An Improved Incremental Queue Association Rules for Mining Mass Text

Wenchuan Yang [1], Lei Hui [2], Dong Zhang [3] and Yimin Fu [4]

Beijing University of Posts and Telecommunications, Beijing,100876, China

[1]yangwenchuan@bupt.edu.cn

***Abstract*—Association rules is an important data analysis and mining method, and the FP-Growth and the traditional FP-Tree algorithm is used in the full confidence of rules. This paper proposes a incremental queue algorithm models based on association rules, which is the improved FP4W-Growth algorithm. It is proposed and applied to the calculation the association text by the correlation of incremental queue. Its feasibility is validated by experiment. After optimization of the algorithm and model, it can find hidden and useful new information and new pattern. And those rules found in text can be potentially used as the scientific decision-making methods.**

***Keywords*—Scientific text, incremental queue, full confidence, association mining**

I. INTRODUCTION

With the rapid development of the Internet and the computer today, in the virtual world of data showing explosive growth, technology of text information is accompanied by the development of information technology to form a large number of data accumulation, followed by how to meet demand or potentially useful data and knowledge from the mass of information. In the face of massive text information processing technology of invalid means, with the development of data mining technology, not only the analysis of massive data put forward theoretically, more technically a theory to practice, resulting in a large number of data mining, association analysis model, in order to better from the data in knowledge provides creative solutions[1].

Algorithm optimization and research of association mining in the field of data mining has become an important research topic. Rakesh Agrawal's first Apriori algorithm is proposed[2], aiming to knowledge discovery using association rule mining area. On this basis, then made a series for the optimization and improvement of the research paper, incremental association rule, text information in the incremental association rule mining is used to discover frequent episodes algorithm[3]. From the information, Lu[4,5] et al proposed the problem of mining multi-dimensional association, also proposed the association rules mining algorithm E-Apriori span and EH-Apriori algorithm. Two algorithms are Apriori algorithm based on EH-Apriori algorithm, which generates the frequent item sets in 1, using the Hashing technology, the frequent itemsets in 1 candidate 2 set, filter out a large number of candidate 2 sets, so its efficiency is higher than that of E-Apriori algorithm. Tung et al. [6] puts forward a kind of association rule mining in the information, then mining association rules incremental information between FITI (First Intra Then Inter) algorithm for [7].

Since the method based on Apriori algorithm, the main performance bottleneck is to generate a lot of itemsets and scan the database many times after. In order to obtain better efficiency of Han[8], a FP-Growth algorithm is proposed based on FP-Tree does not generate candidate, it constructs a highly compressed FP-Tree, compressed the original database. Do not use the Apriori method for the generation of test strategy, while focusing on the frequent pattern (segment, sequence) growth, to avoid the high cost of candidate item, can obtain better efficiency of. Ming Fan, based on the FP-Growth algorithm proposed a new algorithm for mining is not FP-Tree formation condition, greatly improving the efficiency of mining frequent patterns in space-time.

The related technology in this thesis, the incremental association rules efficiently generated as the research background, the mathematical model of incremental mining queue associated text information is established, and the introduction of full confidence metric extends support confidence framework of measurement. Then based on the FP-Growth algorithm does not FP-Tree formation condition of thought, the mining task was improved by using the properties of all confidence, make the algorithm more suitable for the processing technology of information data, and the experimental results the performance of the algorithm are given.

II. DEFINITION AND DESCRIPTION

Definition 1: text feature set

Some type of data as text , text for a given number for S, P for the quantitative text set, the text set feature words produced by weighting interval sets Q={Q1，Q2，Q3，…… Qn}summarized as ordered.

W={Si, W1, W2, W3，……Wn} in the Q reference interval is divided according to the feature set sequence, the Si to generate the feature words given text sets, Wi said the feature words with a weight of interval Q=Q1 collection.

Definition 2: increment term set queue

For the first time a given amount of text of S, W as the feature set to a given sequence, followed by several times, and

978-1-5090-3071-2/16 $31.00 © 2016 IEEE
DOI 10.1109/IS3C.2016.119

each time the number of Si text , feature words produce set sequence for Wi.

Define the increment term set queue K={W1 , W2,W3, ……Wn}.

Definition 3: text increment interval

Based on the given number for each S_i text , the characteristics of the set of words W_i sequence for characteristic words from the number interval set $\sum_{o}^{i}(S_i)$ to $\sum_{o}^{i}(S_{i-1})$.

