International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Developing an Efficient Mining Framework using Dependable High Utility Pattern Mining

Aditya Nellutla*[1], Dr. N. Srinivasan [2]

Submitted: 22/**07**/**2022** **Accepted:** 25/**09**/**2022**

Abstract: One of the most active areas of study is utility mining, which has a wide range of real-world application possibilities. A utility function is used in high utility pattern mining to extract all patterns that are more useful than the minimum. It is an inherent shortcoming of these algorithms because when this threshold is set too low, a large number of patterns are formed. As a result, it may be more difficult to interpret the patterns revealed throughout the mining process, making it less efficient. In addition, many of these patterns are unreliable and difficult to use in making judgments. Adapting the notion of dependability to mine an important sort of pattern known as reliable high utility patterns, this research offered a unique challenge of mining dependable high utility patterns. An effective method known as Dependable High Utility Pattern Mining (DHUPM) is provided to solve this problem. DHUPM presents numerous ways for effectively handling reliable patterns with high utility values and introduces three new metrics for measuring the reliability of utility-based patterns. The experimental results show that up to 98.56% of the patterns found by typical high utility pattern mining methods are unreliable. In comparison, the DHUPM technique yields patterns with an average dependability proportion that is at least 43.6% greater. In addition, the proposed pruning algorithms reduce both the performance and memory consumption of the programme.

Keywords: *Dependable Itemset Mining, Data Mining, Pattern Mining, Utility-based Patterns*

1. Introduction

Considerable study has been done on itemset mining, which is the process of analysing enormous datasets for intriguing relationships between items. An enormous number of transactions are included in the dataset, each of which has a unique collection of data. Itemset mining aims to identify interesting collections of items based on a user-specified measure of interestingness from such a dataset [1]. Various interestingness measures can be selected to extract, for example, Frequent Itemsets (FIs) comprising of commonly co items, Association Rules (ARs) consisting of similar information, High Utility Itemsets (HUIs) consisting of highly valuable ones, and so on, adapted to the needs of the consumer.

An attribute of an image, sensor data, or peptide set might be considered an item in a transaction from a broader perspective. There are a variety of uses for itemset mining, starting with customer transaction analysis. These include picture categorization and healthcare applications, among others. Since there is so much data out there, it is tough to mine items. The search space is defined as the set of all possible itemsets if there are M different items in a dataset [2]. Each of these item sets cannot be examined using a basic technique. Since this is the case, researchers have devised numerous data structures and algorithms that may efficiently reduce the search space by taking into account the kind of target itemset. The "downward closure property" for FIs is the most often used, which states that any subset must be common as well. By discarding itemsets that comprise one or more rare subgroups of items, the Apriori algorithm drastically reduces the search space [3].

Additionally, the FP-tree (Frequent Patten tree) can be constructed using the downward closure attribute to provide efficient hierarchical organisation of things exclusively connected to FIs. Upper bound utilities are established to ensure that the property is preserved among "possible" itemsets that may be HUIs, given that HUIs lack the downward closure property. Using these upper bound utilities, a tree structure is built to keep things connected only to prospective item sets [4]. Itemset mining approaches like the ones described above, on the other hand, are rigid due to the requirement of a certain data structure or algorithm to extract distinct types of itemset. One sort of itemset is specified by an interestingness metric, and there are several of these measures. Developing a database schema or method for all of these

[1] *Research Scholar, Sathyabama Institute of Science and Technology, Chennai 600 119.*
[2] *Professor, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai*
* *Corresponding Author Email: adityaresearch3@gmail.com*

item sets is plainly infeasible. Figure 1 shows the overview of data mining process.

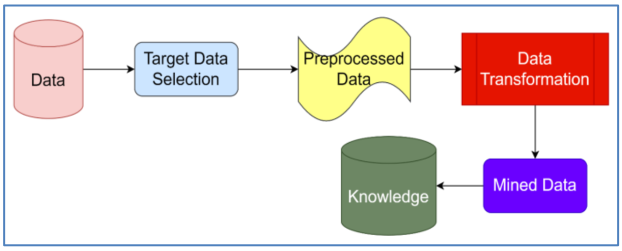

Fig. 1 Data mining process overview

As a result, a wide variety of itemsets remain a mystery since no data structure or technique exists to facilitate their discovery. One of the most important issues in itemset mining is the establishment of a consistent framework for extracting a variety of itemset kinds. Our solution to this problem is to use Reinforcement Learning to mine generic item sets (GIM-RL). An important part of the concept is based on how we humans look for a specific sort of item set in a collection [5]. As a starting point, it is likely that humans will skim over the information to acquire a general idea of what objects seem relevant to the target type, how they are connected, and so on. His next step is to compile a "itemset" of relevant things and see if it fits the goal kind. After that, a human modifies the plausible set by adding or eliminating things depending on the information gained from previous search experiences [6].

Humans are capable of adapting their strategy to making convincing item sets based on the type of target and their prior experiences. In order to provide a cohesive framework for itemset mining, the human itemset search is regarded essential. Based on the principles of behavioural psychology, we focus on Reinforcement Learning (RL), which allows an artificial agent to interact with an uncertain environment to learn the best strategy for sequential decision making. Each time the agent performs an activity, the environment responds by producing a reward. Repetition of this trial-and-error procedure helps to amass rewards from different activities at different environmental conditions. It does this by looking for the best possible policy that allows the operator to decide on a series of behaviours that maximises the total reward [7].

