IT in Industry, Vol. 9, No.2, 2021 Published Online 10-April-2021

Survey on Frequent Itemset Mining Methods and Techniques

Divvela Srinivasa Rao [1]* Dr.V.Sucharita[2]

[1]Research Scholar, Computer Science and Engineering Department,
Koneru Lakshmaiah Educational Foundation, Guntur, Andhra Pradesh, India,

[1]Sr.Assistant Professor, Computer Science and Engineering Department,
Lakireddy Bali Reddy College of Engineering, Mylavaram, Andhra Pradesh, India,

[2]Professor, Computer Science and Engineering Department,
Narayana Engineering College, Gudur, Andhra Pradesh, India,

Email: [1] srinumtechcse2007@gmail.com,[2]jesuchi78@gmail.com

Abstract

In the decision making process the Data Analytics plays an important role. The Insights that are obtained from pattern analysis gives many benefits like cost cutting, good revenue, and better competitive advantage. On the other hand the patterns of frequent itemsets that are hidden consume more time for extraction when data increases over time. However less memory consumption is required for mining the patterns of frequent itemsets because of heavy computation. Therefore, an algorithm required must be efficient for mining the patterns of the frequent itemsets that are hidden which takes less memory with short run time. This paper presents a review of different algorithms for finding Frequent Itemsets so that a more efficient algorithm for finding frequent items sets can be developed.

Keywords: Data mining, Frequent itemset, patterns, association rule mining, Apriori algorithm.

1. Introduction

In the current world all the business organization has good competition to get profits. The business executive makes the strategic decision for getting success. By daily activities a larger amount of data is generated in the areas of business, science and engineering because of advances in computerized techniques. Most of the users are not having any ideas about the various patterns of the data. So, mining the important data from large amounts is required for the process of decision making. Extracting the useful information is the challenging task. Frequently, the data required is extracted first from a data warehouse into database[1][9] as shown in Fig 1. Therefore a data mining system is required to mine different patterns of data that fulfill the requirements of the users. Therefore the process uses different tools for extracting the knowledge from the huge datasets. The term knowledge discovery is used to tell that discovery useful knowledge from huge amounts of data. The steps in the knowledge discovery process are cleaning of data, integration, selection, mining, transformation, and representation of the results to ensure that the useful information is extracted from the data. The noise and irrelevant data are removed in the cleaning phase. Different data sources are combined during integration phase. Relevant data are retrieved during data selection as per the user needs. In the data transformation process the data selected is transformed into the appropriate forms required. In the mining step various methods are applied for extracting the patterns. In the evaluation of the

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

patterns interesting patterns are identified based on measures given. In the final step the knowledge discovered is represented visually to the user.

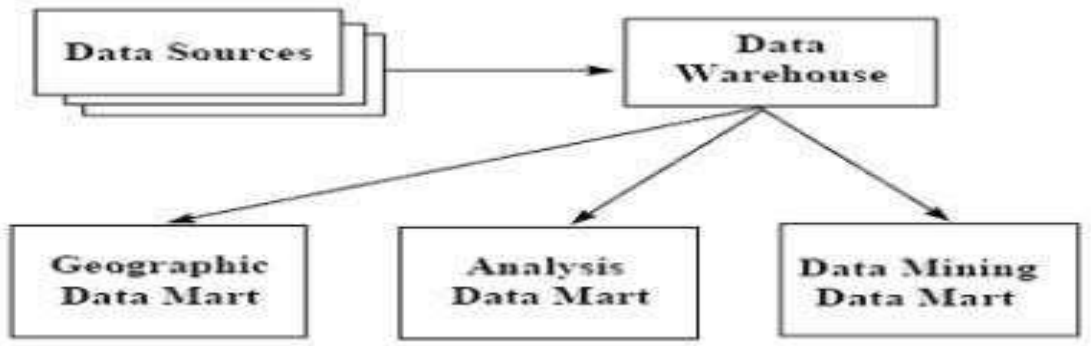


Figure 1. Data Warehouse and its Relations

2. Association Rule Mining

The major techniques in data mining are Association rule mining. It is used for finding associations, patterns that occur frequently from the transactional databases and other repositories. It is mainly used in inventor, retail, agriculture sector, marketing, bioinformatics etc. The major task of finding associations is to search for relationships that are interesting among itemsets from a given database. Many algorithms are proposed to find the frequents patterns[10]. One of them is apriori algorithms and the other is FP growth approach. In apriori method length(p+1) patterns are generated based on the frequent patterns of length p. To generate frequent itemsets which are long many scans of the database is required. Many studies from [4][6][13][21][25][33][29][32] have adopted this apriori method. In Fp Growth method the FP-tree is used to store the database for finding frequent patterns. In this method only two times only database is scanned [15][16]. So It is faster than apriori. There are many extensions and alternatives to FP-growth approach[12][14][17][18][19] [23][24][27] [28][30] [31].

