Research on association rules based on Complex Networks

Yi Zhenwei[1, a], Wei Lingyun[1, b], and Wang Lizhu[1, c]

[1]Beijing University of Posts and Telecommunications, Beijing, P. R. China

[a]yi_zhenwei@163.com, [b]weilingyun2010@sina.com , [c]464665887@qq.com

Keywords: association rules, complex networks, community detection

ABSTRACT. In most cases, the association rules mining (ARM) will generate a large number of rules, most of which are of no interest. In order to find the interesting rules from the large-scale association rules more effectively, this manuscript presents a new method to research the association rules based on Complex Networks. The proposed method first illustrate the association rules by complex networks, and then puts forward a kind of community detection algorithm, which is aim to get the relationship and classification of the result rules. Thus we can locate the interesting patterns more quickly and accurately. Meanwhile, this method also provides a visualization process of the association rules. The experimental results demonstrate that this new method with complex networks can group the result rules and find out the interest patterns more intuitively and efficiently.

1. Introduction

With the advent of the era of big data, association rules mining [1] is becoming more and more important task of data mining techniques. The goal of ARM is to find out the potential relationships among data items by extracting frequent patterns from the given database. In 1993, R. Agrawal et al. proposed the idea of association rules firstly, shortly after that put into practice by the well-known Apriori algorithm. Thereafter, a lot of research on the association rules mining has been conducted. These studies mainly concentrated in improving the efficiency of algorithm, such as the introduction of random sampling, parallel processing. At present, Clustering is one of the main methods to deal with association rules. Sahar S [2] proposed a clustering method, which describes the rules with five attributes, this method cluster rules by using the syntax between item sets. Jorge A [3] puts forward a hierarchical clustering method for large data set specially. These methods reduces the difficulty of analysis rules in a certain extent. However, there are some shortcomings, the clustering method is mainly based on the relationship between the rules, so the relationship between item sets of rules is ignored.

In the manuscript, the method based on complex networks [4] are presented to ARM, which is aim to group result rules and provide a visualization method of association rules, which can solve this problem above. It is proved that the method based on the complex networks is essential to study sophisticated problems practically [5,6]. Good results are obtained when analysis of association rules. The network community structure of complex networks is very important to analyze the features of association rules, discovery the potential patterns, and predict the relationship of the items between the result rules. The approach taken in this manuscript has improved the veracity and efficiency of finding the potential patterns.

2 The formal statement of association rules and complex networks

In order to capture the following work more accurately, we will state the formal model of association rule and complex networks.

Generally speaking, association rules mining [7] can be described in the following way: Let $I=\{i_1,i_2,i_3,\cdots,i_d\}$ be a set of items, and the set of transactions is $T=\{\mathrm{t}_1,\mathrm{t}_2,\mathrm{t}_3,\cdots,\mathrm{t}_N\}$, where the items are the subsets of I, namely $I \supseteq t$. An association rule is defined as $X=>Y$ (the X is named antecedent or left-hand-side (LHS), meanwhile, the Y is named consequent or right-hand-side (RHS)), where $X \subset I, Y \subset I, and \quad X \cap Y = \varnothing$. The rule is generated in the transaction set T by support s and

confidence c, where s is the percentage of transactions in T that contains $X \cup Y$, s is expressed as probability $P(X \cup Y)$, and confidence c is the percentage of transactions in T containing X that also contains Y, c expressed as probability $P(Y|X)$ .we can get that, $s(X => Y) = P(X \cup Y)$ and $c(X \rightarrow Y) = P(Y \mid X)$.

We will express the complex networks [8] model with G (V, E), where the V is the set of all vertices and E is the set of all edges in the graph, namely $V = \{v_1, v_2, v_3, \cdots, v_m\}$, and $E = \{e_1, e_2, e_3, \cdots, e_n\}$. In this manuscript, all the complex networks are generated by association rules. It is appropriate for researching an enormous and complex system through complex networks intuitively and effectively, such as the association rules.

