International Research Journal in Global Engineering and Sciences. (IRJGES)
ISSN : 2456-172X | Vol. 1, No. 3, September, 2016 | Pages 17-25

Big Data Association Rule Mining Approaches

Mrs. Prathima V. R[1], Dr. Fayaz. K[2]

Research Scholar, Department of Computer Science and Technology , Rayalaseema University, Kurnool, A.P, India[1]
Research Supervisor, S.K University, Ananthpur, Andhra Pradesh, India[2]

Abstract: Association rule mining is one of the interesting and challenging data mining techniques which is used to cluster the objects together from large databases. The ultimate aim is to extract the interesting correlation, patterns and relation among huge amount of data. Frequent pattern mining (FPM) is the focussed research topic in data mining with decent number of references in the literature. In this paper we discuss the brief review and analysis of the frequent pattern mining approaches and some of the promising research directions where high dimensional data sets are involved.

Keywords – Association rule mining, Data mining, frequent pattern mining, high dimensional data

I. INTRODUCTION

Data mining has become indispensible for enterprises to make expert decisions and this is the reality of current era. No organization can afford to make manual analysis of business data to take decisions. Organizational growth is based on the intelligent data analysis and usage of information obtained from data mining process. Recently the concept of big data [1] came into existence which reflected data in large quantity. Processing such huge amount of data which is characterized by volume, velocity and variety to derive some of the interesting patterns it is required to know the support of various techniques. Association and Sequential mining are the main tasks involved in the descriptive mining techniques. Association rule mining [2] is one of the important concept treated in KDD and can be defined as extracting the interesting correlation and relation among huge amount of transactions.

II. RELATED WORK

This section describes the brief survey, mainly focussed on our research methods for mining the frequent itemsets and association rules with utility considerations. When ever there is larger data set with high dimensionality, the novel association rule mining algorithms also generates large amount of rules.The execution time involved with such algorithms will be very much high with huge consumption of memory. There are chances of skipping highly valuable information. To address this issue there is a proposal of top-k rules which are having highest support, where the parameter k is set by the user.we also tried to expand the rule by incorporating optimization techniques. Some of the popular association rule mining algorithms are explanied below.

Apriori: Is a fundamental algorithm proposed by [3]. According to this algorithm it scans and searches for the occurrence of frequent item sets in breadth at each and every level of the lattice. It also uses the pruning technique.

AprioriTID: According to AprioriTID the algorithm adopts the candidate item sets used in the previous passes rather than counting the support of the candidate. This reduces the scan time involved with the data base. The details are discussed in [4].

DHP: Direct Hashing and Pruning[DHP] makes use of hashing technique to efficiently generate large itemsets. It attempts to reduce the transaction database size.

CARMA: This is yet another association rule mining algorithm which allows the user to change the support threshold any time. CARMA stands for Continuos Association Rule Mining proposed by Hidber[5].CARMA user both the techniques of Apriori and DIC on low support thresholds.

FP-Growth: FP-growth method [6] is to found that few lately frequent pattern mining methods being effectual and scalable for mining long and short frequent patterns. The **FP-Growth** Algorithm, proposed by Han in, is an efficient and scalable **method** for mining the complete set of frequent patterns by pattern fragment **growth**, using an extended prefix-tree structure for storing compressed and crucial information about frequent patterns named frequent-pattern tree (**FP**-tree).

ECLAT: This algorithm is used for identifying the frequent itemsets occurrence from the active transactional database.ECLAT was proposed by Zaki[7]. According to this technique during the first scan the TID's are built from the frequent (K+1) itemsets which is grown from a previous k-itemset as per the involvement of Apriori property with the depth first computation order as similar to FP Growth.Later on the intersection property is applied to frequent k-itemsets to that of the (k+1) itemsets. The process is repeated for the remaining candidate itemsets.

