Volume 6, Issue 3, July 2019 ISSN: 2348-8565 (Online)
IJMERT
International Journal of Modern Engineering and Research Technology
Website: http://www.ijmert.org Email: editor.ijmert@gmail.com

An Approach for Identifying Weighted Frequent Itemsets in Uncertain Database

Chinapaga Ravi
Research Scholar
Computer Science and Engineering
Jawaharlal Nehru Technological University
Hyderabad, (T. S.), INDIA
Email: ravi.chinapaga@gmail.com

M Bal Raju
Professor
Department of Computer Science and Engineering
Swami Vivekananda Institute of Technology
Hyderabad, (T. S.), INDIA
Email: drraju.jb@gmail.com

N Subhash Chandra
Professor
Department of Computer Science and Engineering
CVR College of Engineering
Hyderabad, (T. S.), INDIA
Email: subhashchandra.n.cse@gmail.com

ABSTRACT

The problem of mining frequent itemsets from uncertain data under a weighted probabilistic framework it consider transactions whose items are associated with existential probabilities and give a formal definition of frequent patterns under such an uncertain data model. The traditional algorithms are not appropriate for mining frequent itemsets on uncertain data model. A Transaction Based Weighted Rating frequent Itemset (TBWRFI) Algorithm is proposed to improve mining efficiency. First it identify the probability of each and every item as well as their weights. Test with the min $Min1_{expsup}$ and $Min2_{expsup}$and prune the items, which are not satisfy the condition. Through extensive experiments, we show that the data weighted frequent itemset mining approach can achieve significant results.

Keywords:— *Frequent patterns, uncertain data, weights, Item probability.*

I. INTRODUCTION

Frequent itemsets mining is analyzed and studied in many previous articles and journals. From the traditional approaches, that shows all items in a transactional data base, which are having equally importance and it will not consider their item count, variable price and occurrence probability and ratings. In transaction database, each item has a different level of importance. For example, in retail market on online applications some products may be much more expensive than the others, and these expensive items may not be present in a large number of transactions. Frequent pattern mining in certain data is much easier than uncertain data, in uncertain data mining face data mining challenges. The dataset is analyzed to discover frequent itemsets among probabilistic items. TBWRFI algorithm has two steps to eliminate the irrelevant items which are not satisfy the minimum threshold values. In first step each item probability identified, P = $\{p_1, p_2, p_3 \ldots p_m\}$ and probability of all items will be calculated with the existential probability P(X) = sum P (x_i) where itemset I = {i_1, i_2, i_3 … i_m}. Then second step weighted probabilistic items will be pruned through $Min2_{expsup}$ and found the relation among the items. Uncertainty consists of noisy, missed values, inconsistency, unstructured. If you want to find out frequent item set from uncertain data, traditional algorithm (or) previous techniques are Inappropriate.

II. Related work

Datasets that are collections of transactional records. Each record contains a set of items that are associated with *existential* probabilities. An itemset is considered *frequent* if it appears in a large-enough portion of the dataset. The occurrence of itemsets and identifying correlation among the frequent itemsets usually based on support count, user given minimum support to prune the low frequency of item. In uncertain data items, which contains probabilistic values so regular support base method should be enhanced in the form of probability.

Chui, et al. [20] have introduced the UApriori algorithm that is based on the computation of expected supports. In uncertain databases, downward closure property also is satisfied. So, we still can prune all the supersets of expected support-based infrequent itemsets. Chui, et al. [20] has proposed decremental pruning methods to improve the efficiency of UApriori. The decremental pruning methods are employed to estimate an upper bound on the expected support of an itemset from the beginning, the performance of UApriori is better than the other mining algorithms in the domain of uncertain data.

III. Proposed Approach

Technique of algorithm used to compute union of set of private subsets and extracts frequent itemsets by using *Transaction Based Weighted frequent Itemset Mining (TBWFIM) Algorithm.* Uncertainty consists of noisy, missed values, inconsistency, unstructured. If u want to find out frequent item set from uncertain data, traditional algorithm (or) previous techniques are Inappropriate.

