iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** IV **Month of publication:** April 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Mining of infrequent itemset From Transactional Weighted Datasets Using Frequent Pattern Growth

J.Jaya[1], S.V.Hemalatha[2]
[1] PG Scholar, Dept. of CSE, KalaignarKarunanidhi institute of Technology, Coimbatore
[2] Asst Prof, Dept of CSE, KalaignarKarunanidhi institute of Technology, Coimbatore

Abstract— Itemset mining is a data mining method extensively used for learning important correlations among data. Initially item sets mining was made on discovering frequent itemsets. Frequent weighted item set characterizes data inwhich items may weight differently through frequent correlations in data's. But, in some situations, for instance certaincost functions need to be minimized for determining rare data correlations.In recent years, the thoughtfulness of the research community has also been focused on the infrequent itemset mining problem, i.e., discovering itemsets whose frequency of occurrence in the analyzed data is less than or equal to a maximum threshold. This grind addresses the discovery of infrequent and weighted itemsets, i.e., the infrequent weighted itemsets, from transactional weighted data sets. To speech this issue, the IWI-support measure is defined as a weighted, frequency of occurrence of an itemset in the analyzed data. Occurrence weights are derived from the weights associated with items in each transaction by applying a given cost function. In particular, we focus our attention on two different IWI-support measures: (i) The IWI-support- min measure, (ii) The IWI-support-max measure. Furthermore, two algorithms that perform IWI and Minimal IWI mining efficiently, driven by the proposed measures, are presented.

Keywords— Association Rule Mining, Infrequent item set, classification, Utility mining

I. INTRODUCTION

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases.

Datamining is the process of revealing non-trivial, previously unknown and potentially useful information from large databases.Datamining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Datamining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified

Technically, Datamining is the process of finding correlations or patterns among dozens of fields in large relational databases.Datamining finds patterns and relationships using data analysis tools and techniques in order to build models. There are two main kinds of models in Datamining. One is predictive models, which uses data with known results to develop a model that can be used explicitly to predict values. Another is descriptive models, which describes the patterns in existing data. Data mining finds its application mainly on Market basket analysis, Risk analysis, Fraud Detection, DNA data analysis, Web Miningetc.

In Datamining Association Rule Mining is a popular and well researched Datamining technique for finding interesting relationship between variables in a large database. Market basket analysis is a good example of this model. An example of an association rule is "customers who buy computers tend to also buy financial software".The extraction of interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories is the main objective of Association rule mining. Association rule mining extracts interesting correlation and relation between large volumes of transactions

This process is divided into two phases. First phase is itemset mining. Second phase is rules construction. Itemset mining was focused on discovering frequent itemset, i.e., patterns whose observed frequency of occurrence in the source data (the support) is above a given threshold. Itemset below the threshold value is referred as Infrequent itemset.

Frequent itemsets mining is a central part of data mining and distinctions of association examination, namely association-rule mining and sequential-pattern mining respectively. From large amount of data, frequent itemset are constructed by concerning some rules or association rule mining algorithms to calculate all the frequent itemsets. In many association investigation methods, frequent itemset extraction is considered as a primary step. An itemset is named as frequent if it is available in a large-enough part of the dataset. This frequent occurrence of item is represented by means of the count of support. Consequently, it

requires complex techniques for hiding or restructuring users' private information through a data construction process. Furthermore, this technique does not yield the accuracy of mining results. Discovering such frequent pattern is termed as a significant position in mining relations, correlations.

However, considerably less consideration has been noticed to mining of infrequent itemsets, even though it has obtained major usage in mining of negative association rules from infrequent itemsets, statistical disclosure risk measurement whereas exceptional patterns in anonymous sample data can direct to statistical disclosure. Then infrequent itemsets is adapted to fraud detection whereas uncommon patterns in financial or tax data might imply unusual action associated with fraudulent behavior and then applied in the field of bioinformatics where unusual patterns in microarray data could imply genetic disorders. Patterns that are rarely established in database are frequently measured to be irrelevant. Such patterns are named as infrequent patterns. Mining infrequent patterns is a challenging attempt since there is huge number of such patterns that can be incorporated from a well-known data set. Generally, the primary issues in infrequent patterns mining are identification of appropriate infrequent patterns and efficiently discovering such patterns in large data sets.

