Research on Method and it's Evaluation for User Focused Frequent Itemset Mining

Zhenya Zhang[1,a], Weili Wang[1,b], Hongmei Cheng[2,c]

[1]Anhui provincial Key Laboratory of Intelligent Building, Anhui Jianzhu University, Hefei, 230022, China

[2]School of Management, Anhui Jianzhu University, Hefei, 230022, China

[a]email:zyzhang@ahjzu.edu.cn, [b]email:ustc301@sina.com, [c]email:ustc017@sina.com

Keywords: Frequent Itemsets; Attention; Association Rule; Log File

Abstract. High frequent network request pattern in office automation system (OAS) is one kind of important network behaviors which can affect the performance of OAS, especially OAS in intelligent building. High frequent network request patterns in one OAS can be mined from network access log file of the OAS. To mine high frequent network request concerned by user, user focused frequent itemset is used to describe high frequent network request concerned by user. According to early selection model of attention on information filter mechanism, attention based user focused high frequent itemset mining method is presented in this paper. To evaluate performance of algorithm for user focused high frequent itemset mining, precision ratio and recall ration are defined. Experimental results show that the performance of our proposed method is better.

Introduction

In intelligent building [1], network management can be optimized according to user's behaviors log file of the network. Because data about user behaviors such as network access request from users in network is recorded in log files of gateway and router of the network for the OAS, user's network request pattern can be mined from those network log files [2][3]. It is easy to judge whether user's network surfing is associated with office affair according to user's network request pattern. Because data in those network log file is massive, data mining technologies such as association rule discovery is usually used in the mining of user's network request pattern from network log file [4][5].

To mine user's network request patterns [5] with association rule discovery technologies quickly and accurately, early selection model for information filter mechanism of attention (usually, the model is called as early selection model of attention) is used to prepare data for association rule technologies in our research. And to evaluate performance of our proposed method on user focused frequent itemset mining, precision ratio and recall ratio are used as evaluation criteria in our search. In this paper, our early selection model of attention based method is shown at part II. Definitions about precision ratio and recall ratio are introduced at part III. And experimental results are shown at part IV. Our future works are drawn at part V.

The Mining of User Focused Frequent Items

Because end user's network request can be recorded into network log file, non-trivial network request patterns of end user can be discovered smartly from network log file with association rule technologies. When end user's network request pattern is mined by association rules technologies, each network request of end user is treated as one item and each transaction is the set of network request of end user in a period of time. Because data in network log file is massive, the number of all items is usually large and the number of elements in transaction set is usually large too. Thus the running time of association rule mining algorithm will be long and provoking when association rule mining algorithm is used to discover non-trivial network request patterns of end user from network log file. To accelerate the running speed of association rule mining algorithm for the mining of end user’s network request patterns, sample data set which is sampled from full data set is often used. Random sampling method is one kind of common method for the construction of sample data.

According to characteristics of cognitive on information filter, not all end user’s network request behavior needs to be considered in the mining of user’s network request pattern for the optimization of network management. Attention, one property of consciousness, can cast some sunshine on fast mining for end user network request pattern. Directivity and concentration are two important features of attention. The directivity feature of attention indicates that selection is one important function of cognition activity. With the early selection model of attention, sample data set with right size can be constructed appropriately.

Let event *e*=<*id*, *name*> where $id \in \Lambda$ is the identification of *e*, Λ is the indicator set and *name* is the name of event *e*. As to association rule discovery, if $I = \{i_1, i_2 ... i_n\}$ is the itemset for all items and i_{j}=<id_{j}, $name_{\mathrm{j}}$>, *j*=1…n, one transaction is one event sequence and transaction set is the set of all event sequences. It is obvious that association rule discovery algorithm can be used to mine non-trivial event pattern if each item is treated as one event. Let $F = \{ f_1, f_2 ... f_m \}$ be event set, f_{j}=<id_{j}, $name_{\mathrm{j}}$>, j=1…m, and each element in *F* required to be pay attention specially, *F* is one attention set.

