AN APPROACH FOR IDENTIFYING WIEGHTED HIGH FREQUENCY ITEM SETS IN UNCERTAIN DISTRIBUTED DATABASE.

1. Chinnapaga Ravi 2. M Bal Raju 3. N Subhash Chandra
1. Research scholar 2. Professor 3.professor
Computer science and engineering, JNTUH, Hyderabad, India

ABSTRACT:

Distribution and uncertainty are considered as the most important design issues in database applications nowadays. A lot of ranking or top-k query processing techniques are introduced to solve the problems of communication cost and centralized processing. On the other hand, many techniques are also developed for modelling and managing uncertain databases. Although these techniques were efficient, they didn't deal with distributed data uncertainty. This paper proposes a framework that deals with both data distribution and uncertainty based on weighted high frequency items set algorithms. Within the proposed framework LMTV &GMTV algorithms are investigated for retrieving the WHFI from uncertain. The main objective of these algorithms is to reduce the n no of items which are not satisfied MTV (minimum threshold value) while extracting frequent itemsets from heterogeneous local sites. Experimental results show that both proposed techniques have a great impact in reducing communication cost, pruning time & Both techniques are efficient but in different situations. The first one is efficient in the case of low number of sites while the other achieves better performance at higher number of sites.

Key words: DDB, local dB, central db. Frequent patterns, heterogeneous, uncertain data

Introduction: we learn about the problems in the horizontally partitioned database by the secure Mining of Association Rules .Holding information on different paths by sharing the same schema is nothing but database. Data Mining is the working process of understand the data from varies representatives and making it to be a useful information. In the database the data which is separated in the different sites at the same schema .The user's wants to use the data which was varied into different path's for to perform some function's .while performing(or) compositing the functions some issue's has occurred .Separated data is become an a new way of solving the problems and a new technology will be arrived and its mostly useful for the businesses and companies sets of their data stored placed as warehouses .One of the privacy preserving data mining in the Association Rule Mining .The database operation which are held is nothing are Horizontally partitioned database .Popular tool used in data mining "Storing the data in the main site" and running some algorithm against the stored data .Main method is to search all the Association Rules with the mining support 'S' and confidence 'C' to the related database .While performing this; we should be careful and should not be contain any results of private data information .For minimizing the problem of private information leakage the Multi-party computation is used as protocol. Existed trusted third party the overcome inputs should be transfer to him .After that function distribution can be progress and resulting output can be shifted to them.

User's wants to run on their own in order to consist the required outputs 'y' at Own-rule .To overcome this the protocol should be necessary to evolve . Before for this; So many programs are utilized for mining of Association Rules in Horizontally partitioned database uncertain database.

Local databases are inputs and list of Association rules in the demand output that for the given threshold 'S' & 'C' and hold confidence for these are not to be small.

PROPOSED APPROACH:

Technique of algorithm used to compute union of set of private subsets and extracts high frequency itemsets by using Weighted High- frequency Items [WHFI] mining algorithm .Actually, the complexity which through from the uncertain data. Previously, In homogenous the easy way of finding the frequent items in the distributed database are simple concept. But, coming into the process of finding from uncertain data in Heterogeneous distributed database a secured manner to further proceeding information of data owner's records.

Homogenous distribution is the holding the serial information at same (or) common location .In homogenous distribution the data which is represented.We can find the frequent patterns .In this we can merge the others data records. But it is a long process of treating the items (or) data to find the frequent pattern. Here the new technique used in this uncertain data from different path.

