A Weighted Frequent Item-Set Mining using WD-FIM Algorithm

Abdulhusein latef Khudhair

Abstract**:** *Smart systems are the one of the most significant inventions of our times. These systems rely on powerful information mining techniques to achieve intelligence in decision making. Frequent item set mining (FIM), has become one of the most significant research area of data mining. The information present in databases is in-general ambiguous and uncertain. In such databases, one should think of weighted FIM to discover item sets which are significant from end user's perspective. Be that as it may, with introduction of weight-factor for FIM makes the weighted continuous item sets may not fulfil the descending conclusion property anymore. Subsequently, the pursuit space of successive item set can't be limited by descending conclusion property which prompts a poor time effectiveness. In this paper, we introduce two properties for FIM, first one is, weight judgment downward closure property (WD-FIM), it is for weighted FIM and the second one is existence property for its subsets. In view of above two properties, the WD-FIM calculation is proposed to limit the looking through space of the weighted regular item sets and improve the time effectiveness. In addition, the culmination and time productivity of WD-FIM calculation are examined hypothetically. At last, the exhibition of the proposed WD-FIM calculation is confirmed on both engineered and genuine data sets.*

Keywords **:** *Frequent item set mining (FIM), Downward closure property (DCP), Weight judgment downward closure property (WD-FIM), Data mining, Decision making.*

I. INTRODUCTION

A smart framework depends on good choice. Information mining has been showing an undeniably significant role in basic choice making exercises. FIM, the most trending research points in data mining, is a significant way to deal with association rules in datasets, i.e. broadly utilized in accuracy promoting, customized recommendation, network streamlining, restorative analysis, etc. Until now, many immaculate and perfect FIM approaches have been used for binary databases. But, with the quick innovations of information processing techniques, different types of complex information have risen, for example uncertain information.

Uncertain information implies the presence of a thing in a transaction, is depicted by a probability measure. If we consider a binary information model, at that point everything in a transaction can only be available or missing. In any case, in the uncertain information model, the presence of a thing in a transaction can be shown by a likelihood, in this manner it enables more data to be caught by the dataset which can prompt progressively exact analytical outcomes. But, each coin has different sides. Uncertain information model also has its downsides. The main limitation is, the size of the dataset, due to the storage of presence likelihood. Another limitation is, the mining calculations for uncertain databases are complex and tedious. Accordingly, creating productive mining approach for uncertain databases has turned into a hot research theme as of late.

Numerous approaches have been created to mine successive item sets in uncertain databases.

Existing investigations consider that every one of the things in uncertain databases have similar significance. But, in real picture, the significances of different items are typically unique to clients. For instance, the benefits of an expensive products and a low-cost living product can't be referenced at the same time. Thus, the mining dependent on just frequencies or presence probabilities without considering significances or estimations of things is inadequate to distinguish valuable and important patterns. To deal with this problem, a good arrangement is to give the clients a chance to allocate various weights to items to demonstrate their relative significances or qualities. The weight of things can be assigned by the clients based on their expertized area or explicit application needs to demonstrate benefits, dangers, costs, etc. In that way, item sets with high significances for the clients will be found. In addition, the idea of implementing weights of things can enormously diminish the frequent item sets. But, the DCP used for mining frequent item sets in uncertain databases would not hold any more due to the various weights, allocated to things.

This implies that a rare item set may have a frequent superset. Subsequently, the search area can't be limited by DCP anymore that will prompt low time proficiency of FIM calculations.

In this paper, based on the weight judgment DCP, the WD-FIM approach is suggested to limit the looking through space of weighted successive item sets and to increase the time proficiency. Subsequently, valuable and important weighted FIM in uncertain databases can be found. The primary goals of this paper are mentioned as following.

1. The weight judgment DCP and the existing property of weighted incessant subsets for uncertain databases are presented and demonstrated.
2. The WD-FIM approach is demonstrated in detail to limit the looking through space of weighted FIM and to increase the time proficiency.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author

Abdulhusein Latef Khudhair*, Shatt Al_Arab University College Iraq-Basrah abdulhussain2002@yahoo.com

3. A lot of investigations are directed to assess the exhibition of the WD-FIM approach based on execution time, number of patterns and memory utilization.