SPADE: This algorithm is used for mining sequential patterns from a sequence database and is once again proposed by Zaki in 2001. SPADE basically uses the combinatorial properties to divide the original problem into smaller subproblems which can be solved independently in the main memory using the effective lattice searching methods.Later on uses the simple join operation to obtain the final result. The overall time required for data scan is reduced.

SPAM: SPAM algorithm also aims for mining sequential patterns and is proposed in 2002 by Ayres et al[8]. It adopts the general depth first search strategy with an involvement of depth first traversal of the search space tree with efficient pruning mechanisms.

Diffset : Proposed by Mohammed J. Zaki et al. [9] in 2003 as a new vertical data depiction. This work proves that diffsets largely exceeds (by orders of magnitude) the extent of memory needed to keep intermediate results.

DSM-FI: Data Stream Mining for Frequent Itemsets is termed as a single-pass algorithm implemented in 2004 by Hua-Fu Li, et al. [9]. It is a novel method which aims at extracting all frequent itemsets over the history of data streams.

PRICES: It is a skilled algorithm developed by Chuan Wang [10] in 2004, It first recognizes all large itemsets which are used to construct association rules.This algorithm helps in decreasing the time of large itemset generation by scanning the database once and then applying the logical operations for furthur process. It is faster than Apriori.

PrefixSpan: PrefixSpan is proposed by Pei et al. [11] in 2004.Here the project recursively splits a sequence database into a set of smaller projected DB's, and sequential patterns are grown in each projected DBs. Prefix span is more optimized than the exisitng apriori based algorithms like GSP, FreeSpan, and SPADE.

Sporadic Rules: Is an algorithm for mining perfectly sporadic association rules proposed by Koh & Rountreel.[12]. e sporadic rules as those rules which have low support but high confidence. Here Apriori inverse is used as a method to obtain the sporasic rules.

IGB: This is called as Informative and generic basis of association rules from a transaction database. This algorithm is proposed by Gasmi et al. [13] in 2005.

GenMax: GenMax proposed by Gouda and Zaki [14] in 2005 is a backtrack search based algorithm for mining max frequent itemsets. It uses technique of optimization to prune the search space with a forward focus to perform maximality checking, and uses diffset algorithm to perform fast FC.

FPMax: FPMax (Frequent Maximal Item Set) is an algorithm proposed by Grahne and Zhu, (2005) [15] based on FP Tree. It uses array based structure than the tree structure . The input is the set of transactional data items from relational data model, with the two interesting measurements Min Support, Min Confidence and then generates Frequent Item Sets with the help of FPTree.

FHARM: It is yet another optimized algorithm developed by M. Sulaiman Khan et al. [16] in 2006. Accordng to this approach, edible attributes are filtered from transactional input data by the kind of rejections and are then converted to Required Daily Allowance (RDA) numeric values. Then the averaged RDA database is then converted to a fuzzy database which contains normalized fuzzy attributes comprising different fuzzy sets.

H-Mine: H-Mine is an algorithm which aims at discovering frequent itemsets from a transaction database developed by Pei et al. [17] in 2007. H stands for hyper-links, H-struct, and a new mining algorithm, Hmine, dynamically adjusts links in the mining process.It has limited and predictable memory cost and runs fastly in memory-based settings. H-Mine can be scaled up to very large databases using DB partitioning.

FHSAR: FHSAR is an algorithm for hiding delicate association rules proposed by Weng et al. [18]. The algorithm can completely hide any given SAR by scanning database at once from left to right, it significantly reduces the execution time. Literature reveals FHSAR is more effective from previous discussed algorithms in terms of execution time

Reverse Apriori: Is a novel algorithm described in 2008 by Kamrul et al. [19] .The described approach generates large frequent itemsets upon the satisfying factor of user specified minimum item support. It then gradually decreases the number of items in the itemsets until it gets the largest frequent itemsets.

