A Proposed Method to Select Potential Item Set for High Utility Item Set Mining using Genetic Algorithm Techniques

Pradeep Sharma
Dept. of Computer Science & Engineering, MIT Ujjain

Ruchika Pachori
Dept. of Information Technology, MIT Ujjain

RajLaxmi Garg
Dept .of Information Technology, MIT Ujjain

ABSTRACT

Utility mining is a technique to prune high utility itemset from the given transactional database on the basis of user-defined minimum utility threshold. Frequent itemset mining, only focus on itemset appear most frequently in the database while in utility mining we concern about utility i.e. importance or profit of itemset according to the user preference. In this paper we are proposing a two-phase algorithm, in the first phase, we are using weighted transaction utility concept to calculate and compare the utility of itemset with minimum utility threshold and then in the second phase, we are proposing genetic algorithm technique to search high utility itemset from the recognized transactional database obtain after the first phase..

Keywords

Data Mining, Weighted Transaction Utility, Utility Mining, Genetic Algorithm.

1. INTRODUCTION

Data mining technique is used to discover hidden pattern from data already stored in a large database. Data mining is a combined technique of database, statistics, Artificial Intelligence and machine learning. Data mining helps users to identify the purchase items and their consumers. Market basket analysis has also been used in data mining techniques for items & consumers. Frequent itemset mining, which is one of the efficient techniques of data mining, identify items appear most frequently in the database but it's not considered that how much items are profitable or important for the user. In utility mining profitability and interest of user related with items taken into consideration. There are so many techniques of utility mining have been purposed for pure high utility itemset means itemset which is more profitable than others with some selection criteria. A genetic algorithm is also an efficient technique of soft computing which works on the concept of the genetic process with some steps of the process like mutation, crossover, selection etc. D Charles Darwin's "The Origin of Species" publication in 1859 brought about genetic algorithm detailing how complex, problem-solving organisms could be created and improved through an evolutionary process of random trials, sexual reproduction, and selection [1]. GAs are used to construct a version of biological evolution on computers.GA have been successfully adopted in a wide range of optimization problems such as control, design, scheduling, robotics, signal processing, game playing and combinatorial optimization [1]. We can use the concept of data mining as an application area of genetic algorithm.

The main contributions of this paper are summarized as follows.

A new method called MHUI_TWU-GA is proposed for search high utility itemset with TWU concept and using genetic algorithm approach. In this proposed method in the first step, we calculate weighted transaction utility of each itemset and then compare it with minimum utility threshold. In the second step, we apply genetic algorithm approach to pruning and generate high utility itemset.
The rest of this paper is organized as follows. Section II describes the basic concepts and definitions of utility mining and genetic algorithm. Section III presents the related works. High utility itemset using genetic algorithm concepts describes in section IV. The proposed approaches are discussed in Section V, Conclusions are finally given in Section V.

2. BASIC CONCEPTS AND DEFINITIONS

2.1 Utility Mining

Utility mining concepts and definitions given in UP-Growth[2] called utility pattern growth are sufficient to study and understand concept of utility mining as well as all essential definition related with utility mining as follows:

Definition1:A frequent itemset is a set of items that appears at least in a pre-specified number of transactions. Formally, let I={i1,i2,. ..,i_m}be a set of items and DB={T1,T2,..., Tn} a set of transactions where every transaction is subset of items(i.e. itemset).

Definition2.The utility of an item is a numerical value defined by the user. It is transaction in dependent and reflects importance(usually profit) of the item. External utilities are stored in an utility table.

Table: 1

TID	Transaction	TU
1	(A,2)(C,2)(D,2)	16
2	(A,2) (C,6) (E,2) (G,5)	27
3	(A,1) (B,2) (C,1) (D,6) (E,1) (F,1)	30
4	(B,4) (C,3) (D,3) (E,1)	20
5	(B,2) (C,2) (E,1) (G,3)	11

Definition3: The utility of an itemset X in a transaction T is denoted by U(X,Ti)&it is calculated as follows. For example ({AC}, T1)=U({A}, T1) +U({C},T1) = 10+ 2 = 12.

