International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 09 | Sep 2020 www.irjet.net p-ISSN: 2395-0072

Frequent Itemset Mining on Large Uncertain Database

A. Periyanayaki[1], S. Shobana[2], E. Jasmine Reena Winsy[3]

[1,2,3]*Dept of Computer Science and Engineering, Kamaraj College of Engineering and Technology, Virudhunagar, Tamilnadu*

---***---

Abstract -*The frequent item set mining meet some challenges by large scale and rapidly expanding datasets. In sensor monitoring system and data integration System the data manipulated is highly ambiguous. Frequent itemset mining from generous uncertain database illustrated under possible world semantics is a crucial dispute. In our project, Approximated algorithm is established to extract the threshold based PFI from generous ambiguous database exceedingly .Incremental frequent itemset algorithm is used to retain the mining sequence. This reduces the need of re-executing the whole mining algorithm on the new database, which is often more expensive and unnecessary. Here both tuple and attribute uncertainty is reinforced. The efficiency of our proposed algorithm is validated by interpreting both real and synthetic dataset.*

***Key Words*: Frequent itemset mining, Incremental mining, Approximation algorithm.**

1. INTRODUCTION

Data mining is defined as a process used to extract usable data from a larger set of any raw data. Data mining is also known as Knowledge Discovery in Data (KDD). It is an interdisciplinary sub field of computer science. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interesting metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. It implies analyzing data patterns in large batches of data using one or more software. Data mining has applications in multiple fields, like science and research. Data mining involves effective data collection and warehousing as well as computer processing. For segmenting the data and evaluating the probability of future events, data mining uses sophisticated mathematical algorithms. The objective of this process is to sort through the large quantities of data and discover new information. The benefit of data mining is to turn this new found knowledge into actionable result, such as increasing a customer's likelihood to buy, or decreasing the number of fraudulent claims. Data Mining is also the search for valuable information in large volumes of data. It is a cooperative effort of humans and computers. Humans, design databases, describe problems and set goals. Computers sift through data, looking for patterns that many these goals. A marketing company using historical response data to build models to predict how will respond to a direct mail or telephone solicitation is using data mining. A manufacturer analyzing sensor data to isolate conditions that lead to unplanned production stoppages is also used in data mining.

2. RELATED WORKS

Sapna Saparia et al[2], The problem of frequent pattern mining with uncertain data they will show how broad classes of algorithms can be extended to the uncertain data setting. In particular, they will study candidate generate-and-test algorithms, hyper-structure algorithms and pattern growth based algorithms. One of their insightful observations is that the experimental behavior of different classes of algorithms is very different in the uncertain case as compared to the Deterministic case. In particular, the hyper-structure and the candidate generate-and-test algorithms perform much better than tree-based algorithms. They will test the approach on a number of real and synthetic data sets, and show the effectiveness of two of our approaches over competitive techniques.

Hong Cheng[3]et al, Frequent itemset mining has been a focused theme in data mining research and an important first step in the analysis of data arising in a broad range of applications. The traditional exact model for frequent itemset requires that every item occur in each supporting transaction. However, real application data is usually subject to random noise or measurement error, which poses new challenges for the efficient discovery of frequent itemset from the noisy data. Mining approximate frequent itemset in the presence of noise involves two key issues: the definition of a noise-tolerant mining model and the design of an efficient mining algorithm. In this chapter, they will give an overview of the approximate itemset mining algorithms in the presence of random noise and examine several noise-tolerant mining approaches.

Ben kao[4] et al, They study the problem of mining frequent itemsets from uncertain data under a probabilistic framework. They consider transactions whose items are associated with existential probabilities and give a formal definition of frequent patterns under such an uncertain data model. They show that traditional algorithms for mining frequent itemsets are either inapplicable or computationally inefficient under such a model. A *data trimming* framework is proposed to improve mining efficiency. Through extensive experiments, They show that the data trimming technique can achieve significant savings in both CPU cost and I/O cost.

Smith Tsang[5] et al, Traditional decision tree classifiers work with data whose values are known and precise. They extend such classifiers to handle data with uncertain information, which originates from measurement or quantization errors, data staleness, multiple repeated measurements, etc. The value uncertainty is represented by multiple values forming a probability distribution function. They discover that the accuracy of a decision tree classifier can be much improved if the whole pdf, rather than a simple statistic, is taken into account. They extend classical decision tree building algorithms to handle data tuples with uncertain values. Since processing pdf's is computationally more costly, they propose a series of pruning techniques that can greatly improve the efficiency of the construction of decision trees.