DTFIM : It is referred as Distributed Trie-based Frequent Itemset Mining is an approach presented in 2008 by Ansari et al. [20]. It can be applicable for multicomputer environment and it is revisited with modified FDM algorithm ideas for candidate generation step. The proposed algorithm proved that Trie data structure can be used for larger databases with distributed association rule mining concept.

GIT-tree: GIT-tree is a tree structure developed in 2009 by [21] to mine frequent itemsets in an organized hierarchical database with the aim to reduce the time involved in mining. Here the algorithm scans database one time only and uses Tidset to compute the support of generalized itemset with a faster pace.

Scaling Apriori: This is developed in 2010 by Prakash & Parvathi [22] .The proposed improved Apriori algorithm aims to minimize the number efficient analysis of association rule and pattern mining approaches of candidate sets. Generation of association rules is adopted by evaluating quantitative information associated with each item that occurs in the given transaction.Here the reduction of number of itemsets generated is aimed with the minimization of the overall execution time of the algorithm.

CMRules: This is an algorithm for mining sequential rules from a sequence database proposed by Fournier-Viger et al. [23] in 2010. The said algorithm proceeds by first finding association rules to prune the search space for items that may occur togetherly in many sequences. It then eliminates association rules which doesnot satisfy the minimum confidence and support thresholds as per the time ordering principle. CMRules is faster and holds the good scalability for low support thresholds.

TopSeqRules: It is an algorithm for mining sequential rules from a sequence database proposed by Fournier-Viger et al. [24] in 2010. The discussed algorithm allows to extract the top-k sequential rules from sequence databases, where k is the number of sequential rules to be found and is set by the user. This new algorithm is proposed, because of the fact that the current algorithms are very slow and can generate extremely large amount of results or generate too few results, skipping valuable information.

Approach based on minimum effort: This work is described by Rajalakshmi et al. (2011) [25].It is also a novel method to generate the maximal frequent itemsets with minimum effort. According to [26], the method uses the concept of partitioning the data source into segments and then mining the segments for maximal frequent itemsets. It reduces the number of scans over the transactional data source to only two. However, the time spent for candidate generation is eliminated. This algorithmic steps are segmentation of the transactional data source,prioritization of the segments and mining of segments.

FPG ARM: Frequent Pattern Growth Association Rule Mining is an approach proposed In 2012 by Rao & Gupta [27] as a novel scheme for extracting association rules having the idea with the number of database scans, memory consumption, the time and the quality of the rule. The discussed algorithm uses a FIS data extracting association algorithm to remove the drawback of APRIORI algorithm which is efficient in terms of the number of database scan and time complexity.

TNR: It is defined as an approximate algorithm developed by Fournier-Viger & S.Tseng [28] in 2012 which mainly aims to mine the top-k non redundant association rules TNR (Top-k Nonredundant Rules). It is a recent approach for generating association rules which can be named as “rule expansions”.It basically adds constraint strategies to avoid generation of redundant rules. An evaluation of the algorithm with datasets used in the history proved that TNR is capable of providing excellent performance and is very well scalable.

ClaSP: Mining frequent closed sequence concept was proposed by Gomariz et al. [29] in 2013. Clasp uses multiple efficient search space related pruning methods in collaboration with a vertical database and its related layout.

MRApriori:MRApriori is a map-reduce apriori proposed by Snehal Ramteke et al. [30] in the year 2016. It makes use of hadoop HDFS for storing larger transactional databases.

III.PERFORMANCE ANALYSIS

This section focuses on the comparative study of the research methods for mining the frequent itemsets, mining association rules, sequential rules and mining sequential patterns. Most of the existing methods paid attention to performance and memory requirements. The methods and various approaches used for the large databases are discussed in the below table 1.

