Methodology for Hiding Sensitive Information and Pruning Infrequent Itemsets for Association Rule Mining

K. Kavitha, PhD
Assistant Professor
Department of Computer Science
Mother Teresa Women's University, Kodaikanal

ABSTRACT
Association Rule Mining plays a major role in current research. This classical algorithm extracts frequent itemsets from large dataset which identifies Correlation between different items in the Transaction. Main issue in this algorithm is doubling the data scanning time. Many algorithms are proposed to find association rule and avoid complexity. This paper highlights two algorithms such as Novel Pruning approach for association rule mining and Hiding of Sensitive Association Rule by using improved Apriori algorithm. Finally, Suggested an integrated approach for Filtering Infrequent Itemsets and hiding Sensitive Association Rules using Same method which removes infrequent itemsets for hiding sensitive items in the Dataset.

Keywords
Association Rule, Apriori, Support, Confidence, Pruning, Hiding

1. INTRODUCTION
Data Mining helps to extract interesting patterns from large databae. It is an important tool to transform this data into information. Main focus of data mining is to reduce unnecessary database scanning time finding frequent itemset.

Mining association rules are particularly useful for extracting relationships among items from large database[4]. Mining process of association rules can be partition into two steps.

i. Frequent itemset generation

ii. Association Rule Generation

In Association Rule Generation, Two Bottlenecks are there in Apriori algorithm such as Complex frequent itemset generation (ie uses most of the time and memory) and Multiple scan of the database.

Association Rule Mining is used to discover the strong rules in the database. Several methods have been proposed to satisfies the user specified minimum support. Association Rule generation contains two steps

1. Minimum Support is used to find all frequent items in the dataset
2. Frequent items and the Minimum Confidence constraint are used to form Rules.

2. PROBLEM FORMULATION
1. Association Rule Mining

Itemset Count I in D and the Dataset size is the number of transactions in D. For two itemsets X and Y where $X \cap Y = \phi$; $X \rightarrow Y$ holds in D if both the following conditions hold, where X and Y are called Precedent and the Consequent respectively.

2. Association Rule Hiding

Let D be the dataset after applying sequence modifications to D. A Strong Rule $X \rightarrow Y$ in D will be hidden in D'.

D – Origin
D' – Modified Database

Strong Rules will be stored in D'.

3. RELATED WORK

3.1 Novel Pruning Approach For Association Rule Mining
Traditional Apriori algorithm consists of three steps such as Joining, Pruning and Verification for solving Association Rule Mining algorithms. Here Pruning steps eliminates the weak candidate itemsets. This algorithm proposed new pruning step named as "Filtration"[2].

Filtration step generates k-frequent and k-infrequent itensets at the same time. This process is based on the following lemma.

Lemma:

If any itemset Z is infrequent then none of it superset can be frequent

Proof:

Let Y be an item containing all items of Z and number of items in Y is greater than Z ie $Z \subset Y$. As if Z is infrequent, it can be stated that

$$Support(Z) < Minsup \quad \text{---------(1)}$$

It is obvious that Support of Y is not greater than any other subset because cardinality of Y is more than Z ie $|Z| < |Y|$. It can be referred as

$$Support(Y) \leq Support(Z) \quad \text{----------(2)}$$

Using (1) and (2) of association can be treated as Support(Y) < MinSup.

In otherwords, Y cannot be frequent if support of Z is infrequent.

Example for Modified Algorithm with six transactions such as $\{T_1, T_2 \ldots T_6\}$ as follows in Table

T_1	1,2,5
T_2	2,4
T_3	2,3
T_4	1,2,4
T_5	1,3,5
T_6	2,3
T_7	1,3
T_8	1,2,3,5
T_9	1,2,3

1 List of Transactions

Let N={ 1,2,..5} be a set of items Customer can purchase. Let D={T_1,T_2…..T_n} be the set of 9 transactions. Let user fix the user defined minimum support and minimum confidence. Frequency of each item in database as shown in the following table.

Item	Frequency
1	6
2	7
3	6
4	2
5	3

1-Itemset Generation

It is easy to generate 1-frequent itemset. Frequent itemsets are extracted based on Minimum Support. Here, MinSup=3 fixed by user. Based on the Frequency, item 4 is infrequent which is smaller than minimumsupport. So it is eliminated initially.