Man Lung Yiu[6] et al, They study the problem of answering spatial queries in databases where objects exist with some uncertainty and they are associated with an existential probability. The goal of a thresholding probabilistic spatial query is to retrieve the objects that qualify the spatial predicates with probability that exceeds a threshold. Accordingly, a ranking probabilistic spatial query selects the objects with the highest probabilities to qualify the spatial predicates. They propose adaptations of spatial access methods and search algorithms for probabilistic versions of range queries, nearest neighbors, spatial skylines, and reverse nearest neighbors and conduct an extensive experimental study, which evaluates the effectiveness of proposed solutions.

3. IMPLEMENTATION

The system architecture of "Efficient mining of frequent itemset in large uncertain database" is shown below through a simplified diagram.

In this architecture diagram, the text document is first taken as an input. The text document is the xml file then the file is converted into text file and then load into the database. For the given data the support value is calculated then the given items are sorted based on the support value. Then the candidate generation of 2 and 3 itemet based on dynamic programming algorithm. Detect the Minimal support. Then threshold based Candidate 2 and 3 itemset based on Model based algorithm is calculated. Calculating the time taken to compute PFI by both the techniques. Calculating the number of PFI obtained by both the techniques. Finally we prove that our method out performs much better than the existing one by graph.

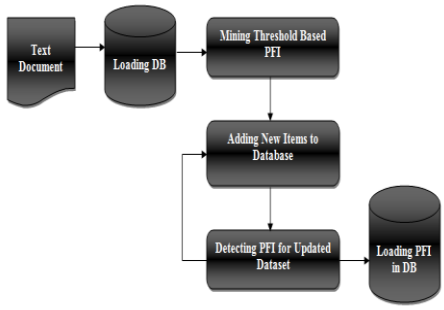


Fig 4.1 System Diagram

4. RESULTS AND DISCUSSION

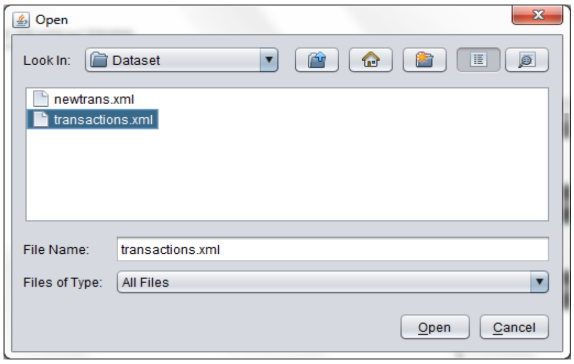


Fig 6.1.1 Opening a xml file

Fig 6.1.1 shows that the input xml file is open from the computer.

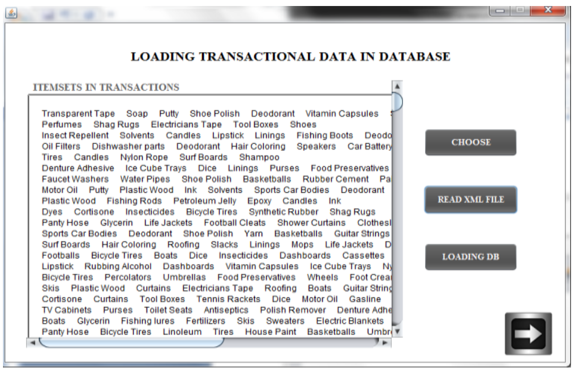


Fig 6.1.2 Loading the data

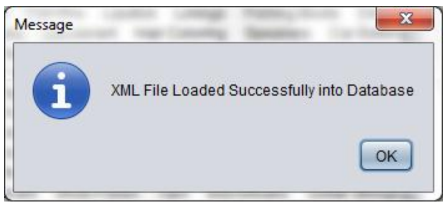


Fig 6.1.2.1

Fig 6.1.2 and fig 6.1.2.1 shows that the xml file is loaded into the database.

