International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 17, 2017 ISSN: 2348-6600
Reference ID: IJCS-238 PAGE NO: 1488-1501
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

Lessen Database Scanning towards Hiding Sensitive Data based on Binary Search Approach

Dr.K.Kavitha
Assistant Professor,
Department of Computer Science,
Mother Teresa University, Kodaikanal.

***Abstract*—- Data mining is an emerging tool for extracting hidden knowledge from huge dataset. Major challenges in data mining are to identify, hidden information with minimum number of database scanning. The main objective is to develop an algorithm to find sensitive patterns such that preserving hidden information. This paper proposes an algorithm that uses pruning step and Binary Search Method for tackling the problems. The problem of association rule mining can be solved by using pruning and binary search method. This paper proposed and algorithm to integrate these two methods for eliminating weak candidate sets and avoiding unnecessary database scanning.**

Keyword: Association Rule, Apriori Algorithm, Minimum Association Rule Mining Support, Minimum Candidate Threshold.

I. INTRODUCTION

Data mining is a technique that helps to extract knowledge data from large database. Data mining is becoming an important tool to transform the data into information. During the data mining process main focus is to reduce unnecessary database scan at the time finding frequent itemset by Apriori algorithm. Apriori is designed to operate on databases containing transactions. Association rule is based on discovering frequent itemsets. The main objective of the project is to perform efficient mining of association rule and hiding of sensitive rule. Mining association rules is particularly useful for discovering relationships among items from

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 17, 2017 ISSN: 2348-6600
Reference ID: IJCS-238 PAGE NO: 1488-1501
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

large databases. A standard association rule is in the form of X → Y which says if X is true of an instance in a database, so is Y true of the same instance, with a certain level of significance as measured by two parameters, support and confidence. Association rule contains a database of transaction and each transaction contains set of items.

The rules thus discovered from the databases can be used to rearrange the related items together or can be used to make new market strategies. Application domain of association rule mining is not only limited to the context of retail application but can also be used in the decision logics. This paper is organized as follows. Section 2 summarizes the related work in this field. In section 3 a modified algorithm is proposed with example and in the last section, conclusion and future directions are stated followed by section of references.

II. RELATED WORK

Association rule is used to identify the association of items and also produces strong rules based on support and confidence. It mainly focuses on

- Minimum support is used to find all frequent itemsets in the datasets.
- Minimum confidence constraint is used to generate frequent items and form rules.

Consider an Itemset I and Dataset D represents the size of the transactions. For two itemsets X and Y where,

$$X \cap Y = \emptyset; X \rightarrow Y \; holds \; in \; D$$

If both the conditions hold then it forms the rules. Here X represents precedent and Y represents consequent respectively.

Association Rule Hiding:

Strong rule can be generated based on the size of the transactions D.

X → Y in D will be in D'.

After performing sequence modifications in D, D' will be generated which holds strong rules.

Apriori Algorithm:

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 17, 2017 ISSN: 2348-6600
Reference ID: IJCS-238 PAGE NO: 1488-1501
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

Classical algorithm in data mining uses three steps such as joining, pruning and verification for solving the difficulties in Association Rule Mining. Here pruning step removes the weak candidate itemsets.

1. **Procedure:** Joining – Generates candidate set

$$C_K = C_K \cup Z$$

Where C_K is the collection of K-frequent candidate itemsets and Z is the new itemset of size k.

2. Pruning – Identifies potential itemsets

$$P_K = P_K \cup Z$$

Where P_K is the collection of K-frequent potential itemsets.

3. Verification – Generates frequent itemset

$$F_K = F_K \cup Z$$

Where F_K is the collection of K-frequent itemsets.