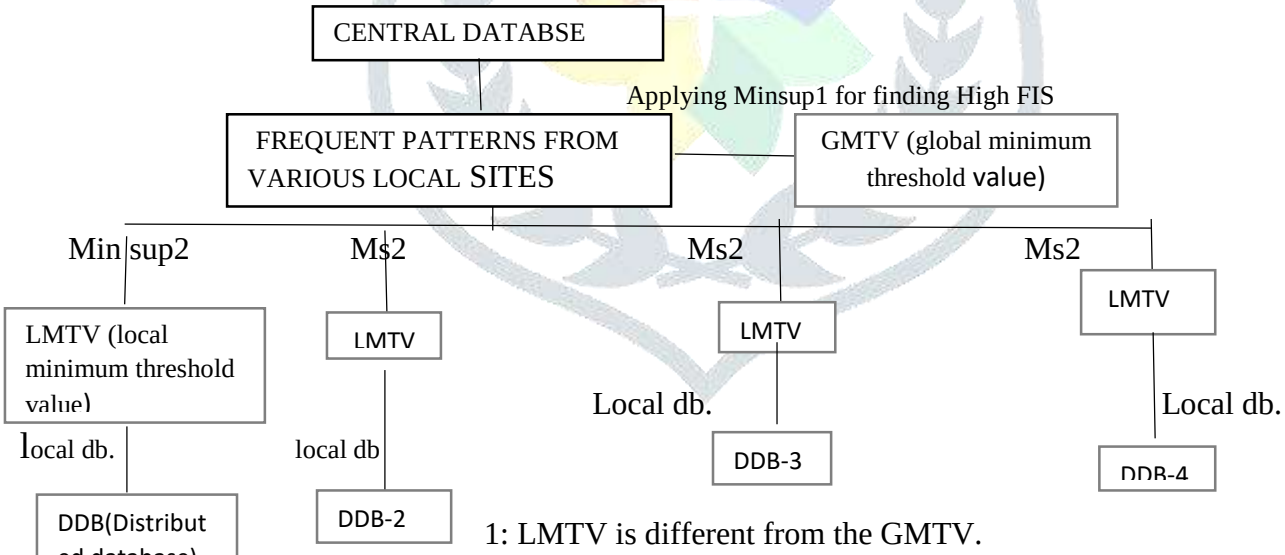
Technically, Heterogeneous data is the data which holding the different information in the local site of different schema. The concept arrived in this finding the frequent pattern from the heterogeneous data.

In distribution data bases .I would like to address the uncertain data.

Uncertainty consists of noisy, missed values, in-consistency, un-structured. If u want to find out frequent item set from uncertain data in the distributed databases.

Traditional algorithm (or) previous techniques are In-appropriate.

Let us take the probability item-set and weighted item set is to be considered for the finding High frequency itemset by using Weighted High-Frequency Items [WHFI] algorithm in uncertain data in distributed data base.



1: LMTV is different from the GMTV.

2 nd out frequent items in heterogeneous distributed database.

That distributed data bases may generate huge no. of item set after transforming frequent patterns into central distribution database.

3: Global MTV (minimum threshold value) has given by the user and GMTV is greater than or equal to LMTV. Thus it can be identify high frequent items.

Transaction-ID	TRANSACTIONS	TRANSACTION-WEIGHT
T1	Pen, pencil, eraser, scale, sharpener	0.42
T2	Pencil, eraser, sharpener	0.15
T3	Pencil, eraser, scale, sharpener	0.423
T4	Eraser	0.7

Transactions minimum weight (WT min) = Average transactions/2

Average transactions= T1+T2+T3+T4/4= 0.42+0.15+0.423+0.7/4=

=2.053/4= 0.42325

WT min = Avg. transactions/2= 0.42325/2= 0.211625

Weighted transactions uncertain database for example:

T-ID	TRANSACTIONS
T1	1. 0.321 2. 0.432 3. 0.313 4. 0.313
T2	2. 0.422 3. 0.303 4. 0.323
T3	3. 0.322 4. 0.432 5. 0.315

Weights Table:-

ITEMS	WEIGHTS
1	0.3
2	0.2
3	0.4
4	0.7
5	0.6

Those weighted transactions are not satisfied minimum weighted (WT min) .That weighted transactions are not considered further pruning, In the pre-processing that transactions are eliminated.

Pruning condition 1:
sup (x) > min sup && R W(x) > W min
Given; min sup 1 =0.5 No. of transactions are '3'
WT min = 0.211625
Min sup 2 =min sup 1/N = 0.5/3
=0.1666

Min sup 2 =0.1666 WT min = 0.2

Sup (1) = 0.321 > 0.1666 && w (1) = 0.3 > 0.2 Frequent
Sup (2) = 0.854 > 0.1666 && w (2) = 0.2 = 0.2 In-frequent
Sup (3) = 0.938 > 0.1666 && w (3) = 0.4 > 0.2 Frequent
Sup (4) = 1.068 > 0.1666 && w (4) = 0.7 > 0.2 Frequent
Sup (5) = 0.315 > 0.1666 && w (5) = 0.6 > 0.2 Frequent

Now, frequent item set's is identified from Local databases.
Taking only frequent item sets from above;
Sup (1) = 0.321 > 0.1666 && w (1) = 0.3 > 0.2
Sup (3) = 0.938 > 0.1666 && w (3) = 0.4 > 0.2
Sup (4) = 1.068 > 0.1666 && w (4) = 0.7 > 0.2
Sup (5) = 0.315 > 0.1666 && w (5) = 0.6 > 0.2

Thus, we can eliminate 'n' no. of itemsets .Unsatisfied itemsets will be eliminated in the uncertain distribution databases. Frequent itemsets are extracted to central database from Local databases .In this process most of the low frequency items is reduced .From each and every local sites frequent itemsets are identified, transferred &stored in the central database. Comparatively, many dis-qualified itemsets are removed ,thus the scanning time will be reduce, performance time will be decreased. Again numerous item sets are appeared in the Central Data Base (CDB).
So we need to find out the High frequency item sets by using Global Minimum Threshold value (GMTV).