Co-occurrence is the only criterion used to extract useful patterns (i.e., itemsets, sequences, rules, etc.) out of a database table using Frequent Pattern Mining (FPM). Many real-world applications rely on the discovery of useful correlations between transactional items, and this technique is designed to accomplish so. While utility mining is the process of finding patterns in transactional databases that meet all or part of these requirements. For example, HUPM is divided into item sets, situations, principles, as well as high-utility sequential patterns based on the intended patterns [8].

Recommender systems, consumer market research, and therapeutic applications have all employed HUIM to derive process models. A series of pruning algorithms and data structures have also been developed to decrease the subspace and greatly increase speed the mining process. For example, researchers have recently devised as well as the Correlated HUIM, Stable HUIM, as well as Enclosed HUIM. Dependable HUIM, on the other hand, is the primary goal of the suggested technique, which aims to identify itemsets that may consistently yield high utility values in the absence of relevant data. Reliability conveys our belief that the utility function will return the same result when applied to new data. Statistical likelihood as a function of time is the most common way to quantify it. Anticipating demand for items, analysing consumer behaviour, are all critical functions of reliability mining [9].

An algorithm for mining itemsets that is dependable and profitable is an important research challenge. Accordingly, existing HUIM techniques have three major drawbacks. When it comes to extracting useful patterns from the data, classic HUIM algorithms rely solely on utility measures, while ignoring interestingness objective metrics. A further issue is that if the minUtil is set too low, it will generate a large number of patterns, which will make analysis more complex and take longer. In addition, the overwhelming majority of the similarities revealed by HUIM systems are not credible. When it comes to making important decisions, the user's ability to trust these results is weakened. As a result, it is difficult to come up with a method for mining trustworthy utility patterns.

As a result, the following are the paper's main contributions:

- Dependable High Utility Pattern Mining (DHUPM) is offered as a new method for discovering valuable itemsets in previously unknown data.
- Algorithm design and development to mine transactional databases for trustworthy high-utility itemsets.
- Pruning algorithms have been devised to minimise the search process and enhance performance when looking for trustworthy and high-utility items.
- Snipped Utility, Internal Reliability Factor, and Utility Coherence Correlation are proposed as relevant metrics for examining the stability of high utility item sets.

2. Literature Survey

An exhaustive and a non-exhaustive approach of itemset mining are both now available. Enumeration techniques for itemsets are included in the first. However, when a dataset grows, the runtime of comprehensive approaches drastically decreases, especially as the number of different objects increases. Using non exhaustive approaches, an approximation of the target type's itemsets can be generated instead. In the sections that follow, we'll go over the shortcomings of other approaches, both exhaustive and non-exhaustive, before highlighting the advantages of GIM-RL over them [10].

A data model or technique specific to a target type is typically used in an exhaustive approach. Apriori method and FP-growth (Sequential Pattern development) technique based on FP-tree have been established for effective enumeration of FIs because of the downward closure feature of FI. Closed FIs, which have no superset maintained by the same set of transactions, may be extracted using a divide-and-conquer strategy based on a binary operation vertical representation of a dataset [11]. An item expansions and dataset reduction in the depth-first order enable the approach to identify all maximum FI instances that are not included in any other instance of the FI model.

FP-growth is expanded by building upper bound ratings, three pruning approaches, and a parallel mining approach to efficiently extract weighted FIs composed of items connected with high weights. For a study of current infrequent itemset mining methods and an up-to-date survey, we extract rare and poorly weighted infrequent weighted itemsets by rewriting FP-growth with a specialised interestingness measure and a methodology of early eliminating unpromising items. It is an augmentation of a weighted FI in that the usefulness of each component in a transaction is computed by taking into account both its weight and its quantity [12]. Researchers have devised a list that makes it easier to expand an item set and calculate its utility, as well as a tree model based on the descending closure feature of upper limit utilities, to quickly extract all of the HUIs.

This is followed by an FP-growth process that converts requirements such "the median weights in a FI must be greater than a threshold" and incorporates them into the FI-growth process. It is common to use pattern sampling to approximate a probability distribution over the search space by assigning probabilities to each set of itemsets that are proportionate to how interesting they are in relation to the target type. As a result, the subset of itemsets that fit the target type is drawn from the entire set based on the probability distribution [13]. For this reason, patterns can only be applied to a small subset of itemset types. For example, an itemset type has to be described by a certain weight function or by XOR restrictions combined. GIM-RL is able to extract a wider range of item sets.

Furthermore, it is not taken into account whether or not an approximate posterior distribution for a source information may be transposed to that of a destination dataset. So, GIM-RL offers a highly flexible method of moving agents between datasets with differing sets of unique things, which it does so using an agent transfer method that is quite versatile. All patterns that meet a minimal utility threshold are to be discovered via utility-based mining [14]. The classical HUPM evaluates the significance of itemsets solely based on utility assessment. Data structures or tactics used to reduce search space, memory usage, and performance are the only variations between these algorithms in terms of input and output.