Example of Apriori Algorithm:

1. Consider the transactional data source having variety of products with transactional id's

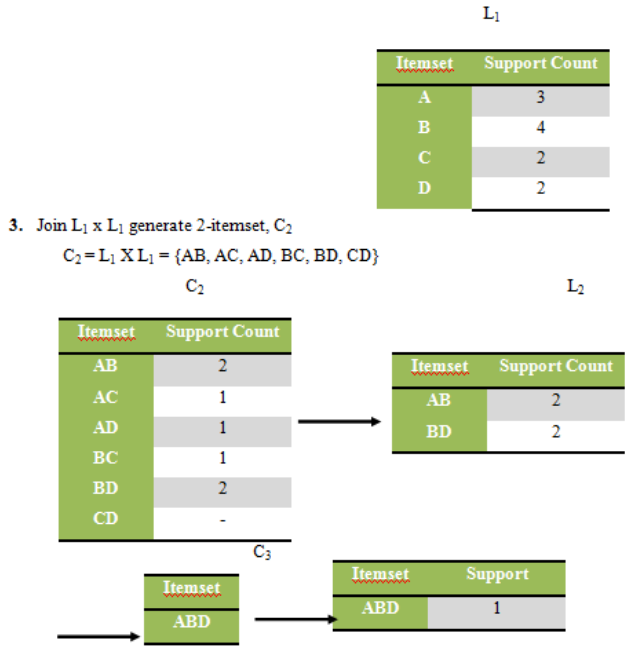
Transactional ID's	List of items
T_1	ABE
T_2	BD
T_3	BC
T_4	ABD
T_5	AC

2. Determine 1 - itemsets then make C_1

Itemset	Support Count
A	3
B	4
C	2
D	2
E	1

→ C_1

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 17, 2017 ISSN: 2348-6600
Reference ID: IJCS-238 PAGE NO: 1488-1501
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com



Novel Pruning Approach for Association Rule Mining:

Filtration step is introduced in this algorithm which is an alternative step for pruning in apriori perquisite for modified Apriori algorithm is same like classical with the following lemma[1].

Lemma:

If any itemset Z is infrequent then of it superset can be frequent

Proof:

Let Y be an itemset containing all items of Z and number of items in set Y which is greater than Z. (ie) Z < Y. it can be stated as

Support(Z) < minsup ---------- (1)

It is clear that support(Y) is not greater than any subset became cardinality(Y) is more than Z. (i.e) |Z|<|Y|. It can be written as,

Support(Y) <= Support(Z) ---------- -(2)

Using (1) and (2) of association can be treated as support(Y) < minsup. Otherwise if support of Z is infrequent then Y cannot be frequent.

Procedure:

1. **Joining**

   Generates candidate set is

   $C_K = C_K \cup Z$

2. **Filtration**

   Removes weak candidate set

   $P_K = P_K - Y, IF_K = IF_K \cup Y$

3. **Verification**

   Generates frequent and infrequent itemset

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 17, 2017 ISSN: 2348-6600
Reference ID: IJCS-238 PAGE NO: 1488-1501
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

$$F_K = F_K \cup Z$$

$$IF_K = IF_K \cup Z$$

Example for modified algorithm with five transaction such as {T_1, T_2, ……T_5} as follows in table.

Transactional ID's	Itemset
T_1	ABE
T_2	BD
T_3	BC
T_4	ABD
T_5	AC

Let D = { T_1, T_2, ……T_5} be the set of five transactions then user fix the threshold limit. Frequency of each item is calculated as shown below

Itemset	Frequency
A	3
B	4
C	2
D	2
E	1

1-Itemset

Frequency itemsets are extracted based on minimum support. Here minsup=1 fixed by user. Based on the frequency, item 'E' is infrequent which is eliminated initially.

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

C_2 ={AB, AC, AD, BC, BD, CD}

Potential itemsets are

P_2 ={AB, AC, AD, BC, BD, CD}

In third step, frequencies are compared with minimum support and following 2-frequent itemsets are generated. Frequent itemsets are {AB, BD}

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 17, 2017 ISSN: 2348-6600
Reference ID: IJCS-238 PAGE NO: 1488-1501
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

Candidate Set	Frequency
AB	2
AC	1
AD	1
BC	1
BD	2
CD	-

2-Itemset

In the second itemsets, 3-frequent itemsets are generated by using former steps.

C_3 ={ABC, ABD, ACD, BCD}

At this point, {AC, BC} are infrequent so it is to be eliminated. Potential 3-itemset dataset is

P_2 ={A B D}

Itemset	Support Count
ABD	1

Finally, then is no 3-itemset.

