International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 4, Issue 09, September-2018

REVIEW ON FREQUENT ITEMSET MINING ALGORITHMS IN DATA MINING

[1]Vikram Rajpoot, [2]Shanu kumar, [3]Sadhna K. Mishra

[1]*Assistant professor, Department of CSE, LNCT College Bhopal*
[2]*Research Scholar, Department of CSE, LNCT College Bhopal*
[3]*Professor& Head, Department of CSE, LNCT College Bhopal*

Abstract—Data mining is the procedure of removingvaluableinfo from this swamped data, which supports in creatinggainfulupcoming decisions in these fields. Frequent item-set mining is an essential step in finding association rules.Association rule mining (ARM) is the important part of data mining, which helps to predict the association among multiple data items.In this paper, studied about different-different efficient algorithm that was designed like Improved Apriori, FP-Growth and combination of both (i.e. Hybrid algo.).Also, a brief study about frequent item mining.

Keywords—Data Mining, Hybrid Algorithm, Frequent Iemset, FP-Growth, Improved Apriori.

I. INTRODUCTION

Nowadays, vast progressive knowledge of information and technology provides users with multiple internet-based services like online shopping, searching & surfing on the web. Service providers improve their way of service and modernize business activities by keeping track of user records and behavior by gathering valuable information. Data Mining is a method of Knowledge Discovery in Databases also termed as KDD. KDD uses multiples computing theories and tools to help people finding valuable knowledge from data. In the data mining arena, association rule extraction is the most widely used exploration technology, and is mainly used to find hidden relationships between data in order to generate classification clusters, wherein data items are combined based on their various granularity levelss.

Data mining skill has been playing anprogressivelyvital role in decision-making events. Frequent itemset mining (FIM), as ansignificant step of association rule analysis, is becoming one of the most significantinvestigation fields in data mining.

Frequent itemset mining is an important step in finding association rules. There are many algorithms for mining frequent itemsets, some are the state of the art algorithms which started a new era in data mining and make the concept of frequent itemset and association rule probable [2].

Association rule mining (ARM) is the important part of data mining, which helps to predict the association among multiple data items. The big challenge of ARM is efficiently extract the knowledge from large size databases of various applications. As per concern of data holder, the main challenge of ARM is to share the accurate information with protection of sensitive information. To achieve this, Privacy preserving ARM plays very important role [3].

The goal for finding association rules came from examine of super market dataset, to find out customer behaviour based on purchased products. Finding of association rules is a criticalproblematic in data mining. Two sub-problems of mining association rules. First find out frequent itemsets from **dataset** and then grow association rules centered on frequent item sets.

The mining of association rules is an important mission in the field of date mining, which aimed at mining meaningful association in the affairs database. The association rules mining from the database becomes more and more necessary with the constantly collecting and storing date.

