IARM with User Specified Constraint and K-Subset Methodology

Sangita Kalmodia and Jitendranath Mungara

Computer Science and Engineering, CMRIT/ VTU-Belgum, India
sanjaychoudhar@gmail.com, jmungara@yahoo.com

Abstract

To considered the problem of discovering of interesting association rule among item sets in data base. Algorithms for mining association rule are practical methods to find interesting rules implied in large database. Proposed an innovative approach, beyond minimum support and minimum confidence framework, some extra measures consider for rule improvement is user interestingness constraint. It use three user defined constraint minimum support, minimum confidence, and interesting item and in addition makes use of k- nonempty subset generation methodology of the item which are user interest. Proposed algorithm fundamentally different from the identified algorithms, a number of algorithm is developed for association rule mining, the identified algorithms go through as of number of scanning of data base, and generate the candidate item set , unnecessary or uninteresting rule . The current method applies the user interesting constraint to generate only interesting association rule in data base. Proposed approach not just reduces the number of scanning of data base but also generated frequent itemset, and mine interesting association rule. Experimental result shows that the number of uninteresting rules can be reduced successfully and the validity of rules which mined are better.

Keywords: *Association rule, Frequent item, Interesting item, Minimum support, Minimum confidence*

1. Introduction

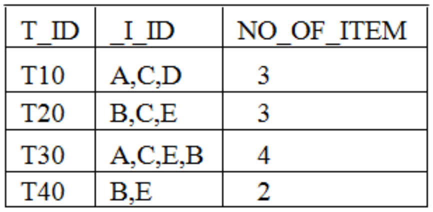
Frequent item set mining lead to the finding of association rule among item in large datasets. With enormous amount of data endlessly being together and stored, many industries are becoming interesting in mining such pattern from their data base. How to efficiently analyze these data and extract patterns representing knowledge implicitly stored in data source becomes more and more important. Even though, simple statistical techniques and machine learning for data analysis re developed long ago. Data mining is a new age group of information processing technology which is the process of extracting previously unknown and useful information from a large data sources. It has been extensively applied to a wide variety of applications like sales analysis, healthcare, manufacturing, financial analysis etc.

A number of studies have been proposed on efficient data mining methods and its relevant applications. The methods of data mining can extract interesting (nontrivial, implied, previously unknown and potentially useful) patterns or knowledge from huge amount of data In general, the data mining research targets on extracting approximate knowledge/patterns by only scanning the data stream once so that the mining results can be obtained as quickly as possible. The traditional concept of association rule mining will lose some information in generating association rules therefore some approach is developed [4]. In this the current research the problem of finding frequent item and mining of association rule in data base is discussed

2. Problem Statements

There are lots of data mining methods, out of this association rules finding might be the most promising one. It was introduced by R. Agrawal, et al., [8-9]. The problem can be stated as follows. Let I={x_1, x_2, x_3………x_m} be a set of items and T be a set of transactions, and each transaction T contain a set of items such that T ⊆I, the Table 1 is shown.

Table 1. Data Base



T_ID	_I_ID	NO_OF_ITEM
T10	A,C,D	3
T20	B,C,E	3
T30	A,C,E,B	4
T40	B,E	2

A unique identifier is associated with each transaction called it TID. Here, transaction T contains A, a set of some items in I, if A ⊆T. An association rule is an implication of the form C => B, where C subset of I, B ⊂I, and C ∩B = φ.

An association rule is expressed as:

C=>B

Where C and B are sets of items. The implication of such a rule is that transactions of the data base which hold C tend to hold B. An illustration of an association rule is “45% of transactions that hold C also hold B, 5% of all transactions hold both C and B”. Here, 45% is called the confidence of the rule, and 5% is the support of the rule.

Mining of association rule can be done in two steps:

First, find all the combination of items or item set whose support is greater than a user specified minimum support. Identify these combinations frequent item set.