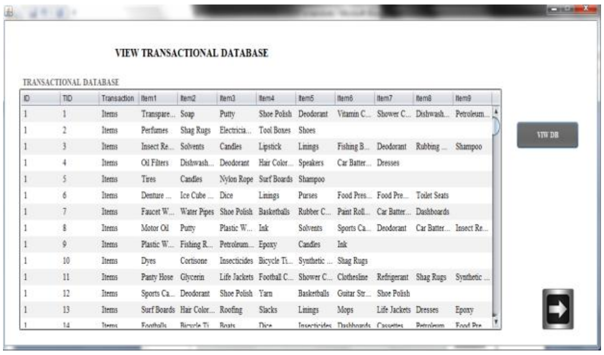


Fig 6.1.3 Transactional database

Fig 6.1.3 display the transactional database to the user.

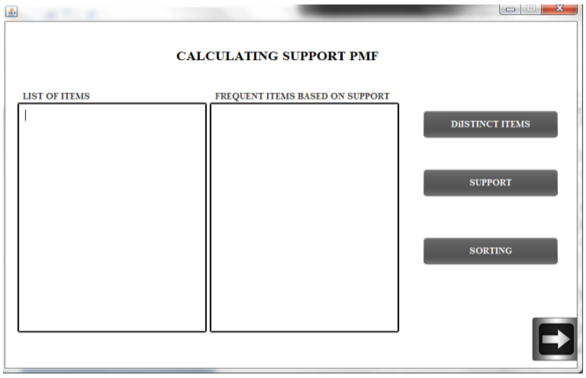


Fig 6.1.4 Calculating the support value

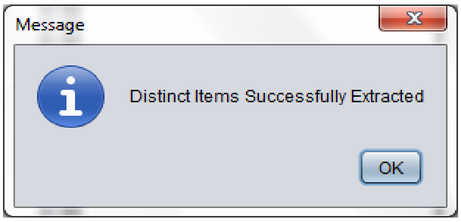


Fig 6.1.4.1

Fig 6.1.4 and fig 6.1.4.1 shows the distinct items that are extracted from the database.

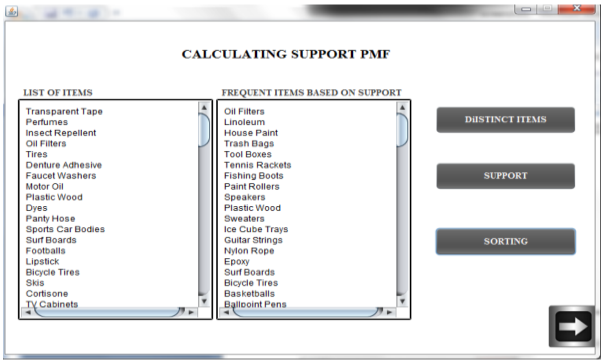


Fig 6.1.5 Sorting the items

Fig 6.1.5 shows the sorting of the items based on the support count.

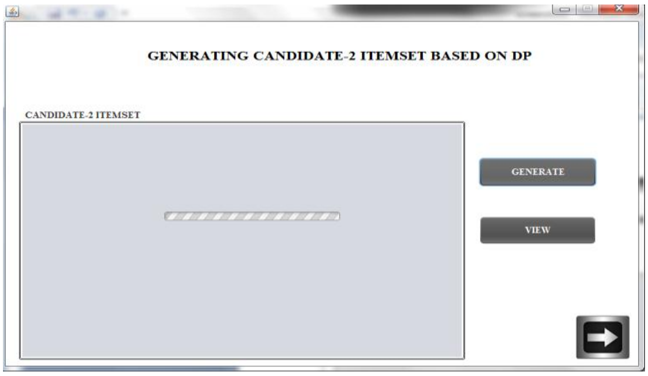


Fig 6.1.6 Generation of candidate-2 itemset based on dp

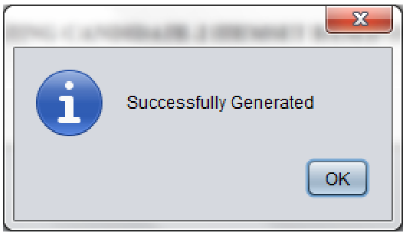


Fig 6.1.6.1

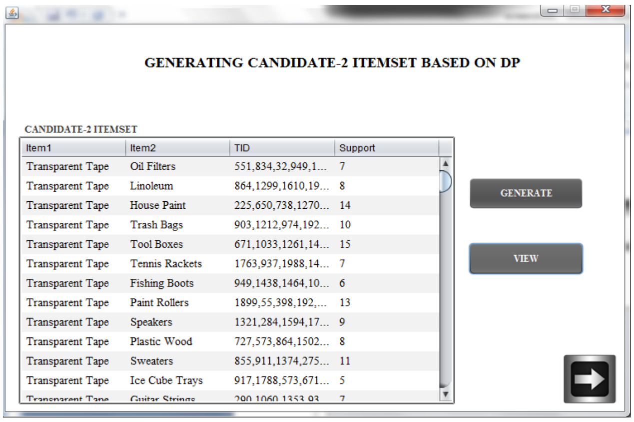


Fig 6.1.6.2 Result

Fig 6.1.6 and fig 6.1.6.2 shows the generation of candidate -2 itemset based on dynamic programming algorithm.