3. Related Work

Agarwall, mielinkski swami algorithm[2][3] is the first one for association rule mining. In this algorithm the itemsets are generated and counted when the database is scanned. candidate itemsets can be generated by extending the itemsets that are larger with the other items. Drawback of this algorithm is multiple scan. In Set oriented mining like in agarwall & mielinkski candidate itemsets are generated by scanning the database and is countedat the end.SQL join operator is being used for itemset generation.

Apriori algorithm is used for finding the frequent itemsets[5]. It is called as the breadth first search algorithm which is used for finding out the frequent itemsets.

Another algorithm called AprioriTid[6] uses the function of apriori candidate generation. The important characteristic of this algorithm is for support counting transactional database is not used. Instead of that it uses CK. When XK is the largest itemset, the entries are of form CK for the transaction.

An algorithm called Improved apriori algorithm is based on apriori algorithm. It used new count for pruning the candidate itemsets and also use record generation to reduce the scans of the database[21]. For the reduction of the scan time and candidate itemsets transaction reduction is used which is improved apriori algorithm[26].

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

As the apriori algorithm is not efficient because more time is taken to scan the database. To avoid this drawback Transaction reduction-Bit array matrix algorithm is used. In this algorithm Scanning is done for all database only once and the data is made available in Bit array matrix[29][34].

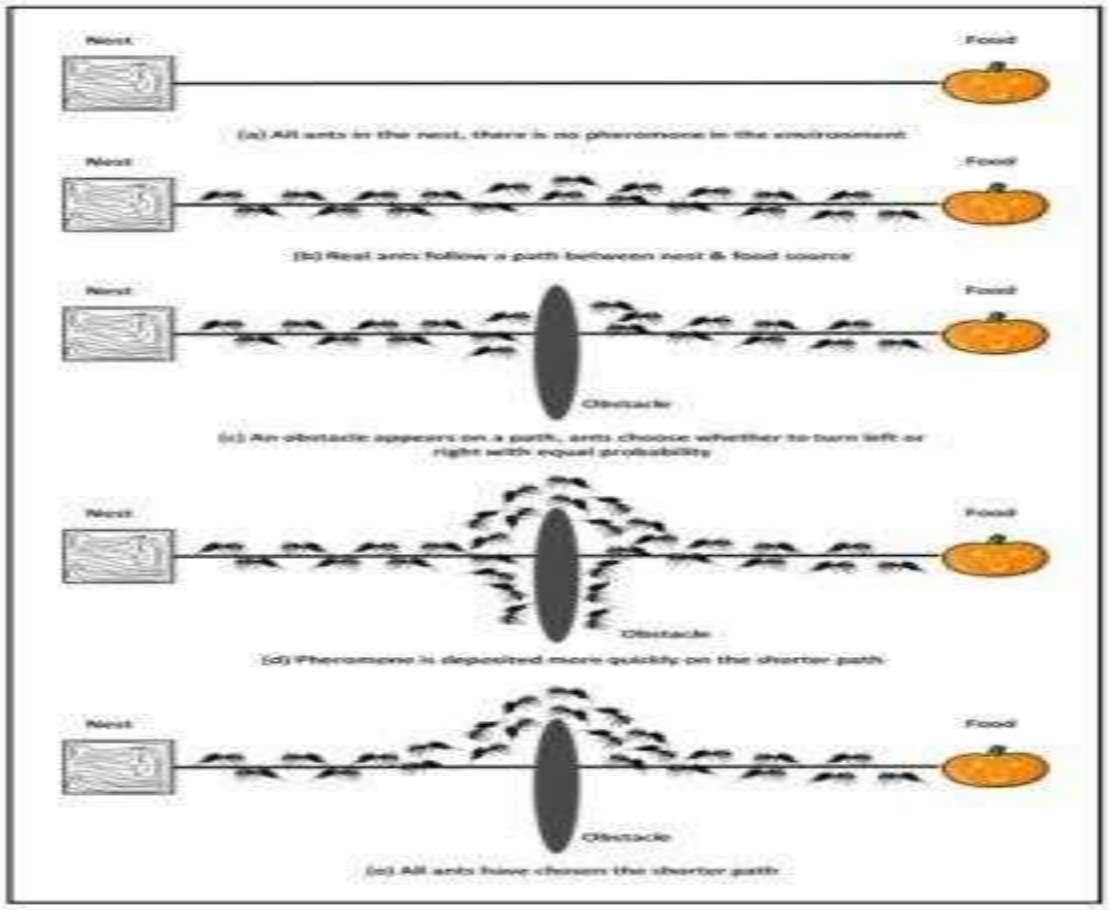
Binary count Table F1 and count table F1 algorithms are used to avoid expensive generation of candidate itemsets.[33].