Frequent Itemset Generation Using Binary Search:

Author suggested binary search method, to generate frequent itemsets. Mid value is calculated by low(minimum itemset number) and high(maximum itemset number) value. Based on the mid value, candidate sets are generated.[16]

Procedure:

- Input dataset and threshold value
- Calculate length of large itemset in the corresponding dataset
- Apply binary search to find frequent itemsets

low = length of small itemset

high = length of large itemset

mid = (low + high) /2

while(low<=high)

if mid>threshold + length

low = mid + 1

high = mid - 1

Based on the mid value, candidate sets are generated.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 17, 2017 ISSN: 2348-6600
Reference ID: IJCS-238 PAGE NO: 1488-1501
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

1. Consider the transactional database having 5 items with their transactional id's.

Transactional ID's	List of Items
T_1	ABE
T_2	BD
T_3	BC
T_4	ABD
T_5	AC

2. Set the minimum threshold value calculate,

   Low = length of minimum itemset = 2

   High = length of maximum itemset = 3

   Mid = (low + high)/2

   = (2 + 3)/2 = 5/2 = 2

3. Create 2 itemsets based on mid value(2).

Itemset	Support Count
AB	2
AC	1
AD	1
AE	1
BC	1
BD	2
BE	1
CD	-
CE	-
DE	-

4. Itemset with greater than threshold value as

Itemset	Support Count
AB	2
BD	2

5. Set low = mid + 1

   low = 2 + 1 = 3

   high = 3

   mid = (3 + 3)/2 = 6/2 = 3

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 17, 2017 ISSN: 2348-6600
Reference ID: IJCS-238 PAGE NO: 1488-1501
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

create 3-itemset based on mid(3).

Itemset	Support Count
ABC	-
ABD	1
ABE	1
ACD	-
ACE	-
BCD	-
BCE	-
CDE	-
DEA	-

Items with count greater then threshold value are not found.

high = mid – 1 = 3 -1 = 2

low = 3

low > high (i.e) 3 > 2

Therefore, terminate the process.

Mining And Hiding Of Sensitive Association Rule Using Improved Aprori Algorithm:

This algorithm focuses improved apriori algorithm to reduce query frequencies and storage resources, without generating new candidate set, frequent itemset are mined using an improved apriori. This algorithm introduces a new method to avoid redundant generation of sub itemset during pruning.[2]

Method:

- Compare minimum support count by minsup * 101
- Produce L_1, candidate set and simultaneously constust TID set for each item during scanning
- Produce L_1, candidate set from joining $L_1 * L_1$
- Delete the patterns whose frequencies are not satisfy minimum support based on threshold limit
- In order to produce L_K join items which satisfy the rule
- Compare with minimum confidence for selecting the hidden rule

For hiding sensitive rule, mark the sensitive association rule where the confidence level is greater then minimum confidence. Transaction database D as follows

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 17, 2017 ISSN: 2348-6600
Reference ID: IJCS-238 PAGE NO: 1488-1501
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

TID	Items
1	ABE
2	BD
3	BC
4	ABD
5	AC

Let TID-set denote the set of transactions which contain A in D. so the number of transactions are exactly defined by minimum support count. Similarly, 1-itemset as follows

Item	TID_Set	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

Item	TID_Set	Support
AB	2	1,4
AC	1	5
AD	1	4
BC	1	3
BD	2	2,4
CD	-	-

Here, E's frequency is less than threshold value. So, it is eliminated then candidate set L_2 is generated as follows2-Itemset

2-itemset generation as Follows

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 17, 2017 ISSN: 2348-6600
Reference ID: IJCS-238 PAGE NO: 1488-1501
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

Itemset	Support
AB	2
BD	2

Frequent 2-itemset

Candidate set L_3 is generated for 3-itemset.