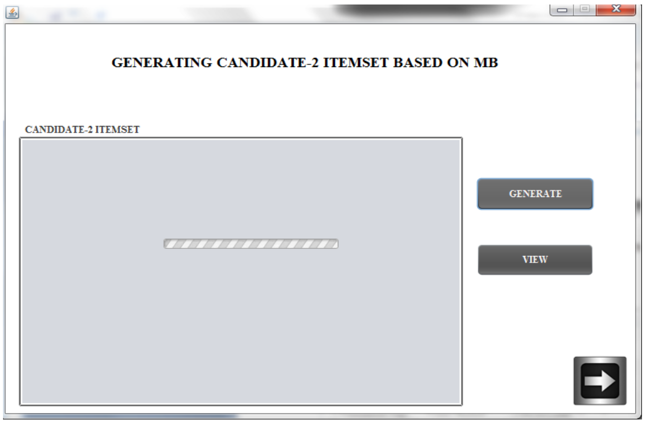


Fig 6.1.7 Generation of candidate-2 itemset based on mb

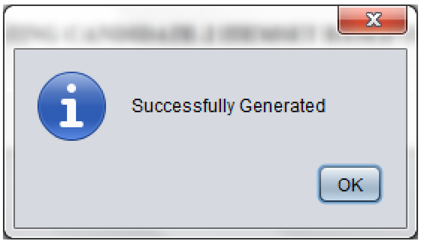
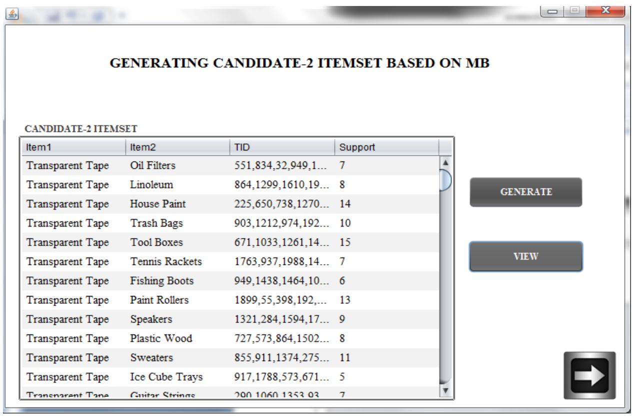


Fig 6.1.7.1

Fig 6.1.7 and fig 6.1.7.2 shows the generation of candidate-2 itemset based on model based algorithm.



Item1	Item2	TID	Support
Transparent Tape	Oil Filters	551,834,32,949,1...	7
Transparent Tape	Linoleum	864,1299,1610,19...	8
Transparent Tape	House Paint	225,650,738,1270...	14
Transparent Tape	Trash Bags	903,1212,974,192...	10
Transparent Tape	Tool Boxes	671,1033,1261,14...	15
Transparent Tape	Tennis Rackets	1763,937,1988,14...	7
Transparent Tape	Fishing Boots	949,1438,1464,10...	6
Transparent Tape	Paint Rollers	1899,55,398,192,...	13
Transparent Tape	Speakers	1321,284,1594,17...	9
Transparent Tape	Plastic Wood	727,573,864,1502...	8
Transparent Tape	Sweaters	855,911,1374,275...	11
Transparent Tape	Ice Cube Trays	917,1788,573,671...	5