3 Association rules research method based on complex networks

Compared with the traditional clustering method of association rules, the method in the manuscript will constitute the rules of the item sets with network graph, which show all the result rules directly and conducive to the association rules between topological structure analysis. Meanwhile, with the introduction of community detection algorithm, the visualization and classification of association rules is more intuitive and effective. We will adopt multiple community detection algorithms to the association networks generated by association rules, and the modularity is the index we used to measure the performance of these algorithms. Finally, we can get the classifications of the rules, and get the potential association rules that interest us.

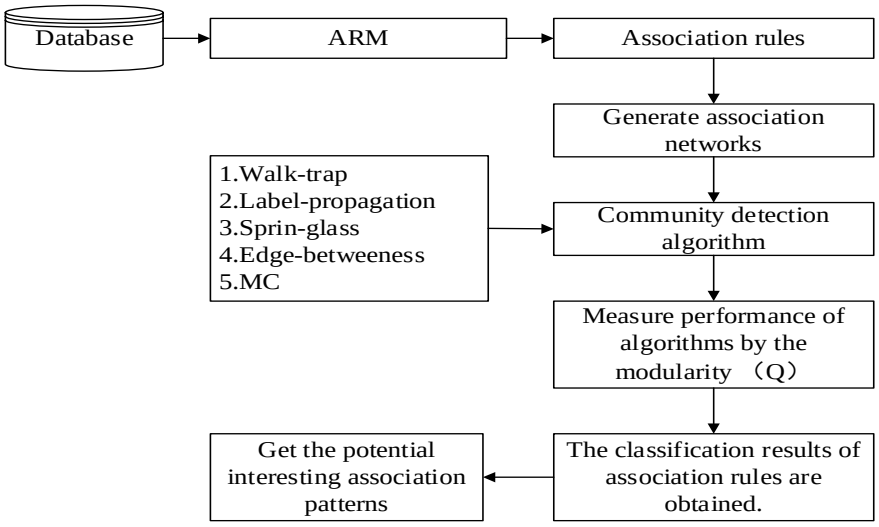

Fig. 1. Structure of ARM results research method based on complex networks

The main structure of the association rules research method based on complex networks can be shown in Fig.1. For introducing the method better, in the remaining part, firstly, we will definite the association networks and the importance of itemsets. Secondly, we introduce a new community detection algorithm for the association networks we have defined. Finally, we apply this method to tow databases, and analyze the results we get, verify the validity and superiority of the method we raised.

3.1 Definition of association networks

We think there are two main shortages of the traditional research method of association rules. Firstly, the expression of rules is not intuitive enough; secondly, the relationship of itemsets, which is the basic elements of association rules is ignored. So it is important to find a reasonable way to display all of the association rules. As the method structure mentioned above, we introduce the association networks as the first step of the method we proposed in this manuscript. The association networks is made up of itemsets and the edges between them. Because of the non-homogenous topological structure, the itemsets have different importance in the association networks. The importance of the itemsets in the networks are determined by 2 main factors. One is the location of the itemsets, another is the connection ability of the itemsets, which is determined by the shortest path of it. According to the

tow factors, we can definite an importance formula, which is aim to measure the importance of itemsets in the association networks [9].

Let $\lambda_{ki} = \{I_j \mid v_j \in V, a_{ij} = 1, \text{where } j = 1, 2, \cdots, n\}$ be a neighborhood of the itemsets I_j .The key degree of item set I_j can be computed as follows:

$$K(i) = \frac{T(i)}{T(i) + UT(i)} \tag{1}$$

Where $T(i)$ is the number of the shortest path that pass the item set I_j , $UT(i)$ is the number of the shortest path that don't pass the item set I_j .

And the closeness of item set I_j can be computed as follows:

$$C(i) = \frac{1}{\sum_{j=1}^{n} d(i_i, i_j)} \tag{2}$$

Where $d(i_i, i_j)$ is the shortest distance between i_i and i_j , and $j \neq i$.

Then we can define the importance of the itemsets as below:

$$I(i) = K(i)C(i) = \frac{T(i)}{[\sum_{j=1}^{n} d(v_i, v_j)][T(i) + UT(i)]} \tag{3}$$

3.2 New community detection algorithm MC

In order to get the potential rules from the association networks we have introduced above, it is necessary to prune and partition the itemsets in the association networks by community detection algorithm. The new community detection algorithm of association networks (MC) we proposed in this manuscript can be decomposed into four step:

Step 1: Calculate the importance of the itemsets in the association networks by the Eq.3, then sort the results in descending order, so we can get a set of results, $V_l = \{v_{\max}, v_j, \cdots, v_{\min}\}$.