The rest of the paper is composed as: Section 2 contains the study of related works. In segment 3, the related properties are presented and demonstrated hypothetically, and the WD-FIM approach is portrayed in detail. Moreover, the time effectiveness of WD-FIM is also examined in this area. Exploratory outcomes are demonstrated in Section 4. At last, section 5 concludes the work and also discusses about future direction.

II. RELATED WORK

With improvement in data storage and retrieval, data management has become simple process, that why many organisation nowadays store this data into their database so that they can use it in later stage. Growing data mining applications is the best example of it. Using this data and machine learning algorithm, prediction and hidden pattern matching becomes possible. This leads to utilization of daily, real-time datasets, this dataset is not structure, this data is collected from various sources like satellite or sensors and then it is stored. Most of this data is unknown and it is present in the un-structured form. This kind of data where there is no certain pattern or structure, is known as uncertain datasets. Handling uncertain datasets required more complex data mining algorithm, so that proper weightage can be applied to data so that they can align in similar plane and hidden patterns can be observed. Many researches has been carried out on Mining information from uncertain data set and the accuracy rate of this algorithm is also good. Many FIM algorithms has also been proposed using which market-basket analysis has come across positive result statics.

There are basically 2 types of FIM algorithm which is classified as

1. candidate generate-test based uncertain frequent item set mining 2. pattern-growth mining

U-Apriori algorithm which is proposed by Chui et al. uses generate-test based uncertain frequent item set mining for FIM purpose. Using U- Apriori algorithm frequent data set is identified [6]. U-

Apriori is the first algorithm which uses uncertain data set for mining the hidden information. Like Apriori algorithm, this also needs to take all the dataset into count, make pair of it and then frequent dataset is decided, this algorithm generates lot of pairs and hence time and space complexity issues can be arrived. Pruning technique which is quite famous in classification algorithm cab be applied to FIM so that not required or least required data sets can be ruled out from the analysis. This technique is known as decremented pruning, proposed by Chui and Kao. It reduces the issue present with time and space complexity in U-Apriori algorithm. After this many research were carried out regrading approximation methods it includes MBP approximation methods proposed by L. Wang, D. W.-L. Cheung in 2012. It uses approximation and statistical based techniques to identify hidden pattern in FIM. X. Sun, L. Lim, and S. Wang in 2012 proposed a research – ‘An approximation algorithm of mining frequent item sets from uncertain dataset’, in which IMBP was introduced which has improvement in space and time efficiency over MBP algorithm, the major concern here was is accuracy of underlying algorithm get reduced to some extent.

