Formation of Association Rules on Hubness Using Data Mining

Vidisha H. Zodape (MTech. III[rd]sem, student), Prof. Leena H. Patil (Lecturer)

***ABSTRACT* - Discovery of association rules among the huge data sets is considered as an important feature of data mining. Every day it is going to be very essential to find pattern from large data which enhances the association rule mining. There are many algorithms and techniques have discovered for determining association rules. Among all algorithms and techniques Apriori algorithm and FT tree growth are at the top of the list. But there are some disadvantages of both the algorithms. What if we merge these two algorithms and get most appropriate results? In this paper, both algorithm are used in combination so that we can get the optimized result and get exact what the user is looking for from the huge database.**

***Index Terms*: Data Mining, Association rules, Apriori algorithm, FP Tree algorithm.**

INTRODUCTION

Every day bulk of information is stored into the database making task of data mining [9] more and more difficult. Users wish for the tool/search engine which will provide relevant information. Service providers will have to find the techniques to create the web site by minimizing the load to best serve the site to the different users. Business analyst wants the tool to analyse the behaviour of consumer needs. Mining is the process of finding out what users are looking for on the internet, some are interested in document file, and some users are interested in media file or images. Various data mining techniques [11] such as, decision trees, association rules, and neural networks are proposed earlier and has grabbed attention for several years. Association rule mining [5] is the most efficient technique for data mining which finds the correlation among items from huge datasets.

Vidisha H. Zodape (MTech. III[rd]sem, student)
CSE Department, PIET
Nagpur, Maharashtra, India

Prof. Leena H. Patil (Lecturer)
CSE Department, PIET
Nagpur, Maharashtra, India

Association rule mining gives us the interrelation among the data that whether they occur simultaneously with other data. This paper presents association rule mining technique with combined apriori and FP tree growth algorithm approach on vertically distributed database. The advantage of this approach is we get most frequent items from large amount of dataset. The main idea is we have firstly applied apriori algorithm on main database then on the output of apriori algorithm. FP tree algorithm [14] is applied which is giving most frequent items. The Apriori Algorithm is powerful algorithm for mining frequent pattern [4] for boolean association rules. It is also an algorithm for frequent item set mining and association rule learning [9] over transactional databases. It identifies the frequent individual items in the database and expands them to larger item sets so that the item set appear most in database. Apriori uses breadth first search to extract the occurrence of individual item. But there are disadvantages of apriori that it generate number of candidate set and repeatedly scans database. On the other hand, FP growth is used to construct FP tree which is the mining of frequent pattern. FP tree provides compressed dataset. It also avoids repeatedly database scanning and candidate set generation. FP tree also have disadvantage that it cannot generate the candidate set. So when we apply FP tree on the output of apriori, FP gets candidate set for processing. Data mining techniques have been introduced successfully to retrieve knowledge in order to support a variety of domains marketing, weather forecasting, medical diagnosis, and national security. But it is still a challenge to mine the data by protecting the private database of user. Most organizations want information about individuals for their own specific needs. However, different units within an organization themselves share the information. In such cases, in each they must be sure that the privacy of the individual is not violated or that sensitive business information is not revealed. In order to provide security, records can be modified before the records are shared with

ISSN: 2278 – 1323

anyone who is not permitted directly to access the data. This can be done by deleting from the dataset some identity fields, such as name and passport number in passenger information record.

RELATED WORK

In paper [1], Research work focused on web use mining [5] and specifically keeps watch on running across the web utilization examples of sites from the server log records. The memory usage and time taken for performance is compared by means of Apriori algorithm and improved Frequent Pattern Tree algorithm. The central part of this method is the usage frequent-pattern tree (FP-tree), which keeps the piece set association information.

In simple words, this algorithm works as follows:

- It compresses the input database and gets an FPtree instance to represent common items.
- It then divides the compressed database into a set of conditional databases, each one associated with one common pattern.
- Eventually, each database is extract separately.