Example:-
Min sup 2 = 0.1666 WT min = 0.2

Sup (1) =0.321 >Min sup 2 && w (1) = 0.3 > WT min
Sup (3) =0.938 > Minsup2 && w (3) = 0.4> WT min
Sup (4) =1.068 > Min sup2 && w (4) =0.7 >WT min
Sup (5) =0.315 > Min sup 2 && W (5) =0.6 > WT min

Apply GMTV =0.211 WT min =0.2

W sup (1) =0.321 > 0.211 && w (1) = 0.3 > 0.2
W sup (3) =0.938> 0.211 && w (3) = 0.4 > 0.2
W sup (4) =1.068 > 0.211 && w (4) = 0.7 > 0.2
W sup (5) =0.315 > 0.211 && w (5) = 0.6 > 0.2

W sup (1) = 0.321*0.3 = 0.0963 < 0.211 In-frequent
W sup (3) = 0.938*0.4 = 0.3752 > 0.211 frequent

W sup (4) = 1.068*0.7 = 0.7476 > 0.211 frequent
W sup (5) =0.315*0.6 = 0.189 < 0.211 In-frequent

W sup (3) , W sup (4) High frequent item set .
WIEGHTED HIGH FREQUENCY ITEM SETS IN UNCERTAIN DISTRIBUTED DATABASE:
WHIF Algorithm:
Input: - A transaction set T with $|T| = N$, - A set I of items in the transactions, - $minsup1$, $minsup2$, - Weights of the items $w1,2,\ldots,wm$, - Minimum weight threshold: min_weight.

Output: All the Uncertain Weighted High Frequent Itemsets.
Algorithm: Step 1: Pre-Processing GMTV=min sup1 , $minsup2$= min sup1/N,
Min weight (wt)=average transactions Tavg/2
Step 2: /* Find the uncertain weighted frequent items 1.Scan through the database to find weighted frequent items x that do not satisfy both pruning conditions.
Condition 1: sup (x) > min sup 2 && R W(x) > W min
2. Update the database by deleting all infrequent items in local database that satisfy pruning conditions
3. (sup(x)*w(x)>minsup2 || w(x) > wt min)
4. Scan through the central database to find uncertain weighted High frequent-itemsets x that do not satisfy both pruning conditions.

Conclusion:-
This paper involves extracting weighted High frequency itemsets (WHFI) from uncertain distributed databases (UDDB) .Threshold based WHFI from Local sites in first level at pruning by min sup2. From multiple local sites (computer) reduce the infrequent uncertain itemsets .By using WHFI algorithms efficiently than the previous algorithm.
In order to overcome the problem, this paper propose new algorithm WHFI. Remove the infrequent itemsets in UDDB in local sites by using min sup2.Then index the frequent itemsets in central database and identify the High frequent itemsets in central UDDB. Comparing WHFI algorithm with dynamic programing WHF is move efficient. it reduce no of scans automatically reduce the Time complexity.

REFERENCES:

1. R. Agrawal, T. Imielinski, and A. N. Swami, “Mining association rules between sets of items in large databases”, Proceedings of ACM SIGMOD International Conference on Management of Data, ACM Press, Washington DC, pp.207-216, May 1993.
2. P. N. Santhosh Kumar, C. Sunil Kumar, C. Venugopal, “Improving Association Rule based Data Mining Algorithms with Agents Technology in Distributed Environment”, Proceedings of The Intl. Conf. on Information, Engineering, Management and Security 2014 [ICIEMS 2014].
3. M. H. Dunham, Y. Xiao, L. Gruenwald and Z. Hossain, “A Survey Of Association Rules”. International Journel of Computer TheoryAnd Engineering, vol.4, No.2 June 2003
4. Z. Qiankun, S. B. Sourav, “Association Rule Mining: A Survey”, Technical Report, Center for Advanced Information Systems (CAIS), Nanyang Technological University, Singapore, 2003.
5. R. Agrawal, R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases”, Proceedings of the Twentieth International Conference on Very Large Databases, pp. 487-499, Santiago, Chile, 1994.
6. A. Savasere, E. Omiecinski, and S. B. Navathe, “An Efficient Algorithm for Mining