Approach	Year	Data Source	Sequential Pattern Mining	Sequential Rule Mining	Frequent Item Set Mining	Association Rule Mining
Apriori	1994	Transaction DB	-	-	Yes	-
Apriori TID	1994	Transactional DB	-	-	Yes	-
DHP	1995	Transactional DB	-	-	Yes	-
FDM	1996	Transactional DB	-	-	Yes	-
GSP	1996	Sequential DB	Yes	-	-	-
DIC	1997	Transactional DB	-	-	Yes	-
Pincer Search	1998	Transactional DB	-	-	Yes	-
CARMA	1999	Transactional DB	-	-	Yes (closed)	-
CHARM	1999	Transactional DB	-	-	Yes (maximal)	-
Depth-Project	2000	Transactional DB	-	-	Yes	-
ECLAT	2000	Transactional DB	-	-	Yes	-
SPAD	2001	Sequential DB	Yes	-	-	-
SPAM	2002	Sequential DB	Yes	-	-	-
Diffset	2003	Transactional DB	-	-	Yes	-
FP-Growth	2004	Transactional DB	-	-	Yes	FP-growth
DSM-F1	2004	Transactional DB	-	-	Yes	-
PRICES	2004	Transactional DB	-	-	Yes	-
PrefixSpan	2004	Sequential DB	Yes	-	-	-
Sporadic Rules	2005	Transactional DB	-	-	-	Yes
IGB	2005	Transactional DB	-	-	-	Yes
GenMax	2005	Transactional DB	-	-	Yes(maximal)	-
FPMax	2005	Transactional DB	-	-	Yes(maximal)	-
FHARM	2006	Transactional DB	-	-	Yes	-

Approach	**Year**	**Data Source**	**Sequential Pattern Mining**	**Sequential Rule Mining**	**Frequent Item Set Mining**	**Association Rule Mining**
H-Mine	2007	Transactional DB	-	-	Yes	-
FHSAR	2008	Transactional DB	-	-	-	Yes
Reverse Apriori	2008	Transactional DB	-	-	Yes(maximal)	-
DTFIM	2008	Transactional DB	-	-	Yes	-
GIT-Tree	2009	Transactional DB	-	-	Yes	-
Scaling Apriori	2010	Transactional DB	-	-	Yes	-
CMRules	2010	Sequential DB	-	Yes	-	-
MinimumEffort	2011	Transactional DB	-	-	Yes(maximal)	-
TopSeqRules	2011	Sequential DB	-	Yes	-	-
FPG ARM	2012	Transactional DB	-	-	Yes	-
TNR	2012	Transactional DB	-	-	-	Yes
ClasP	2013	Sequential DB	Yes (closed)	-	-	-
MRApriori	2016	Transactional DB (Support for larger databases with Hadoop HDFS)	-	-	Yes	-

Table 1: Survey of Association Rule Mining Methods

IV.CONCLUSION

In this paper we have presented a study of algorithms and related methods that existed for association analysis. How ever the major tasks of frequent pattern mining approaches are: itemset mining, sequential pattern mining, sequential rule mining and association rule mining.We have performed the detailed study on almost all the methods and the comparitive study of various approaches are highlighted in our paper. A deep peep into the literature has left several critical issues for future like the high dimensional larger databases , mining for multilevel association rules without the threshold and so on.

REFERENCES

[1] Cios K.J., Pedrycz W, Swiniarski RW, & Kurgan LA. Data mining: A knowledge discovery approach. New York, NY: Springer, 2012.

[2] Marek Wo, Krzysztof Ga, Krzysztof Ga. "Concurrent Processing of Frequent Itemset Queries Using FP-Growth Algorithm", Proc. of the 1st ADBIS Workshop on Data Mining and Knowledge Discovery (ADMKD'05), 2005,Tallinn, Estonia.

[3] Agrawal R. & Srikant R. Fast Algorithms for Mining Association Rules. In Proc. 20th Int. Conf. Very Large Data Bases (VLDB), 1994: 487-499.

[4] Park J.S., Chen M.S. & Yu P.S. An Effective Hashbased Algorithm for Mining Association Rules. In Proc. 1995 ACM SIGMOD International Conference on Management of Data, 1995, 175-186.