In case of pattern growth researches, a smaller number of candidates gets generated as compare to candidate generate based algorithm [7]. This algorithm is highly based on tree or hyperlinked structure. UH-mine to mine frequent patterns which is proposed by Aggarwal et al. is the based example of hyperlinked based structure [8]. Similar to hyperlink, Leung et al. proposed a model which is based on tree structure, in this research tree structure is used to store the information or nodes of uncertain datasets [9]. FP-growth algorithm also uses hyperlink-based model to store this dataset. UFP-growth and CUF growth algorithms were proposed by, Aggarwal et al[10] and Leung and Tanbeer [11] which uses tree pruning which reduces the tree size after compilation this helps to reduce the time complexity to great extent. CUF tree were advanced version of UFP growth tree. Later on Leung and Tanbeer proposed PUF-growth algorithm for mining uncertain dataset. TPC – growth tree is advancement in PUF, in which upper bound is reduced so that time complexity can increase but on counterpart it slightly reduces the accuracy. Time required for PUF is less than CUF tree. C. W. Lin and T. P. Hong proposed CUPF tree. In CUFP tree recursive calls are not present which means it takes count of only exact frequent patterns [13]. But as the dataset size increases the performance of CUFP-mine decreased. L. Wang, L. Feng proposed AT-Mine algorithm [12] which is based on tree model, but it reduces the fatal issues which were present in CUFP-Mine. Accuracy of CUFP-mine algorithm is way better than CUFP algorithm but still there exist an issue of time and space complexity. G. Lee and U. Yun proposed a data structure less pattern even though it accuracy is not greater than other algorithm but the false positive rate optimization is far better than previous pattern based algorithms [14]. W. Wang, J. Yang proposed UWFI model which is tree based and uses Wight into account which increases the accuracy of the algorithm but it has similar dis-advantages as described above. All this algorithm tries to produces better result, but as FIM usages large dataset constrains, data size in each iteration become large and the data mining algorithm becomes complex. LUNA algorithm was proposed by Lee and Yun in 2014, this algorithm uses list as a data structure and it also applies pruning methods. This algorithm produces complete set of frequent data sets, and their no pattern loss. This is a traditional approach of data mining, and in case of FIM, size and complexity are a major issue, as if the pruning is avoided then there is a problem of data accuracy and if the data structure is allowed to grow then accuracy increases with time complexity. To increase accuracy further, a real time value of item from item set can be considered. In technical terms we call this real time value as weight. Weight is applied to each item set so that its value can be normalized. There exist many weights based FIM algorithms which works on certain dataset. This are Weighted Frequent Item set Mining (WFIM), Weighted Maximal Frequent Pattern mining over data streams based on Sliding Window model (WMFP-SW), Weighted Association Rules (WAR), WSpan algorithm, Weighted Association Rule Mining (WARM), Weighted Erasable Patterns (WEP), Maximal frequent pattern mining with Weight conditions over data Streams (MWS). Now this technology of weight is applied in mining un-certain datasets. Uncertain Mining of Weighted Frequent Item sets was proposed by Lee et al. [15]. In this algorithm tree based approached is used to find out uncertain frequent dataset. Result analysis of this research prove that, not only matrix accuracy but the real time accuracy and existential probability detection of this algorithm has good accuracy rate. High Expected Weighted Item set (HEWI-U-Apriori) Algorithm was proposed by Lin et al., it has upper bound and it also uses downward closure property to reduce the search space so that time required for item set generation is minimized [16]. Still the required time is greater than expected one, that's why further research is needed to reduce time complexity.

III. PROPOSED MODEL

In this paper, we proposed an algorithm for mining which is based on weight of the frequent item set, instead of using item weight, weight of item set is considered. In this research candidate generate and test paradigm model is used for generating frequent data sets. User defined weight value is applied to the algorithm. Weight and support of frequent item set is the key of this algorithm then Apriori_gen function from HEWI-UApriori is used for finding out. Candidate weighted frequent item set, their count is used to find out reduced Candidate weighted frequent item set. This algorithm reduces the searching space for finding weighted frequent item set. Because of this reduced searching space, time optimization in proposed algorithm is observed.

Phase 1: calculation of weight and expSupport of item sets:

In this phase the first step is to collect the weight from user, instead of using the calculated, stored or any pre-defined values, user defined weight is used so that items can be normalized to more realistic value. Wight of the item is nothing but the importance of an item. Weight of the item i is denoted as w(i).

In the proposed algorithm instead of using weight of each item, weight of item set is used in calculation. It is nothing but an average of weight of all items present in the specific item set. Weight of item set is calculated as,

$$w(x) = \frac{\sum \quad \mathrm{W(X)}}{|\mathrm{K}|} I\ \varepsilon\ x$$

Where, W(X) is the weight of item set x,

N is the total number of items present in item set X.

Existential probability is needed to calculated, the support of item set in the given data set (DS). It is also called as transaction item set probability. Existential probability in each transaction is required for finding support, it is denote as $p(X, T_m)$ where, X is the item set for in which item is present and T_m is that specific transaction. Finally using summation Existential probability of data set is computed. Formula for Existential probability of Dataset X is:

$$\mathrm{P(X, T)} = \prod p(I, j)$$

Iε X

Now to minimize the search space in the first space, expected support value is used. Expected support is denoted as expSupport(X). It is calculated by summation of existential probability of all item sets which contain item X.

$$\mathrm{expSupport(X)} = \prod_{\mathrm{X\ cTq}\ \Lambda^{\mathrm{Tq}}\ \text{€}DS} p(X, T)$$

Phase 2: finding weighted frequent item set

Frequent item set is nothing but an item set which has the expected support of dataset X greater than or equal to minimum expected support

*expSupport(X) >= δ * |DS|*

Where δ is threshold value of minimum support.

|DS| is total count of transaction present in DS.