Overall, the HUPM algorithms have been well explored in terms of their efficiency and scalability. However, the HUPM approach's quality cannot just be gauged by looking at how long it takes to execute or how much memory it uses. The utility of newly found patterns is directly correlated to the efficiency of the algorithms used to find them. One of the performance metrics used to determine an item's long-term significance is reliability. As a result of this, the trustworthiness of pattern mining has received substantial attention [15]. There have been a slew of algorithms and models created by experts looking into the chances of anything being important by accident.

As a result, a number of statistical probabilities and information retrieval measures have been proposed to evaluate dependability. The dependability of analysing models of regularity, recurrence, coherence, or consistency has been demonstrated via the presentation and discussion of several studies. It is argued in the following sections that using these novel criteria will lead to more consistent results. According to Geng and Hamilton, one of the primary objective metrics of pattern interestingness is dependability, which is based on the likelihood that a pattern will occur. It is also considered dependable if the pattern's interestingness measures occur in a high proportion or routinely appear in a series of occurrences when it occurs frequently.

IS, weighted comparative correctness, Recall, and Jaccard are only few of the generality-based measures used to verify the quality of association rules. In big datasets, Prajapati et al. defined consistent association rules as those that are both locally and globally common. A sequential pattern is considered trustworthy if the user-specified minimum threshold for inter-arrival time between successive items is met. As a result, the SPP-Growth method is presented to uncover more predictable patterns by detecting all stable regular structures in a transaction database with timestamps using a novel periodicity measurement known as Lability. Stability is included into the TSPIN algorithm in order to uncover periodic patterns that have a stable periodic behaviour.

A more practical approach would be to look at how useful the patterns are rather than just the intriguing things they can tell us. Periodic high-utility pattern discovery mining, developed by Fournier-Viger et al., integrates utility measures with regularity evaluations to prevent itemsets with inconsistent periods from having periods that fluctuate greatly inside an algorithm known as PHM. In certain research, a correlation factor such as the lift or additional value is utilised to reflect the reliability of the association rules. As a result, a number of studies have used utility and correlation measures to extract more similar trends from quintessential HUPM methodologies while attempting to avoid completely pointless or non-discriminative patterns that occur by chance.

For example, the following measurements were used: all strength, relationship, bandwidth attachment. There are a number of reasons why all-confidence is becoming increasingly popular as a way to uncover associated patterns. The periodicity of correlated pattern mining is also strengthened, making it more practical in real-world applications. For example, a new measurement known as periodic-all-confidence combines the all-confidence and periodicity measurements to sign the predictive behaviour of consumers' purchases. While frequency and correlation can have a significant impact in a variety of fields, they don't encompass the entire concept of dependability.

Reliability is a measure of how consistent and error-free a product or service is. In other words, the dependable pattern should be: Overall, it can be stated that most recommended pattern mining books supplied appropriate metrics and utilised them to derive more useful patterns. As such, they might be considered as inaccurate or imprecise in terms of defining dependable patterns. It would be ideal if utility-based patterns could be discovered that are both internally trustworthy and dependable over time, as well as being error-free in practise.

3. Proposed Model

Considering a set of transactions that constitutes a transaction database that can be represented as

$$DB = \{TD_1, TD_2, TD_3 \dots TD_m\}. \qquad (1)$$

Individual transactions are identified by their own unique "TD" number. It's important to note that each transaction is composed up of a set of items but every item j has an intrinsic utility and an exterior utility: $r(j, T_d)$. It is undesirable to include sequences that have not shown at least once in successive transactions with scope length P_{max} when determining the size of the window P_{max}. According to the concept of dependability presented, a Pattern Y is termed a reliable utility-based pattern if three requirements are satisfied in Y. First, when high outlier values are excluded, $v(Y)$ is equal to or greater than U_{min}. A second consideration is whether or not Y is internally consistent, or whether or not its degree of correlation meets a user-specified minimum positive correlation criterion. Third, $v(Y)$ is regenerated or has a consistency degree no just under a preset minimal stability criterion provided by the user. Predictive performance is a regularly used methodology for assessing the effectiveness of the categorization rule.

$$Perf_{pred} = \frac{TP}{TP+FP} \qquad (2)$$

where, TP represents true positive and FP represents false positive according to the transaction performance predictor that is represented by $Perf_{pred}$.

The utility evaluation is used as a classification data to simulate the model that successfully predicts the model in unseen data in this study. It is known as "True Positives" when a model is able to construct a pattern with a high utility cost from training data and identify a pattern with a high utility cost inside the test data. False Positives are the result of a system that develops a design with a high utility barrier from either the training dataset, but the detected pattern will not have a large expected utility in the testing dataset. Utility-based Reliability Classical utility hypothesis and the idea of reliability are used to generate patterns that are both dependable and profitable.

Measures of the dependability of high-utility itemsets are presented in the first part. Secondly, effective pruning techniques for seeking dependable high utility items are described, while the last part provides an algorithm that utilises the suggested metrics to locate RUPs. Numerous companies have found success using DHUPM, which employs numerous measurements on previously recorded data in order to get more precise findings and foresee product scenarios in the future. To make this technique innovative, new pruning procedures are proposed, and then new metrics are introduced to check various elements of dependability.