Definition 4: The utility of an itemset X in D is denoted by U(X) & it is calculated as follows For example, U({AD})= U({AD}, T1)+ U({AD}, T3) = 14+ 17= 31.

Definition5. An itemset called high utility itemset if its transactional weight utility is higher than minimum utility threshold otherwise it is called item with low utility value.

Table: 2

Item	A	B	C	D	E	F	G
Profit	5	2	1	2	5	2	1

Definition 6.The transaction utility of a transaction T is sum of utility(internal utility × external utility) of each item present in transaction T, denoted as TU(Td) and defined as u(X, Td). For example, TU(T1)= u({ACD},T1) = 8.

Apart of all above definitions we have already purposed a new concept to discover high utility itemset mining [3] in which we introduce a new term Weighted Transaction Utility (WTU) for calculating utility of each items from Transaction Database and profit table .for example WTU of item A is 20 because A is present in three transaction and its profit value is 5.in this paper we are using concept of WTU to find out potential itemset comparing WTU with minimum utility as WTU≥min_uty.

2.2 Genetic Algorithm

Genetic algorithm is a heuristic search technique used to generate or optimization useful solutions for the given problem. Genetic algorithm process under basic concept of natural selection start with initial population , calculate the fitness of chromosomes in the initial population and repeat this process for new generated offspring. After this evolution technique performs including selection, crossover, and mutation. In selection process of picking effective chromosomes from the population performed. Crossover is also known as recombination, process to taking to individual

from the population and generate new individual. Fitness value of each individual calculates for survival of the fitness. Mutation is a technique for selecting new chromosome from two or more valuable individual from initial population. In table 3 represent terminologies used in Genetic Algorithm.

Table 3: Terminologies used in Genetic Algorithm

Biological Term	Genetic Algorithm Term
Chromosome or Genotype	Coded design Vector
Gene	Every Bit
Population	A Number of Coded design Vector
Generation	Population of design vectors which are obtained after one computation
Locus	A particular position on the string.
Phenotype	Parameter Set
Fitness function	It is a measure associated with the collective objective functions that indicate thefitness of a particular chromosome.
Chromosome or Genotype	Coded design Vector
Gene	Every Bit
Survival of the fittest	The fittest individuals are preserved and reproduce, which is referred to as survivalof the fittest
Selection	The process of picking effective chromosomes from the population for a later
Crossover	breeding is called as selection.
Mutation	The process of creating a new chromosome by mating two or more valuable

3. RELATEDWORK

Many researchers have published their research paper or study in the field of utility mining. Basic of utility mining is association rule mining and frequent item set mining. One of the well-known algorithm is Apriori algorithm [4],which is the fundamental for association rule mining to select items which are related with each other in term of x y. Then frequent pattern growth is proposed for item sets occur frequently in the transactional database based on their support value higher than minimum support count. A tree base concept to identify potential itemset from the first phase FP-Growth [5] was afterward proposed. After comparison it has been evaluated that FP-Growth provide better result than Apriori-based approaches because it scans database twice without generating candidate itemsets.

