# Random Response Privacy Data Mining Based on Cloud Computing Resource Association Rules

**Hui Baofeng, Jia Guoqing, Chen Shanji**

Qinghai University for Nationalities, Xining, Qinghai, China

**Keywords:** association rules, cloud computing, data resource, random response

**Abstract:** To strengthen data privacy protection, and improve data mining accuracy, random response mode is adopted to design privacy protection mining method based on association rules. Granular computing method and technology are applied to mining fields of association rules in data mining, and mining to association rules is researched in more extensive way from another perspective in this paper. Firstly, partial concealing mode is adopted to conceal and transform original privacy data and improve data security; secondly, associated frequent item set is utilized to construct simple and efficient privacy protection mining algorithm; finally, algorithm proposed is verified to have higher privacy and accuracy through theoretical analysis and experimental verification. After classical association rules mining algorithm is analyzed and researched in detail with its characteristics and restrictions summarized through examples in this paper, association rules mining model based on granular computing is proposed, which makes theoretical preparation for proposal and construction of association rules pick-up algorithm based on granular computing. Experimental result shows that association rules mining method based on granular computing is feasible and effective.

## 1. Introduction

In recent years, data mining technology has been widely concerned by information industry circle, which is the inexorable outcome of paradoxical movement between rapidly increasing data quantity and increasingly poor information amount. Systematic and intensive research to data mining technology is objective requirement of global information-based development. Data mining technology includes many research fields, of which association rule is an important research direction, having vitally important application value in business decision. This topic mainly makes related research to association rules mining. Traditional association rules mining algorithm, such as Apriori algorithm and its improved algorithm etc., makes mining to certain and accurate concept, and it is difficult to mine non-accurate or blurry concept. Through experiment, it can be found that main calculation to search frequent item set lies in frequent 2-item set generation, and frequent 2-item set generation process is Apriori algorithm mining bottleneck, and therefore, a kind of new association rules mining algorithm based on fuzzy sets is proposed in this paper, fuzzy set theory and semantic association rule concept are introduced in the algorithm, reasonable and non-accurate semantic translation is made to numerical attribute of database, and algorithm efficiency is improved by improving size of item set of pruning part scanned, which avoids exponential growth tendency of length of set scanned. Because core problem of Apriori algorithm is to find the maximum item set, the process to find the maximum item set is global search process, and genetic algorithm is a kind of global optimization algorithm, and avoids local optimization in search process. Therefore, truly useful rules can be found by applying genetic algorithm to rule finding and extraction. Therefore, a kind of association rules mining algorithm based on genetic algorithm is proposed in this paper, which mainly mines quantitative association rules, and algorithm mainly includes association rules coding method design, fitness function construction and genetic operator improvement etc. According to 2 kinds of association rules mining algorithm that are proposed and designed in this paper and that are based on computational intelligence, we extract association rules respectively by taking medical database and student database as mining prototype, and make experimental analysis; experimental result verifies effectiveness of 2 kinds of algorithm and also

 

illustrates wide application prospect of association rules mining.

## 2. PFP Algorithm

PFP algorithm is a kind of parallel FP-Growth algorithm that was proposed by Li et al. of Google Beijing Research Institute in 2008 and that is based on MapReduce frame. Because of high expansibility and high fault tolerance of MapReduce frame, algorithm can process big data in relatively good way.

Basic thought of algorithm is to transform transaction database into a new “intra-group dependency transaction” database and distribute it to corresponding Reducer, and in the process of recursive construction of Reducer to FP-Tree, local FP-Tree generated through different "intra-group dependency transaction” is mutually independent.

## 3. PFP-P Algorithm

Basic starting point of PFP-P algorithm is to replace mining to all frequent item sets with mining to closed frequent item set. Different from mining to all frequent item sets, mining to closed frequent item set does require “intra-group dependency transaction” of all items in transaction data, but “intra-group dependency transaction” of partial suffix items to reduce data transmission quantity. In addition, another advantage of mining to closed frequent item set is that mining result data can be significantly reduced under the premise that information completeness is guaranteed, which is convenient for storage and further processing.