Figure 1 depicts a high-level perspective of the DHUPM approach's architecture, showing the approach's core parts. Six processes in the first section generate probable reliable high utility structures by executing the recommended pruning techniques, while the last three mechanisms show how to use the provided measures to find reliable utility-based itemets. Reliability and utility are combined in the proposed DHUPM technique to uncover RUPs, a form of pattern that can be found. In order to be classified as a RUP, a pattern must have both high utility and acceptable dependability. No one test, however, can accurately assess the dependability of any product.

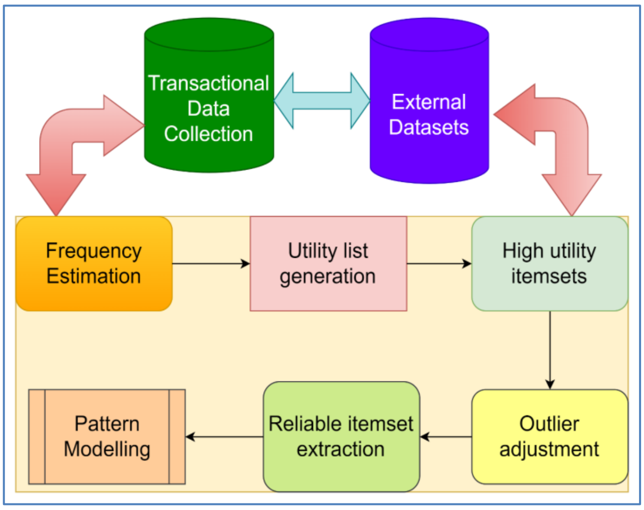

Fig. 2. Proposed System architecture

In order to create RUPs, there are three objective measurements: Utility Sheaths Trimmed Measurement is recommended in order to minimise the influence of outliers. The Internal Consistency Coefficient seeks to evaluate the internal consistency of the data. Introducing the Utility Consistency Coefficient is a way of gauging how utility changes over time. Natural catastrophes, holidays, and other once-in-a-lifetime occurrences were among the causes of the outlier figures. It's difficult to reproduce an outlier's high utility value from a random sample of data. Proposed system architecture is shown in Figure 2.

The Interquartile Range Method (IQR) may be used to find outliers in the statistical literature. Because of this, $v(j, U)$ is regarded an exception if its value is greater than or less than the lower quartile or first quartile. Employing linear interpolation of nearby (non-outlier) data to replace high outlier values, the Snipped Utility of a sequence may be calculated. The Snipped Utility of Y has been represented as $uv(Y) = v(Y) - \partial,$ and ∂ defines the finite different values of the linear model and the outlier neighbouring points. Also, uv is a parameter to identify the less influenced pattern by the presence of deviations relative to the predicted utility value.

Algorithm: Mine and Search - DHUPM

```
input: $DB = \{TD_1, TD_2, TD_3 \dots TD_m\}, U_{min}, I_{min}, P_{max}$
output: dependable utility patterns from itemsets
1. Read $DB$ and retrieve $TTW(j)$ for all the components $j \in J$.
2. Estimate for all $j$ , $TTW(j) \geq U_{min}$ find $P_{max}$ $for\ j$.
3.Arrange in ascending order for the obtained values of $j$.
4. Traverse $DB$ with the item $j$;
5. for each $j$ from $Q \in itemset$ do
        a. $Q$ be the n-th itemset according to value of $j$.
        b. if ($TTW(Q) \geq U_{min} and\ P_{max} < Q_{max}$
            for each transaction
                if ($v(Q, u)$) is an exceptional value then
                        $\alpha + = v(Q, u) - \vartheta$
                        $\vartheta = interpolation\ values$
                endif
            endfor
        c. if ($v(Q) - \alpha \geq U_{min}$) then
            $\alpha += \varepsilon Q_m Q_n$
6. print: dependable itemsets having higher $Q$ values.
```

Internal consistency/dependability across items is the second stage in our suggested strategy to affirm reliability. Internal consistency is characterized as the connectedness among the subitems. However, most of the present assessments fail to extract precise patterns effectively. Statistical studies have shown, however, that the Spearman-Brown prediction equation is better at forecasting dependability. A correlation matrix's average correlation coefficient may be used to assess the set's internal dependability using the prediction in the second step. To see how each item in the itemset connects to the others and to the itemset as a whole, we may use Pearson's correlation analysis to create a covariance matrix.

	x	y	z	xyz
x	1			
y	0.37	1		
z	0.26	-0.33	1	
xyz	0.41	0.21	0.88	1

Fig. 3. A sample itemset {x,y,z} and its corresponding correlation matrix

A less dangerous approach is to base choices on a regular pattern instead of an uncommon one. The reason is because a regular intervals design will be more repeatable dependable over time since its periods will not fluctuate considerably. Therefore, the non-periodic elements and everything their isometric exercises can be eliminated. Through using utility-list structure shown in Figure 1, the following algorithms will keep track of the usefulness of each item and then use the DHUPM technique to find RUPs. A sample itemset {x,y,z} and its corresponding correlation matrix has been depicted in Figure 3.