But in the frequent item set mining [4,5], the importance of item to user is not taken into consideration that is the unit profit related with items and purchased quantities not consider. Thus, some new algorithms or research study purposed for mining high utility itemset from the databases, such as UMining[6],Two-Phase[7], IIDS[8] and IHUP[2].UMining algorithm[9] proposed by Yao et al. Each method consider space and time to improve efficiency for purne high utility itemset.Two-Phase algorithm[7] proposed by Liu et al. consists of two phases. In phase I, breath first search technique is used to generate high utility item sets .It generate candidate itemset compare with minimum utility threshold for length first and then length second and so on. after that it compare it with TWDC property .In each pass, to generate candidate item sets, each item or item set compare with its TWU value for length one to n-1 length which is very time and space consuming process because in each pass we have to calculate potential itemset and store it.To overcome this problem, Li et al.[8] proposed an isolated items discarding strategy, abbreviated as IIDS, to reduce the number of candidates. By pruning isolated itemsusing depth wise search then number of potential item sets reduces significantly. This method is better than previous methods but this methods still perform multiple scan over transactional database which is still time and space consuming .To avoid scanning database multiple times, Ahmed et al.[9] proposed a tree-based algorithm, called IHUP, formining high utility item sets. They use an IHUP-Tree to maintain the information of high utility item sets and transactions. Every node in IHUP-Tree consists of an item name, as support count, and a TWU value. In this algorithm three steps are used first is construction of IHUP-Tree; second is the generation of high weighted transaction utility potential candidate itemsets and third identification of high utility itemsets In step of IHUP, first candidate are arranged again in lexicographic order, support descending order or TWU descending order, Then, the rearranged transactions are inserted into the IHUP-Tree. In step2,high transactional weighted utility itemsets generated from IHUP tree applies for the FP-Growth[5] for final processing.Note that IHUP and Two-Phase produce the same number of HTWUIs in phaseI since they use transaction-weighted utilizationmining model [7]. FP-growth algorithm also play significant role with two novel steps and FP-Tree and then an effective algorithm called UP-Growth[2](Utility Pattern Growth)with four steps,two steps of FP-Growth and two new steps with UP-Tree for effectively purne high utility itemset from transactional database.UP-Growth is a novel algorithm and perform better than all previous algorithm in term of time and space.A new concept to discover high utility itemset mining[3] in which we introduce a new term Weighted Transaction Utility (WTU) for calculating utility of each items from Transaction Database and profit table.

4. HIGH UTILITY ITEMSETS USING GENETIC ALGORITHM CONCEPTS

4.1 Encoding:

Let I= $\{i_{1,},i_{2,},\ldots,i_{n,}\}$is a set of items,

D=$\{T_{1,},T_{2,},\ldots,T_{n,}\}$ be a transaction database where each transaction is a subset of I. An itemset X is a high utility itemset if it satisfies the minUtil threshold, i.e. minUtil is a threshold which is defined by the user.

This section of high utility itemsets is based on genetic algorithm used in the proposed works. Encoding: Different types of encoding techniques used in genetic algorithm like binary encoding ,hexadecimal encoding, octal encoding , real number encoding, integer or literal permutation encoding and tree encoding etc .Here in our problem we are using binary encoding technique to encode the solution of our problem into chromosomes. In this coding technique 1 represent the presence of item in transactional database and 0 represent absence of item. Chromosome length is equal to the number of distinct items of transactional database and it is fixed. Example:

The representation of a chromosome is shown in fig.1

1 0 1 0 1 1 0 1

Fig.1. Chromosome representation for the itemset {1, 3, 5, 6,8}

4.2 Population Initialization:

If we let N is population size and M is a binary string space dimension than at generation time t a list of binary. The algorithm for population initialization is given in Figure 2. string is denoted by

B_t

4.3 Fitness function:

The main goal this work is to generate the high utility itemsets from the transaction database. Hence, the high utility with minimum threshold value using GA, we use Yao et al.'s [14] utility measure u(X) as the fitness function. Fitness function is essential for determining the chromosome (itemset) which satisfy minUtil threshold.

4.4 Genetic Operators

Mainly there are three genetic operators, selection, crossover and mutation in generic algorithm.

1. Selection: Different types of selection methods have used in genetic algorithm like Roulette Wheel Selection, Random Selection, Rank Selection, Tournament Selection, Steady State Selection. . In this work, roulette wheel selection [10] is used. After decoding we have to decide how to perform selection that is select individuals from the population to create new individual for next generation and how many new offspring each can create. The selection of individual focus on individual with higher fitness value.