ALgorithm1: TS: early selection model of attention

Input: attentionSet, item //attentionSet is one attention set and item is one event.

Output: judgeFlag:// if item in attentionSet judgeFlag=true else judgeFlag=false.

```
1)    if isEmpty(attentionSet) tmpFlag=flase
2)    elese if currentNode(attentionSet)<item tmpFlag=TS(attentionSet.LChild, item);
3)              else if currentNode(attentionSet)>item tmpFlag=TS(attentionSet.RChild, item);
4)                      else tmpFlag=true;
5)    judgeFlag=tmpFlag
```

Algotithm2: early selection model of attention based high frequency itemset mining

Input: OT, minS, attentionSet, TS() //OT is the initial transaction set OT, minS is the minimum support, attentionSet is the attention set and TS() is the attention function

Output: freqItemSet //high frequent itemset in OT

```
1) Let T be the sample transaction set and I be the itemset of all items.
2) I=Φ,T=Φ;
3) For each element s in OT
4)      If TS(attentionSet, s) is true then T=T+s
5) For each item i in T
6)      I=I+ i;
7) Let T be the transaction set, I be all the itemSet, minS be the minimum support.
   Mine high frequent itemset in T to FI with apriori algorithm.
8) freqItemSet = FI
```

In Algorithm 2, the function TS() is used to determine whether the selected items is need to be paid attention. Clearly, the TS() is for the implementation of the attention based pre-filter. In Algorithm 1, function isEmpty() determined whether the current attention set is empty, and currentNode() determined whether the current node in the binary tree is empty. In algorithm2, if the sample transaction set is constructed from raw data in log file, function TS() should be redesigned with data or data pattern which is focused by attention as input.

The Evaluation of Method for User’s Network Request Pattern Mining

Definitio1: Let *Φ* be one association rule mining algorithm, i_{j}=<id_{j}, $name_{\mathrm{j}}$> be *j*th event, $I = \{i_1, i_2 ... i_n\}$ be the itemset of all items, $F = \{ f_1, f_2 ... f_m \}$ be the event set where each event in *F* should be paid attention by user, $FS = \{ fs_1, fs_2 ... fs_r \}$ be frequent itemsets discovered by *Φ*. If $fs_j \cap F \neq \phi$, fs_{j} is one user focused frequent itemset.

Definition2: Let *Φ* be one association rule mining algorithm, *A* be the set of all high frequent itemset mined by *Φ* from transaction set T, S_{a} be minimum support. And let *B* be the set of all high frequent itemset in transaction set *T* with S_b as the minimum support. If *recall*(Φ, S_a, S_b)= $\frac{|A \cap B|}{|B|}$, *recall*(Φ, S_a, S_b) be the recall ratio of Φ.

Definition3: Let Φ be one association rule mining algorithm, *C* be set of all user focused high frequent itemset mined by Φ from transaction T with S_c as the minimum support, *A* be set of all high frequent itemset mined by Φ from transaction T with S_a as the minimum support. If *precision*(Φ ,S_c , S_a)= $\frac{|C|}{|A|}$, *precision*(Φ ,S_c , S_a) is the precision ratio of Φ.

Recall ratio and precision ratio defined in definition2 and definition3 can be used to evaluate performance of user focused frequent itetmset mining algorithm. On the one hand, to draw end user's network request behavior in detail, more user focused frequent itemsets needs. And more user focused frequent itemsets discovered by association rule mining algorithm Φ means that the size of transaction set is more large. If the size of transaction set used by Φ is larger, more high frequent itemsets discovered by Φ may be not concerned by user. The phenomenon that more high frequent itemsets discovered by Φ may be not concerned by user mean that the running time of Φ may be longer meanwhile the precision of Φ may be decreased. On the other hand, if the size of transaction set is decreased as possible, high frequent itemsets which is not concerned by user may be less. But if the size of transaction set is decreased as possible, the recall ratio of Φ may be decreased much even if the precision of Φ may be increased more. Because the precision and recall ratio of Φ are restricted each other, it is appropriate that the transaction set used by Φ can be sample from original transaction set. If the size of transaction set mine is appropriate, Φ can be run in stipulated time meanwhile the balance between precision and recall rate of Φ can be achieved.

