging Social Networks—Sina Weibo," 2012 32*nd International Conference on Distributed Computing Systems Workshops* (*ICDCSW*), 2012, pp. 147-154.

[7] C. Wang, X. Guan, T. Qin and W. Li, "Who Are Active? An In-Depth Measurement on User Activity Characteristics in Sina Microblogging," *Global Communications Conference* (*GLOBECOM*), 2012, pp. 2083-2088.

[8] Sina Open API. http://open.weibo.com/wiki/

[9] J. Han, M. Kamber and J. Pei, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2006.

[10] J. M. Juran and A. B. Godfrey, "Juran's Quality Handbook (Vol.2)," McGraw Hill, New York, 1999.

[11] I. Frohne and R. J. Hyndman, "Sample Quantiles," R Project, 2009.

[12] J. Feng, "Romancing the Internet: Producing and Consuming Chinese Web Romance," Brill, 2013. http://dx.doi.org/10.1163/9789004259720