Roulette Wheel Selection:

In roulette wheel selection we provide fitness to possible solutions by fitness function .candidate with less fitness value to be eliminated. the advantage of this method that weaker solution may also survive for the selection process.

```
(1) At Time t=0 compute inilial population $B_{0.}$
(2) If condition not fulfilled Compute initial population
    for i =1 to N.
(3) Select $b_i$ at time t = t+1 from B.
(4) For i= 1 with probability pc perform crossover of
    $b_i$ and $b_i$+1 at time t = t+1;
(5) For i=1 with probability pm eventually mutate $b_i$ at
    t=t+1.
(6) Increase time t for next step.
(7) End
```

Figure 2: Population Initialization

Less fitness value to be eliminated. The advantage of this method is that weaker solution may also survive for the selection process.

2. Crossover: crossover also have different variant like one point crossover, two point crossover, multi-point crossover, and random multipoint crossover. Using crossover technique we produce new individual which is different from parents. Crossover mates chromosomes in the mating pool by pairs and generates candidate offspring by crossing over the mated pairs with probability.

Single -point crossover: In Single -point crossover two parent chromosomes are interchanged at a randomly selected point thus creating two children.

Before Crossover:

1	0	1	1	1	1	0	0	0	1	P_1

0	1	0	1	0	1	1	1	0	P_2

After Crossover:

1	0	1	1	1	1	1	1	0	0	C_1

0	1	0	1	0	1	0	0	0	C_2

Figure 3: Single Point Crossover

Two (Multi) point crossovers: In two (Multi) point crossovers, two crossover points are selected instead of just one crossover point.

Before Crossover:

1	0	1	1	1	1	0	0	0	1	P_1

0	1	0	1	0	1	1	1	0	P_2

After Crossover:

1	0	1	1	1	1	1	1	0	1	C_1

0	1	0	1	0	1	0	0	0	C_2

Figure 4: Two Point Crossover

3. Mutation

In mutation process after section and crossover we take some of the individual for mutation. The most common technique used in mutation is to alter or flip bit from chromosome with some predefine probability.

There are mainly two types of mutation are perform single point mutation and multipoint mutation .mutation is also used to produce new best individual from parents which improve performance significantly.

Before Mutation:

1	0	1	1	1	1	0	0	0	1	C_1

After Mutation: Point of Mutation

1	0	1	1	1	0	0	0	0	C_2

Figure 5. Mutation

4. Evaluation

Evaluation step intends to select the chromosomes for next generation. In this work, elitist selection [10] method is used. This method copies the chromosome with higher fitness value to new population.

5. Termination criteria:

Termination Criteria Conditions of termination criteria are used to decide whether to continue search or stop, which are as follows:

1) Fixed number of generations reached
2) The solution's fitness with highest ranking at a fixed number of generations.
3) Interrupt solution
4) Combinations of the above three steps

MHUI_{TWU}-GA:

Phase I:

1) Scan Transaction Database to compute weighted Transaction Utility (WTU) of item and itemset.
2) Compare WTU with minimum utility threshold and remove unpromising item set from transaction database to get recognized transaction database.
3) Get the number of distinct item from recognized transaction database and set chromosome length (CL).

Phase II:

1) Genrate a chromosome length (CL).
2) Calculate fitness value (fv) for each indivisual if fv≥min_uty then goto step 6 otherwise goto step 4.
3) Check the population size p_size≥N goto step 7 otherwise goto step 4.
4) If termination conditions are fulfilled get output otherwise continue.
5) Select parents using roulette whell selection for next generation.
6) Perform crossover and mutation and again calculate if fv≥min_uty and p_size≥N then goto step 10 otherwise goto step 8.
7) Evaluvate new individual from new and old population for next generation.