In [2] the Genetic algorithm (GA) is applied on large data sets to discover the frequent itemsets. We first load the sample of records from the transaction database that fits into memory. The genetic learning starts as follows. An initial population is created consisting of randomly generated transactions. Each transaction can be represented by a string of bits. In proposed genetic algorithm based method for finding frequent itemsets repeatedly transforms the population by executing the following steps:
Fitness Evaluation: The fitness (i.e., an objective function) is calculated for each individual.
Selection: Individuals are chosen from the current population as parents to be involved in recombination.
Recombination: New individuals (called offspring) are produced from the parents by applying genetic operators such as crossover and mutation.
Replacement: Some of the offspring are replaced with some individuals (usually with their parents).

The working of genetics algorithm is as follows:

- An initial population is created. It is a group of individuals and represents a candidate solution. A Chromosome i.e. individual is a string of genes.
- Select chromosomes with higher fitness.
- Crossover between the selected chromosomes to produce new offspring with better higher fitness
- Evolve the new chromosomes if needed.
- Terminate when an optimum solution is found. This generational process is repeated until a termination condition has been reached. Common terminating conditions are:
  1. A solution is found that satisfies minimum criteria
  2. Fixed number of generations reached
  3. Allocated budget (computation time/money) reached
  4. The highest ranking solution's fitness is reaching or has reached a plateau such that successive iterations no longer produce better results
  5. Manual inspection
  6. Combinations of the above

The main obstacle of FP growth is, it generates a huge number of conditional FP tree. In [3] proposed work focuses on a new and improved FP tree with a table and a new algorithm for mining association rules. Without generating the conditional FP tree the proposed algorithm mines all possible frequent item set. It also provides the frequency of frequent items, which is used to evaluate the desired association rules. In proposed FP-Tree consists of mainly two elements- the tree and a table. The tree represents the link among the items more specifically and table is used to store the spare items. We called it spare table which has two columnitem_name and frequency. Item_name is the name of the items and frequency means how many times it occurs in Stable. The main reason to discover the spare table is, in traditional FPtree a lot of branches are created and the same item appears in more than one node.

In [4] a text mining system is proposed to extract and use the information in radiology reports. The system consists of three main modules: medical finding extractor, report and image retriever. The medical finding extraction module is used to automatically retrieve medical findings and associated modifiers to construct radiology reports. The constructing of the free text reports viaduct the gap between users and report database and make it available to access. It also serves as median result to other components of the system. The retrieval module examine user's query and returns the reports and images that match the query.

In [5] Considered the applications on business environment, its benefits are defined by collaboration, team efforts and partnership, rather than individual efforts. So the collaboration is most important because it brings mutual benefit. Sometimes, collaboration even occurs among

ISSN: 2278 – 1323

competitors, or among companies that will give them an advantage over other competitors.

PROPOSED WORK

In proposed work two algorithms are combined to get the most relevant information from the database are apriori algorithm and FP tree algorithm. In proposed system apriori algorithm is applied on the database and on the output of apriori algorithm FP tree is applied and we get the relevant output.

The database used is the crime dataset which contains data about communities in US. The dataset contains 125 attributes called predictors and 18 potential goals about number of different types of crime in each community from 1995 for 48 states.

I. Apriori algorithm:

The primary key is crime description, which on selecting retrieves all the entries present in the database and shows in tabular form. Then support and confidence is calculated based on field serial no, IUCR, district, location, FBI code, community area and primary type that is description.

For example consider market basket:

Transaction ID	Items Bought
T1	{Mango, Onion, Nintendo, Key-chain, Eggs, Yo-Yo}
T2	{Doll, Onion, Nintendo, Key-chain, Eggs, Yo-Yo}
T3	{Mango, Apples, Key-chain, Eggs}
T4	{Mango, Umbrella, Corns, Key-chain, Yo-Yo}
T5	{Corn, Onion, Onion, Key-chain, Ice-cream, Eggs}

For simplicity
M = Mango
O = Onion
And so on
So the table becomes

Transaction ID	Items Bought
T1	{M,O,N,K,E,Y}
T2	{D,O,N,K,E,Y}
T3	{M,A,K,E}
T4	{M,U,C,K,Y}
T5	{C,O,O,K,I,E}