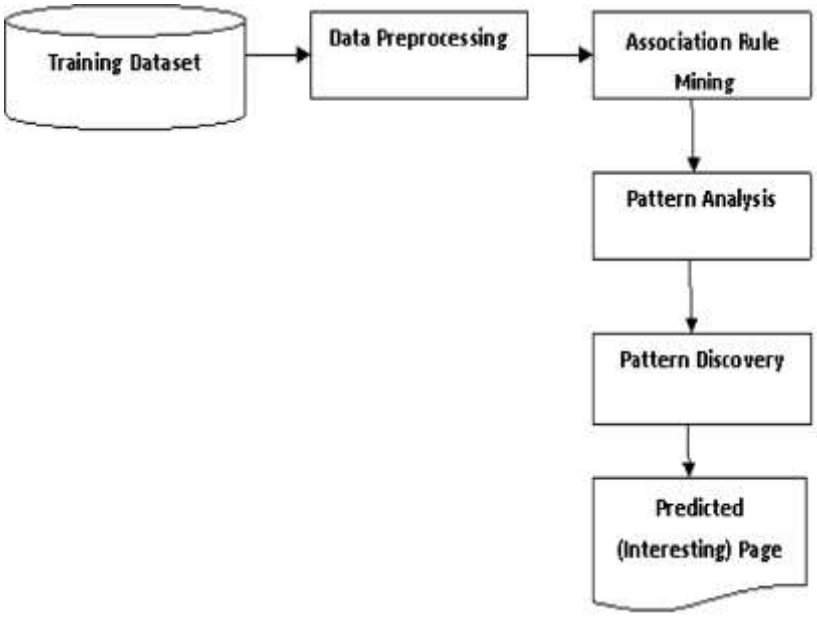

Fig. 1:Steps in Data Mining

II. FREQUENT ITEMSET MINING

Frequent itemset mining is one of the significant domains in pattern mining. This deals with mining the frequent itemsets that occur in the dataset. Frequent itemsets are mined for framing association rules. Other than framing association rules, mining frequent itemsets leads to effective classification, clustering and predictive analysis. The commonly used algorithms are Apriori, FPGrowth and Eclat. Researches are still an ongoing process in this area. So far variousalgo have been planned for mining frequent itemsets.

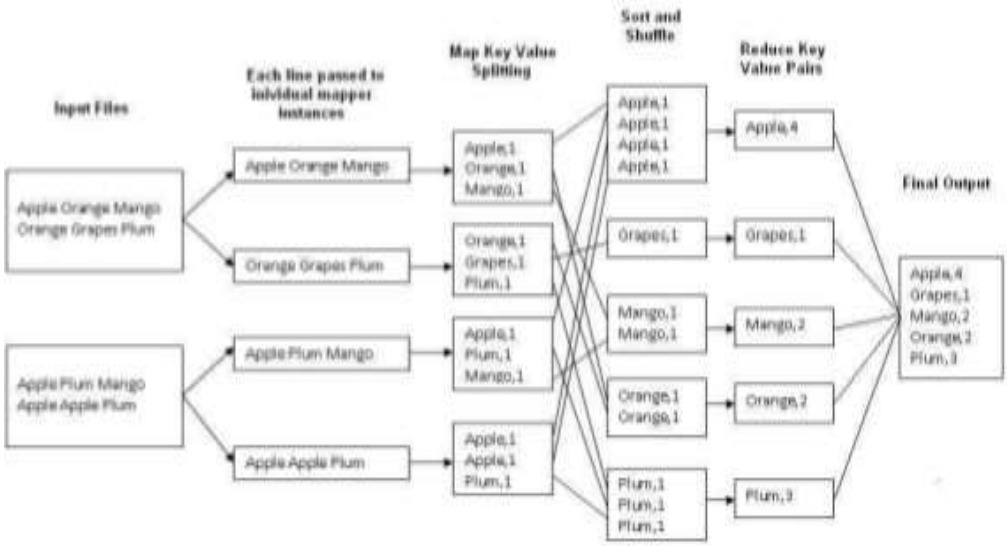

Fig. 2:FrequentItemset Mining

As the basicmethod of applying association rules mining, frequent itemsets mining is used to excerptcommon itemsets from items in a large databank of transactions. Frequent itemset mining (FIM) finds potentially interesting patterns which called frequent itemset in large transactions dataset. The measure of frequent itemset is minimal support which represents the frequency threshold of this itemsetoccurrence.

Table I. Comparison Between Different Frequent Itemset Mining Algorithms

Algorithm	Methodology	Strength	Limitations
Apriori	Join and prune	State of the art algorithm	More memory and time consumption
DHP	Hashing technique	Small execution time	Consume more space
Partitioning	Partitioning technique	Utilize less memory due to partitioning	Require more time to find local than global frequent itemset
Sampling	Picking random sample for checking the frequency of the whole database at lower threshold support	Less memory utilization and execution time	Sample selection is difficult
DIC	Dynamic insertion of candidate items	Small execution time	Require different amount of memory at different point
Improved Apriori	Forward and reverse scan	Less memory utilization and small execution time	Useful if the maximum frequent itemsets cannot be found fast
Eclat	Intersection of ids list is used for generating candidate itemsets	Less memory allocation if itemsets are small in number with small execution time	Performance is not feasible
FP-growth	Conditional frequent pattern tree	Consume less memory	Execution time is high