### 3.1 Suffix item list

For the convenience of discussion, this paper maps transaction data as transaction mode, which means that transaction data ranked according to sequence of items in F-List after non-frequent items in transaction data are deleted is expressed as T; item in transaction T is expressed as I; support degree of mode T and submode is expressed as sup (), and the minimum support degree meeting frequent mode is expressed as sup_min.

Definition 2.1 Closed frequent mode. Assumed that submode X=I1I2…Ik, sup(X)=u≥sup_min; if ∀Ip, where p ≥k+1, sup(I1I2…Ik…Ip)<u, then X is called as closed frequent mode of item Ik, and Ik is called as suffix of X. Closed frequent mode corresponds to closed frequent item set one by one.

Definition 2.2 Suffix item. For submode X=I1I2…Ik of T, if sup (I1)= sup(I2)=…= sup(Ik)=u is met, then:

Definition 2.2.1 If ∀Ip, where p ≥k+1 and sup(Ip)<u, then Ik is called as u support degree suffix item of T.

Definition 2.2.2 For any item Ii and any suffix item I in X (Ii≠I), if sup(IiI)<u, then Ii is called as u support degree suffix item of T.

Theorem 2.1 Suffix of closed frequent mode must be suffix item.

Prove. Prove through proof by contradiction. Assumed that submode X=I1I2…Ik of T is closed frequent mode, sup(X)=u and Ik is not suffix item, then according to definition 2.2.1, item Ip must exist in T, where p ≥k+1 and sup(Ip) ≥u, because support degree of item in T is monotonous and does not increase, sup(Ip) =u; according to definition 2.2.2, sup(IkIp)=u; therefore, a submode X’=I1I2…IkIp must exist in T, to make sup(X’)=u; according to definition 2.1, X is not closed frequent mode, which conflicts with assumption, so original conclusion is verified.

According to theorem 2.1, when Mapper distributes data to Reducer, it just need to distribute data to suffix item to obtain enough information to construct closed frequent item set, and it does not need to distribute data to non-suffix item. List consisting of all suffix items in transaction data is called as suffix item list, and it is a subset of F-List.

### 3.2 PFP-P algorithm analysis

This section analyzes data communication complexity of PFP-P. Quantity of data transmitted in

step 1 of algorithm is the same with that of PFP, being MDB. Quantity of data needing to be transmitted in step 2 to construct P-List is also MDB.

Data transmission quantity of step 3 is:

$$M_{Dup} = M_{DB} + \sum_{i=1}^{n}\sum_{j=1}^{l_i-1} jf(I_j I_{l_i})g(I_j) \quad (1)$$

Where function g(I) is defined as follows:

$$g(I) = \begin{cases} 0 & I \text{ is non-suffix item.} \\ 1 & I \text{ is suffix item.} \end{cases} \quad (2)$$

Add data transmission quantity of 3 steps together to obtain data transmission quantity of algorithm in the whole process:

$$M = M_{DB} + M_{Dup} = 3M_{DB} + \sum_{i=1}^{n}\sum_{j=1}^{l_i-1} jf(I_j I_{l_i})g(I_j) \quad (3)$$

For the convenience of discussion, assumed that average length of transaction data is l, the maximum value of grouping is obtained, and average total data transmission quantity of PFP and PFP-P can be obtained according to formula 1-4 and 2-3:

$$\overline{M_1} = (3/2)nl + (1/2)nl^2 \quad (4)$$

$$\overline{M_2} = 3nl + n\sum_{j=1}^{l-1} jg(I_j) \quad (5)$$

According to formula 2-4 and formula 2-5:

$$\overline{M_2} - \overline{M_1} = n\sum_{j=1}^{l-1} jg(I_j) - n(1/2)l(1-l) + nl \quad (6)$$

When all g(I) is 1, which means that all items of transaction data are suffix items, value of formula 2-6 is nl of the third item, which means that PFP-P transmits one more MDB data than PFP. When g(I) is item of 0, which means that sum of coefficient j of non-suffix item is greater than average length l of transaction, data transmission quantity of PFP-P will be lower than that of PFP. When average length l increases, effect of nl of the third item in formula 2-6 on total transmission quantity will decrease rapidly, and what plays a decisive role will be former 2 items in the formula. When l is relatively great, only few non-suffix items are required to make sum of coefficient exceed 1.