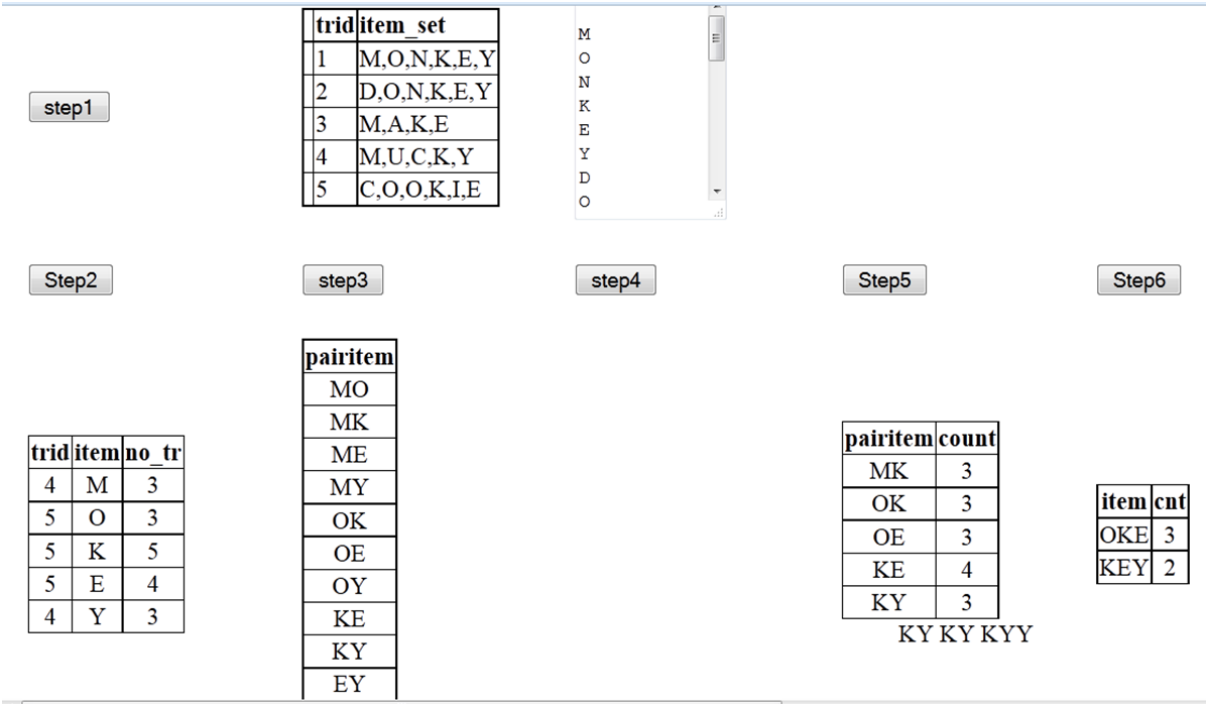


Figure1: working of apriori algorithm (snap 1)