Expected weighted support is calculated by multiplying weight of item set X and expSupport of X. for Analysis purpose it is defined as expwSupport(X).

expWSupport(X) = w(X) * expSupport(X)

Finally, in phase 2, weighted frequent item set is prepared. If the expected weight support is equal or greater than minimum expwSupport(X), then it is included in frequent item set.

Minimum expected weight support = ε *|DS|

Where ε is weight support threshold value

And |DS| is total count of transaction present in DS.

Condition for weighted frequent item set is:

expwSupport(X) >= Minimum expected weight support

Phase 3: WD-FIM algorithm

This Algorithm uses the Apriori Gen function of U-Apriori algorithm. U-Apriori algorithm. U-

Apriori algorithm is based on a thesis that all superset of infrequent item set is not necessarily frequent. It is an iterative and bottom up approach.

In U-Apriori algorithm, Apriori Gen Function creates k-candidate C_k after that Sub-Set function is used to check their support count. All candidate item sets which has support greater than specified value, is selected as frequent. Then this K-item set is used in Apriori gen function to get candidate for next iteration. When Ck+1 becomes empty, this algorithm stops.

In this algorithm gen function of U-Apriori algorithm is used to find out frequent item set using iteration. U-Apriori algorithm is used to find out frequent item set in uncertain datasets, while WDFIM algorithm is used to find out weighted frequent item set in uncertain datasets.

Pseudo code of WD-FIM algorithm:

Input: data of uncertain dataset

ε is min expected weighted support – (user-specified).

Output: frequent weighted itemset.

Code:

```
Data pre-processing
for each item I in complete Data set

  {
        Scan dataset and compute expwSupport(I)
         If(expwSupport(X) >= Minimum expected weight support)
          {
             $WFIS_1 = WFIS_1 U \{I\}$
           }
       }
  $WFIS = WFIS U WFIS_1$
  $CWFIS_1 = I$
  $SWFIS_1 = sort (CWFIS_1)$ by weight desc
  Z = 2
  While ($WFIS_{k-1}$ != empty)
  {
     $CWFIS_k = genFunction(WFIS_{k-1} , CWFIS_1)$
     $NCWFIS_k = WeightedgenFunction(WFIS_{k-1}-WFIS_{k-1}),SCWFIS_1)$
     $RCWFIS_k = CWFIS_k - NCWFIS_k$
     Foreach (X itemset in $RCWFIS_k$ )
      {
         Compute expWSupport(X)
         If(expWSupport(X>= Minimum expected weight support))
        {
                                  $WFIS_k = WFIS_k U \{X\}$        }
      }
    $WFIS = WFIS U WFIS_k$
  }
  Return WFIS
```

In this algorithm first the data pre-processing is carried out, to remove the noise form the data set. This step is important and significant in uncertain data set. Along the way there is high probability that few item are unrelated to others, in such case this item should be maintained in training.

After pre-processing step, data set DS is initialized. expWsupport for each item which is present in data set is computed . This is iterative phase in which expwSuuport of each item is compared with minimum expected weight support. If the expWsupport value is greater, then item set is included in weighted frequent item set. After this set weighted frequent set is get generated. Now to reduce the time complexity, NCWFISk and RCWFISk is Calculated. Now for each item set one more iteration process is carried out, and stopping condition for this iteration is WFIS = empty. In the next step genFunction() from U-Apriori algorithm is used, after applying this candidate weighted frequent item set is generated. WeightedGenFunction is applied on the CWFIS output to get number of candidate weighted frequent item set. This values are important in finding reduced candidate weighted frequent item set. This reduced version of candidate weighted frequent item set narrow down the search space of the algorithm. After this iteration weighted frequent item set is returned.