Core strategies for identifying reliable high-utility item sets are described in Algorithm while reliability measurements are used as recommended in Algorithm to identify prospective RHUIs. Suggested trimming algorithms are used to identify the most likely high-utility item sets without exploring the whole search field. The TTW and frequency of items are acquired during the first search of the database. Pruning involves removing the less promising elements and arranging the promising ones in TTW increasing order. Although the database's second action generates 1-itemset utility lists, the utility lists for k-itemset are obtained through a process known as utility lists overlaying.

The pseudo-code of the key DHUPM algorithms, whose input is the possible dependable high utility itemsets, is provided to understand our technique. As a preliminary step, select the datasets with Trimmed Utility less than U_{min} by subtracting the outlier value from the linear interpolation of nearby nonoutlier values. The second phase calculates coefficient values for all subgroups of candidate itemsets to recognise the itemsets possessing an Internal Consistency Coefficient no less than I_{min}. The itemsets with a Utility Consistency Constant no less than U_{min} are later found in the third set, utilising Average Window Value and Global Mean Utility. Eventually, the method computes the RUPs.

Moreover, the suggested method needed the consumers to select four criteria, which was unclear. Therefore, by assessing the experimental data comparable with the statistical independence, the least reasonable range of I_{min} is 0.58. While in the majority of situations, U_{min} values greater than 0.502 are associated with higher levels of predictive precision. Furthermore, parameters settings U_{min} and maxPer were set as per the unique scenario of each dataset. If U_{min} is set to 2000, for example, this is regarded too low and will result in a significant number of itemsets in huge datasets. Still, it is regarded excessively high in other datasets. As a result, U_{min} is substituted overall duration of this project. Almost the similar approach, P_{max} is specified as a proportion of the overall population of communications. Generally, U_{min} and P_{max} are determined by the user which based on demands of the operations.

4. Results and Discussion

For the purpose of evaluating the efficiency of three techniques representing various utility mining methodologies, this study utilised four datasets. The SPMF Repository provides all datasets. A Windows 10 PC with a 64-bit Intel Core i7-6820HQ CPU @ 2.70GHz and 16 GB of RAM is used to perform all of the algorithms. HUIM algorithms are tested by dividing the dataset into two parts: 80% for training as well as 20% for testing.

The training data set is used to identify high-utility item sets, and the findings are used to predict high-utility item sets in a test dataset, as mentioned above. The high-utility itemsets selected the testing dataset were selected to examine the correctness of the algorithms in light of the major predictable item sets. Expected performance measures are used to evaluate the following methodologies: Period HUIM (PHM), Causally linked HUIM (FCHM all-confidence), and the suggested reliable high utility pattern mining (DHUPM).

Setting the parameter values to U_{min} and P_{max} in accordance with the specific circumstances of each dataset should assure the production of a wide variety of candidate itemsets in these tests. The FCHM all-confidence algorithm may be used to deal with the problem of mining associated high utility item sets (CoHUIs) by simply setting the lowest all-confidence threshold as 0.45 for all experiments. The PHM method can also be used to tackle the problem of mining PHUIs (periodic high utility itemsets). For datasets Retail, Mushroom, and Fruithut, the P_{min} and A_{min} values are each set to 0, while the P_{max} and $mean_{max}$ limits are each set to 500,1000,1500. However, the range of probable circumstances shown by the relative U_{min} varies greatly. Sample frequency distribution for data itemset has been listed in Table 1.

Table 1. Sample frequency distribution for data itemset

Symbol	**v(u,T1)**	**v(u,T2)**
1	3.24	4.91
2	4.89	3.89
3	5.23	5.12
4	2.35	6.32
5	6.35	2.34
6	5.49	4.66
7	2.13	5.24
8	1.25	6.15
9	6.32	5.32
10	3.65	4.31

Using a dataset with I_{min} as 0.5 and U_{min} as 0.3, these tests exhibit the prediction accuracy values of an algorithm on a dataset with various values of U_{min}. Analyzing experimental data yields the following conclusions: When it comes to discovering reliable patterns, the suggested DHUPM technique surpasses the correlation and regular HUIM approaches since it incorporates all components of dependability. Another advantage of the FCHM all-confidence method is that it beats the periodic-based approach in datasets with fewer transactions, such as Ecommerce and Mushroom. There is a rationale for this, and it has to do with the interrelatedness of the sub-items in minor transactions.