BeApriori which is improved algorithm is mainly based on optimization of pruning . In this, thefrequent itemsets will be reduced and therefore running time is also reduced.[32]

The frequent item sets are generated from huge sets of data sets by applying various algorithms like Apriori, border algorithm, Partition, Incremental etc that take much of the computer time for computing different frequent item set. This scenario can be improved by using Genetic Algorithm. The advantage of using genetic algorithms is performing global search for finding frequent item sets. The time complexity is less when compared to other algorithms[7].

One of the promising technique to find the hidden pattern in transactional databases is association rule mining. Ant colony optimization algorithm is used for optimizing the asocial rules. It generates the quality rules that can be mined from the apriori algorithm.[8]. ACO follows the behavior of the real ants. Ants put the pheromone on the ground for marking favorable way which should be followed by remaining member of the colony. Ants find their shortest path between their nest and the food sources as shown in fig2.

Fig 2. Concept of Ant Colony



ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

To overcome usually the problems encountered in various apriori like algorithms with respect to candidate set generation a novel algorithm for finding frequent patterns of length n with out the generation of candidate sets called FPgrowth id developed by Jiawei han[10]. This algorithm compresses the data sets by using a small data structure refered as FP tree. FP tree is explore in a bottom up manner for finding the frequent patterns.

Diffset[11] which is novel algorithm based on vertical format has been proposed. In this algorithm maintains the differences in the transaction ids of the candidate items. It will not store Transaction ids. That is why this algorithm reduces the memory size and hence performance increases.

Another algorithm called PPV i.e prepost vertical which is proposed by Zhihong for finding frequent itemsets very quickly. It uses a data structure called node list. A Tree is constructed for storing the database and traverses in the postorder and preorder[20].

4. Comparative study of various techniques of frequent item set mining

S.no	Authors	Method	Advantages	Drawbacks
1	Agarwal at el	AIS	When we scan the database, the itemsets are created.	It takes numerous scans on the database, for creating and counting so many applicants pointlessly
2	Agarwal at el	SETM	The candidate itemsets are generated by using SQL join operator and the support count is determined at the end	It takes more no scans on the database
3	R.Srikanth at el	Apriori	We can reduce the amount of applicants created by using the applicant's creation method to analysis to get advanced.	It needs a lot of time for scanning if the size of the database is big.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

4	R.Srikanth at el	AprioriTid	'K' as the primary database cannot help in counting	More memory
5	Huan Wu at el	IAA	Less amount of time required for scanning. Overall performance is good	It requires additional amount of time if the database is huge
6	Jaishree singh at el	Transaction Reduction	The amount of time for scanning and I/O is greatly reduced in the database by using Size of Transaction(SOT).	The amount of memory is greatly increased each time as the database is constantly updated.
7	Ashok at el	Improved genetic algorithm	Time complexity is less	Accuracy rate has to be increased and longer running time.
8	Ritu Waliaura	Ant colony optimization algorithm	Relatively efficient	Convergence is not guaranteed.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

9	V.Vijayalakshmi at eal	TR – BAM	The repeated data is shown using the Repetition Count(RC) column and the number of non-zero values are presented with sum.	It does not requires any storage space.
10	Zhuang Chen at el	BE-Apriori	As the repetitive itemsets are reduced the run time is greatly reduced. Transactional database is greatly reduced by compressing the no of Transactional to be scanned	Operating cost of maintaining the temporary table is high.
11	Jiawei at el	FP-Growth	As the repetitive itemsets are reduced the run time is greatly reduced. Transactional database is greatly reduced by compressing the no of Transactional to be scanned	It is not easy to construct the FP-tree in a database as it requires large database.
12	Mohammed at el	Diffset	By only managing to know the types of the Tid's, we can reduce the memory required for the database	It requires more execution time as number of comparisons are vast.
13	Mingjun at el	Transaction Mapping	Transaction tree is constructed for all Transaction Id's and those are compressed into transaction intervals. This compression reduces the intersection time.	It requires the construction of transaction tree

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

14	Zhihong Deng, at el	PPV	The Node list is more active because transactions with general prefix share the identical nodes.	It requires large amount of storage space because of the PCC tree manages the item, frequency of the item, preorder code and post order code.