IV. RESULTS AND DISCUSSION

In this section performance of WD-FIM algorithm is compared with different algorithm, time required for the item set generation is compared among different algorithms. It is observed that accuracy and other factors of U- Apriori, HEWI-U-Apriori and WD-FIM algorithm is good. U-Apriori algorithm is most famous algorithm. In this section the comparison between this algorithms is analysed so that a best approach can be selected to reduce the time required for frequent item set generation of uncertain data set. Along with time complexity, the dependencies of Minimum Expected Weighted Support Threshold and relationship of dataset count with frequent item set, with all these algorithms is compared. Finally memory factor is also compared for these algorithms.

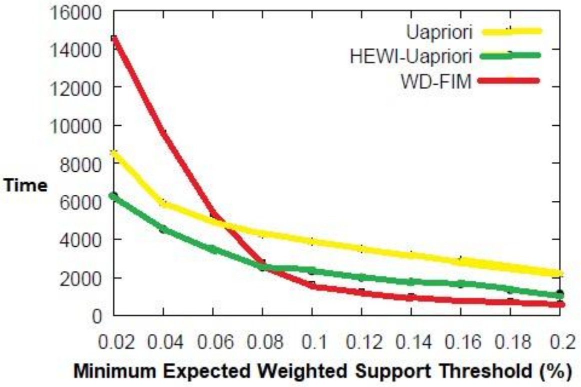

Figure 1: Runtime analysis

Figure 1, shows the result analysis of U-Apriori, HEWI-U-Apriori, and WD-FIM algorithm , it is observed that first the time required for WD-FIM algorithm is high, but as the value of Minimum expected weighted support threshold is decreases the complexity of WD-FIM algorithm improves and after some extend the time complexity is much better than U-Apriori and HEWI-U-Apriori algorithm.

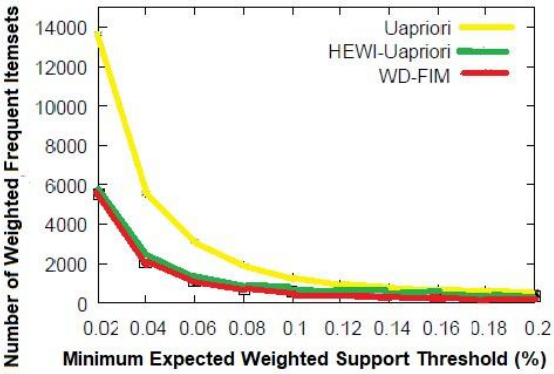

Figure 2: Count of weighted frequency item set

From above figure it is observed that as the Number of Weighted Frequent Item sets, Minimum Expected Weighted Support Threshold value get decrease or vice-versa, this analysis shows that all these algorithm is highly depended on Minimum Expected Weighted Support Threshold. It is observed that as minimum expected weighted support threshold value increases he pattern discovery count reduces. Initially, the drop is too heavy but as threshold it adjusted, count is level off.

It is also observed that count of frequent item set in case of U-Apriori algorithm is always greater than HEWI-Uapriori and WD-FIM Algorithm. The count of discovered pattern by HEWI-U-Apriori and WD-FIM is almost equal.

Memory analysis

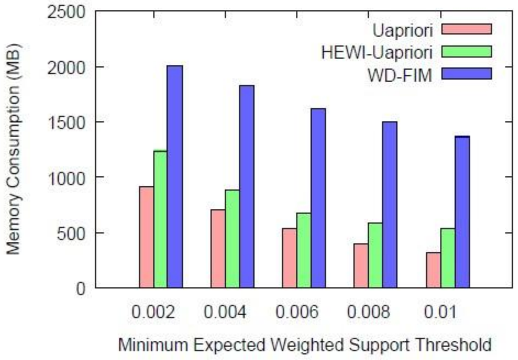

Figure 3: Memory consumption analysis

From the above figure it is observed that memory consumption from U-Apriori algorithm is less than HEWI-U-Apriori and WD-FIM. The reason behind this is U-Apriori algorithm uses downward closure property to reduce the frequent item set count. In WD-FIM algorithm downward closure property is applied on weighted frequent item set and the value of candidate weighted frequent item set is always need to be stored in the memory throughout the process.