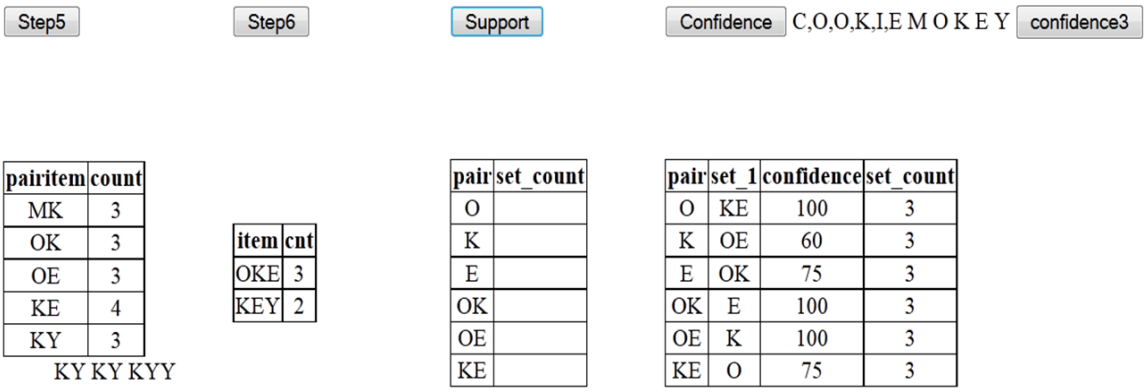
ISSN: 2278 – 1323

Step 1: Count the number of transactions in which each item occurs.

Now following a simple **golden rule**: an item/itemset is frequently bought if it is bought at least 60% of times. So for here it should be bought at least 3 times.

Step 2: So in this step we remove all the items that are bought less than 3 times from the table.

Step 3: Now we have to find a pair of items that are bought frequently.
Step 4: Now we count how many times each pair is bought together.
Step 5: Again applyingGolden rule. Remove all the item pairs with number of transactions less than three.
Step 6: Now we have to find a set of three items that are brought together. To make the set of three items we need one more rule (self-join).



pairitem	count
MK	3
OK	3
OE	3
KE	4
KY	3

item	cnt
OKE	3
KEY	2

pair	set_count
O	
K	
E	
OK	
OE	
KE	

pair	set_1	confidence	set_count
O	KE	100	3
K	OE	60	3
E	OK	75	3
OK	E	100	3
OE	K	100	3
KE	O	75	3

Figure 2: working of apriori algorithm (snap 2)

Now support and confidence is find.

a) Support: The rule X ⇒Y holds with support s if s% of transactions in D (i.e. database) contains X ∪ Y. Rules that have as greater than a user-specified support is said to have minimum support.

b) Confidence: The rule X⇒Y holds with confidence c if c% of the transactions in D (i.e. database) that contains X also contain Y. Rules that have a c greater than a user-specified confidence is said to have minimum confidence.

$$\text{conf}(R) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}$$

In proposed work for calculating support and count each row is compared with the main string and the support is calculated on the rule specified above. Same method is repeated for calculating the confidence using the formula above. We have given that confidence should be 53% at least. Only those items are considered. Each attributed is then counted and count is increased accordingly. And the result is printed on the screen.

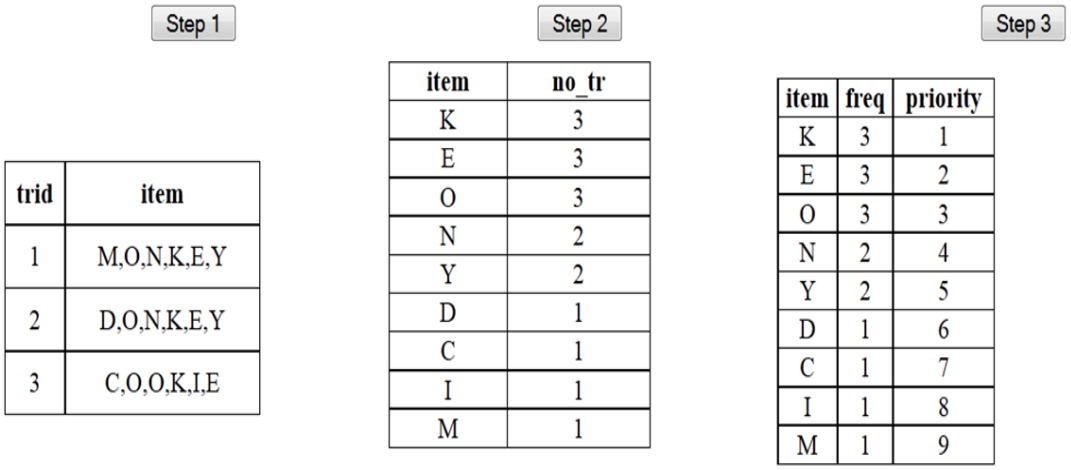
II. FP tree algorithm:

FP growth is used to construct FP tree which is the mining of frequent pattern. FP tree provides compressed dataset. It also avoids repeatedly database scanning. Firstly it scans database and finds the support for each item. Then items are removed which are not frequent. Sort other items in descending order based on counter value.