Fig 6.1.7.2 Result

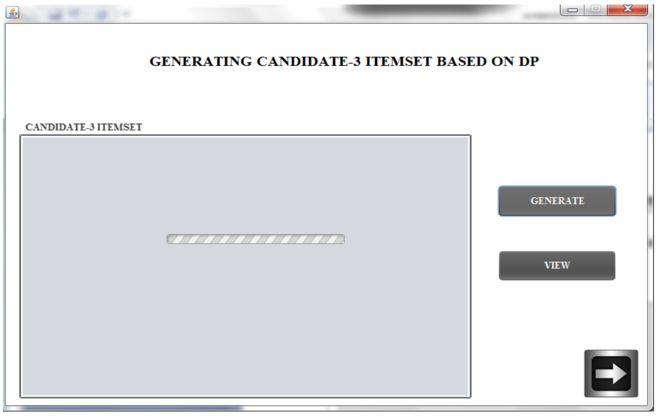


Fig 6.1.8 Generation of candidate-3 itemset based on dp

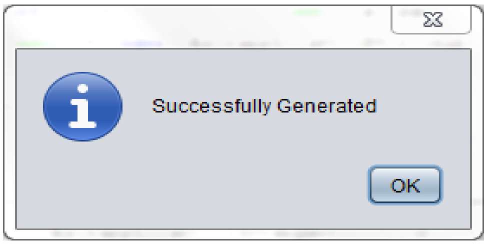
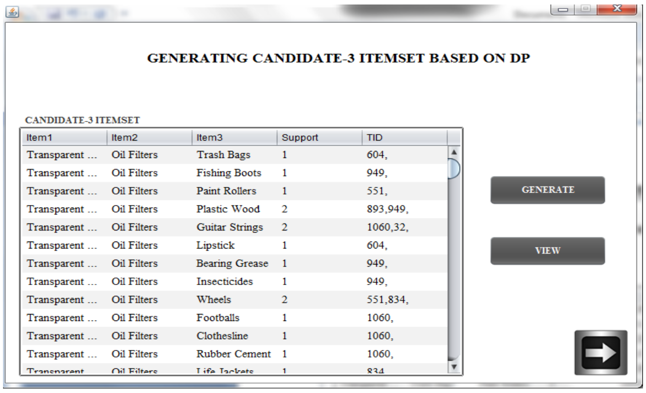


Fig 6.1.8.1



Item1	Item2	Item3	Support	TID
Transparent ...	Oil Filters	Trash Bags	1	604,
Transparent ...	Oil Filters	Fishing Boots	1	949,
Transparent ...	Oil Filters	Paint Rollers	1	551,
Transparent ...	Oil Filters	Plastic Wood	2	893,949,
Transparent ...	Oil Filters	Guitar Strings	2	1060,32,
Transparent ...	Oil Filters	Lipstick	1	604,
Transparent ...	Oil Filters	Bearing Grease	1	949,
Transparent ...	Oil Filters	Insecticides	1	949,
Transparent ...	Oil Filters	Wheels	2	551,834,
Transparent ...	Oil Filters	Footballs	1	1060,
Transparent ...	Oil Filters	Clothesline	1	1060,
Transparent ...	Oil Filters	Rubber Cement	1	1060,

Fig 6.1.8.2 Result

Fig 6.1.8 and fig 6.1.8.2 shows the generation of candidate -3 itemset based on dynamic programming algorithm.