Step 2: Determine the communities' center set from $V_l = \{v_{\max}, v_j, \cdots, v_{\min}\}$ we get from the step 1, then we get set $V_c = \{v_{\max}, v_j, \cdots, v_k\}$.

Step 3: Calculate the distance $d(v_i, v_j)$ between the other items and the center items.

Step 4; According to the distance $d(v_i, v_j)$, we get the community detection of all the items.

After the community detection, we hope that there is an indicator that can be used to judge the performance of the algorithm. When we divide the itemsets into different communities, we hope that the association among the nodes of the same community is as much as possible, and that the association among nodes in different communities is as little as possible. So we define an index named modularity to measure the degree of community level of each itemset.

Let $P_k = \{(V_1, E_1), \cdots, (V_k, E_k)\}$ be a set of community. Formally, the modularity formula is given as follows:

$$Q = \sum_{i=1}^{m} (e_{ii} - (\sum_{j=1}^{m} e_{ij})^2) \tag{4}$$

Where the n is the number of the lines in the graph, and, the l_i is the number of lines in the community P_k , $e_{ij} = d_i / 2m$, where the d_i is the total degree of all nodes in community P_k . The greater of the Q value, the better of the algorithm results.

4 Examples

In order to verify the method we have defined above, we get tow data sets from the Frequent Itemset Mining Dataset Repository. The data sets are described in the Table 1 below.

Table 1 Data sets description

Dataset	transaction	items	Min-Support	Min-confidence	rules
Retail	59310	14973	0.04	0.25	173
Accident	7503	305	0.9	0.9	73

By association rule mining algorithm Apriori (Apriori is a well-known algorithm for mining frequent itemsets for Boolean rules), 175 and 73 association rules were obtained from the two data sets above. Part of the association rules are showed in the Table 2 below.

Table 2 Part of rules generated from data set retail

rules	support	confidence	lift
{38} => {39}	0.116625	0.660334	1.158094
{39} => {38}	0.116625	0.204536	1.158094
{32} => {48}	0.091013	0.525557	1.114477
{48} => {32}	0.091013	0.192999	1.114477
{39} => {41}	0.135862	0.238275	1.33903
{39} => {48}	0.32374	0.567775	1.204001
{38,48} => {39}	0.067864	0.762311	1.33694
{38,39} => {48}	0.067864	0.5819	1.233954
{39,48} => {38}	0.067864	0.209624	1.186905
{32,48} => {39}	0.06058	0.665617	1.167359
{32,39} => {48}	0.06058	0.628917	1.333656
{39,48} => {32}	0.06058	0.187126	1.080559

As Fig.2 showed, we generate tow association networks of the result rules we get from the datasets retail and accident separately. (a) is the association networks of the retail, (b) is the association networks of the accident. In order to evaluate the performance of the community detection algorithm we proposed in this manuscript, several comparing done with the different detection algorithm, these algorithm include Walk-trap, Spin-glass, Edge-between-ness, Label-propagation. Fig.3, Fig.5 show the community detection results of the tow databases by using different algorithm.