Association Rules in Large Databases", Proceedings of the 21nd International Conference on Very Large Databases, pp. 432-444, Zurich, Swizerland, 1995
7. J. Han, J. Pei,Y. Yin. "Mining Frequent Patterns without Candidate Generation". Proc of ACM-SIGMOD, 2000.
8. J. Han, J. Pei, "Mining frequent patterns by pattern-growth: methodology and implications", ACM SIGKDD Explorations Newsletter 2, 2, 14-20.
9. J. S. Park, M.-S. Chen, and P. S. Yu, "Efficient Parallel Data Mining for Association Rules", Proceedings of the International Conference on Information and Knowledge Management, pp. 31-36, Baltimore, Maryland, 22-25 May 1995.
10. M. J. Zaki, M. Ogihara, S. Parthasarathy, and W. Li, "Parallel Data Mining for Association Rules on Shared-Memory Multiprocessors", Technical Report TR 618, University of Rochester, Computer Science Department, May 1996.
11. D. W.-L. Cheung, J. Han, V. Ng, A. W.-C. Fu, and Y. Fu," A Fast Distributed Algorithm for Mining Association Rules", Proceedings of PDIS, 1996.
12. T. Shintani and M. Kitsuregawa, "Hash Based Parallel Algorithms for Mining Association Rules", Proceedings of PDIS, 1996.
13. E.-H. Han, G.e Karypis, and V. Kumar, "Scalable Parallel Data Mining For Association Rules", Proceedings of the ACM SIGMOD Conference, pp. 277-288, 1997.
14. M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, "New Parallel Algorithms for Fast Discovery of Association Rules", Data Mining and Knowledge Discovery, Vol. 1, No. 4, pp. 343-373, December 1997.
20
15. 40. C. C. Aggarwal, "Managing and Mining Uncertain Data", Kluwer Press, 2009.
16. C. C. Aggarwal, Y. Li, J. Wang, and J. Wang, "Frequent pattern mining with uncertain data", KDD, pp. 29–38, 2009.
17. C. C. Aggarwal, and P. S. Yu, "A survey of uncertain data algorithms and applications", IEEE Transactions on Knowledge and Data Eng., 21(5): pp.609–623, 2009.
18. T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zufle, "Probabilistic frequent itemset mining in uncertain databases", KDD, pp.119–128, 2009.
19. Y. Tong, L. Chen, Y. Cheng, and P. S. Yu, "Mining Frequent Itemsets over Uncertain Databases", Proc. VLDB Conference, PVLDB, Vol. 5, 2012.
20. C. K. Chui, B. Kao, and E. Hung, "Mining frequent itemsets from uncertain data", The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), pp.47-58, 2007.
21. C. K.-S. Leung, M. A. F. Mateo, and D. A. Brajczuk, "A tree-based approach for frequent pattern mining from uncertain data", PAKDD, pp. 653-661, 2008.

22. Y. Liu, K. Liu, and M. Li, "Passive diagnosis for wireless sensor networks", IEEE/ACM Trans. Netw.,18(4):1132–1144, 2010
23. S. Suthram, T. Shlomi, E. Ruppin, R. Sharan, and T. Ideker, "A direct comparison of protein interaction confidence assignment schemes", BMC Bioinformatics, 7:360, 2006.
24. C.C. Aggarwal, "On Unifying Privacy and Uncertain Data Models," Proc. 24th IEEE Int'l Conf. Data Eng. (ICDE), 2008 A. Motro, P. Smets, "Uncertainty Management in Information Systems", ISBN 978-1-4615-6245-0, 1997.
25. C. H. Cai, A. W. Chee Fu, C. H. Cheng, and W. W. Kwong. "Mining Association Rules with Weighted Items," Proceedings of the Sixth International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2005), July 1998.
26. F. Tao, "Weighted Association Rule Mining Using Weighted Support and Significant Framework," Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 661-666, Aug. 2003
27. W. Wang, J. Yang, and P. S. Yu, "Efficient Mining of Weighted Association Rules

(WAR)," Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 270-274, Aug. 2000.
28. U. Yun, and J. J. Leggett, "WFIM: Weighted Frequent Itemset Mining with a Weight Range and a Minimum Weight," Proceedings of the Fourth SIAM International Conference on Data Mining, pp. 636-640, April 2005.
29. U. Yun, "Efficient Mining of weighted interesting patterns with a strong weight and/or support affinity". Information Sciences 177, 3477–3499 (2007).