For example, consider the same example of market basket.

Transaction ID	Items Bought
T1	{Mango, Onion, Nintendo, Key-chain, Eggs, Yo-Yo}
T2	{Doll, Onion, Nintendo, Key-chain, Eggs, Yo-Yo}
T3	{Mango, Apples, Key-chain, Eggs}
T4	{Mango, Umbrella, Corns, Key-chain, Yo-Yo}
T5	{Corn, Onion, Onion, Key-chain, Ice-cream, Eggs}

ISSN: 2278 – 1323



trid	item
1	M,O,N,K,E,Y
2	D,O,N,K,E,Y
3	C,O,O,K,I,E

item	no_tr
K	3
E	3
O	3
N	2
Y	2
D	1
C	1
I	1
M	1

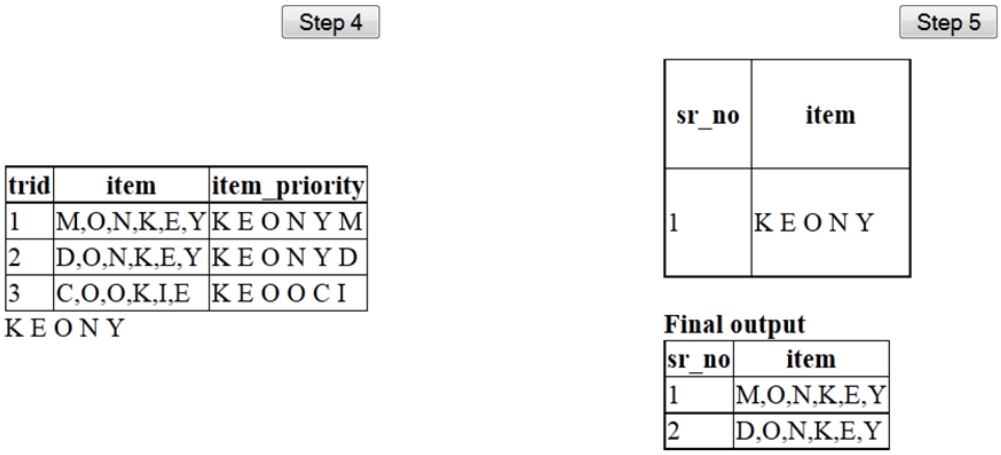
item	freq	priority
K	3	1
E	3	2
O	3	3
N	2	4
Y	2	5
D	1	6
C	1	7
I	1	8
M	1	9

Figure 3: applying FP on Apriori (snap 1)

Step 1: In this step final output of apriori algorithm in example is considered.

Step 2: The number of transaction is counted.

Step 3: The priority is counted.



trid	item	item_priority
1	M,O,N,K,E,Y	K E O N Y M
2	D,O,N,K,E,Y	K E O N Y D
3	C,O,O,K,I,E	K E O O C I

K E O N Y

sr_no	item
1	K E O N Y

Final output

sr_no	item
1	M,O,N,K,E,Y
2	D,O,N,K,E,Y

Figure 4: Applying FP on Apriori (snap 2)

Step 4: Items are set according to the priority.

Step 5: Frequent items are extracted and the final output is printed.