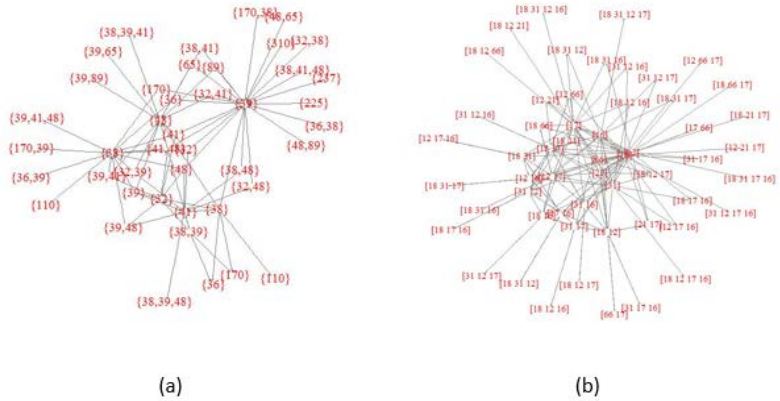

Fig. 2. Association networks of the result rules (a) Retail (b) Accident

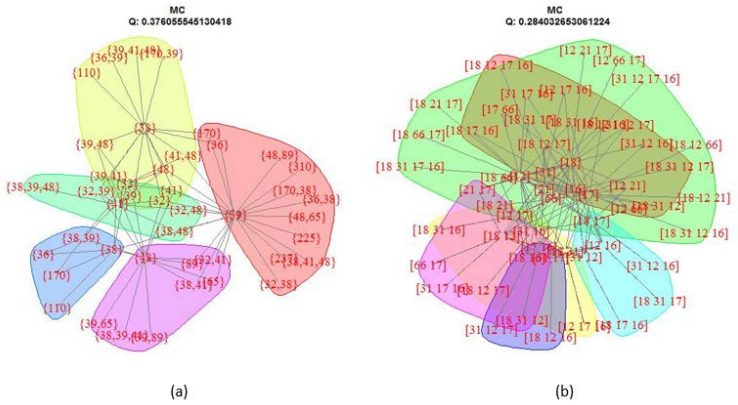

Fig.3. Community detection results (a) Retail (b) Accident

Table 3 and Fig.4 presents the results of all the community detection algorithms. According to the results showed in the Table 3, we observe that in the database retail, the Q of the algorithm Label-propagation is 0, the number of result communities is one, so the Label-propagation is of the worest performance to the retail, the max Q is belong to the algorithm MC, which gets 6 communities results as the Fig.3 (a) showed. In the database accident, the Edge-between-ness has the min Q, and the same as the database retial, the MC has the besest performance, the association worknets is divided into 5 different communities as the Fig.3 (b) showed. The tow databases both prove that the community detection that proposed in this manuscript has better performance. So we chose the results of MC algorithm as the final results as the Fig.3 showed.