IV. CONCLUSION

To find out the frequent item set in the uncertain dataset, weight based approach is used. Weight downward property is used to narrow the search space of WD-FIM algorithm. Finally the comparison with respect to time, number of frequent item set and memory consumption is carried out with U-Apriori, HEWI-U-Apriori and WD-FIM algorithm. It is observed that the time complexity of WD-FIM is better than other algorithm, memory consumption of WD-FIM algorithm is found out it be more than U-Aprioir , further research can be made on WD-FIM algorithm to reduce the memory consumption by optimising the handing value of candidate weighted frequent item set.

REFERENCES

1. Xuejian Zhao, Xihui Zhang, Pan Wang, Songle Chen and Zhixin Sun, “A weighted frequent itemset mining algorithm,” 2016.
2. C. K.-S. Leung, M. A. F. Mateo, and D. A. Brajczuk, “A tree-based approach for frequent pattern mining from uncertain data,” in Proc.PAKDD, 2008, pp. 653 6621.
3. R. Ishita and A. Rathod, “Frequent Itemset Mining in Data Mining: A Survey,” International Journal of Computer Applications, vol. 139, no. 9, pp.15-18, April 2016.
4. L. Yue, “Review of Algorithm for Mining Frequent Patterns,” International Journal of Computer Science and Network Security, vol. 15, no.6, pp.17-21, June 2015.
5. Wang, Le & Feng, Lin & Wu, Mingfei, “AT-Mine: An Efficient Algorithm of Frequent Itemset Mining on Uncertain Dataset.” Journal of Computers, 2013.
6. C. K. Chui, B. Kao, and E. Hung, “Mining frequent itemsets fromuncertain data,” in Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2007, pp. 47–58.
7. J. Pei, J. Han, and W.Wang. “Constraint-based Sequential Pattern Mining:The Pattern-Growth Methods,” Journal of Intelligent Information, April 2007.
8. C. C. Aggarwal, Y. Li, J. Wang, and J. Wang, “Frequent pattern mining with uncertaindata,” in Proceedings of the ACM KDD 2009, 2009
9. K. S. Leung, M. A. F. Mateo, and D. A. Brajczuk, “A tree-based approach for frequentpattern mining from uncertain data,” in Proceedings of the PAKDD 2008, 2008
10. C. C. Aggarwal, Y. Li, J. Wang, and J. Wang, “Frequent pattern mining with uncertaindata,” in Proceedings of the ACM KDD 2009.
11. K. S. Leung and S. K. Tanbeer, “Fast tree-based mining of frequent itemsets fromuncertain data,” in Proceedings of International Conference on Database Systems for Advanced Applications, 2012.
12. C. K. Leung, R. K. Mackinnon, and S. K. Tanbeer, “Tightening upper bounds to the expected support for uncertain frequent pattern mining,” in the 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, 2014
13. C. W. Lin and T. P. Hong, “A new mining approach for uncertain databases using CUFP trees,” Expert Systems with Applications, March 2012.
14. G. Lee, U. Yun, and H. Ryang, “An uncertainty-based approach: Frequent itemsetmining from uncertain data with different item importance,” Knowledge-Based Systems, vol. 90, Dec. 2015.
15. G. Lee, U. Yun, and H. Ryang, “An uncertainty-based approach: Frequent itemset mining from uncertain data with different item importance,” Knowledge-Based Systems, vol. 90, , Dec. 2015.
16. P. Fournier-Viger, T. P. Hong, and V. S. Tseng, “Weighted frequent itemset mining over uncertain databases,” Applied Intelligence, vol. 44, Jan. 2016

AUTHORS PROFILE

Abdulhusein latef Khudhair Shatt al_arab university college Iraq-Basrah abdulhussain2002@yahoo.com