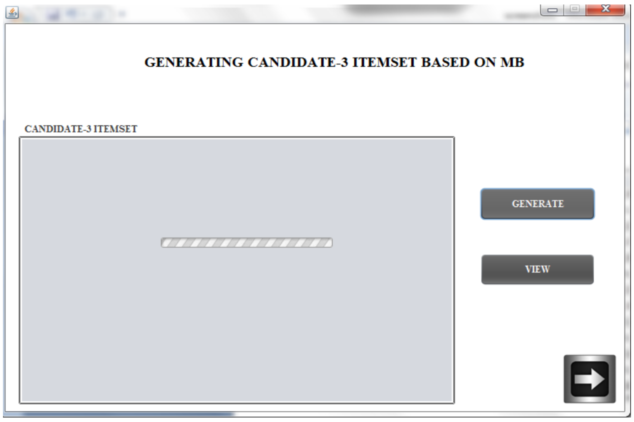


Fig 6.1.9 Generation of candidate-3 itemset based on mb

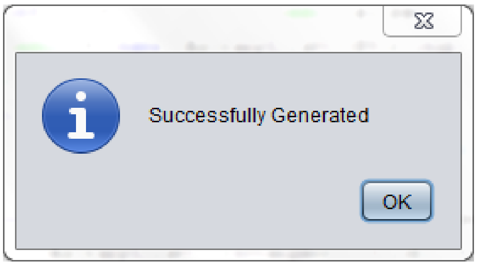
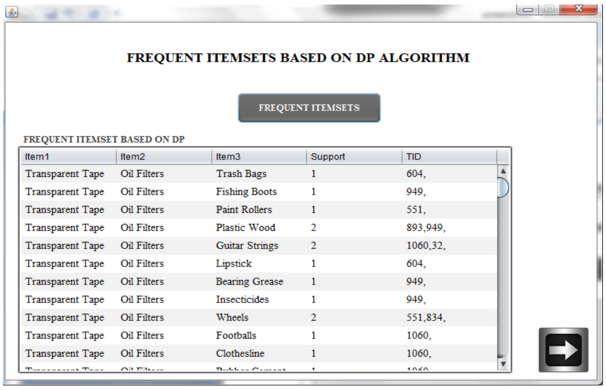


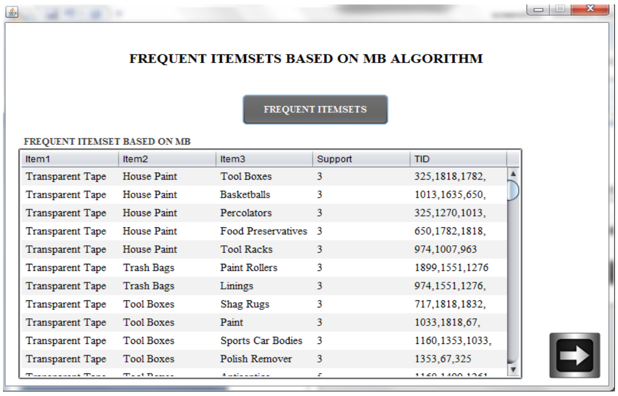
Fig 6.1.9.1

Fig 6.1.9 and fig 6.1.9.2 shows the generation of candidate -3 itemset based on model based algorithm.



Item1	Item2	Item3	Support	TID
Transparent Tape	Oil Filters	Trash Bags	1	604,
Transparent Tape	Oil Filters	Fishing Boots	1	949,
Transparent Tape	Oil Filters	Paint Rollers	1	551,
Transparent Tape	Oil Filters	Plastic Wood	2	893,949,
Transparent Tape	Oil Filters	Guitar Strings	2	1060,32,
Transparent Tape	Oil Filters	Lipstick	1	604,
Transparent Tape	Oil Filters	Bearing Grease	1	949,
Transparent Tape	Oil Filters	Insecticides	1	949,
Transparent Tape	Oil Filters	Wheels	2	551,834,
Transparent Tape	Oil Filters	Footballs	1	1060,
Transparent Tape	Oil Filters	Clothesline	1	1060,

Fig 6.1.10 Frequent itemset based on dp



Item1	Item2	Item3	Support	TID
Transparent Tape	House Paint	Tool Boxes	3	325,1818,1782,
Transparent Tape	House Paint	Basketballs	3	1013,1635,650,
Transparent Tape	House Paint	Percolators	3	325,1270,1013,
Transparent Tape	House Paint	Food Preservatives	3	650,1782,1818,
Transparent Tape	House Paint	Tool Racks	3	974,1007,963
Transparent Tape	Trash Bags	Paint Rollers	3	1899,1551,1276
Transparent Tape	Trash Bags	Linings	3	974,1551,1276,
Transparent Tape	Tool Boxes	Shag Rugs	3	717,1818,1832,
Transparent Tape	Tool Boxes	Paint	3	1033,1818,67,
Transparent Tape	Tool Boxes	Sports Car Bodies	3	1160,1353,1033,
Transparent Tape	Tool Boxes	Polish Remover	3	1353,67,325

Fig 6.1.11 Frequent itemset based on mb

Fig 6.1.10 and fig 6.1.11 shows generation of frequent itemset based on dynamic programming and model based algorithm.