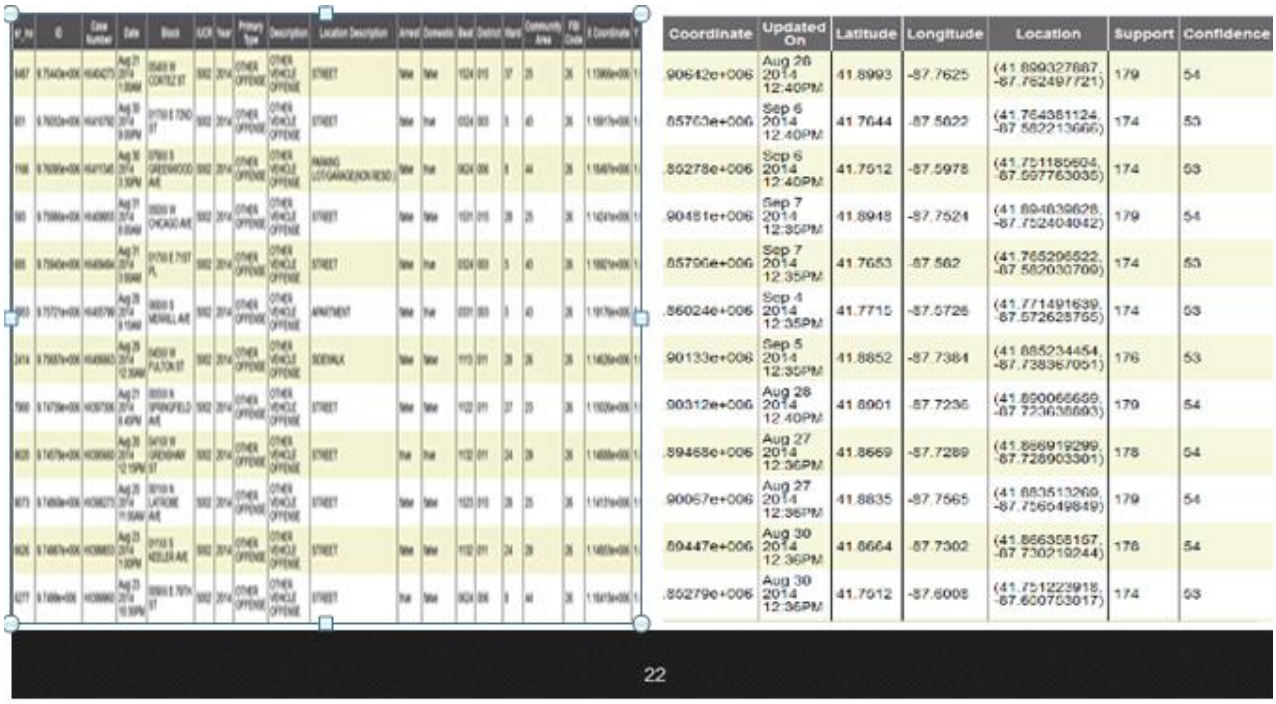
In proposed work after getting the result FP tree is applied. FP tree works on six field location, beat, district, ward, community area, update on. Firstly priority is set for this six field and arranged in descending order. The data is rearranged according to the priority. The frequency is calculated by the means of total rows dived by t_row (t_row is the row from the six fields we have considered) and after calculating frequency if it is greater than 5 then the row is inserted in the main table. Then each field is compared with each row to find the match. This searching is done based on priority. And then the final result is displayed.

EXPERIMENTAL RESULTS

The experimental result shows that the proposed system gives the most relevant result than the plain apriori algorithm (figure 5) and FP tree algorithm (figure 6).

For example in description other vehicle offence shown in graph, it results 55 records which are present in main database.On applying apriori on this 55 records we are getting 22 records as relevant output. And finally on applying FP tree we are getting 5 records as most relevant output. This shows that the proposed system is giving us the better results than the existing algorithms. Here we are getting candidate set for the FP Tree. As we know that FP Tree cannot generate good candidate set. So this is the advantage for the FP tree algorithm. After applying the merged approach of apriori and FP tree algorithm, the graph is designed to summaries the result.