According to suffix item list construction algorithm, assumed that the number of item meeting support degree sup is ksup, then probability of non-suffix item in item with support degree being sup can be expressed as:

$$P(\exists I \notin PList) = Min\left(1, \frac{k_{\sup} - 1}{n(n-1)(n-2)...(n-\sup+1)}\right) \quad (7)$$

Assumed that the minimum support degree of frequent item in transaction data is minsup, and the maximum support degree of item is maxsup, then

$$M_{DB} = \sum_{i=1}^{n} l_i = \sum_{i=\min\sup}^{\max\sup} ik_i \quad (8)$$

Assumed that n and support degree of each item are kept unchanged in transaction data, according to formula 2-8, k will increase with increase of l; according to formula 2-7, probability that item in transaction data is non-suffix item also increases with increase of it. Therefore, for data with relatively great average transaction length, PFP-P can reduce average transmission quantity of data effectively.

Taking data {T1=(a1,a2,…a50),T2=(a1,a2,…a100)} as example, assumed that the minimum support degree threshold is 1 with the maximum grouping adopted, then in PFP, data shall be divided into 100 groups, and total transmission quantity of data is 150*(3/2)+(1/2)*(50*50+100*100)=6475. But in PFP-P, a50 is suffix item of support degree 2, while a100 is suffix item of support degree 1, and other items are non-suffix items with division of

2 groups (a50 and a100), and total transmission quantity of data is 150*3+49+99=589.

## 4. Conclusion

Based on PFP algorithm, a kind of parallel closed frequent item set mining algorithm PFP-P based on suffix item list is proposed in this paper. This algorithm replaces mining to all frequent items in original algorithm with mining to closed frequent item set to improve mining efficiency; aimed at features of closed frequent item set, suffix item list is introduced in mining process to reduce transmission quantity of grouped data in mining process, and lower internal consumption of system. Experiment shows that the algorithm is superior to original algorithm in average performance, and in decrease of the minimum support degree threshold, it can shorten mining time effectively; the algorithm can lower consumption of communication between nodes effectively with good speedup quality; compared with processing to low-dimension dataset, the algorithm has more advantages in processing high-dimension dataset, and therefore, the algorithm is more applicable to mining task of massive high-dimension data.

## Acknowledgement

This work was supported by the International Science & Technology Cooperation Project of Qinghai (2013-H-811, 2014-HZ-821) and Application Basic Research Project of Qinghai (2015-ZJ-721).

## References

[1] Weisen Pan, Shizhan Chen, Zhiyong Feng. Investigating the Collaborative Intention and Semantic Structure among Co-occurring Tags using Graph Theory. International Enterprise Distributed Object Computing Conference, 2012.

[2] Jennifer W. Chan, Yingyue Zhang, and Kathryn E. Uhrich. Amphiphilic Macromolecule Self-Assembled Monolayers Suppress Smooth Muscle Cell Proliferation, Bioconjugate Chemistry, 2015.

[3] Yingyue Zhang, Evan Mintzer, and Kathryn E. Uhrich. Synthesis and Characterization of PEGylated Bolaamphiphiles with Enhanced Retention in Liposomes, Journal of Colloid and Interface Science, 2016.

[4] Jonathan J. Faig, Alysha Moretti, Laurie B. Joseph, Yingyue Zhang, Mary Joy Nova, Kervin Smith, and Kathryn E. Uhrich, Biodegradable Kojic Acid-Based Polymers: Controlled Delivery of Bioactives for Melanogenesis Inhibition, Biomacromolecules, 2017.

[5] Lv, Z., Halawani, A., Feng, S., Li, H., & Réhman, S. U. Multimodal hand and foot gesture interaction for handheld devices. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2014.

[6] Yizheng Chen, Fujian Tang, Yi Bao, Yan Tang, *Genda Chen. A Fe-C coated long period fiber grating sensor for corrosion induced mass loss measurement. Optics letters, 2016.