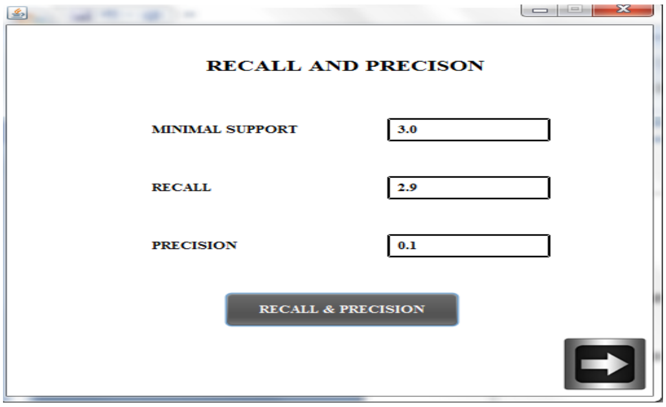


Fig 6.1.10 Recall and precision

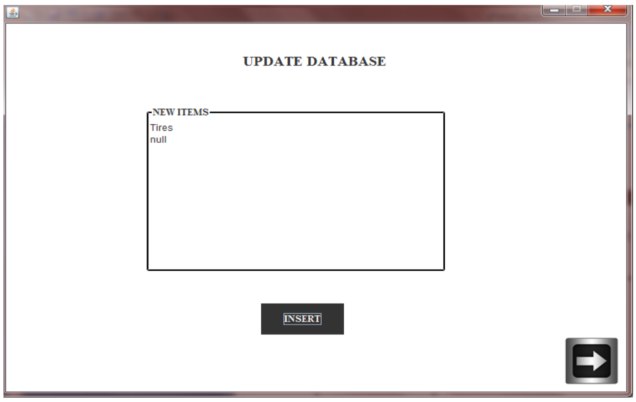


Fig 6.1.11 Inserting the new items

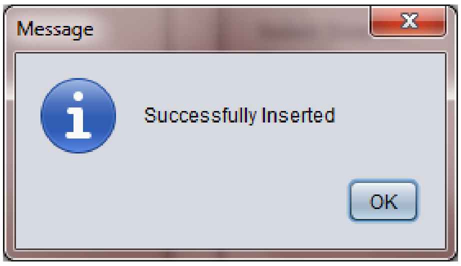
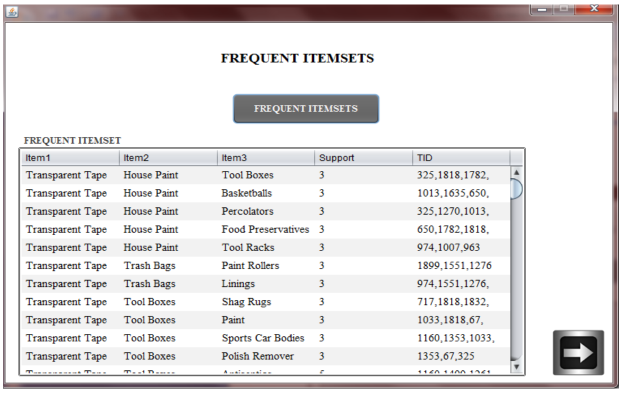


Fig 6.1.11.1

Fig 6.1.11 and fig 6.1.11.1 shows the insertion of new items into the database.



Item1	Item2	Item3	Support	TID
Transparent Tape	House Paint	Tool Boxes	3	325,1818,1782,
Transparent Tape	House Paint	Basketballs	3	1013,1635,650,
Transparent Tape	House Paint	Percolators	3	325,1270,1013,
Transparent Tape	House Paint	Food Preservatives	3	650,1782,1818,
Transparent Tape	House Paint	Tool Racks	3	974,1007,963
Transparent Tape	Trash Bags	Paint Rollers	3	1899,1551,1276
Transparent Tape	Trash Bags	Linings	3	974,1551,1276,
Transparent Tape	Tool Boxes	Shag Rugs	3	717,1818,1832,
Transparent Tape	Tool Boxes	Paint	3	1033,1818,67,
Transparent Tape	Tool Boxes	Sports Car Bodies	3	1160,1353,1033,
Transparent Tape	Tool Boxes	Polish Remover	3	1353,67,325