[5] Cheung C., Han J., Ng V.T., Fu A.W. & Fu Y. A Fast Distributed Algorithm for Mining Association Rules. In Proc. of 1996 Int'l Conf. on Parallel and Distributed Information Systems (PDIS'96), 1996, Miami Beach, Florida, USA.

[6] Srikant R, Agrawal R. Mining sequential patterns: generalizations and performance improvements. In: Proceeding of the 5th international conference on extending database technology (EDBT'96), 1996, Avignon, France: 3– 17.

[7] Brin S., Motwani R., Ullman J.D., and Tsur S. Dynamic itemset counting and implication rules for market basket data. In Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, 1997, 26(2): 255–264.

[8] Lin D. & Kedem Z. M. Pincer Search : A New Algorithm for Discovering the Maximum Frequent Set. In Proc. Int. Conf. on Extending Database Technology,1998.

[9] Hidber C. "Online association rule mining". In Proc. of the 1999 ACM SIGMOD International Conference on Management of Data, 1999, 28(2): 145–156.

[10] Zaki M. J. and Hsiao C.-J. "CHARM: An efficient algorithm for closed association rule mining". Computer Science Dept., Rensselaer Polytechnic Institute, Technical Report, 1999: 99-10.

[11] Agrawal R.C., Aggarwal C.C. & Prasad V.V.V. Depth First Generation of Long Patterns. In Proc. of the 6th Int. Conf. on Knowledge Discovery and Data Mining, 2000: 108- 118.

[12] Han J., Pei J. & Yin Y. Mining Frequent Patterns without Candidate Generation. In Proc. 2000 ACM SIGMOD Intl. Conference on Management of Data, 2000.

[13] Zaki MJ. Scalable algorithms for association mining. IEEE Trans Knowl Data Eng 12, 2000:372–390. [14] Mohammed J. Zaki. SPADE: An efficient algorithm for mining frequent sequences, Machine Learning, 2001: 31— 60.

[15] Jay Ayres, Johannes Gehrke, Tomi Yiu and Jason Flannick. Sequential Pattern Mining using A Bitmap Representation. ACM Press, 2002:429—435.

[16] Zaki M.J., Gouda K. "Fast Vertical Mining Using Diffsets", Proc. Ninth ACM SIGKDD Int‟l Conf. Knowledge Discovery and Data Mining, 2003: 326-335.

[17] Hua-Fu Li, Suh-Yin Lee, and Man-Kwan Shan. An Efficient Algorithm for Mining Frequent Itemests over the Entire History of Data Streams. The Proceedings of First International Workshop on Knowledge Discovery in Data Streams, 2004.

[18] Chuan Wang, Christos Tjortjis. PRICES: An efficient algorithm for mining association rules, in Lecture Notes Computer Science vol. 2004, 3177: 352-358, ISSN: 0302- 9743.

[19] Pei J, Han J, Mortazavi-AslB,Wang J, PintoH, ChenQ,DayalU, HsuM-C. Mining sequential patterns by pattern-growth: the prefixspan approach. IEEETransKnowl Data Eng 16, 2004:1424–1440.

[20] Yun SK and Nathan Ro. Finding sporadic rules using apriori-inverse. InProceedings of the 9th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining (PAKDD'05), Tu Bao Ho, David Cheung, and Huan Liu (Eds.). Springer-Verlag, Berlin, Heidelberg, 2005: 97- 106.

[21] GASMI Ghada, Ben Yahia S., Mephu Nguifo Engelbert, Slimani Y., IGB: une nouvelle base générique informative des règles d'association, dans Information-Interaction-Intelligence (Revue I3), 6(1), CEPADUES Edition, octobre 2006: 31-67.

[22] Gouda, K. and Zaki,M.J. GenMax : An Efficient Algorithm for Mining Maximal Frequent Itemsets', Data Mining and Knowledge Discovery, 2005, 11: 1-20.