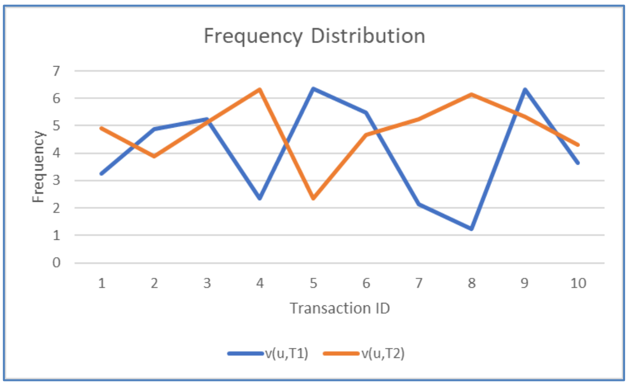

Fig. 4. Frequency distribution for itemset

As a result, correlation, internal consistency, and symmetric measurement should all be considered when assessing dependability. Another benefit of using datasets with a high number of transactions is that the PHM method appears to be more reliable than the correlation-based technique. In order to assess a pattern's long-term profitability, it is necessary to take into account the pattern's stability, consistency over time, or symmetrical measure. These four sets of datasets each provide a count of the number of acceptable high utility itemsets, abbreviated as HUI, PHUI, CoHUI, and RHUI, that were reported. Firstly, there are two points to note: There are a few reasons why the number of RHUIs may be fewer than the number of HUIs or CoHUIs. Table 2 lists the memory usage in Megabytes for different datasets.

Table 2. Memory usage analysis

Datasets	Memory usage analysis		
	PHM	**FCHM**	**DHUM (Proposed)**
Mushroom	241	325	215
Ecommerce	324	389	300
Retail	487	598	625

Overall U_{min} value as 0.04 percent yields 7833 HUIs on Fruithut. Only 55 of the 64 itemsets created by the DHUPM algorithm, which have a high possibility of reproducibility, were in fact replicated with a high utility value in the testing set. It's possible that the non-reproducibility of the final HUIs, which had 4516 itemsets, would mislead dictionary writers. The PHM algorithm, on the other hand, detects 901 non-repeatable patterns, comprising 66 itemsets, in the testing set. The FCHM all-confidence approach, on the other hand, found 286, of which 55 itemsets were not replicable in the testing set. It also examines the effect of a change in the U_{min} on the dependability proportion of HUIM patterns generated by the algorithm. In Figure 5, various datasets were applied with different algorithms and the memory utilization has been compared.

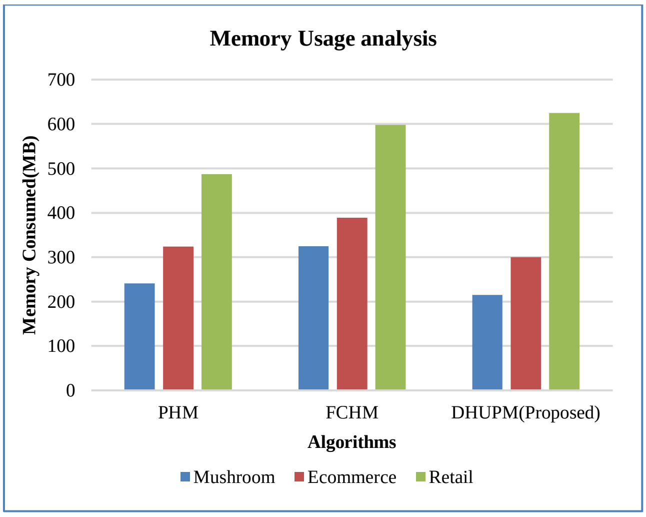

Fig.5 Memory utilization Comparison

According to our findings, reducing U_{min} decreases the dependability ratio by expanding the search field. Real-world datasets contain numerous unreliable patterns that the DHUPM algorithm can successfully filter out. There are too many possible 1-itemsets to be able to exclude those that have anomalous, non-periodic, or significantly declining behaviour. As well as being superior in terms of forecast accuracy, it offers a major competitive edge over the competition. The recommended pruning algorithms for each dataset were also evaluated in terms of memory usage. The Java API was used to do the memory measurements. Results suggest that DHUPM may effectively minimise memory use, particularly in retail and e-commerce environments. FCHM all-confidence, on the other hand, Table 3 displays utility value distribution sample for a sample itemset {x,y,z}.

Table 3. Utility value distribution for itemset {x,y,z}

ID	**v(x)**	**v(y)**	**v(z)**	**v({x,y,z})**
1	-0.09825	-0.98054	0.211661	2.112413
2	-0.98427	-0.68047	2.120449	1.465969
3	-0.869	-0.91807	1.87213	1.977836
4	0.711473	0.036806	-1.53276	-0.07929
5	0.066765	0.718465	-0.14383	-1.54782
6	-0.71259	-0.99863	1.535165	2.151384
7	0.847678	-0.86401	-1.82619	1.861376
8	0.948985	-0.13279	-2.04444	0.286079
9	0.036806	-0.82101	-0.07929	1.768744
10	-0.48679	-0.92013	1.048704	1.982269

There are just a few activities in Mushroom, yet each transaction has a high average quantity of things. A single item with a high utility is immediately produced since the measurement equals 1. The PHM technique outperforms the competition for datasets like Fruithut, which have a high amount of transactions. Utility distribution analysis is shown in Figure 6.