5. EXPRIMENT EVALUATION & RESULT:

Example: Let us consider another example of transactional data base T 4 as follows:

Table 4: Transactional Database T4

TID	Items	Item utility value for this transaction
t 1	2 4 1 3 5	3 1 5 10 6
t 2	2 4 1 5	3 1 5 6
t 3	3 2 5	10 3 6
t 4	2 5 4 7	3 6 1 12
t 5	2 5 1 6	3 6 5 9

Table 5: Transaction Database with TU value

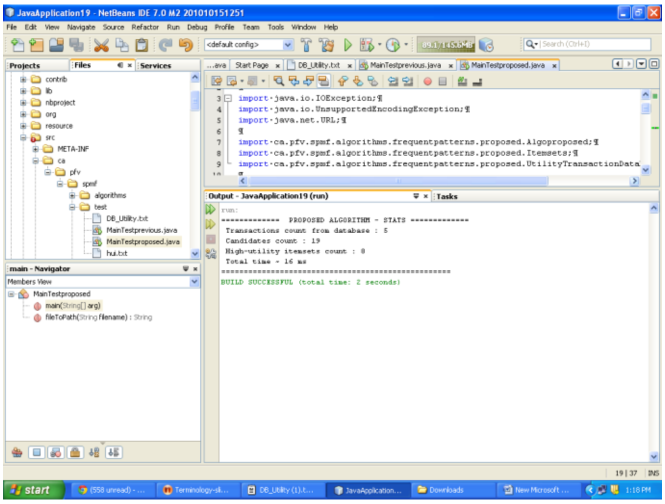
TID	Items	Transaction utility value
t1	2 4 1 3 5	25
t2	2 4 1 5	15
t3	3 2 5	19
t4	2 5 4 7	22
t5	2 5 1 6	20

Table 6: candidate itemset for T4

TID	Items	Weighted transaction utility
t1	5	30
t2	1 5	33
t3	2 5	45
t4	3 5	32
t5	1 2 5	42
t6	2 3 5	38
t7	2 4 5	30
t 8	1 2 4 5	30

6. CONCLUSION

In this paper we used proposed a novel approach of Utility itemset mining using concepts of genetic algorithm. This proposed method would be very effective especially when transaction database contain many distinct items because in each method of utility mining memory requirement and execution time are the main factors for efficient mining.To overcome this problem we proposed a method in which itemset are selected based on Transaction Weighted Utility (TWU) and using Genetic Algorithm (GA) technique named MHUI_{TWU}-GA. Experiment evaluation and result of this proposed method on different transactional database shows that time time taken by this algorithm is less than previous algorithm.



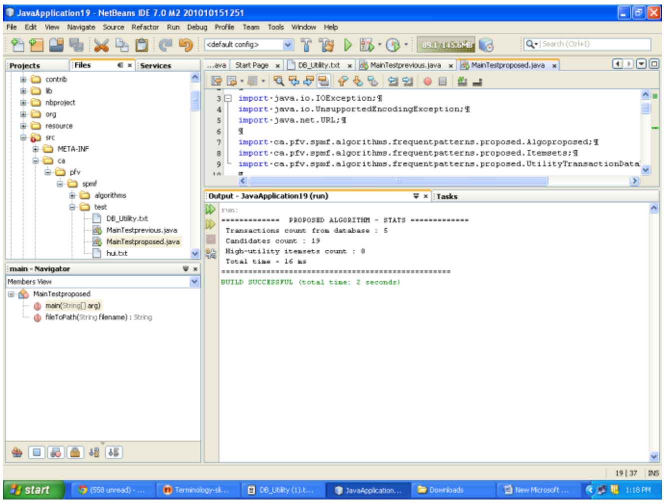
```
import java.io.IOException;
import java.io.UnsupportedEncodingException;
import java.net.URL;

import ca.pfv.spmf.algorithms.frequentpatterns.proposed.Algoproposed;
import ca.pfv.spmf.algorithms.frequentpatterns.proposed.Itemsets;
import ca.pfv.spmf.algorithms.frequentpatterns.proposed.UtilityTransactionData
```

```
run:
=============  PROPOSED ALGORITHM - STATS =============
 Transactions count from database : 5
 Candidates count : 19
 High-utility itemsets count : 0
 Total time ~ 16 ms
===================================================
BUILD SUCCESSFUL (total time: 2 seconds)
```