Fig 6.1.12 Frequent itemset

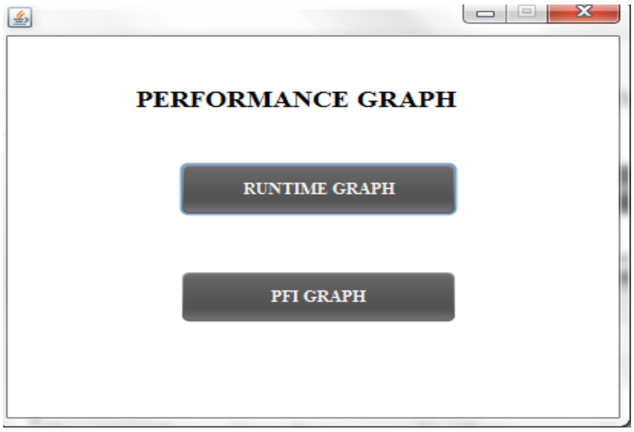


Fig 6.1.13 Performance graph

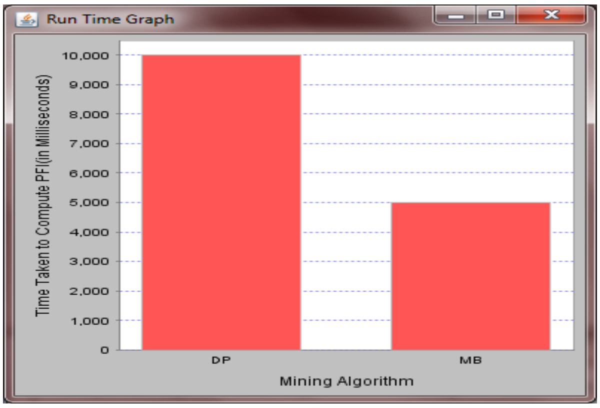


Fig 6.1.13.1 Time taken graph

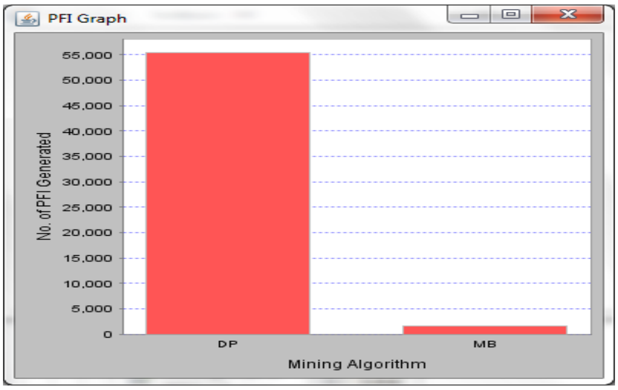


Fig 6.1.13.2 PFI generation graph

Fig 6.1.13.1 and 6.1.13.2 shows the comparion of both algorithm based on their computation time and number of PFI generation.

5. CONCLUSION AND FUTURE WORK

The proposed system is used to find the frequent itemset of large uncertain database. A Model-based approach is adduced to excerpt threshold-based Probabilistic Frequent Itemset (PFI).Support Probabilistic Mass Function is approximated by using model based approach to justify the PFI expeditiously. Incremental Mining Algorithm is analyzed to fetch PFIs from progressing database. The efficiency and accuracy of Incremental mining algorithm is proved. They support both Tuple and Attribute ambiguity in uncertain data. In future this mining model solves the update and deletes operations based itemset mining problems and Rule Mining Algorithm can be used to extract frequent item set from large uncertain database.

REFERENCES

[1] Liang Wang, David Wai-Lok Cheung, Reynold Cheng, Member, IEEE, Sau Dan Lee, and Xuan S. Yang,"Efficient mining of frequent itemset",(2012), vol 24.

[2] Sapna Saparia, Prof. Madhushree B, "A Review Paper on Frequent Pattern Mining with Uncertain Data", (2015).

[3] H. Cheng, P. Yu, and J. Han. "Approximate frequent itemset mining in the presence of random noise" Soft Computing for Knowledge Discovery and Data Mining,(2008)

[4] Chun-Kit Chui1, Ben Kao1, and Edward Hung2," Mining Frequent Itemsets from Uncertain Data",(2007).[5] Smith Tsang, Ben Kao, Kevin Y. Yip, Wai-Shing Ho, and Sau Dan Lee," Decision Trees for Uncertain Data",(2011), vol 3.

[5] Man Lung Yiu ,Nikos Mamoulis, Xiangyuan Dai,Yufei Tao, "Efficient Evaluation of Probabilistic Advanced Spatial Queries on Existentially Uncertain Data",(2008), vol 21

[6] T. Bernecker, H. Kriegel, M. Renz, F. Verhein, and A. Zuefle."Probabilistic frequent itemset mining in uncertain databases",In KDD,(2009).

[7] L. Wang, R. Cheng, S. D. Lee, and Cheung,"Accelerating probabilistic frequent itemset mining: A model-based approach", In CIKM, (2010).

[8] Q. Zhang, F. Li, and K. Yi." Finding frequent items in probabilistic data. In SIGMOD",(2008).