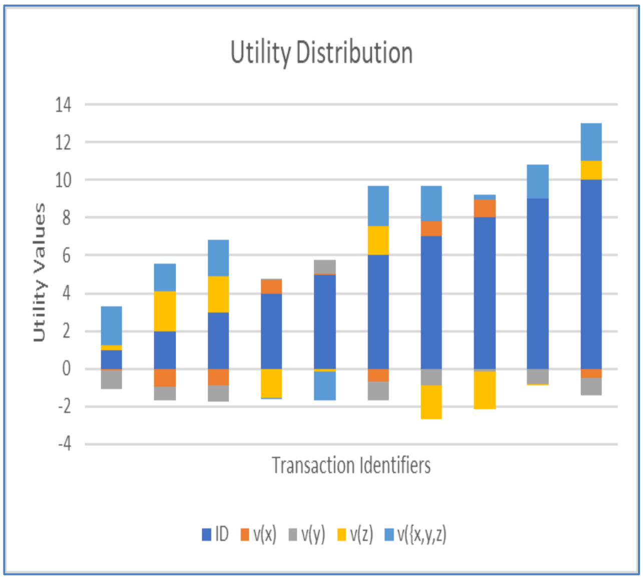

Fig. 6. Utility distribution analysis

P_{max} thresholds are used to remove a significant number of non-periodic entries. While DHUPM is best for accuracy, it may also save memory in datasets with high numbers of items, like Retail and Ecommerce. As a result of the DHUPM pruning algorithms excluding items with little potential for replication, this occurs. In addition, it tries to calculate the largest number of nodes possible to avoid losing fulfilled patterns. To begin, the user's choice of U_{min} threshold affects the reliability value. For example, in the Ecommerce dataset, the classic HUIM approach has a dependability percentage of 0.43 percent when U_{min} is 0.01 percent and P_{max} is 3000. As a result, relying on resulting patterns in making decisions is extremely challenging. Although the dependability percentage will be 90% in this situation, utilising the DHUPM technique with the parameters. I_{min} as 0.7 and U_{min}as 0. Execution time analysis results are tabulated in Table 4.

Table 4. Execution time analysis

Datasets	Execution Time (ms)		
	PHM	**FCHM**	**DHUM (Proposed)**
Mushroom	28.3652	45.25412	26.354
Ecommerce	34.5243	49.35478	32.6547
Retail	30.2547	57.3254	29.3654

A second series of experiments has shown how the DHUPM pruning algorithms may eliminate the search area including unpromising patterns and conserve approximately to 98% of the search space for datasets like Retail, E-commerce, Mushroom and Fruithut respectively. Finally, testing results reveal that the DHUPM algorithm's resultant patterns have higher average dependability proportions in the datasets Retail, Ecommerce, Mushroom, and Fruithut than competitive algorithms by up to 19.45%, 32.64%, 45.02%, and 14.66%, respectively. Experimental results show that the suggested technique yields more reliable patterns than utilising comparison techniques in mining, and the provided measures give adequate criteria for estimating utility-based patterns' reliability. Figure 7 depicts execution time comparison for various approaches with different datasets.

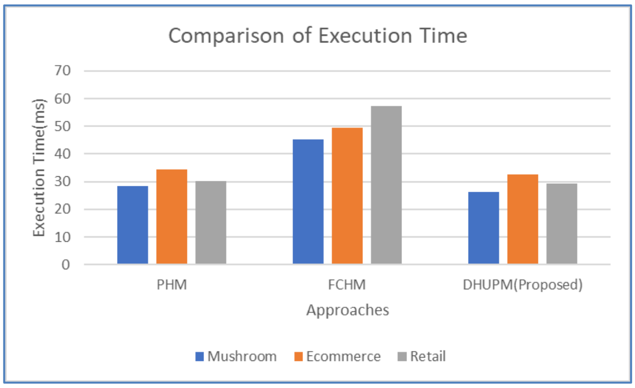

Fig. 7. Execution time comparison

5. Conclusion

As a new field of study, the discovery of patterns with high usefulness has emerged. However, the great majority of the patterns that have been established are not dependable in any way shape or form. There are substantial restrictions to the mining process for dependable patterns with high utility values. The reliability domain of the existing pattern mining literature is quite inaccurate since it includes error-free or bias-free, internal reliability, and time-dependent reliability. DHUPM, or Dependable High Utility Pattern Mining, is a new technique to mining dependable patterns having high utility values that applies the reliability principle to a major new pattern type: dependable high utility patterns. As a result of this innovative technique, the search area may be drastically reduced and a set of three metrics is provided to avoid the disadvantages of utility-based pattern mining approaches. The RUPM algorithm has been tested on a variety of datasets to establish its usefulness and efficiency. It has been shown that the patterns revealed using utility-based patterns data mining methods are typically incorrect, especially when working with large datasets. It appears that the suggested measures give enough criteria for estimating the trustworthiness of utility-based patterns after extensive trials on sparse and dense, synthetic and actual data. As a result, by altering the minimal internal consistency and minimum utility consistency criteria, decision-makers can obtain exceptionally dependable high utility item sets. The findings of the studies show that the RUPM algorithm is up to 41% and 47% more reproducible than the scheduled high utility itemset mining as well as the associated high utility itemset mining algorithms. To remove boring patterns, you may also utilise the given pruning techniques, however exploratory datasets will give you at least 98% of the search agent.

References

[1] Ahmed, J. C.-W. Lin, G. Srivastava, R. Yasin, and Y. Djenouri, ‘‘An evolutionary model to mine high expected utility patterns from uncertain databases,’’ IEEE Trans. Emerg. Topics Comput. Intell., vol. 5, no. 1, pp. 19–28, Feb. 2021, doi: 10.1109/TETCI.2020.3000224.