ISSN: 2278 – 1323



Coordinate	Updated On	Latitude	Longitude	Location	Support	Confidence
90642e+006	Aug 26 2014 12:40PM	41.8993	-87.7625	(41.899327887, -87.762497721)	179	54
85763e+006	Sep 6 2014 12:40PM	41.7644	-87.5822	(41.764381124, -87.582213666)	174	53
85275e+006	Sep 6 2014 12:40PM	41.7512	-87.5978	(41.751185604, -87.597763035)	174	53
90481e+006	Sep 7 2014 12:35PM	41.8948	-87.7524	(41.894839628, -87.752404842)	179	54
85796e+006	Sep 7 2014 12:35PM	41.7653	-87.582	(41.765296522, -87.582030709)	174	53
86024e+006	Sep 4 2014 12:35PM	41.7715	-87.5726	(41.771491639, -87.572628765)	174	53
90133e+006	Sep 5 2014 12:35PM	41.8852	-87.7384	(41.885234454, -87.738367051)	176	53
90312e+006	Aug 28 2014 12:40PM	41.8901	-87.7236	(41.890066659, -87.723636893)	179	54
89468e+006	Aug 27 2014 12:35PM	41.8869	-87.7289	(41.886919299, -87.728903301)	178	54
90067e+006	Aug 27 2014 12:35PM	41.8835	-87.7565	(41.883513269, -87.756049849)	179	54
89447e+006	Aug 30 2014 12:35PM	41.8664	-87.7302	(41.866358167, -87.730219244)	176	54
85279e+006	Aug 30 2014 12:35PM	41.7512	-87.6008	(41.751223918, -87.600703017)	174	53

22

Figure 5: Output of apriori algorithm on crime dataset.

id	location	Beat	District	Ward	Community_Area	update_on	t_count	match1
7	OTHER VEHICLE OFFENSE	false	1113	011	28	26	100	72
10	OTHER VEHICLE OFFENSE	false	1523	015	28	26	100	70
11	OTHER VEHICLE OFFENSE	false	1132	011	24	26	100	72
15	OTHER VEHICLE OFFENSE	false	1135	011	2	26	100	70
16	OTHER VEHICLE OFFENSE	false	1135	011	28	26	100	73

5

Figure 6: Output of FP Tree on result of apriori.

For the aim of comparison is done with the plain algorithms itself i.e. apriori algorithm and FP tree growth algorithm (figure 7).