Definition 4: incremental association rules

According to the above definition, increment term set queue K, each feature word set sequence of W corresponding to an increment of S. We use a section of S long word set sequence for the origin, if in the interval, word set sequences in W feature words, then the feature words marking in W, and recorded as R_iw_j.

Multiple time series inter transaction association rules to support S_p, reliability C_f defined as:

$$Sp=\frac{C_{AB}}{n}, \quad Cf=\frac{C_{AB}}{C_A} \tag{1}$$

The number of A is C_A, the number of $A \bigcup B$ for C_{AB}, n for the feature words set sequence number.

Incremental association rules text information to meet the following conditions:

(1)$A \subset K, B \subset K$，$X \bigcap Y=\Phi$

(2)$\exists$ $R_iw_j \in A, 1 \le j \le n$

(3) $\exists$ $R_iw_j \in A$, $1 \le j \le n$,$((i=j) \wedge (1 \le i< S_i)) \vee ((i \ne j) \wedge (0 \le i<S_i))$

(4)$\exists$ $R_iw_j \in B$, $1 \le j \le n$, $max(j)<i \le S_i$

At the same time, the incremental association rules containing type is $A \Rightarrow B$.

Definition 5: full confidence

Support degree and confidence in the previous model, the massive text information incremental superposition model, especially in the support degree is relatively low, the useless rules seem to feel helpless to exclude. At the same time, the phenomenon of negative correlation processing text information data is not very good. Mining association rules mining quality is the soul, in order to make more reliable quality, we take full confidence the way to reconstruct the model.

We define the confidence with Acf said, according to the definition of 4, for a given set of $D=A \bigcup B$, D full confidence:

$$Acf(D)=\frac{Sup(D)}{max\_item\_sup(D)}=\frac{Sup(A \bigcup B)}{max\_item\_sup(A \bigcup B)}$$
$$=\frac{Sup(A \bigcup B)}{\max\{Sup(R_i) \mid \forall R_i \in (A \bigcup B)\}} \tag{2}$$

Among them, $\max\{sup(R_i) \mid \forall R_i \in (A \bigcup B))\}$ is the largest of all the items in a $A \bigcup B$ (single) item support.

All confidence has two properties:

1 zero invariance. Its value is not affected by the air data.

2 similar to Apriori's downward closure . That is, if a pattern is full of confidence, then each of its sub mode is also full confidence. Conversely, if a pattern is not full confidence, further growth of this mode also does not meet the minimum all confidence threshold . During the full confidence of mining growth, can help cut does not meet the conditions of the model, to improve the efficiency of the algorithm.

Definition 6: a frequent pattern F is full of confidence, two conditions must be met:

(1) $Acf(F) \ge min\_\theta$

(2) $Sup(F) \ge min\_\theta$

Lemma 1: counting space pruning rules

Set $\theta=K_1, K_2, \ldots\ldots K_n$, θ in the constrained sub tree in the , to meet the full confidence of the F frequent pattern

$$Sup(F) \le \frac{Sup(\theta)}{min\_\theta} \tag{3}$$

Proof: let θF be the full confidence of frequent pattern,

$$all\_conf(\theta F) \ge min\_\theta \Rightarrow \frac{Sup(\theta F)}{max\_item\_sup(\theta F)} \ge min\_\theta$$
$$\Rightarrow max\_item\_sup(\theta F) \le \frac{Sup(\theta F)}{min\_\theta} \tag{4}$$

Here $\because |Sup(\theta)| \ge |Sup(\theta F)| \Rightarrow max\_item\_sup(\theta F) \le \frac{Sup(\theta)}{min\_\theta}$ (5)

Hence,$\because max\_item\_sup(F) \le max\_item\_sup(\theta F) \Rightarrow$

$$max\_item\_sup(F) \le \frac{Sup(\theta)}{min\_\theta} \tag{6}$$

In the AFP-Growth algorithm, using the pruning rules, can reduce the constrained tree size. Therefore, the traversal of FP-Tree, can reduce the number of access nodes.

III. ALGORITHM FOR INCREMENTAL QUEUE

FP4W-Growth algorithm, mainly based on Ming Fan's not condition tree structure idea of FP-Tree, the process of FP-Tree was improved, the mining of Association for the text incremental queue. At the same time, in order to reduce the overhead, the closed frequent sets are constructed from FP-Tree. In the process of FPmine, by using lemma 1, improved pruning rules, finally generate association rules availability high use full confidence new model.