[2] Deepak Mathur, N. K. V. . (2022). Analysis &amp; Prediction of Road Accident Data for NH-19/44. International Journal on Recent Technologies in Mechanical and Electrical Engineering, 9(2), 13–33. https://doi.org/10.17762/ijrmee.v9i2.366

[3] Djenouri, J. C.-W. Lin, K. Nłrvåg, H. Ramampiaro, and P. S. Yu, ‘‘Exploring decomposition for solving pattern mining problems,’’ ACM Trans. Manage. Inf. Syst., vol. 12, no. 2, pp. 1–36, Jun. 2021, doi: 10.1145/3439771.

[4] Fournier-viger, Y. Wu, D. Dinh, W. Song, and J. C. Lin, ‘‘Discovering periodic high utility itemsets in a discrete sequence,’’ in Periodic Pattern Mining. Singapore: Springer, 2021, pp. 133–151.

[5] Fournier-viger, Y. Wang, P. Yang, J. C. Lin, U. Yun, and R. Uday, ‘‘TSPIN: Mining top-k stable periodic patterns,’’ Appl. Intell., vol. 104, pp. 1–22, Feb. 2021.

[6] Alaria, S. K., A. . Raj, V. Sharma, and V. Kumar. “Simulation and Analysis of Hand Gesture Recognition for Indian Sign Language Using CNN”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 4, Apr. 2022, pp. 10-14, doi:10.17762/ijritcc.v10i4.5556.

[7] Gan, J. C.-W. Lin, P. Fournier-Viger, H.-C. Chao, and

P. S. Yu, ‘‘Beyond frequency: Utility mining with varied item-specific minimum utility,’’ ACM Trans. Internet Technol., vol. 21, no. 1, pp. 1–32, Feb. 2021.

[8] Geng and H. J. Hamilton, ‘‘Interestingness measures for data mining: A survey,’’ ACM Comput. Surv., vol. 38, no. 3, p. 3, 2006.

[9] Krishnamoorthy, ‘‘Pruning strategies for mining high utility itemsets,’’ Expert Syst. Appl., vol. 42, no. 5, pp. 2371–2381, 2015, doi: 10.1016/j.eswa.2014.11.001.

[10] Liu and J. Qu, ‘‘Mining high utility itemsets without candidate generation,’’ in Proc. 21st ACM Int. Conf. Inf. Knowl. Manag., pp. 55–64, 2012..

[11] Tume-Bruce, B. A. A. ., A. . Delgado, and E. L. . Huamaní. “Implementation of a Web System for the Improvement in Sales and in the Application of Digital Marketing in the Company Selcom”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 5, May 2022, pp. 48-59, doi:10.17762/ijritcc.v10i5.5553.

[12] Prajapati, S. Garg, and N. C. Chauhan, ‘‘Interesting association rule mining with consistent and inconsistent rule detection from big sales data in distributed environment,’’ Future Comput. Inform. J., vol. 2, no. 1, pp. 19–30, Jun. 2017.

[13] Shyur, C. Jou, and K. Chang, ‘‘A data mining approach to discovering reliable sequential patterns,’’ J. Syst. Softw., vol. 86, no. 8, pp. 2196–2203, Aug. 2013.

[14] Vet, L. B. Mokkink, D. G. Mosmuller, and C. B. Terwee, ‘‘Spearman–Brown prophecy formula and Cronbach’s alpha: Different faces of reliability and opportunities for new applications,’’ J. Clin. Epidemiol., vol. 85, pp. 45–49, May 2017.

[15] Kose, O., & Oktay, T. (2022). Hexarotor Yaw Flight Control with SPSA, PID Algorithm and Morphing. International Journal of Intelligent Systems and Applications in Engineering, 10(2), 216–221. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/1879

[16] Wei, B. Wang, Y. Zhang, K. Hu, Y. Yao, and H. Liu, ‘‘FCHUIM: Efficient frequent and closed high-utility itemsets mining,’’ IEEE Access, vol. 8, pp. 109928–109939, 2020, doi: 10.1109/ACCESS.2020.3001975.

[17] Zhang, W. Fang, J. Sun, and Q. Wang, ‘‘Improved genetic algorithm for high-utility itemset mining,’’ IEEE Access, vol. 7, pp. 176799–176813, 2019.

[18] Zhang, M. Han, R. Sun, S. Du, and M. Shen, ‘‘A survey of key technologies for high utility patterns mining,’’ IEEE Access, vol. 8, pp. 55798–55814, 2020, doi: 10.1109/ACCESS.2020.2981962.

[19] Gupta, D. J. . (2022). A Study on Various Cloud Computing Technologies, Implementation Process, Categories and Application Use in Organisation. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(1), 09–12. https://doi.org/10.17762/ijfrcsce.v8i1.2064

[20] Zida, P. Fournier-Viger, J. C.-W. Lin, C.-W. Wu, and V. S. Tseng, ‘‘EFIM: A fast and memory efficient algorithm for high-utility itemset mining,’’ Knowl. Inf. Syst., vol. 51, no. 2, pp. 595–625, 2017.