5. Conclusion

This research paper briefly gives the complete view of the algorithms for the frequent pattern mining. This analysis presents that various methods will have both advantages and disadvantages. occurred. Mining of Frequent item-sets can be done using Apriori, FPtree, ant colony and so many algorithms. In this research paper various algorithms have been explored comprehensively for finding frequent patterns and with the reason to find how the algorithms explored can be used to get frequent patterns from various datasets. In this digital world frequent item set mining is an important task for Business executives. As the algorithms reviewed has got several drawbacks. A new hybrid model is proposed for extracting frequent item set which will be processed in less execution time, high accuracy rates, low error rate.

6. References

[1] Rupesh Panwar, Abhishek Raghuvanshi " A Literature survey of modern techniques used for frequent item set mining", ISSN: 2455-2631 © April 2016 IJSDR | Volume 1, Issue 4

[2]Han.J, Kamber.M, "Data Mining:Concepts and Techniques", Morgankaufmann Publishers, Book, 2000.

[3] R. Agrawal, T. Imielinski, A. Swami,"Mining associations between sets of items in large databases, Proceedings of the ACM SIGMOD 1993 Conference Washington DC, USA, May 1993.

[4] R. Srikant and R. Agarwal, " Mining quantitative association rules in large relation tables" proceedings of the 1996 SIGMOD, pp. 1-12, 1996.

[5] T. Karthikeyan and N. RaviKumar, "A Survey on Association rule mining",International Journal of Advanced Research in Computer and Communication Engg(IJARCCE),PP.5223-227,2014.

[6] R. Agarwal and R. Srikant, " Fast algorithm for mining association ules",Proceedings of the 20th international conference on very large databases ,Margunkaufmann , PP. 487-499.

[7] D. Ashok Kumar , T. A. usha An Analytical Study of Genetic Algorithm for Generating Frequent Itemset and Framing Association Rules At Various Support LevelsIOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 4, Ver. VI (Jul.-Aug. 2016), PP 11-17

[8]Ritu Walia, "ARM using Ant colny Optimization Algorithm – A review". International journal for Multidiciplilnary Engineering and Business Management, vol 3, Issue 1, Jan- March 2015.

[9] J. Han, M.Kamber, "Data Mining:Concepts and Techniques", The Morgan kaufmann Series in Data

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Management Systems, Champaign: CS 497JH, fall 2001.

[10] Jiawei Han, Jianpei, and Yiwenyini,"Mininig Frequent Patterns without Candidate Generation", Proceedings of the ACM SIGMOD International Conference on Management of Data Pages , PP. 1-12,2000.

[11] Mohammed J. Zaki and Karam Gouda,"Fast Vertical Mining using Diffsets"Proceedings of the ASM SIGKDD '03 Washiton, DC, USA, Aug-2003.

[12] Mingjun Song, and SanguthevarRajasekaran, "A transaction mapping algorithm for requent itemset mining ", in IEEE transactions on knowledge and Data Engg.

[13] M. J. Zaki, S. Parthasarathy, M. Ogihara,and whi, "New Algorithms for Fast Discovery of Association Rules"Proceedings of KDD – 97, pp. 983-286.

[14] Jianyong Wang, Jiawei Han, " TFP: An Efficient Alogirithm for Mining Top-K Frequent Closed Itemsets" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17,NO. 5, MAY 2005.

[15] M.J. Zaki and C.J Hsiao, "CHARM: An Efficient Algorithm for Closed Itemset Mining." Proc. 2002 SIAM Int' I Conf. Data Mining[SDM '02), pp. 457-473. Apr.2002.

[16] I. Wang, J. Han, and J. Pei, "CLOSET+ : Searching for the Best Startagesies for Mining Frequently Closed Itemsets, " Prc.2003 ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining(KDD '03), pp. 236-245, Aug.2003.

[17] Yew KwongWoon, Wee Keong Ng, Ee-Peng Lim, "A Support-Ordered Trie for Fast Frequent Itemset Discovery", IEEE transactions on Knowledge and Data Engineering , PP. 875-879, Aug. 2004.

[18]Divvela. Srinivasa Rao, Dr.V.Sucharita, published a paper "Reduction of Frequent Itemset Mining in Big Data with the Help of FP Algorithm and Msegt-Tree" in International Journal of Innovative Technology and Exploring Engineering ISSN: 2278-3075, Volume-9 Issue-4, February 2020,2169-2172

[19] Y.K. Woon, W.K,Ng, and A. Das, "Fast Online Dynamic Association Rule Mining", Proc. Secont Int'l Copnf. Web Informatoion Systems Eng., pp. 278-287,2001.