ISSN: 2278 – 1323

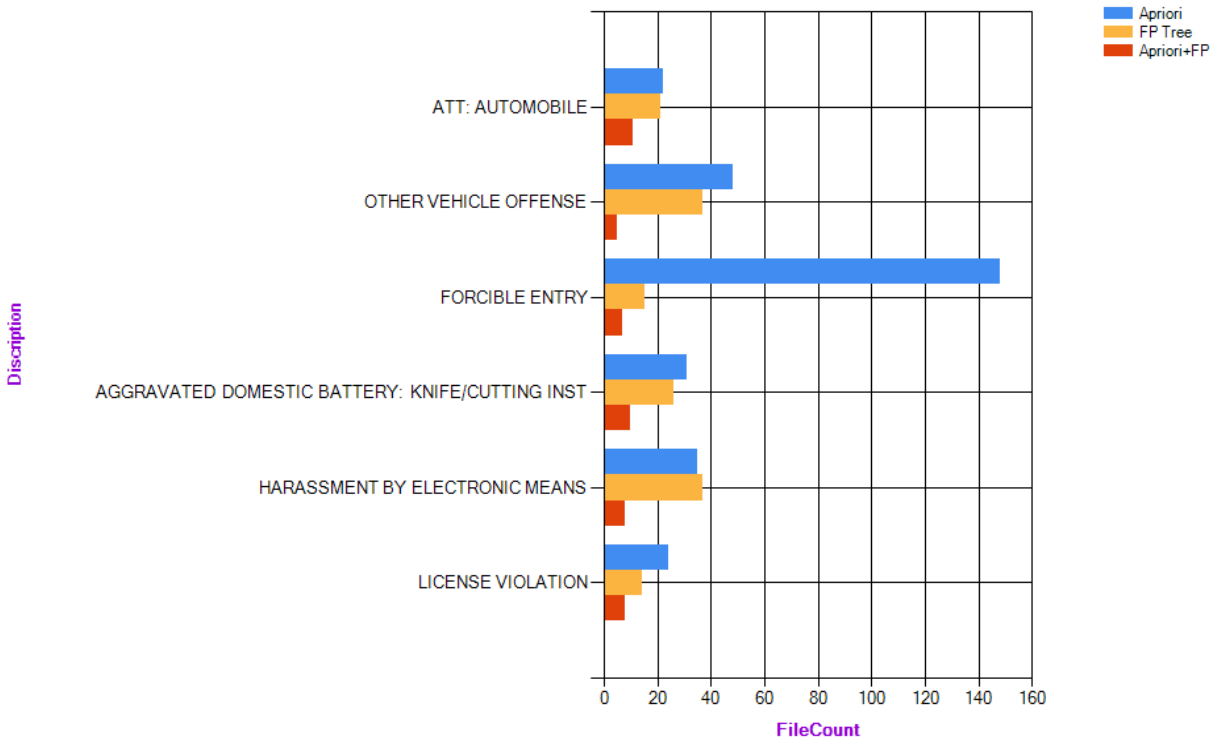


Figure 7: Comparison graph.

CONCLUSION

Association rule mining is the procedure of finding out the correlation among items and which users are looking for. This research work generated more efficient association rule [8] using combined apriori and FP tree algorithm. The comparison is shown with the base algorithm apriori algorithm and FP tree algorithm which shows that the combined approach is giving more efficient result. The proposed work is applied on crime database and can also be used for the organisational database, web based database. In future the work can be extended to extract the information from image files and web mining.

ACKNOWLEDGEMENT

Prof. L. H. Patil has guided me along the way for the completion of this paper and the project. With their support and expert advice on the subject matter. I was able to complete the paper successfully. I thank them for their valuable input and time.

REFERENCES

[1] Ms.Monalsaxena "Association rules Mining Using Improved Frequent Pattern Tree Algorithm", International Journal of Computing, Communications and Networking, Volume 2, No.4, October - December 2013

[2] "Identifying Best Association Rules and Their Optimization Using Genetic Algorithm", International Journal of Emerging Science and Engineering (IJESE) Volume-1, Issue-7, May 2013

[3] "An Improved Frequent Pattern Tree Based Association Rule Mining Technique", IEEE Trans. Knowledge and Data Eng 2011

[4] Han J "Mining frequent patterns without candidate rules mining technique," in the national seminar of the international web of data, ACM Press, pp. 4-11-2004

[5] M.J. Freedman, K. Nissim, and B. Pinkas,"Privacy Preserved Collaborative Secure Multiparty Data Mining," Proc. Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT), pp. 1-19, 2012.

[6] E-H. Han, G. Caryopsis "Scalable Data web mining for Association web Rules," IEEE Trans. Eng.,vol. 12, no. 3, July 2012.

[7] J. H and M. Kaber, "association mining:" 2014.

[8] Tan, P. N., M. St., V. Kumar, "Introduction to web Mining", Addison-Wesley, 2013, 769pp.

[9] MannilaH,"Efficient algorithms for discovering association rules mining." conference Knowledge Discovery in Databases (SIGKDD). 181-83.

[10] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation, 2nd ed. San Mateo

[11] Huang, H., Wu, X..Association analysis with one scans of web data bases. Paper submitted at the IEEE On Data Mining, Japan.

[12] Masseglia F., "Using Data Mining Techniques on Web Access Logs to Dynamically Improve Hypertext Structure language", In ACCM Web Letters, Vol. 10 No. 9, pp.13-19, 2011

[13] T. Tassa and E. Gudes, "Secure Mining of Association Rules in Horizontally Distributed Databases," IEEE Trans. Database Systems, vol. 37, 2014.

[14] R. Jin "An Efficient Implementation of Apriori Association web mining," Proc. Workshop on High Performance Data webMining, Apr. 2011.

[15] "ADMiner: An Incremental Data Mining Approach Using a Compressed FP-tree", *Journal of Software*, Vol 8, No 8 (2013), 2095-2103, Aug 2013

ISSN: 2278 – 1323