Step 1: Expsup (x) >=$Min1_{expsup}$

R(x)* WT(x) >= WTR_{min}

Step 2: Expsup (x)* WTR(x)

>= Min2expsup

Let us take minimum Expected support1 should be greater than minimum expected support2.

In a transactional uncertain database items have to convert in probabilistic database by using Conditional probability. Consider the following transaction table which consists of certain Transactions and items.

Example database:

Transactions	Items
T1	I1 (2), I3 (3), I4 (1), I5 (2)
T2	I1 (3), I2 (2), I5 (1)
T3	I2 (2), I3 (2), I5 (2)
T4	I2 (4), I3 (1), I4 (2)

By taking quantity of an item for each transaction we will be calculating probabilities for each item in a transaction. Minsup = 0.007, $WT_{min} = 0.4$, $Min1_{expsup} = N * Minsup$,

$Min2_{expsup} = Min1_{expsup} / 2$
Transaction *T1:*

Taking quantity for each item is as follows,

Find the probability for all items in transaction T1,

I1=1/4*2/8=0.062

I3=1/4*3/8=0.093
I4=1/4*1/8=0.031
I5=1/4*2/8=0.062,
In this approach all items have its probabilities final probabilistic database is:

TID	Probabilistic Items
T1	0.062(I1), 0.093(I3), 0.031(I4), 0.062(I5)
T2	0.125(I1), 0.083(I2), 0.041(I5)
T3	0.083(I2), 0.083(I3), 0.083(I5)
T4	0.142(I2), 0.035(I3), 0.071(I4)

Weights Calculation:

WT (1) =0.81>0.4

WT (2) =0.736>0.4

WT (3) =0.816>0.4

WT (4) =0.44>0.4

WT (5) =0.728>0.4

Rating* Calculation*:

Rating	1	2	3	4	5	
I1	100	200	200	500	1000	=4.05
I2	50	50	100	500	100	=3.68
I3	50	100	75	200	500	=4.08
I4	10	50	10	400	100	=3.92
I5	50	75	80	300	150	=3.64

I1 = 1*100+2*200+3*200+4*500+5*1000/100 +200+200+500+1000

= 100+400+600+2000+5000/2000

= 8100/2000= 4.05

Similarly, all ratings calculated in above table and rating ratio will be calculated in below

Rating Ratio:

I1 = 4.05/5

= 0.81

I2 = 3.68/5

= 0.736

I3 = 4.08/5

= 0.816

I4 = 9.92/5

= 0.784

I5 = 3.64/5

= 0.728

From the above two tables profit and rating multiplied in respective items, resultant values for individual items considered as a weight to that item. Weights for the profit and rating is multiplied is $WT_p(x)$ * $WT_r(x)$. The resultant weighted rating values pruned through WTR_{min}

The probability and weighted profit rating itemset is to be considered for the finding weighted Frequency Items from uncertain data. Thus, it can pruned and unsatisfied itemsets will be eliminated from the uncertain dataset. In this process most of the low frequency items is reduced. Comparatively, many disqualified itemsets are removed, thus the scanning time will be reduce, and performance time will be decreased. So we need to find out the High frequency item sets

IV. Conclusion

This paper involves extracting weighted frequent itemsets (TBWRFI) from uncertain databases (UDB) on transaction based layout. Identify the weights and probability for each and every item in a transactional database. Pruning the itemsets in two steps individual items probability and weights and the multiplication of both probability and weights for finding uncertain weighted rating frequent items. These experimental results compared with u–apriori algorithm, that provided efficient than the previous algorithm. By using TBWRFI algorithms efficiently than the previous algorithm. Remove the infrequent itemsets in UDB by using min sup2. Then index the frequent itemsets in central database and identify the weighted frequent itemsets in central UDB. Comparing UApriori algorithm with dynamic programming TBWRFI is more 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 International. Conference 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 Journal of Computer Theory and Engineering, *vol.4, No.2 June 2003*

[4] Z. Qiankun, S. B. Sourav, "Association Rule Mining: A Survey", Technical Report, Centre 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, Switzerland, 1995.

[7] J. Han, J. Pei, Y. Yin. "Mining Frequent Patterns without Candidate Generation". Proceedings 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] 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 International Conference. 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 onIntelligent 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)*.

* * * * *