This work reports the discovery of infrequent and weighted itemsets, i.e., the infrequent weighted itemsets, from transactional weighted data sets. To address this issue, the IWI-support measure is defined as a weighted frequency of occurrence of an itemset in the analyzed data. Occurrence weights are derived from the weights associated with items in each transaction by applying a given cost function. In particular, we focus our attention on two different IWI-support measures: (i) The IWI-support- min measure, which relies on a minimum cost function, i.e., the occurrence of an itemset in a given transaction is weighted by the weight of its least interesting item, (ii) The IWI-support-max measure, which relies on a maximum cost function, i.e., the occurrence of an itemset in a given transaction is weighted by the weight of the most interesting item. Note that, when dealing with optimization problems, minimum and maximum are the most commonly used cost functions. Hence, they are deemed suitable for driving the selection of a worthwhile subset of infrequent weighted data correlations.

II. PREVIOUS WORK

In[1]RAgarwal introduces Frequent itemset mining which is widely used data mining technique. Here, the rules are framed based on the itemset mined which is said to be frequent. Those itemset satisfying minimum support and confidence are taken as frequent and is used for framing association rules. Most approaches to association rule mining assume that all items within a dataset have a uniform distribution with respect to support. The main problem with this is items in a transaction are treated equally

In[2]W.Wang introduces the concept of weight to be assigned for item in each transaction which reflects the intensity or the importance of the item within the transaction. The main problem with this is that weights are introduced only during the rule generation step not used for the mining purposes.

In[3]Feng Tao et.al presents Weighted Association Rule Mining for frequent itemset mining. In this work the limitationof the conventional Association Rule Mining model is avoided specifically its inability for treating units differently.The presented method uses weights which can be incorporated in the mining process to resolve this difficulty. Then thechallenge is solved when doing enhancement towards using weight, especially the invalidation of downward closureproperty. In order to adapt weighting in the new setting, a set of new concepts are used. With this weighted downwardclosure term is used as a substitute of the unique downward closure property. At last this method is confirmed assuitable and gives reason for the efficient mining scheme in the new construction of weighted support. By learning thesimulation of the lattice building, solution is suggested that weight can be utilized to guide the mining focus to thosesignificant itemsets with high degree of consequence. *Transaction weight* is a type of itemset weight. It is a valueattached to each of the transactions. Usually the higher a transaction weight, the more it contributes to the miningresult. However weights are to be priorly assigned which is difficult in real life cases.

In [4]C.K.Chui addresses the issue of relating the weight to probability of occurrence but in most cases both of themare uncorrelated. For example an item which is very likely to occur in a transaction may seem to be of least importance.

In[5]Thomas Bernecker et.al presented Probabilistic Frequent Itemset Mining for mining uncertain transactional databases. This probabilistic method brings new probabilistic mechanism of frequent itemset which is based on probable world semantics. In this probabilistic circumstance, an itemset is said to be frequent if the probability thatitemset happens in at least minSup transactions is higher than a given threshold. Considerably this is the said to be firstmethod deals with the problem under probable world's semantics. In addition to the probabilistic mechanisms, aframework is presented in which it has the capability to solve the Probabilistic Frequent Itemset Mining (PFIM)problem proficiently

In[6] Jiawei Han et.al presented novel frequent pattern tree (FP-tree) structure, which is an widened prefix-tree construction for storing compressed, critical information about frequent patterns, and expands an effective FP-treebasedmining system, FP-growth, for mining the absolute set of frequent patterns by pattern fragment growth.Effectiveness of mining is attained with three methods:1) a huge database is compressed into a largely reduced. 2)thepresented FP-tree-based mining approves a pattern

fragment growth process to eliminate the costly generation of ahuge number of candidate sets. 3)Finally a partition-based method known as divide-and-conquer system is used todivide the mining job into a set of minor tasks for mining detained patterns in conditional databases, where the searchspace is reduced appropriately