Experimental Results and Analysis

In our research on user focused frequent itemset mining, raw data is from network log file in gateway for Anhui Provincial Key Laboratory of Intelligent Building. With the log file, original transaction set with 76026 items and 31658 transactions is constructed. The table for those transactions storing has 4261578 records. Because the goal of our research is to find high frequent network request from end user in our laboratory, elements in our attention set are network requests with occurring frequency in top 200.

Table1 performance of algorithm2

minimum support	running time(ms)	precision ratio	recall ratio
10	213140	100%	63.13%
20	155062	100%	86.98%
30	71765	100%	88.78%
40	29796	100%	92.12%
50	22296	100%	94.44%
60	17531	100%	93.92%
70	17406	100%	89.09%
80	13562	100%	91.11%
90	10843	100%	96.15%
100	10937	100%	100%

Some information based on precision ration and recall ratio of algorithm2 is shown at table1 where minimum support is varied from 10% to 100%. According to the definition of precision ratio and the principle of algorithm2, the precision ratio of algorithm2 for the mining of user focused frequent itemset should be 100% and the conclusion is verified with data in precision column of table1. According to recall ratio column in table1, the recall ratio of algorithm2 is higher and the value of recall ratio is not less than 85% when the minimum support is increased from 10% to 100%. As comparison, the recall ratio of random sampling based method is only 72.44% with 30% as the minimum support meanwhile the recall ratio of algorithm2 is 88.78% when the minimum support is 30% too. It is obvious that more user focused high frequent itemsets can be mined by algorithm2 than random sampling based method if the minimum support is same.

Conclusion and Future Work

Non-trivial network request pattern in network log file can be mined by association rule discovery algorithm if network request pattern is treated as itemset in association rule discovery. Non-trivial network request pattern in OAS can be used to instruct the optimization process of network management. With the optimization of network configuration, the abuse of OAS's network can be monitored and limited. With those limitation tactics for network management, the efficiency of OAS can be promoted. To mine non trivial network request pattern in network log file of OAS with association rule discovery algorithm quickly, early selection model of attention for information filter is used in the construction of sample transaction set in our research. With our proposed method, efficiency of process for non trivial network request pattern mining is promoted meanwhile the performance of our proposed method is better.

To discover user focused high frequent itemset, attention set is used to describe features about user's attention. Usually, item/event concerned by user can be appointed by user or specialist. Method for the construction of full attention set based on little items/events as initial user's attention appointed by specialist/user is one of our future researches on attention based data mining method. The generalization and migration characteristic of attention will be two important factor in our future method on attention based data mining.

Acknowledgement

This work is partially supported by the National Natural Science Foundation of China (No.61340030, 61300060), Special Foundation for Young Scientists of Anhui Province (2013SQRL043ZD) and Science and technology project of Ministry Housing and Urban-Rural Development Department in (2014-K8-061).

References

[1] Wong J K W, Li H, Wang S W. Intelligent building research: a review. Automation in Construction, 2005, 14(1): 143-159.

[2] Cooley R W. Web usage mining: discovery and application of interesting patterns from web data. University of Minnesota, 2000.

[3] Wang Y, Xiang Y, Zhou W, et al. Generating regular expression signatures for network traffic classification in trusted network management. Journal of Network and Computer Applications, 2012, 35(3): 992-1000.

[4] Grace L K, Maheswari V, Nagamalai D. Analysis of web logs and web user in web mining. arXiv preprint arXiv:1101.5668, 2011.

[5] Lin K C, Liao I E, Chen Z S. An improved frequent pattern growth method for mining association rules. Expert Systems with Applications, 2011, 38(5): 5154-5161.