III. LITERATURE SURVEY

Xuejian Zhao et al [2018] In this paper, the weight judgment downward closure property for the weighted frequent itemsets and the existence property of weighted frequent subsets are introduced and proved first. In light of these two properties, the Weight judgment descending conclusion property-based FIM (WD-FIM) algo is proposed to limit the looking space of the weighted frequent itemsets and improve the time efficiency. Moreover, the completeness and time efficiency of WD-FIM algorithm are analyzed theoretically. At long last, the executionof the proposed WD-FIM algo is confirmed on both manufactured and genuine-informational databank.

NF Zulkurnain[2017] In this paper, a proficient hybrid algo was composed utilizing a bringing together procedure of the calculations improved Apriori and FP-Growth. Results indicate that the proposed hybrid algorithm, albeit more complex, consumes fewer memory resources and faster execution time. [8].

MadhuraKaranjikar And S. V. Kedar [2017] This paper presents the privacy preserving ARM over partitioned databases named as vertical partitioning of databases. In this, Bi-Eclat algorithm is used to partition the database vertically and then identify the frequent item sets in all partition to mine the association rules. Further the research is enhanced by providing the security over mined association rules by using cryptographic techniques[3].

Vid Podpeˇcanet al. [2016] This paper introduces a novel proficient algo FUFM (Fast Utility-Frequent Mining) which discovers entire utility-frequent item-sets inside the given utility and support constraints threshold. It is quicker and straightforward than the first 2P-UF algo (2 Phase Utility-Frequent), as it depends on proficient techniques for frequent item-set mining. Trial assessment on artificial data-sets demonstrate that, interestingly with 2P-UF, our algo can likewise be connected to mine vast data-bases.

Bin Pei et al [2016] Because of instrument mistakes, loose of sensor checking frameworks & so on,certifiable information have a tendency to be numerical info with innate vulnerability. To manage these circumstances, we propose a FP development-based mining algo PNFP-development to proficiently discover association rules from probabilistic numerical info, where each numerical thing in the exchanges is related with an existential likelihood. In addition, to deal with big data situation, we also introduce a parallelized PNFPGrowth in the MapReduce framework, which scales well with the size of the dataset while minimizing data replication and communication cost.

SlimaneOulad-Naoui [2015] In this paper, present another demonstrating for the Frequent item-set(FI) mining issue. To be sure we encode the item-sets as words over an arranged letter set, and express this issue by a formal arrangement over the semiring $(\mathbb N,+,\times,0,1)$, whose support constitutes the item-sets & the coefficients their frequencies. This formalism offers numerous points of interest in both crucial and handy perspectives: The presentation of a reasonable and bound together hyphothetical system, which we can demonstrate the proportionality of FI-algo, the probability of their speculation to mine other more complex objects, and their incrementalization as well as parallelization; by and by, we clarify how this issue can be viewed as that of word acknowledgement by an automaton, permiting a usage in $O(|Q|)$ memory and $O(|\mathcal{F}_M||Q|])$ time, where Q is the set of states of the automaton used for representing the data, and $\mathcal{F}_M$ the set of maximal FI [11].

Qin LX.et al. [2005] In this paper, we exhibit a algo, CFPmine, that is propelled by a few past works. CFPminealgo joins a few prefernces of existing methods. One is utilizing compelled sub-trees of a conservative FP-tree to mine frequent pattern, so it is doesn't have to build contingent FP-trees in the mining procedure. Second is utilizing a array-based method to decrease the traverse time to the CFP-tree.Whats more a memory administration is additionally executed in the algo. The trial assessment demonstrate that CFPminealgo is a superior performance algo. It outperforms Apriori, Eclat and FP-development and requires less memory than FP-development.