Specific algorithm idea is as follows:

Construction of FP-Tree:

Step 1: find the text all meet the minimum support threshold in frequent 1- itemsets.

Step 2: frame based on divide and rule. For each reference reference point in time to meet the conditions of the Riw0, perform the following operations:

(1) first found in Riw0 appear to be the case in the queue, incremental frequent 1- itemsets Fre_Item_S1

(2) will be in accordance with Fre_Item_S1 (Ri+1w0,⋯ Rnw0, R1w1,⋯Rnw0,⋯, R1wm-1,⋯, Rnwm-1) order.

(3) scans to increment term set queue K is read into memory, each increment queue as a transaction, in the framework of divide and rule, the merger pruning method, find the Riw0 appears, frequent itemset Clo_FiSet closed the transaction.

(4) the items in the Clo_FiSet to support count in descending order, get every item number, and production order conversion table.

(5) create a branch for each incremental queue, building FP-Tree.

Improved algorithm for mining association:

In the longer the confidence of frequent pattern mining process is recursive, there has been a length of K full confidence of frequent patterns in {F1,⋯Fk}, the serial number k, if Sup(Fk) ≤ Sup{ F1,⋯Fk }/min_θ and all confidence of Fk all_conf ≤ min_θ ,have longer full confidence of frequent pattern.

Finally, in the {Sup,Conf,Acf} framework, the frequent itemsets to generate and output the association rules, such as:

A ⇒ B[support,confidence,all_confidence].

The algorithm for FP4W-Growth is as,

Algorithm 1. FP4W- Growth

```
Procedure FP4W-Growth (ST(k1,…,km)){
(1)   for i=km-1 downto 1 do{
(2)   if(ST(k1,…,km).sup[i]<=sup(ST(k1,…,km))/
        min_α and ST(k1,…,km).conf[i]>= min_α{
(3)   FP[++length]=item(i);
(4)  output FP and its support ST(k1,…,km).count[i]/n;
(5)   build ST(k1,…,km,i) based on ST(k1,…,km);
(6)   If(there is an non-root node in ST(k1,…,km,i))
(7)   mine(ST(k1,…,km,i));
(8)   length--;
```

IV. Simulation

According to a Research Institute of science and technology of text information in the 110000 data, a total of about 3000 words, more than 200 useful words. In a laboratory environment, were compared on the performance of FP4W-Growth, FP-Growth and E-Apriori three kinds of algorithms.

110000 data for the incremental queue length, incremental scale for 10000 data. We set the minimum all confidence is 80%, total confidence frequent maximum length is 8, the lowest confidence level is 80%, the minimum support degree 5%.

Comparison of association rules of simulation results for FP4W-Growth, FP-Growth, E-Apriori and EH-Apriori two produced by the algorithm, as well as time and space performance comparison of FP4W-Growth, FP-Growth, E-Apriori and EH-Apriori of four algorithms, as shown below.

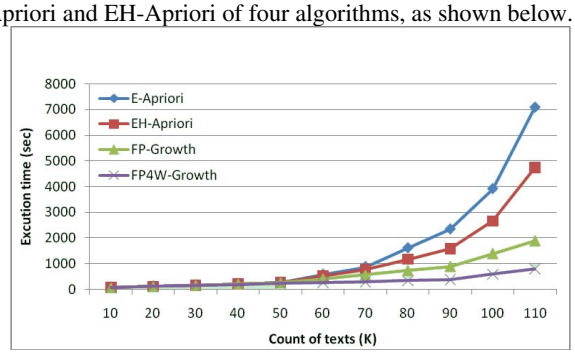

Fig. 1 . Performance comparison of FP4W-Growth with E-Apriori, EH-Apriori, FP-Growth in time consume

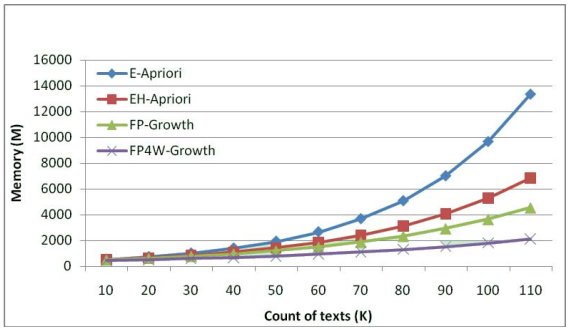

Fig. 2 . Performance comparison of different algorithms in space consume

From Fig.1, 2 it is clear that E-Apriori, and the FP-Growth algorithm, in the amount of information under article 50000, the execution efficiency reached the general level. But with the increasing amount of data, E-Apriori algorithm and FP-Growth algorithm memory usage increased dramatically, and exceeded the range of the main memory, virtual memory, put a lot of time spent in I/O operation.