Thereafter, Apriori algorithm generates 2-itemset and also pruning is applied for generating potential 2-itemsets.

C_2= [(1,2) (1,3) (1,5) (2,3) (2,5) (3,5)]

Item-4 is infrequent so in the pruning stage it is eliminated and retains the same candidate set

P_2=[(1,2) (1,3) (1,5) (2,3) (2,5) (3,5)]

In third step, Frequencies are compared with minimum support and following 2-frequent itemsets are generated.

Candidate Set	Frequency
(1,2)	4
(1,3)	4
(1,5)	3
(2,3)	4
(2,5)	1
(3,5)	2

2-Itemset Generation

Frequent Itemsets are (1,2) (1,3) (1,5) (2,3)

In the Second iteration, following 3-frequent itemsets are generated by using former steps

C_3=[(1,2,3) **(1,2,5) (1,3,5) (2,3,5)**]

At this point, it is to be noted that the bolded 3-frequent itemsets are useless to generate because itemset (2,5) and (3,5) are not 2-frequent itemset. By eliminating the items **(1,2,5) (1,3,5) (2,3,5)** from 3-ietmsets,Potential 3-frequent itemsets are generated as follows

P_3=[(1,2,3)]

3.2 Mining And Hiding Of Sensitive Association Rule By Using Improved Apriori Algorithm

The main aim for this algorithm is to reduce query frequencies and storage resources. Without generating new candidate set frequent itemset are mined using an improved apriori algorithm. This algorithm adopts a new method to avoid reducndant generation of sub itemsets during Pruning. It proposed the following improvements to the traditional Apriori Algorithm[1].

1. Compute Minimum Support Count by min-sup * |D|
2. Produce L_1 candidate and simultaneously construct TID set for each item during database scanning
3. Produce L_1 candidate set from joining L_1 * L_1
4. Delete the patterns whose frequencies are not satisfy minimum support count and find L_2
5. In order to produce L_k, Join items which satisfy the rule
6. Compare with minimum Confidence for selecting hidden rules.

For hiding Sensitive Rule, Mark the Sensitive Association Rule where the confidence level is greater than the minimum Confidence.

Transaction Database D

TID	Items
1	A,B,E
2	B,D
3	B,C
4	A,B,D
5	A,C

Sample Dataset

Let TID-Set(A) denote the set of transactions which contain A in D, so the number of transactions are exactly defined as Sup-Count(A).

Similarly, transaction have A and B means which is defined by the intersection of TID-Set(A) and TID-Set(B) and Sup-Count(A→B).

Item X	TID-Set(x)	Support
A	1,4,5	3
B	1,2,3,4	4
C	3,5	2
D	2,4	2
E	1	1

1-Itemset with Transaction ID -- Sup-Count=4

E's Support is less than Threshold value. So it is eliminated. Then Candidate Set L_2 is generated as follows

Item X	TID-Set(x)	Support
AB	2	1,4
AC	1	5
AD	1	4
BC	1	3
BD	2	2,4
CD	-	-

2-Itemset Generation

Sup-Count=2 which satisfies the following candidate sets

Items	Support
AB	2
BD	2

Frequent 2-Itemset

Candidate set L_3 is generated for 3-itemset

Items	Support
ABD	1

Frequent 3-Itemset

Itemset does not satisfy the predefined threshold value. So it is eliminated.

3.3 An Integrated Approach For Pruning Infrequent And Hiding Sensitive Rules

This approach integrates the concept of above two algorithms for reducing the rule generation. This proposed idea improves classical apriori algorithm and perform pruning and hiding effectively. The proposed work consists of 4 steps named as

1. Joining – Generate Candidate Sets
2. Pruning – Identify Infrequent Itemsets
3. Verification – Extract Frequent Itemsets based on minimum support
4. Hiding – Extract Sensitive Rules based on minimum Confidence

Procedure for integrated approach as follows

Initialize D(Set of Transactions), minsup(Minimum Support specified by User), F_k(Collection of 1-itemset)