IV. FREQUENT ITEMSET MINING ALGORITHMS

A. Hybrid Algorithm[8]

The hybrid algorithm using a unifying process to combine Improved Apriori and FP-growth.The HYBRID algorithm included the property of the Apriori that non-void subsets of the frequent itemsets are also frequentthe HYBRID algorithm included the property of the Apriori that non-void subsets of the regular item-sets are additionally frequent.

In the first part of the algorithm, the Improved Apriori property was used to discover all the maximal frequent itemsets which are repeating in the transactional database with a support esteem equivalent to or more noteworthy than the least support specified. There are yet numerous item-sets which are frequent-1 but yet excluded in the maximal frequent item-sets. So the database which contains frequent-1 elements are pruned but there are no maximal frequent itemsets which make the database smaller and easy to traverse. The pruned database becomes the input in the second part of the algorithm which discovers all the frequent-1 itemsets and removes all the infrequent-1 itemsets from the transaction. Then, the FP-Tree algorithm was implemented by constructing an FP-Tree from the pruned transactions. This part of the algorithm assists in discovering all the frequent itemsets remained from the first procedure.

B. Improved Apriori Algorithm[13]

In order to improve the efficiency of mining of frequent itemsets, contrapose the two key problems of reducing the times of scaning the value based database and decreasing the quantity of candidate item-sets, an inhancedalgo is introducedin light of the great Apriorialgo.

Apriori utilizes an iterative strategy called searching step by step, k-item set used to investigate (k +1) - itemsets. In order to improve the efficiency of genetaring frequent itemsets step by step, we can use Apriori's nature to compress the search space, namely: all non-empty subset of frequent item-sets are should likewise be frequent.

The inhancedalgo in the examining of the original exchange databank to build-upa 1-itemsets as the key component in the set, T_{set} implies the new database table structure of item-set that contains its elements' transaction ordinal, in the calculation of the candidates for the support of the collection, and through such fundamentalactivities to decrease discovered frequently sets the unpredictability of the figuring procedure, in order to inhance the execution of the algo, and furthermore can enormously decrease the space possessing rate. The fundamentalstrides of algo: Scanning the exchange databank one by one, resulting in 1-itemset candidate set C_1, when examining of every exchange, not just count each item, but also record the contained transaction identifier T_{ID}. Subsequent to checking thedatabase once, in the candidate set C_1, every item-set contains a rundown of corresponded exchange identifier. The structure of C_1 as follows: (itemset Items, supports several sup, transaction identifier set T_{set}). Evecuate the item-sets that don't meet the least support from C_1,and get L_1. L_{k-1} self-connected, generates C_k, in which the affair identifier set of C_k is equal to the intersection of its two L_{k-1}'saffair identifier sets. We can get the count of each item-set in C_k through computing the quantity of T_{ID} that in the undertaking identifier set relating to the item-sets in C_k. The algo is being inhanced mostlycontrapose the phase of discovering frequent set.

C. FP-Growth

FP- Growth permits frequent item-set discovery without candidate item-set generation. Two stages approach:
Stage 1: Build a smaller info structure called the FP-tree Built utilizing 2 passes over the databank.
Stage 2: Extracts frequent item-sets straightforwardly from the FP-tree Traversal through FP-Tree.

The FP-Growth algorithm (PFP) based on map-reduce solves the problem of communication overhead, but no improvement has achieved did the algorithm[15].

When the size handled date collection increases to a certain degree, the following problems exist in FP-Growth algorithm [16]:

(1): One by one and repeatedly scans results in the cost of time and space direct proportional to the size of database, which seriously affect the speed of diagnosis.
(2): When the data size reaches a certain degree, if there exists more or longer branches, it will construct a large number of conditions FP-tree, which is time-consuming &memory-wasting.
(3): The algorithm recursively generate conditional database and FP-tree, where FP-tree generates from the top to the bottom, and the pattern mining generates in a opposite direction. Recursively construct FP-tree mining, which results in a large number of frequent patterns group. Because of repeatedly scanning the same paths, the both iterative times and pointer increase, which would use a larger space. The longer average affairs path, the worse the adaptability of algorithm.