Items	Support
ABD	1

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

IV. PROPOSED METHODOLOGY

This approach integrates the concept of pruning, and binary search method. This proposed idea improves classical apriori algorithm and removes infrequent itemset with minimum number of database scanning. It consists of 5 steps named as

1. Joining – Generate candidate set
2. Pruning – Identify infrequent itemsets
3. Binary Search – Candidate set generation
4. Verification – Extract frequent itemsets based on minimum support
5. Hiding – Extract sensitive rules based on minimum confidence

Procedure for integrated approach as follows:

Initialize

Set of transactions(D), minimum support count(minsup), F_K(frequent itemset)

1. **Joining**

   Generate the candidate sets based on association and join the sets $L_1 * L_1$ (i.e) $C_K = C_K \cup Z$

2. **Filtration**

   Identify infrequent itemsets based on frequencies from previous candidate set generation and remove it from potential itemsets (i.e) $P_K = P_K - I_F$

3. **Binary Search**

   Identify the itemset generation based on mid value. (i.e) mid = (low + high)/2

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 17, 2017 ISSN: 2348-6600
Reference ID: IJCS-238 PAGE NO: 1488-1501
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

4. Verification

Compare each potential itemset with minsup. If frequency is larger than equal to minsup then it refers frequent itemset.

$F_K = F_K \cup Z$ → Frequent item

else

$IF_K = IF_K \cup Z$ → Infrequent item

5. Hiding Sensitive Rule

Hiding sensitive rule based on minimum confidence. Calculate the confidence for each frequent itemset Z and compare it with minimum confidence. If frequency is greater than minimum confidence then include this to sensitive rule dataset.

$S_K = S_K \cup Z$

Example,

TID	Items
T_1	ABE
T_2	BD
T_3	BC
T_4	ABD
T_5	AC

Let TID denotes the set of transactions and 5 items are available.

Itemset	Supcount
A	3
B	4
C	2
D	2
E	1

1-itemset

User specified threshold value is 2. So value 'E' is eliminated.

After eliminating 'E', binary search method applies in the transactional database.

low = 2 high = 3

mid = (2 + 3)/2 = 5/2 = 2

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 17, 2017 ISSN: 2348-6600
Reference ID: IJCS-238 PAGE NO: 1488-1501
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

Candidate set generation based on mid value(2)

Itemset	Supcount
AB	2
AC	1
AD	1
BC	1
BD	2
CD	-

→

Itemset	Supcount
AB	2
BD	2

Frequent 2-itemset

Sensitive itemsets are generated based on minimum confidence value.

Itemset	Supcount	Confidence
AB	2	0.67
BD	2	0.5

→ Sensitive Item

Conclusion

This paper suggested an approach for removing infrequent itemsets and hiding sensitive rules based on pruning and binary search method towards reduces the database scanning drastically. By using this method, some candidate K-frequent itemsets are generated by apriori pruning and binary search method. It is clearly stated that the proposed method would be better satisfaction than existing approach. In future, this approach will be implemented in real time application.

V. REFERENCES

[1] Shintre Sonali Sambhaji, Kalyanakar Pravin P. "Mining and Hiding of Sensitive Association Rule by using Improved Apriori Algorithm", Proceedings 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 Assocation 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

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 17, 2017 ISSN: 2348-6600
Reference ID: IJCS-238 PAGE NO: 1488-1501
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

International Conference on Very large databases, New York: IEEE press, PP. 487-499,1994

[5] S. Brin, etc – Dynamic itemset counting and implication rules for market based analysis, Proceedings ACM SIGMOD International Conference of Data, PP. 255-264, 1994.

[6] M. J. Zaki, "Scalable algorithms for assoiction mining", IEEE transaction on knowledge and data engineering, pp.372-390, 2000.

[7] J. Han, J. Pei, Y. Yin- Mining 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.Hsiano, "Charm: An efficient algorithm for closed itemset mining", Proc SIAM international conference 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. Intl conf. knowledge discovery and data mining, pp. 236-245, 2003.

[13] G. Grahne, and J.Zhu "Effciently 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.

[16] Manisha Kundal and Dr. Parminder Kaur, "Various Frequent Item Set on Data Mining

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 17, 2017 ISSN: 2348-6600
Reference ID: IJCS-238 PAGE NO: 1488-1501
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

Technique", Global Journal of Computer Science and Technology Software and Data Engineering, Vol 15 issue 4 version 1.0, ISSN: 0975-4172, 2015