Second, use the generated frequent item set in first step to create the desired rules. In general the idea is that, if ABC is frequent item set, then easily it can determine whether the rule AB=>C holds by computing the ratio

r = support (ABC)/support (AB) (1□)

The rule holds only if r >= minimum confidence (2)

Where minimum support (C=>B): s=P (C, B)

And minimum confidence (C=>B): c=P (B/C)

The most fundamental algorithm i.e. APRIORI for finding the association rule is focus on finding the association rules among all items in the database that satisfy user specified minimum confidence and minimum support. To minimize the number of frequent item sets and scanning of database some algorithm [3] are also proposed, such as SETM, PARTITION, FUP, FUFIA etc. But to generate frequent item the first sub step is still used. Many problems observe in finding association rule by standard algorithm. In practice, user is often concerned in finding association rules have only some specified items rather than all items. The call for to change the minimum confidence and support to obtain rule according to user specified constrain. So here a novel approach proposed to make several consecutive queries with different interested items, minimum confidence and support for potential rules. After the rules set is generated, user have to identify the interesting rules from the large rules include many uninteresting rule.

Using the existing association rule mining algorithms will have the following drawbacks:

First, many irrelevant or uninteresting rules will be generated. For example, data base to mining association rules in a transactional data as shown as in Table I. While using Apriori [8] to mine association rules, some rules such as {B, C} => {E} in the results. In this case, a lot of association rules for all item sets in only partial of which are really interested based on user preference Table I data Base

Second to generate the candidate frequent item sets, the whole database to be scanned more than a few times. It's very inefficient in allowing for the big overhead of reading the large database even though only partial items are of user interest. In practice, the weight of different item and transaction of record is different and it's not necessary to scan all item and transaction several times.

Consider the above mention drawbacks, the current approach present a new association rule mining algorithm that can discover the interesting frequent item and interesting association rules involving specified items.

3. Interestingness Constraints

In this section, a new approach proposes to solve the problem defined in Section 1. Database contains many transactions with many item, a few of them are interesting to user and also may change time to time based on the requirement.

Hence, the proposed approach use the interesting constraint to estimate whether an association rule is interesting or not, the estimation expression is

$$I\,(r) = f\,(IA(r), s(r), c(r)) \qquad (3)$$

In this expression, s(r) is the minimum support of association rule r, c(r) is the minimum confidence.

$$IA(r) = \begin{cases} 1, \text{if rule includes user specified item}, \\ \qquad\qquad 0, \text{else} \end{cases} \qquad (4)$$

When a generated rule satisfies s(r), c(r) and IA(r), it's IA = 1, therefore it can be considered as an interesting rule. Otherwise it will be discarded.

4. Algorithm with Interestingness Constraint

Based on user interestingness constraint an interesting association rule mining (IARM) algorithm with constraint on item is designed for generation of frequent item, closed, maximal and mining interesting association rule. In this algorithm user specify the interesting items and minimum support for frequent item set generation and minimum confidence to generate the association rule. The algorithm is shown in Fig1.

Algorithm

Input: transactional database *D*, the set of interesting item K, minimum confidence c, minimum support s

Output: the set of rules satisfy user constraint.

```
1. Find unique item id in D
2. Generate subset (non empty) from unique item set
3. For each sub set do
4.     If subset ^ interesting item $\neq Ø$ then
5.         (Insert the subset in intermediate data structure with number of item available in subset)
6.     End if
7. for each transaction do
8.     If interesting ^ item ids in Trans.! =Ø then
9.     for each subset in intermediate DB do
10.        If No of item in subset >No of item in transaction and If interesting t. constrains ^ item ids in Trans. $\neq Ø$ Then
11.            Increase the count of subset by one
12.    End
13. End
14. Set the flag='F' for all the subset which is having count $\geq$Min. Support.
15. for each set L having flag='F' do
16.     for each subset $\alpha$ of set do
17.     If support count (L) / support count ($\alpha$) $\geq$c then
18            generate a rule as $\alpha$ => (L-$\alpha$),s,c
19. Insert the new rule into result
```