1. **Joining**
   Generate Candidate Sets based on association and join the sets $L_1 * L_1$ ie) $C_k = C_k \cup Z$
2. **Filtration (or) Pruning**
   Identify InFrequent Itemsets based on frequencies from former candidate set generation and remove it from potential itemsets
   ie $P_k = P_k - IF$
3. **Verification**
   Compare each potential itemset with minimum support ie) minsup. If it is larger than equal to minsup then it is frequent itemset.
   $F_k = F_k \cup Z$
   Else
   $IF_k = IF_k \cup Z$
4. **Hiding Sensitive Rule**
   Hiding sensitive Association Rule based on minimum confidence ie) minconf. Calculate the confidence for each frequent itemset Z and compare it with minconf. If it is greater than minconf then include this rule to sensitive rule.
   $S_k = S_k \cup Z$

4. CONCLUSION

This paper proposed an idea for removing infrequent itemsets and hiding sensitive rules. An integrated method for pruning frequent itemset and hiding sensitive rules of Apriori algorithm is suggested. One of these methods expressed as filtration for joining operation and the rest is hiding the sensitive rule. It is observed that the proposed approach generates frequent rules effectively based on Filtration and Hiding methods. It is clearly derived that the suggested method would be the better approach for satisfying the existing problems. In future, this approach will be implemented and applied in real time applications and checks the performance.

5. REFERENCES

[1] Shintre Sonali Sambhaji, Kalyanankar Pravin P. " Mining and Hiding of Sensitive Association Rule by using Improved Apriori Algorithm", Proceddings of 18th International Conference, Jan 2015 ISBN: 978-93-84209-82-7.

[2] Lalit Mohan Goyal, M.M.Sufyan Beg and Tanvir Ahmad, " A Novel Approach for Association Rule Mining", International Journal of Information Technology, Vol 7, Issn 0973-5658, Jan 2015.

[3] Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216.

[4] Agrawal R, Srikant R, ―Fast algorithms for mining association rules, Proceedings of the 20th Int'l Conference on Very Large Databases[C], New York:IEEE press ,pp. 487-499, 1994.

[5] S. Brin, etc, ―Dynamic Itemset Counting and Implication rules for Market Basket Analysis‖, Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 255-264, 1997.

[6] M. J. Zaki, "Scalable algorithms for association mining," IEEE Transaction on Knowledge and Data Engineering, pp.372–390, 2000.

[7] J. Han, J. Pei, Y. Yin.: Minining frequent patterns without candidate generation. Proceedings of SIGMOD, 2000.

[8] F. C. Tseng and C. C. Hsu, ―Generating frequent patterns with the frequent pattern list‖, Proc. 5th Pacific-Asia Conf. on Knowledge Discovery and Data Mining , pp.376-386, April 2001.

[9] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed itemsets for association rules. In Proc. ICDT '99, pages 398–416, 1999.

[10] M.J. Zaki, and C. Hsiao., "Charm: An efficient algorithm for closed itemset mining",Proc. SIAM Int'l Conf. Data Mining, PP. 457-473, 2002.

[11] J. Pei, J. Han, and R. Mao, " CLOSET: An efficient Algorithm for mining frequent closed itemsets", ACM

SIGMOD workshop research issue in Data mining and knowledge Discovery, PP. 21-30, 2000.

[12] J. Wang, J. Han, and J. Pei, "CLOSET+: Searching for the best strategies for mining frequent closed itemsets", proc. Int'l Conf. Knowledge Discovery and Data Mining, PP. 236-245, 2003.

[13] G. Grahne, and J. Zhu, "Efficiently using prefix-trees in mining frequent itemsets", IEEE ICDM Workshop on Frequent Itemset Mining Implementations, 2003.

[14] C. Lucchese, S. Orlando, and R. Perego, "DCI_CLOSED: a Fast and Memory Efficient Algorithm to Mine Frequent Closed Itemsets", IEEE Transaction On Knowledge And Data Engineering, Vol. 18, No. 1, PP. 21-35, 2006.

[15] D.Lin, Z. M. Kedem, Pincer-Search: A New Algorithm for Discovering the Maximum Frequent Itemset, EDBT Conference Proceedings, 1998, pages 105-110.