In[7]David et.al presented a new algorithm of MINIT, for finding minimal τ -infrequent or minimal τ -concurrent itemsets. Firstly, a ranking of items is organized by estimating the need of each of the items and then generating a record ofitems in rising order of support. Minimal τ –infrequent itemsets are determined by using each item in rank orderiteratively calling MINIT on the maintained set of the dataset with regard to items using only those items with superiorrank than current items , after that checking each candidate of minimal infrequent items (MII) against the originaldataset is performed. A system that can be utilized to judge only superior-ranking items in the iteration is to preserve a"liveness" vector representing which items stay feasible at each level of the iteration

In[8]Laszlo et.al presented generation of rare association rules for mining of infrequent itemsets. This work presenteda method to taking out rare association rules that stay hidden for traditional frequent itemset mining algorithms. Whencompared with other method the presented method finds strong but rare associations that are local regularities in thedata are found. These rules are said to be "mRI rules" .Apriori computes the support of minimal rare itemsets (mRIs),i.e. rare itemsets such that all proper subsets are frequent. Instead of pruning the mRIs, they are retained. In addition, itis shown that the mRIs form a generator set of rare itemsets, i.e. all rare itemsets can be restored from the set of mRIswhich have two merits. Initially, they are highly informative in the case that they have an antecedent which is aproducer itemset while adding up the resultant to give ways for a closed itemset. Secondly, the amount of these rules isminimal, that is the mRG rules comprise a dense illustration of all largely confident associations that can be taken fromthe least rare itemsets.

In[9]Ashish Gupta e.al presented pattern-growth paradigm to discover minimally infrequent itemsets.They
recommend a new algorithm based on the pattern-growth paradigm to find minimally infrequent itemsets. It has nosubset which is also infrequent. This work uses novel algorithm of IFP min for mining minimally infrequent itemsets.Then the residual tree concept has been incorporated by using a variant of the FP-Tree structure which is known asinverse FP-tree. In order to mine the minimally infrequent itemsets, optimization of Apriori algorithm is performed.Finally the presented tree are used for mining of frequent itemset as well.

III.PROBLEM STATEMENT

Let $I=\{i1,i2,\ldots.i_m\}$ be the set of data items. A transactional data set $T=\{t1,t2\ldots.t_n\}$ is a set of transactions, where each transaction is a set of items in I and is characterized by a transaction ID(tid).

An itemset I is a set of data items. More specifically, we denote as k-itemset a set of k items in I. The support (or occurrence frequency) of an itemset is the number of transactions containing I in T. An itemset I is infrequent if its support is less than or equal to a predefined maximum support threshold ξ. Otherwise, it is said to be frequent. An infrequent itemset is said to be minimal if none of its subsets is infrequent. Given a transactional data set T and a maximum support threshold ξ, the infrequent (minimal) itemset mining problem entails discovering all infrequent (minimal) itemsets from T.

a. Weighted Transactional Data Set construction

The traditional support measure for driving the itemset mining process entails treating items and transactions equally, even if they do not have the same relevance in the analyzed data set. To treat items differently within each transaction we introduce the concept of weighted item as a pair $<i_k, w_{kq}>$, where i_k I is an item contained in tq T, while w_{kq} is the weight associated with i_k that characterizes its local interest/intensity in tq. Concepts of weighted transaction and weighted transactional data set are defined accordingly as sets of weighted items and weighted transactions, respectively

1) Procedure for Minimum Weighting Function

Step 1:Wref = lowest among weights in the original transaction

Iterative :

Step 2:Assign the Wref value to each item.

Step 3:NewWref=next lowest weight in the original transaction-(sum of the previous Wref value)

Step4:The process is continued until S is empty.