Figure 1. IARM Algorithm

5. Illustration

Consider for example, a sample database given in Table I. Use the next mask to generate the subset of unique item in intermediate data structure, if, the set of n elements, a valid mask would be an array of n Boolean (true/false; 1/0) elements. When apply a mask to a set, check each element (e) in the set and the corresponding one in the mask (m): if m is true (1), add e to the result, otherwise, ignore it. After applying the mask (0, 1, 0, 0, 1) to {A, B, C, D, E}, {B, E} is the generated subset. In the same way S(I)={A},{B},{C},{D},{E},{AB},{AC},{AD},{AE},{BC},{BD},{BE},{CD},{CE},{DE}

,{A,B,C}{A,B,D},{A,B,E},{A,C,D},{A,C,D},{A,D,E},{B,C,D},{B,D,E},{C,D,E},{A, B,C,D},

{A,B,C,E},{A,B,D,E},{A,C,D,E},{B,C,D,E}, {A, B, C, D, E}

All the subset S (I) is generate.

Now suppose in worst case all are interesting item. Then S1 (I) are obtained by using the following procedure

Subset ^ interesting item ≠Ø then the result will be

S_1(I)={{A},{B},{C},{D},{E},{AB},{AC},{AD},{AE},{BC},{BD},{BE},{CD},{CE},{ DE},

{A,B,C},{A,B,D},{A,B,E},{A,C,D},{A,C,D},{A,D,E},{B,C,D},{B,D,E},{C,D,E},{A,B, C,D},

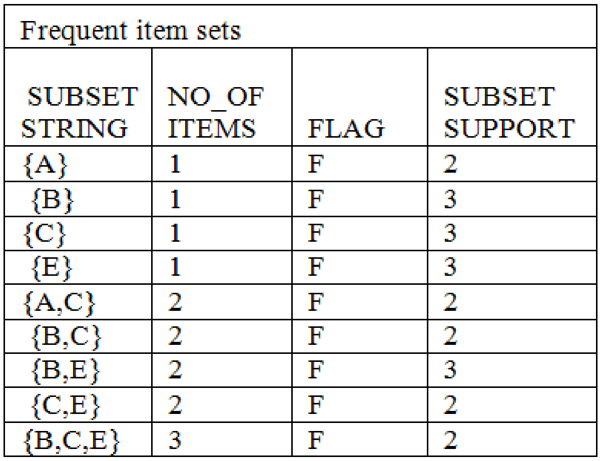
{A,B,C,E},{A,B,D,E},{A,C,D,E},{B,C,D,E},{A, B, C, D, E}

Now to check for frequent item minimum support is 50%,

$$\text{Support}=\frac{\text{Number of transaction contain the items}}{\text{Number of transaction}}$$

Generated frequent item set L is shown in the table II and stored this in some intermediate datastruter.

Table 2. Generated Frequent Item



Frequent item sets			
SUBSET STRING	NO_OF ITEMS	FLAG	SUBSET SUPPORT
{A}	1	F	2
{B}	1	F	3
{C}	1	F	3
{E}	1	F	3
{A,C}	2	F	2
{B,C}	2	F	2
{B,E}	2	F	3
{C,E}	2	F	2
{B,C,E}	3	F	2

Expression to calculate the minimum confidence value.

For every subset α of frequent item L if

$$\frac{\text{Support}_{\text{count}(L)}}{\text{Support}_{\text{count}(\alpha)}} \geq c \qquad (5)$$

Then, Insert the rule α → L in the result.

If minimum confidence is 80%. Then the generated rule is

$$\{A\}\rightarrow\{C\}$$
$$\{B\} \rightarrow \{E\}$$
$$\{E\} \rightarrow \{B\}$$
$$\{B, C\}\rightarrow \{E\}$$
$$\{C, E\} \rightarrow \{B\}$$

6. Performance Evaluations

The performance of association rule generation in the terms of time and the number rule generated by varying the minimum support, interesting item and minimum confidence is evaluated and resultant graph is shown below.