[20] Jie Dong, Min Han "BitTableFI: An efficient mining frequent itemsets algorithm", Knowlwdge-Based Systems,vol.20, pp.329-335, 2007.

[21] D.Burdick, M. Calimlim, J. Flannick, J.Gehrke, T.M. Yiu. "MAFIA: a maximalfrequent itemset algorithm", IEEE Transactions on Knowledge and Data Engineering 17(11) (2005) 1490-1504.

[22] Huan Wu, Zhigang Lu, Lin Pan,RongshengXu, "An Improved Aprioribased Algorithm for Association Rules Mining", Sixth International Conference on Fuzzy Systems and Knowledge Discovery, pp. 51-55, 2009.

[23] Lei Ji, Baowen Zhang, and Jianhua Li, "A New Improvement on Apirori Algorithm", Computational Intelligence and Security,2006 International Conference on Volume 1, Nov 2006, pp 840-844.

[24] Zhihong Deng, Zhonghui Wang, " A New Fast Vertical Methof for Mining Frequent Patterns" International Journal of Computational Intelligence Systems, Vol 3, No. 6,PP. 733-744, Dec 2010.

[25] JieminZheng, Defu Zhang, Stephen C. H.Leung, Xiyue Zhou, "An efficient algorithm for frequent itemsets in data mining", International Conference on Advances in Signal Processing and Communication (ICSSSM), PP. 1-6, June 2010.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

[26] Jaishree Singh, Hari Ram, Dr. J.S.Sodhi, "Improving Efficiency of Apriori Algorithm Using Transaction Reduction",International Journal of Scientific and Research Publications, Vol.3, 2013.

[27] Wei Song, Bingru Yang, ZhangyanXu, "Index-BiTTableFI: An improved algorithm for mining frequent itemsets"Knowledge-Based Systems, Vol. 21,PP. 507-513, 2008.

[28] Deng. Z., Wang. Z., & Jiang, J.(2012)."A new algorithm for fast mining frequent itemsets using N-lists" , SCIENCE CHINA Information Sciences , Vol. 55,PP. 2008-2030, 2012.

[29] Zubair Khan, NeetuFaujdar,Prashantkumar sing, Tarifabbas , "Modified Bit Aopriori Algorithm: An Intelligent Approach for Mining Frequent Item-Set" Proc of Int. Conf. on Advances in Signal Processing and Communication,PP.813-819, 2013.

[30] V. Vijayalakshmi, Dr. A Pethalakshmi,"Mining of Frequent Itemsets with an Enhanced Apriori Algorithm"International Journal of Computer Applications(0975-8887) Volume 81 –No. 4. November 2013.

[31] Divvela. Srinivasa Rao, Dr.V.Sucharita, published a paper "Implementing Frequent Item set Mining by Overcoming Over-Scan Problems", in International Journal of Engineering and Advanced Technology, ISSN: 2249 – 8958, Volume-8 Issue-4, April 2019,816-819

[32] Bay Vo, Tuong Le, FransCoenen, Tzung-Pei Hong , " A Hybrid approach for Mining Frequent Itemsets" Systems, Man,and Cybernatics, IEEE, PP.4647-4651,2013.

[33] Zhi-Hong Deng, Sheng-Long Lv, " Fast mining frequent itemsets using Nodesets"Expert Systems with Applications 41(2014) 4505- 4512.

[34] Zhuang Chen, Shiban Cai, Qiulin Song,and Chonglai Zhu, " An Improved Apriori Algorithm Based on Pruning Optimization and Transaction Reduction", Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), PP.1908-19011, Aug-2011.

[35] Marghny .H, Mohamed .M, and Darwieesh, " Efficient Mining Frequent Itemset Algorithms ", International Journalof Machine Learninbg and Cybernatics,Vol. 5, PP. 823-833, 2013.

[36] Divvela. Srinivasa Rao, Dr.V.Sucharita, published a paper, "Efficient Algorithm using Big Data for Frequent Itemsets Mining", in International Journal of Innovative Technology and Exploring Engineering, ISSN: 2278-3075, Volume-8 Issue-4, February 2019,394-396 .

[37] V. Vijayalakshmi, Dr. A Pethalakshmi,"An Efficient Count Based Transaction Reduction Approach For Mining Frequent Patterns", Procedia Computer Science,Vol.47, PP. 52-61, 2015.

[38] Bay Vo, Tuong Le, Frans Coenen, and Tzung-pei Hong, "Mining frequent itemsets using the N-list and subsume concepts", International Journal of Machine Learning & Cybernatics, PP. 1-13, Apr 2014.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731