While the EH-Apriori algorithm by using Hashing technology, the increase in the amount of data, although the amount of memory used to main memory limit, but much less than that of the E-Apriori algorithm, so not to waste a lot of time in the I/O operation, so the time efficiency than the E-Apriori algorithm. While the FP4W-Growth algorithm by using a number of optimization techniques, the required memory is very low, the required time grows slowly, growth trend also gently more. Thus, the FP4W-Growth algorithm in the time / space efficiency than EH-Apriori algorithm and FP-Growth algorithm.

See from Fig.3, the number of association rules to generate by E-Apriori, EH-Apriori and FP-Growth algorithm is much greater than AFP-Growth algorithm, of which a considerable part that the researchers are not interested in. While the FP4W-Growth algorithm by using the full confidence to filter out a lot of irrelevant, low interest and misleading. However, reducing the number of rules generated, enhance the quality of generated rules.

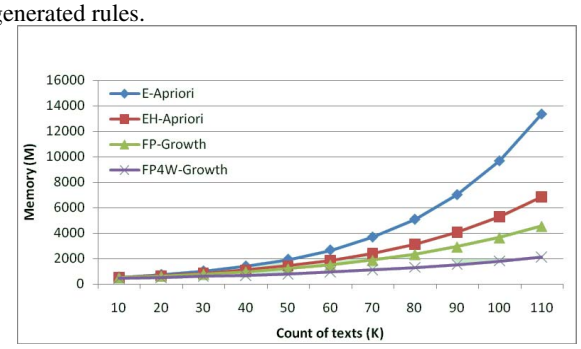

Fig. 3 . Comparison of generated total number for E-Apriori, EH-Apriori, FP-Growth and FP4W-Growth

V. Conclusions

In this paper, according to the characteristics of mass scientific text information and data, the research of association rules mining incremental queue, to predict and find mutual knowledge between application and the trend of development, proposed the incremental queue mass text information in the whole set technology based on the FP4W-Growth algorithm for reliability. The algorithm uses the nature of full confidence, more effectively in searching and pruning on no condition subtree FP-Tree, time and space efficiency is greatly improved, and dig out the interest degree higher rules. In the last part of the dissertation is an incremental updating algorithm for mining association rules based on the simulation and evaluation, proved the feasibility and optimization of the algorithm.

The work of the paper for classifying massive scientific text information prediction and analysis of data mining techniques, such as huge science and technology information classification management, multidisciplinary analysis, promote scientific and technological knowledge with different application has certain guidance and reference significance.

ACKNOWLEDGMENT

This paper is supported by the National Natural Science Foundation of China (No.61471060, No. 61571064).

REFERENCES

[1] K. Hatonen, Knowledge Discovery from Telecommunication Network Alarm Database. ICDE,pp:123-131,1996.

[2] D. Pan, J. Y. Shen. Similarity Discovery Techniques in Temporal Data Mining, Journal of software, 18(2), pp:246-258, 2007

[3] X. B. Li, J. L. Wu, Y. S. Xue, W. Weng. Research of Improvement and Optimization on Association Rules Mining Algorithm, Journal of Xiamen University(Natural Science),14(2),pp：71-74, 2005.

[4] H. Lu, J. Han, I. Feng. Stock movement and n-dimensional inter-transaction association rules, In: Pro of the SIGMOD Workshop on Research Issues on Data Ming and Knowledge Discovery,pp:322-330,1998.

[5] H. Lu, J. Han, I. Feng. J.Beyond, Intra-Transaction Association rules, ACM Transactions on Information Systems,18(4),pp:423-454, 2000.

[6] A. Tung, H. Lu, J. Han, Breaking the Barrier of Transactions: Mining Inter-Transaction association rules, In:Proc of the Knowledge Discovery and Data Ming, 21(3),pp:464-471,1999.

[7] L. X. Qin, Z. Z. Shi, Research on Multiple Time Series Inter-transactional Association Analysis, Computer Engineering and Applications, 27(41), pp:110-117, 2005.

[8] J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation, In:Dunham M,Naughton J,Chen W eds.Proc of 2000 ACM-SIGMOD Int'l Conf on Management of Data(SIGMOD'00), Dallas, TX,New York: ACM Press,pp:1-12,2000.