TABLE 1 EQUIVALENT WEIGHTED TRANSACTION

Tid	Original Transaction	Tid	Equivalent Weighted Transaction
1	<a,0><b,100><c,57><d,71>	1.a	<a,0><b,0><c,0><d,0>
		1.b	<b,57><c,57><d,57>
		1.c	<b,14><d,14>
		1.d	<b,29>

Similarly for maximum weighting function.

b. Infrequent Weighted Itemset Miner Algorithm

Given a weighted transactional data set and a maximum IWI-support (IWI-support-min or IWI-support-max) threshold ξ, the Infrequent Weighted Itemset Miner algorithm extracts all IWIs whose IWI-support satisfies ξ .Since the IWI Miner mining steps are the same by enforcing either IWI-support-min or IWI-support-max thresholds, we will not distinguish between the two IWI support measure types

1) *FP Tree:* IWI Miner is a FP-growth-like mining algorithm that performs projection-based itemset mining. Hence, it performs the main FP-growth mining steps: (a) FP-tree creation and (b) recursive itemset mining from the FPtree index. Unlike FP-Growth, IWI Miner discovers infrequent weighted itemsets instead of frequent (unweighted) ones. To accomplish this task, the following main modifications with respect to FP-growth have been introduced: (i) A novel pruning strategy for pruning part of the search space early and (ii) a slightly modified FPtree structure, which allows storing the IWI-support value associated with each node.

FP-tree is a frequent pattern tree . Formally, FP-tree is a tree structure defined below:

1. One root labeled as “null", a set of *item prefix sub-trees* as the children of the root, and a *header table*.
2. Each node in *the item prefix sub-trees* has three fields:
   - item-name : register which item this node represents,
   - weight, the importance of item in transaction,
   - node-link that links to the next node in the FP-tree carrying the same item-name, or null if there is none.
3. Each entry in the *header table* has two fields,
   - item-name, and
     weight value.

2) Pruning Strategy: To reduce the complexity of the mining process, IWI Miner adopts an FP-tree node pruning strategy to early discard items (nodes) that could never belong to any itemset satisfying the IWI-support threshold ξ. In particular, since the IWI-support value of an itemset is at least equal to the one associated with the leaf node of each of its covered paths, then the IWI-support value stored in each leaf node is a lower bound IWI-support estimate for all itemsets covering the same paths. Hence, an item (i.e., its associated nodes) is pruned if it appears only in tree paths from the root to a leaf node characterized by IWI-support value greater than ξ.

Algorithm 1 IWI-Miner(T,ξ)
Input:T,a weighted transactional dataset
Input:ξ,a maximum IWI-support thersold
Output:F,the set of IWIs satisfying ξ
1:F=ϕ
2: Count item IWI (T)
3:Initially FP-Tree is constructed
4:for all weighted transaction
5: Calculate Equivalent transaction
6: For all transaction create and insert into FP tree
7: end for
9:end for

10: IWIMining(Tree,ξ,null)
11:return F

Algorithm 2 IWIMining(Tree,ξ,prefix)
Input: Tree,a FP –tree
Input:ξ,a maximum IWI-support thershold
Input: prefix,the set of items
Output: F,the set of IWIs extending prefix
1.Items belonging to the header table associated with the input FP-tree are *iteratively* considered.
2. Initially, each item is combined with the *current prefix* to generate a new itemset I .
3. If I is infrequent, then it is stored in the output IWI set F .
4. Then, the FP-tree projected with respect to I is generated and the IWIMining procedure is *recursively* applied on the projected tree to mine all infrequent extensions of I .
5. Unlike traditional FP-Growth-like algorithms , IWI Miner adopts a different *pruning* strategy

c. Minimal Infrequent Weighted Itemset Miner Algorithm

Given a weighted transactional data set and a maximum IWI-support (IWI-support-min or IWI-support-max) threshold, the Minimal Infrequent Weighted Itemset Miner algorithm extracts all the MIWIs that satisfy . The pseudocode of the MIWI Miner algorithm is similar to the one of IWI Miner. However, since MIWI Miner focuses on generating only minimal infrequent patterns, the recursive extraction in the MIWIMining procedure is stopped as soon as an infrequent itemset occurs. In fact, whenever an infrequent itemset I is discovered, all its extensions are not minimal.

if the subsets of the itemsets are *not minimal*
Then the itemset is said to be *Minimal*
otherwise
not minimal