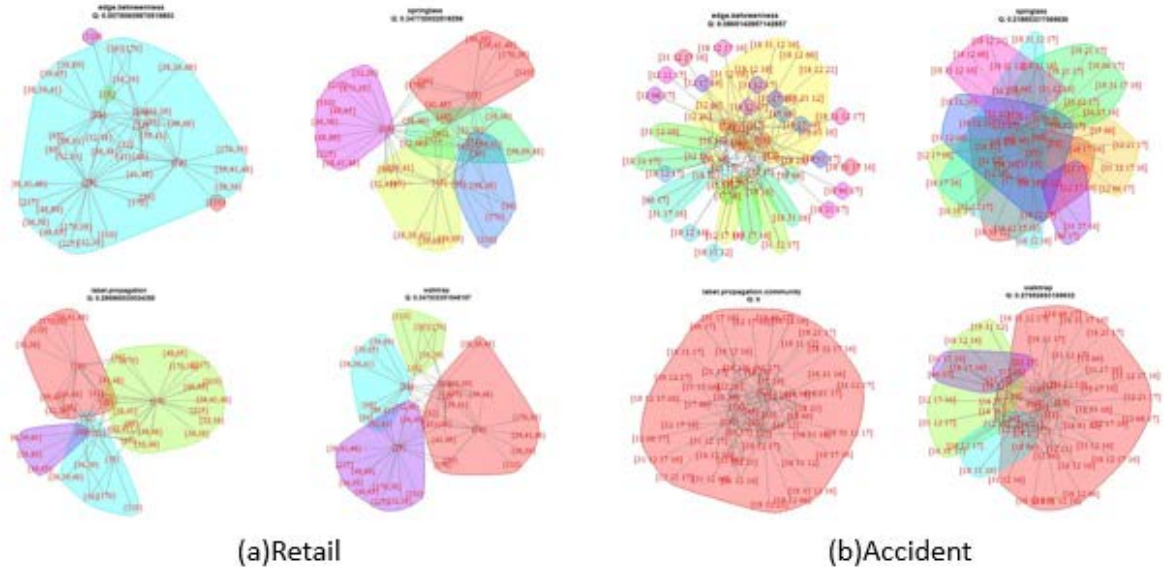

Fig. 4 Community detection results generated by different algorithms

. **Table 3** Results of the community detection

Databases	Retail		Accident	
algorithm	Q	Number of communities	Q	Number of communities
Edge-between-ness	0.007	2	0.086	11
Spin-glass	0.348	5	0.212	9
Label-propagation	0.297	4	0	1
Walk-trap	0.347	4	0.279	4
MC	0.376	5	0.284	6

From Fig.4 and Table 3, we get these results as follows. In Retail, we get 5 groups of the association rules, the group 1 center is the itemset {39}, the rules have the same RHS; the group 2 has the same LHS {38}, we can judge that when people buy the {38} in the retail , they will buy

{110},{41,48},{36,39} with greate probility; the group 3, group 4 and group 5 have the similar structure. In Accident, we get 6 groups of the association rules in the end, in the group 1, the center itemsets are {16},{31}, when {16},{31} happens, {12,17,18} will very likely to happen, and them are all LHS. Group 3 is including six rules which are consisted of four itemsets. Group 4,5,6 have the same structure. Finally, we get the classfication of all result rules and the potentail patterns we may interested in.

5. Conclusions

In this manuscript, we have introduced a new method which aimed to help us to find potential interest patterns from association rules. As we know, association rules mining may generate large quantity of rules, most of which are of no interest to the researchers. Our method takes into account topological structure of the rules, and offer a visualization way to the rules. First of all, we use graph to represent all the association rules, then we use the center degree of graph to evaluate the importance of the itemsets in the complex networks, so we can find the import rules from the result rules or isolated rules. Community detection algorithm is used to group the result rules in the complex networks, these algorithm include Edge-between-ness, Spin-glass, Walk-trap and Label-propagation. We define a modular metrics to measure the performance of these algorithms community partitioning results. This method can improve the efficiency and veracity of finding out the potential rules from large databases. And the MC algorithm has better performance when dealing with the association networks then the other algorithm adapted in this manuscript.

Acknowledgements

This work was supported by National Key Technology R&D Program of the Ministry of Science and Technology of China (No.2014BAH24F02), this work is also supported by Engineering Research Center of Information Networks, Ministry of Education.

REFERENCES

[1] Agrawal R, Imielinski Tomasz, et al. Mining association rules between sets of items in large databases [J]. Acm Sigmod Record, 1993, 22(2):207-216.
[2] Sahar S. Exploring Interestingness Through Clustering: A Framework[C]// IEEE International Conference on Data Mining. IEEE Computer Society, 2002:677-677.
[3] YUAN Sen Miao CHENG Xiao Qing Department of Computer Science and Engineering, Jilin University of Technology, Jilin University of Technology, et al. Clustering Method for Mining Quantitative Association Rules [J]. Chinese Journal of Computers, 2000.
[4] Jorge A. Hierarchical Clustering for Thematic Browsing and Summarization of Large Sets of Association Rules.[C]// Siam Sdm, Data Mining Conference. 2004.
[5] Ai X. Inferring a Drive-Response Network from Time Series of Topological Measures in Complex Networks with Transfer Entropy [J]. Entropy, 2014, 16(11):5753-5776.
[6] Dorogovtsev S N, Mendes J F F. Evolution of Networks [M]. Springer US, 2012.
[7] Geng L, Hamilton H J. Choosing the Right Lens: Finding What is Interesting in Data Mining [M]// Quality Measures in Data Mining. Springer Berlin Heidelberg, 2007:3-24.
[8] L. da F. Costa, F. A. Rodrigues, G. Travieso& P. R. Villas Boas, Travieso G, Boas P R V. Characterization of complex networks: a survey of measurements. Adv Phy[J]. Advances in Physics, 2005, 56(1):167-242.
[9] Jing C, Sun L. Evaluation of Node Importance in Complex Networks[J]. Xinan Jiaotong Daxue Xuebao/journal of Southwest Jiaotong University, 2009, 44(3):426-429.