JavaApplication19 - NetBeans IDE 7.0 M2 201010151251
File Edit View Navigate Source Refactor Run Debug Profile Team Tools Window Help
Projects Files Services
Start Page
DB_Utility.txt
MainTestprevious.java
MainTestproposed.java
Output - JavaApplication19 (run)
Tasks
main - Navigator
Members View
start

Fig.6: candidate itemset previous algorithm



```
import java.io.IOException;
import java.io.UnsupportedEncodingException;
import java.net.URL;

import ca.pfv.spmf.algorithms.frequentpatterns.proposed.Algoproposed;
import ca.pfv.spmf.algorithms.frequentpatterns.proposed.Itemsets;
import ca.pfv.spmf.algorithms.frequentpatterns.proposed.UtilityTransactionData
```

```
run:
=============  PROPOSED ALGORITHM - STATS =============
 Transactions count from database : 5
 Candidates count : 19
 High-utility itemsets count : 0
 Total time ~ 16 ms
===================================================
BUILD SUCCESSFUL (total time: 2 seconds)
```

JavaApplication19 - NetBeans IDE 7.0 M2 201010151251
File Edit View Navigate Source Refactor Run Debug Profile Team Tools Window Help
Projects Files Services
Start Page
DB_Utility.txt
MainTestprevious.java
MainTestproposed.java
Output - JavaApplication19 (run)
Tasks
main - Navigator
Members View
start

Fig.7: candidate itemset proposed algorithm

7. REFERENCES

[1] S. Kannimuthu, Dr. K .Premalatha, Discovery of High Utility Itemsets Using Genetic Algorithm, International Journal of Engineering and Technology (IJET), Vol 5 No 6 Dec 2013-Jan 2014.

[2] VincentS. Tseng, Cheng-Wei Wu, Bai-En Shie, and PhilipS.Yu.UP-Growth: An Efficient Algorithm for High Utility Itemset Mining. InKDD'10, July25–28, 2010, Washington, DC, USA.2010ACM.

[3] Pradeepk. Sharma, Abhishe k Raghuvanshi, An Efficient Methodfor Mining High Utility Data fromaDataSet, in International Journal of Advanced Research in Computer Science and Software Engineering, Volume3, Issue11, November2013.

[4] R.Agrawal and R.Srikant.Fast algorithms for mining association rules.InProc.ofthe20thInt'lConf.onVery Large Data Bases, pp.487-499, 1994.

[5] J.Han,J.Pei,andY.Yin .Mining frequent patterns without candidate generation.InProc.of the ACM-SIGMOD Int'l Conf. on Management of Data, pp.1-12,2000.

[6] H.Yao,H.J.Hamilton,L.Geng, A unified framework for utility-based measures for mining itemsets. In Proc.of ACM SIGKDD 2nd Workshop on Utility-Based Data

Mining, pp.28-37, USA,Aug., 2006.

[7] Y.Liu, W.Liao, and A.Choudhary.A fast high utility itemsets mining algorithm. InProc. ofthe Utility-Based Data Mining Workshop,2005.

[8] Y.-C.Li,J.-S.Yeh,andC.-C.Chang.isolated items discarding strategy for discovering high utility itemsets, In Data &Knowledge Engineering, Vol. 64,Issue1, pp.198-217, Jan., 2008.

[9] C.F.Ahmed,S.K.Tanbeer,B.-S.Jeong,andY.-K.Lee.Efficient tree structures for high utility pattern mining in incremental databases.In IEEE Transactionson Knowledge and Data Engineering,Vol.21, Issue12, pp.1708-1721, 2009.

[10] Yu-Chiang Li, Jieh-Shan Yeh and Chin-Chen Chang, "Isolated items discarding strategy for discovering high utility itemsets", Data and Knowledge Engineering, Elsevier Journal, Vol. 64, pp. 198-217, 2008.[10] Yu-Chiang Li, Jieh-Shan Yeh and Chin-Chen Chang, "Isolated items discarding strategy for discovering high utility itemsets", Data and Knowledge Engineering, Elsevier Journal, Vol. 64, pp. 198-217, 2008.