Taking the minimum support on X-axis and time (millisecond) on Y-axis, by varying the value of minimum support, the resultant graph is shown in Figure 1.

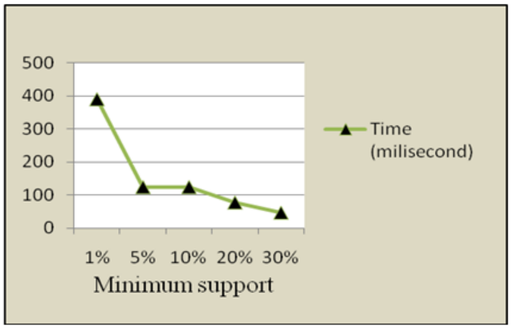


Figure 1. Performance Evaluation of Time Consumption

Hence it can conclude by increasing the value of minimum support time taken is reduced.

Tacking the Minimum support on X-axis and number of rule on Y-axis, by varying the Minimum support the number of rule generated is given by graph is shown in Figure 2.

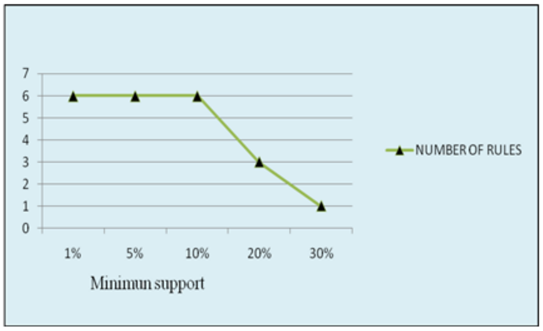


Figure 2. Number of Rule

Taking the constant value of minimum support and minimum confidence and varying the value of interesting item, the resultant graph is show in Figure 3 As indicate by the graph the number of rule generated by varying the interesting item is varying from 32 to 47 which is a significant difference in small data base hence if the algorithm is applied to large database with user interest the number of rule generated will be strong and less in number as compare to other algorithm.

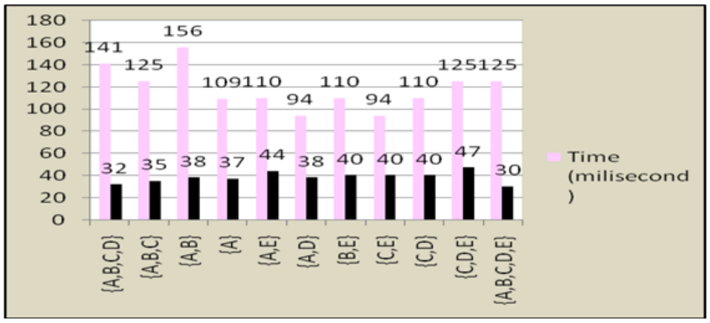


Figure 3. Performance Evaluation of Number of Rule and Time with varying Interesting Item

Number of transaction verses time (millisecond) taken, with increases the number of transaction the time taken is approximately linearly increases. Resultant graph is shown in Figure 4.

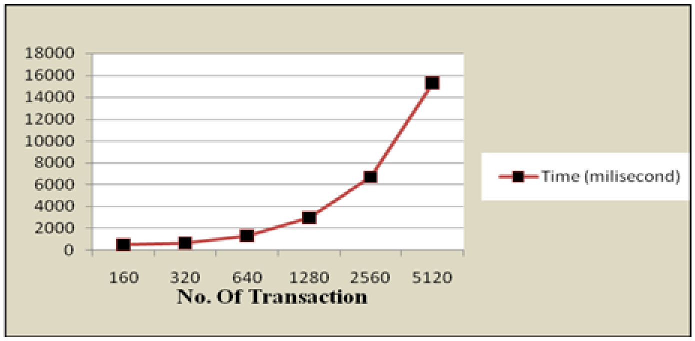


Figure 4. Performance Evaluation of Time (millisecond) With Varying Number of Transaction

7. Conclusion

In this work an algorithm for interesting association rule has been developed in data mining .Interesting, constraints is introduced because many uninteresting rule will generate and that is May not according to the user interest therefore to generate only those rule that satisfy user specification.