IV. RESULTS AND DISCUSSION

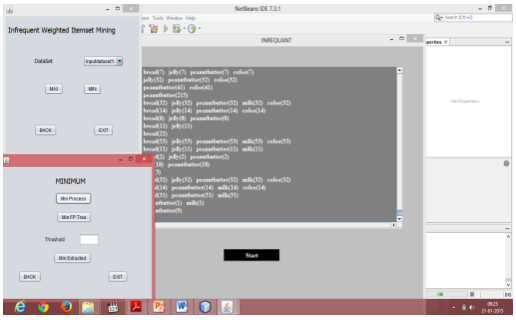


Fig 1:Weighted Transaction Equivalence

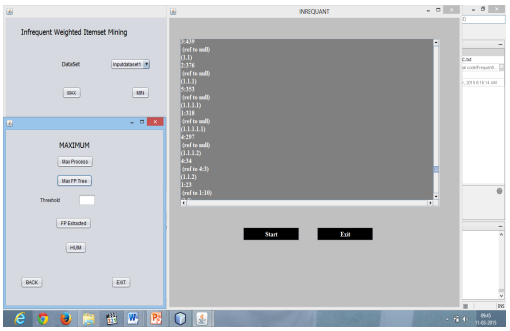


Fig 2:FP Tree construction

Volume 3 Issue IV, April 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

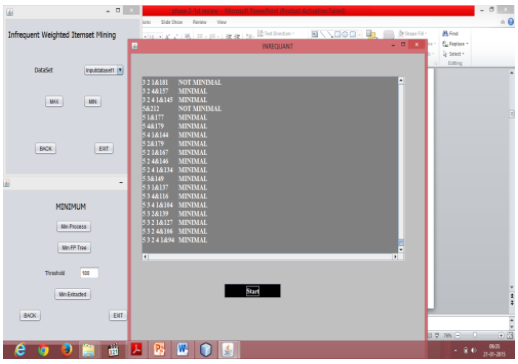


Fig 3: IWI-Miner

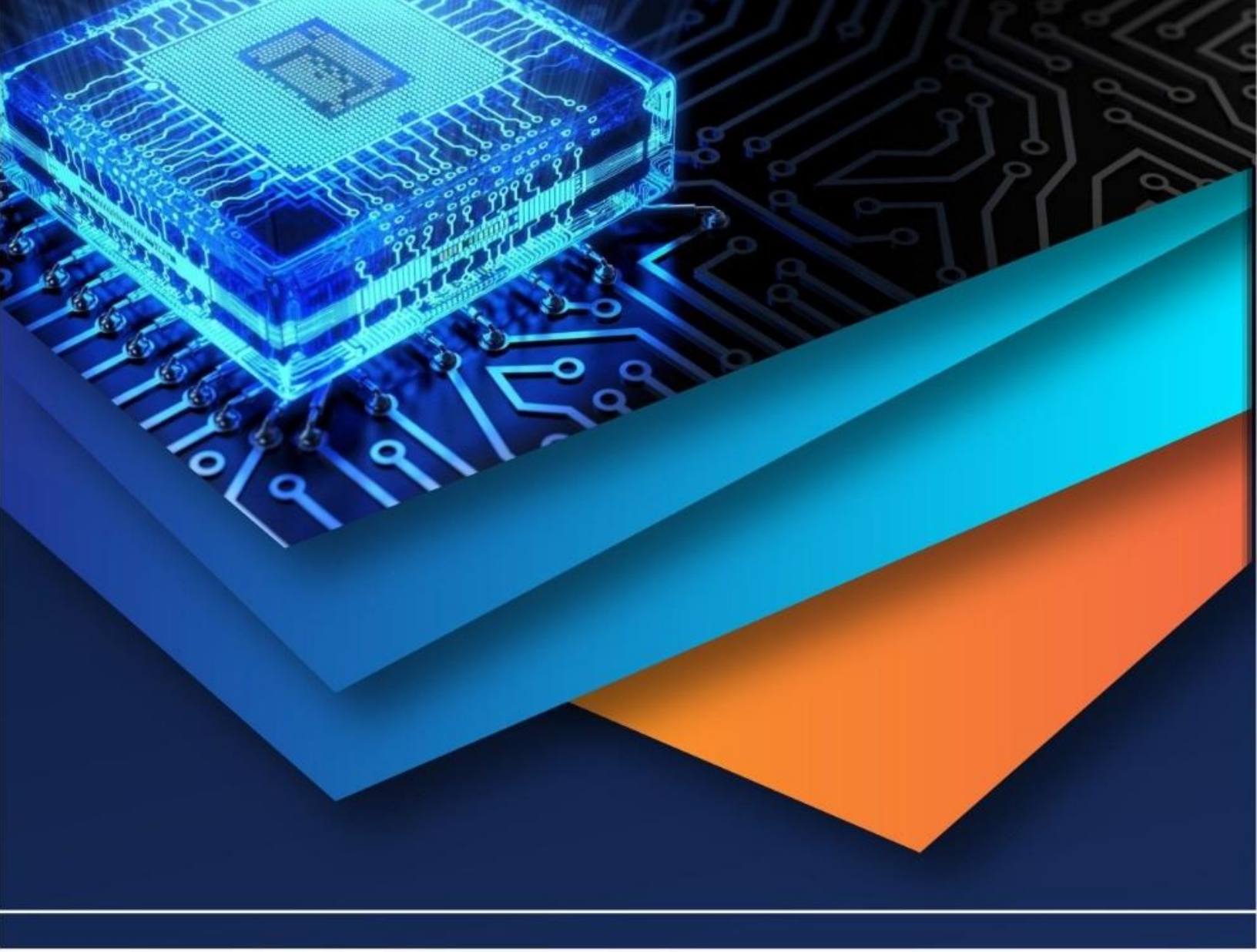
V. CONCLUSION AND FUTURE WORK

Frequent weighted item sets represent correlations frequently holding in data in which items may weight differently.Weights associated with the item is used instead of occurrence of item in the transaction.FP growth algorithm along with pruning techniques are efficiently used. However, in some contexts, e.g., when the need is to minimize a certain cost function, discovering rare data correlations is more interesting than mining frequent ones.. Two FPGrowth-like algorithms that accomplish IWI and MIWImining efficiently are also proposed.

Our future plan, we are analyzing the utility parameters of the infrequent weighted itemsets. We will consider both the individual profit of each item in a database and the bought quantity of each one in a transaction simultaneously

REFERENCES

[1] R. Agrawal, T. Imielinski, and Swami, “Mining Association Rules between Sets of Items in Large Databases,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’93), pp. 207-216, 1993.
[2] W. Wang, J. Yang, and P.S. Yu, “Efficient Mining of Weighted Association Rules (WAR),” Proc. Sixth ACM SIGKDD Int’l Conf.Knowledge Discovery and data Mining (KDD ’00), pp. 270-274, 2000.