[23] Grahne G. and Zhu G. Fast Algorithms for frequent itemset mining using FP-trees, in IEEE transactions on knowledge and Data engineering, 2005,17(10):1347-1362.

[24] Sulaiman Khan M., Maybin Muyeba, Christos Tjortjis, Frans Coenen. "An effective Fuzzy Healthy Association Rule Mining Algorithm (FHARM), In Lecture Notes Computer Science, 2006, 4224:1014-1022, ISSN: 0302-9743.

[25] Jian Pei, Jiawei Han, Hongjun Lu, Shojiro Nishio, Shiwei Tang and Dongqing Yang. HMine: Fast and spacepreserving frequent pattern mining in large databases, IIE Transactions, 2007, 39(6):593-605.

[26] Chih-Chia Weng, Shan-Tai Chen, Hung-Che Lo, A Novel Algorithm for Completely Hiding Sensitive Association Rules, Eighth International Conference on Intelligent Systems Design and Applications, 2008.

[27] Kamrul Shah, Mohammad Khandakar, Hasnain Abu. Reverse Apriori Algorithm for Frequent Pattern Mining, Asian Journal of Information Technology, 2008, :524-530, ISSN: 1682-3915.

[28] Ansari E., Dastghaibfard G.H., Keshtkaran M., Kaabi H. "Distributed Frequent Itemset Mining using Trie Data Structure", 2008, IAENG, vol.35:3.

[29] Bay Vo , Huy Nguyen , Tu Bao Ho , Bac Le, Parallel Method for Mining High Utility Itemsets from Vertically Partitioned Distributed Databases, Proceedings of the 13th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems: Part I, September 28-30, 2009, Santiago, Chile.

[30] Praksh S., Parvathi R.M.S. An enhanced Scalling Apriori for Association Rule Mining Efficiency, European Journal of Scientific Research, 2010, 39:257-264, ISSN: 1450-216X.

[31] Fournier-Viger, P., Faghihi, U., Nkambou, R. and Mephu Nguifo, E. CMRules: An Efficient Algorithm for Mining Sequential Rules Common to Several Sequences. In the Proceedings of 23th Intern. Florida Artificial Intelligence Research Society Conference, Daytona, USA, May 19--21, 2010, AAAI Press:410-415.

[32] Fournier-Viger, P. and Tseng, V. S. Mining Top-K Sequential Rules. In Proc. ADMA 2011 (Beijing, China, December 17--19, 2011). Springer, 2011, 180--194.

[33] Rajalakshmi, M., Purusothaman, T., Nedunchezhian, R. International Journal of Database Management Systems (IJDMS), 3(3), August 2011: 19-32.

[34] Lin D.-I and Kedem Z. Pincer Search: An efficient algorithm for discovering the maximum frequent set. IEEE Transactions on Database and Knowledge Engineering, 2002, 14 (3): 553 – 566.

[35] Rao, S., Gupta, P. Implementing Improved Algorithm Over Apriori Data Mining Association Rule Algorithm. IJCST.2012, 3 (1), 489-493.

[36] Philippe Fournier-Viger and Vincent S. Tseng. Mining top-K non-redundant association rules. In Proceedings of the 20th international onference on Foundations of Intelligent Systems (ISMIS'12), Li Chen, Alexander Felfernig, Jiming Liu, and Zbigniew W. Raś (Eds.). Springer-Verlag, Berlin, Heidelberg, 2012, 31-40.

[37] Antonio Gomariz, Manuel Campos, Roque Marín, Bart Goethals. ClaSP: An Efficient Algorithm for Mining Frequent Closed Sequences. PAKDD, 2013: 50-61.

[38] Holsheimer M, Kersten M, Mannila H, Toivonen H (1995) A perspective on databases and data mining. In Proceeding of the 1995 international conference on knowledge discovery and data mining (KDD’95), Montreal, Canada: 1995, 150–155.

[39] Washio T, Motoda H (2003) State of the art of graphbased data mining. SIGKDD Explor 5:59–68.