In comparison to other Association rule generation algorithm (Apriori) interesting constraints have the following benefits:

- **Less and strong Association rule**: Due to user specified interesting item to generate both frequent item set and association rule it generate only those rule and frequent item set that contain this user specified item set and the number of association rule are less as compare to Apriori algorithm.
- **Less number of scanning of data base:** The known Algorithm, by Apriori, the scanning of transactional data base is more because each time to generate next higher set of item it has to scan the whole data base.
- **No lexicographical order:** In the other known Algorithm the order of item id should be in lexicographical but in this algorithm item can be in any order.

Acknowledgements

I would especially like to thank my research project supervisor, **Dr. Jitendranath Mungara Ph.D (Electronics), (Ph.D) Ad hoc Networks** for his invaluable guidance and tremendous motivation that he provided at every step of this work. His enthusiastic support and untiring interest in the subject is deeply appreciated. I am particularly grateful to my family for providing me with a source of strength and encouragement, and giving me the best possible education, and imbibing in me a thirst for learning.

References

[1] V. Umarani and M. Punithavalli, **"**A Study on EfctiveMining Association from Huge DataBase", International Journal of Computer Science and Research, vol. 1, Issue 1, **(2010)**.

[2] **U.** Fayyad, S. Chaudhuri and P. Bradley, "Data mining and its role in database systems", vol. 5, no. 6, **(1993)**, pp. 914. 925.

[3] B. Kalpana and R. Nadarajan, "Optimizing Search Space Pruning in Frequent Itemset Mining With Hybrid Traversal Strategies-A Comparative Performance on Different Data Organizations", International Journal of Computer Science, vol. 34, no. 1, IJCS_34_1_13.

[4] V. R. Vedula and S. Thatavarti, "Binary Association Rule Mining Using Bayesian Network", IPCSIT, vol. 4 **(2011)**.

[5] P. Adriaans and D. Zantinge, "Data mining", Addison-Wesley, **(1999)**.

[6] J. Han, M. Kamber and J. Pei, "Data Mining: Concepts and Techniques", 3rd Edition, ISBN-9780123814791.

[7] M. Zakrzewicz, "Efficient Constraint-Based Sequential Pattern Mining Using Dataset Filtering Techniques", Proc. of the 5th International Baltic Conference on Databases and Information Systems, **(2002)**.

[8] R. Agrawal, Imielinski and A. Swami, "Mining association rules betweensets of items in large databases", In Proc. of the ACM SIGMOD Conference Management of Data, Washington D.C., **(1993)** May.

[9] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases", Prof. 20th Int'l Conf. Very Large Data Bases, **(1994)**, pp. 478-499.

[10] A. Savasere, E. Omiecinski and S. Navathe, "An Efficient Algorithm for Mining Association Rules in Large Databases", Proceedings of the 21st VLDB Conference Zurich, Swizerland, **(1995)**.

Authors

Sangita Kalmodia

M.Tech in Computer Science and Engineering in 2012 from Visvesvaraya Technological University Belgaum, B.TECH in Computer Science and Engineering in 2008 from Rajiv Gandhi Technological University Bhopal. Her field of study is Theory of Computation, Data Mining, and Data Base Management System.

Dr. Jitendranath Mungara

Double Doctorate in electronics and computer science and system engineering. He is working as a prof. and dean of computer science and engineering department in CMRIT Bangalore. He has published 45 papers in International Journals and conference. He is guiding three P.HD students in Mobile Adhoc Network.