CONCLUSION

Data mining is an essential technique to discover meaningful information for many objectives. Frequent Item-set mining is mostly used in many fields like as retail, financial and media transmission industry. The significant worry of these enterprises is speedier preparing of a lot of info. Frequent item sets are those items which are as often as frequently occurred. So, we utilize distinctive sorts of algo for this reason. Frequent item-set mining can be performed Apriori, FP-tree and so forth algo. In this paper, we have explored comprehensively used algo for finding frequent patterns with the reason for finding how these algo can be utilized to get frequent patterns from big retail dataset.

REFERENCES

[1] Hong-Yi Chang et al, "A Novel Incremental Data Mining Algorithm based on FP-Growth for Big Data", International Conference on Networking and Network Applications, pp. 375-378, 2016.
[2] Song. M and Rajasekaran, A Transaction Mapping Algorithm for Frequent Itemsets Mining, IEEE Transactions on knowledge and Data Engineering, vol. 18, No. 4, 2006.
[3] Madhura Karanjikar And S. V. Kedar,"Secure Association Rule Mining Using Bi-Eclat Algorithm On Vertically Partitioned Databases", International Conference On Intelligent Sustainable Systems (Iciss), Pp. 176-181, 2017.

[4] Patel Harshit and Jayesh Chaudhary, "A Study of Frequent Pattern Mining Methods", Research Journal of Computer and Information Technology Sciences, Vol. 2, Issue 1, pp. 1-3, April 2014.

[5] Ramah Sivakumar; J. G. R. Sathiaseelan , "A performance based empirical study of the frequent itemset miningalgorithms", IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), pp. 1627-1631, 2017.

[6] Yuxin Wang et al, "D2P-Apriori: A deep parallel frequent itemset mining algorithm with dynamic queue", Tenth International Conference on Advanced Computational Intelligence (ICACI), pp. 649-654, 2018.

[7] Xuejian Zhao et al, "A Weighted Frequent Itemset Mining Algorithm for Intelligent Decision in Smart Systems", IEEE Access, Volume: 6, Pp. 29271 – 29282, 2018.

[8] NF Zulkurnain and Ahmad Shah "HYBRID: An Efficient Unifying Process to Mine Frequent Itemsets",IEEE 3rd International Conference on Engineering Technologies and Social Sciences (ICETSS), pp. 1-5, 2017.

[9] Vid Podpe˘can et al., "A Fast Algorithm for Mining Utility-Frequent Itemsets", DTAI, pp. 10-20, 2007.

[10] Bin Pei et al, "Parallelization of FP-growth Algorithm for Mining Probabilistic Numerical Data based on MapReduce", 9th International Symposium on Computational Intelligence and Design, vol. 2, pp. 223-226, 2016.

[11] Slimane Oulad-Naoui et al,"Mining Frequent Itemsets: a Formal Unification", arXIV, 2015.

[12] Qin LX., Luo P., Shi ZZ. (2005) "Efficiently Mining Frequent Itemsets with Compact FP-Tree",Intelligent Information Processing II. IIP 2004. IFIP International Federation for Information Processing, vol. 163, Springer.

[13] Gu J., Wang B., Zhang F., Wang W., Gao M., "An Improved Apriori Algorithm", Applied Informatics and Communication. ICAIC 2011. Communications in Computer and Information Science, vol 224. Springer, pp.127-133, 2011.

[14] Florian Verhein, "Frequent Pattern Growth (FP-Growth) Algorithm", School of Information Technologies, The University of Sydney, Australia, 2008.

[15] Shandong Ji, Dengyin Zhang and Liu Zhang, "Paths sharing based FP-Growth data mining algorithms", 8th International Conference on Wireless Communications & Signal Processing (WCSP), pp. 1 – 4, 2016.

[16] Sidhu S, Kumar Meena U, Nawani A, et al. FP Growth Algorithm Implementation[J]. International Journal of Computer Applications, 2014, 93(8):6-10.