[3] F. Tao, F. Murtagh, and M. Farid, “Weighted Association Rule Mining Using Weighted Support and Significance Framework,”Proc. nineth ACM SIGKDD Int’l Conf. Knowledge Discovery and DataMining (KDD ’03), pp. 661-666, 2003.
[4]C.-K. Chui, B. Kao, and E. Hung, “Mining Frequent Itemsets fromUncertain Data,” Proc. 11th Pacific-Asia Conf. Advances in KnowledgeDiscovery and Data Mining (PAKDD ’07), pp. 47-58, 2007.
[5]. T. Bernecker, H.-P.Kriegel, M. Renz, F. Verhein, and A. Zuefle,“Probabilistic Frequent Itemset Mining in Uncertain Databases,”Proc. 15th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD ’09), pp. 119-128, 2009.
[6]J. Han, J. Pei, and Y. Yin, “Mining Frequent Patterns without CandidateGeneration,” Proc. ACM SIGMOD Int’l Conf. Management ofData, pp. 1-12, 2000.
[7] D.J. Haglin and A.M. Manning, “On Minimal Infrequent Itemset Mining,” Proc. Int’l Conf. Data Mining (DMIN ’07), pp. 141-147,2007.
[8]Laszlo Szathmary, PetkoValtchev, and Amedeo Napoli,” Finding Minimal Rare Itemsets and Rare Association Rules” Proceedings of the4th International Conference on Knowledge Science, Engineering and Management (KSEM 2010)
[9]A. Gupta, A. Mittal, and A. Bhattacharya, “Minimally InfrequentItemset Mining Using Pattern-Growth Paradigm and ResidualTrees,” Proc. Int’l Conf. Management of Data (COMAD), pp. 57-68, 2011



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY