Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 7, July 2014, pg.350 – 357

RESEARCH ARTICLE

EFFICIENT ALGORITHMS SYSTOLIC TREE WITH ABC BASED PATTERN MINING ALGORITHM FOR HIGH UTILITY ITEMSETS FROM TRANSACTIONAL DATABASES

Mr. S. Vignesh[1], Mrs. K. Sujana Banu[2], Mr. K. Manoj Kumar[3]

[1]Assistant Professor, Department of Information Technology, Sri Eshwar College of Engineering, Kinathukadavu, Coimbatore-641202, TamilNadu, India, vigneshbtech76@gmail.com

[2]Assistant Professor, Department of Information Technology, Sri Eshwar College of Engineering, Kinathukadavu, Coimbatore-641202, TamilNadu, India, sujanakadhar2007@gmail.com

[3]Senior Software Engineer, Angler Technologies Pvt.Ltd, TamilNadu, India, ureachmano@gmail.com

ABSTRACT: In transactional database mining high and efficient type of utility itemset plays major role to analysis the properties of the profits in the transaction. Several number of the algorithm have been proposed in earlier work to analysis the results of mined database in the larger transactional database ,in recent work they occurs the problem of the generation of candidate itemsets for high utility itemsets in database. It degrades the performance of the system in terms of the execution time and memory space occupancy in database. In order to solve the problem of the mining high utility itemset in the transaction database and the educational database, in this paper presents an novel software based tree algorithm such as systolic tree algorithm, it becomes faster than the frequent pattern and utility pattern algorithm for high dataset .The proposed algorithm calculate the weight values to each and every generated rules in the association rule mining. The weight values of each and every item in the database are automatically calculated based on the automatic weight estimation methods it becomes complex in order to overcome these problem in this work use an artificial bee colony based optimization algorithm to derive weight values of the each items s it is assign the weight values based on the count and sequence values of the item in the transactional database and the medical database. The performance of the proposed systolic tree algorithm for high utility itemset mined results is compared with the earlier methods such as UP-Growth and UPGrowth+ methods in terms of the parameters like time, memory space, runtime for each and every number of transaction and educational dataset.

Index Terms: Candidate pruning, Systolic tree, frequent itemset, high utility itemset, utility mining, data mining, artificial bee colony algorithm (ABC), swarm intelligence methods

1. INTRODUCTION

Data mining is the process of mining illuminating non-trivial, formerly extraordinary and potentially valuable information beginning large databases. Mining useful and valuable patterns from larger database plays an imperative role in the several number of the datamining task such as pattern mining of frequent itemset in the, larger transactional database ,weighted frequent pattern mining and utility mining itemset for pattern mining , frequent itemset generation of the candidate values based on the association rule mining methods like apriori [1], [2], it has major issues since it takes long time to execution of the multiple database scans and creates more number of candidate itemsets for larger dataset ,it is overcome by creation of the FP growth [3-6] algorithm by reducing the number of steps by creation of prefix-tree without taking into the account of candidate itemset results ,but it has two major issues such as it considers all the items as same by assignment of the weight values equally and, each item in the transaction database appears in the binary (0/1) form, i.e., either current or lacking.

The major objective of the frequent itemset mining is to discover the frequently occurred itemset results in the transaction database and educational database without consideration of the weight values to each itemset in transactional database. Though, capacity and mass are important for deal with real globe assessment harms that need make best use of the effectiveness in an association. In earlier work Yao et al [7-8] developed a novel a structure designed for mining high utility itemset. It enhances the results of high utility itemset mining results. Two Phase algorithm also proposed by Liu et al. [9] to extract high utility itemsets, with a transaction weighted utility (TWU) assess to reduce the exploration break. Their algorithm is appropriate designed for thin data sets through short patterns.

Mining high utility itemsets beginning superior databases is referred as discovering the frequent itemsets through elevated profits. At this point, the significance of itemset effectiveness is interestingness, significance or effectiveness of an item to user. It mines high utility of itemset from larger database and it is used in larger area of the applications such as website relate to stream investigation [10], commerce encouragement in procession hypermarkets, cross-marketing in put on the market provisions [11], online e-commerce administration, mobile business location development [11] and smooth judgment significant prototype in biomedical purpose [5]. Occurrence is not adequate to respond question, such as whether an itemset is extremely cost-effective, or whether an itemset has a well-built force. In order to solve these problems propose an utility mining algorithm to solve the problem of mining high utility itemset in the larger database, the major objective of this paper is to mine high utility itemset in the transactional database based on the use of the systolic tree algorithm, it becomes faster than the existing utility itemset results since it takes the weight values for each and every items in the dataset based on the count and occurrence values of the data.

In this work the weight values of the items is generated using Artificial bee colony optimization algorithm (ABC) ,If the counts value of the items is high is considered as most important items for pattern mining with high utility and less number of scans only required for each itemset in the transactional database ,the running time of the STA takes less time when compare to existing tree algorithm .The performance accuracy of the proposed system is high with less number of memory space occupied and more faster than the earlier methods for educational dataset and other types of the larger dataset also, in the following section discuss about the detail of the existing pattern mining algorithm for itemset mining and drawbacks of the algorithm ,section 3 discuss detail about the proposed STA and major contribution of the work ,section 4 analysis the results of the proposed and existing methods with frequent itemset mining results and finally concludes the work in section 5 and their remarks or issues also mentioned in the section 5.

2. BACKGROUND STUDY

In recent years several number of the works such as [1], [2] propose an association rule mining apriori algorithm to solve the frequent itemset problem in the candidate itemset generation phase for transactional database, but it doesn’t solve the problem to reduce the number of scan for each itemset mining in the transactional database scans. Since for each and every scan it requires to keep existing itemset as same for each generation of the candidate if the frequent itemset becomes larger it becomes hard to scan the larger database frequent itemset and so on.

This above mentioned problem is solved by using frequent pattern growth algorithm in [3] without consideration of the candidate generation for each and every scanning of the process in the larger transactional database, since for entire database it need or wants only two scans, in order to reduce the complexity of the FP growth algorithm the FP-array [4] system was anticipated to decrease the FP-tree traversals and it professionally workings particularly in sparse datasets. In order to measure the results of the proposed FP tree based on the measure h-confidence [13] was to categorize well-built maintain similarity frequent patterns. A number of additional investigate [14], [15], has been completed intended for frequent pattern mining.

The number of the mined itemset can be shared between one user to another user in the transaction database based on the itemset sharing methods in earlier years. It consists of information about the frequency level of the items or count value of the each item in the transaction database .The share value of the each item in the transaction database is defined as the percentage value for each item in the itemset or mined pattern results. In this work the author uses a support measures to demonstrate and provides the information of the each item in the itemset that relates to each transaction database, they use an heuristic based optimization methods to discover the number of the frequent itemset along with distribute values which is greater than the specified threshold value for each item in the database or dataset

Top most k number of the frequent itemset patterns with high utility patterns also mined in the earlier work [12] for different types of the transaction patterns in the educational and transactional database, medical database with different levels of the patients with the same disease affected by each patients in the dataset will have different levels of effectiveness. The existing work cannot use the downward closure property and it performs pruning methods to remove irrelevant itemsets in the database. The detailed explanation of the utility itemset mining is mentioned in the [17] with mining with expected utility (MEU) to decide whether an itemset should be well thought-out as a contestant itemset or not . The number of itemset becomes large it becomes complex to analysis and generates the candidate patterns for larger transactional database .These problem is solved by using the methods such as UMining and UMining H [18], to compute the elevated effectiveness patterns.

In this method the UMining, is one of the pruning method to reduce the number of frequent itemset patterns in the dataset based on the utility upper bound property. UMining H is another pruning methods based on the threshold value from heuristic method. On the other hand, a number of high usefulness itemsets may possibly be inaccurately reduce through their heuristic method. Furthermore, these methods do not assure the descending conclusion belongings of Apriori, and consequently, miscalculate moreover many patterns. They furthermore suffer from unnecessary candidate generations results for each frequent itemset and produces poor utility mined itemset results.

CTU-Mine [19] anticipated to solve the difficulty of the utility itemset mining problem all the way through the use of the specific partition trees. The isolated items discarding strategy (IIDS) [20] designed for determine high effectiveness itemsets be developed to reduce the number of irrelevant patterns in the dataset for each and every scan of the process in the transaction itemset. It shows that the proposed system have better itemset mining results with reduced memory space and more time complexity than the existing itemset mined results .It mines frequent itemset results based on the calculation of the frequency values to each items in the transaction database with calculation of the profit value for each items also in transaction section. Though, their algorithms still go through level-wise candidate set creation, and necessitate numerous database examinations.

3. DATA PROPOSED SYSTOLIC TREE WITH ABC BASED PATTERN MINING ALGORITHM METHODOLOGY

Utility mining plays major imperative role in the itemset mining results in the data mining applications such as web transaction database, educational data mining and medical databases etc, mining high usefulness of the data corresponds to improve the market profit results of the each organization .At this point, the importance of itemset usefulness is interestingness, significance, or effectiveness of an item to customer. It consists of two major objectives to perform mining task for transaction database such as

- The significance of distinctive items, which is named external utility.
- The significance of items in business, which is named as internal utility.

It is defined based the above mentioned two characteristic properties. In earlier work propose a two itemset mining algorithm to find high utility itemset in the transactional databases. Frequent itemsets mining the occurrence is not adequate to respond question, such as whether an itemset is extremely cost-effective, or whether an itemset has a well-built force. In order to solve these problems propose a utility mining algorithm to solve the problem of mining high utility itemset in the larger database.

The major aim of this algorithm is to improve the utility itemset mining results for quicker for large transaction dataset using systolic tree algorithm with automatic weight generation values from ABC. In order to perform this process the user interface are created using the JAVA based implementation for larger database . The design of the user interface must satisfy the user functionality and their elements are performed accurately for all the database information.

A new systolic tree-based move toward to extract frequent item sets in the transaction database with frequently purchased items in the database .The size of the tree must be small so it can be divided the transaction database into small number of the transaction to accurately mine the itemset in the database .The mining speed of the proposed STA-ABC are high since the weight values of the each and every items in the database are created based on the procedure of ABC with count values are calculated to each items in the database for each number of the frequent patterns .

In this work presents a systolic tree algorithm to improve the accuracy results of the frequent itemset mining results when compare to existing frequent itemset mining results. It calculates the weight values automatically based on the request counts and the sequence value for each item in the transactional database. Automatic assignment of the weight values becomes more challenge in the existing work in order to overcome these problems in this work presents an Artificial Bee Colony (ABC) based optimization algorithm to calculate weight values for each item in the transaction database. Systolic tree is the collection of pipelined processing elements (PE) in a multidimensional tree pattern. The position of the systolic tree as plan in FPGA hardware is subsequently equivalent to the FP-tree as second-hand in software. The transaction items are simplified addicted in the direction of the systolic tree through candidate item matching and count up update operations.

It is used to arrange the candidate itemset based on the frequency values for each itemset in the transaction database. Appropriate to the imperfect dimension of the systolic tree, a transactional database should be predictable addicted to less important ones every one of which can be extract in professionally. The responsibility of the systolic tree as illustration in FPGA hardware is then equivalent to the FP-tree as second-hand in software. The prescribed description of FP-tree can be established in complementary.

The propose law of the WRITE form algorithm is with the intention of the developed systolic tree must contain a comparable explain through the FP-tree specified the identical transactional database. After a number of clock cycle, the systolic tree launch the support calculation of the applicant sample backside to the software. The software evaluates the support count through the support based threshold value and makes a decision whether the candidate pattern is recurrent or not. After each and every one of the candidate pattern are checkered through the support based threshold value in the software finally efficient pattern mining algorithm.

In ABC algorithm [21-22], there are three types of the bees plays major important role to make a decision or solve the problems such a colony of bees are : employed bees, onlookers and scouts. In this paper initially the frequency and count values of the each item with randomly generated weight values is considered as bees .Initially the half of the frequent itemset in the transaction with weight values is considered employed artificial bees and the other half of the half of the frequent itemset in the transaction with weight values is considered onlooker bees. The total number of selected frequent itemset weight values is equal to the number of the best randomly generated weight values food sources around the hive. The result of the onlooker bees is abandoned through the scout bees if the utility mining results is not achieved within the maximum number of iteration stage in the ABC procedure ,or else it is kept as same ,the values of the current frequent itemset weight values are search the next best weight values by updating the location of the each population value in the ABC algorithm through position update formula .In first step the weight value of the each and every transaction itemset are randomly create a population through size of the population SN solution $x_i(i = 1,2,\dots,SN)$.

After initialization of the weight value then perform the automatic weight generation process based on the Bee procedure with maximum number of the iterations stage $C = 1,2,\dots,MCN,$ An employed bee produces a weight updating results based on the occurrence and count values of the items in the frequent itemset is considered as fitness values in this paper. If the fitness value of the current created weight value is higher than the earlier one weight value for same frequent itemset the currently generated weight value is considered as best weight value for each item in the transaction database ,or else kept earlier weight as best value in the memory of the ABC results ,update their position value in the employee bee stage ,then these results are sent to onlooker bee stage then highest probability value are calculated to each employee bee generated weight value based on the dance area and checks location of each weight values the probability value of them each item with weight values p_i , calculated by the following expression :

$$p_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n}$$

where fit_i is the fitness value of each weight value for each items *i* which is proportional to the automatic weight value generation results ,the best weight values are stored in the memory, based the following expression,

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj})$$

where $k \in \{1, 2, \ldots, SN\}$ and $j \in \{1, 2, \ldots, D\}$ are randomly chosen indexes. Although k is determined indiscriminately, it have to be dissimilar from *i*. ϕ_{ij} is a indiscriminate amount among [-1, 1]. It manages the assembly of neighbour provisions source approximatelyx_{ij}. If the value of the parameter is not exceeding in the threshold then it is selected as best weight value generation then it is used for itemset mining. If the particular results are not until the maximum number of the iterations specified in the ABC method then current position value is moved to next state which is controlled by the parameter called *limit* designed for rejection. This process be able to be distinct as in,

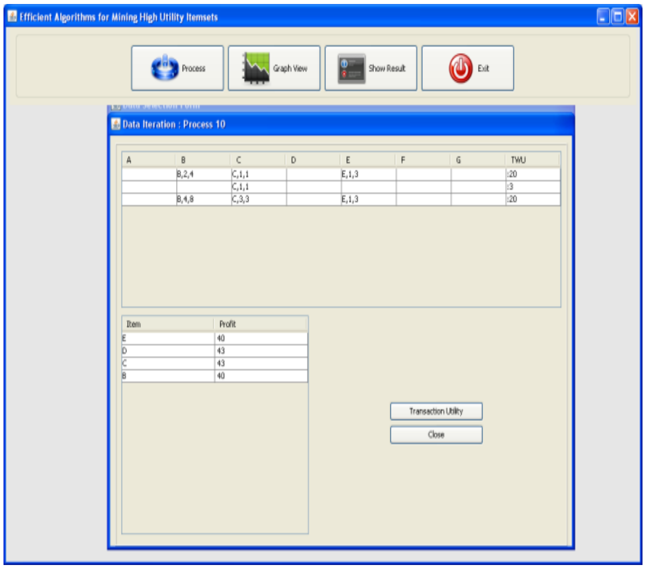
$$x_i^j = x_{min}^j + rand(0,1)(x_{max}^j - x_{min}^j)$$

In this work each and every bee is considered as weight values of the each items ,best candidate bee value are found based on the position values of the each bee with their velocity values assigned to each bee and it can be compared with previous bee weight values in the earlier results .If the current weight value of the frequent itemset is mined the best utility pattern results or equal to the best results it is replaced with earlier weight value results in the bee food source position ,or else kept the elderly generated weight values as same in the bee memory ,then greedy selection process is performed between the earlier generated weight value for frequent itemset and currently generated weight value to each frequent itemset in the transaction database ,it perform the ABC process based on the four parameters like total number of the weight values population (bee size) (*SN*), the limit value to control the searching process of weight value generation *limit*, the maximum cycle number (*MCN*) to complete the weight generation process in ABC. Detailed pseudo-code of the ABC algorithm is given below:

```
Initialize the population of solutions $X_{ij}(i = 1, \ldots SN, j = 1, , \ldots D)$
Evaluate the population
cycle=1
repeat
Produce new solutions $v_{ij}$ for the employed bees by using  and evaluate them
Apply the greedy selection process
Calculate the probability values $P_{ij}$ for the solutions $x_{ij}$
Produce the new solutions $v_{ij}$ for the onlookers from the solutions $x_{ij}$ selected depending on $P_{ij}$ and evaluate them
Apply the greedy selection process
Determine the abandoned solution for the scout, if exists, and replace it with a new randomly produced solution $x_{ij}$
Memorize the best solution achieved so far
cycle=cycle+1
until cycle=MCN
```

4. EXPERIMENTAL RESULTS

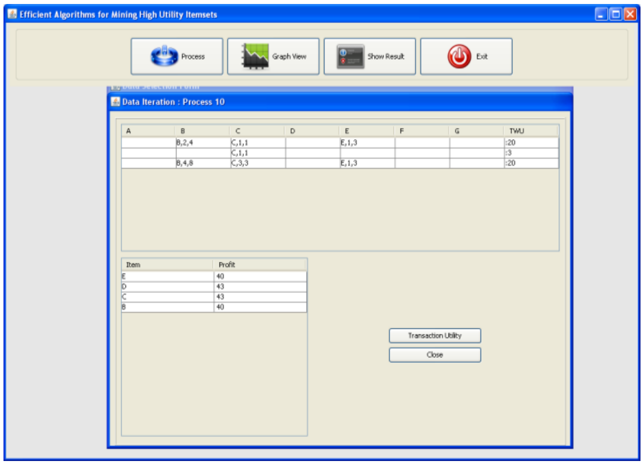
To assess the performance of our existing and proposed STA-ABC based tree algorithm have been done to several number of real and synthetically generated datasets. The Proposed algorithm accuracy is compared with existing methods with the parameter such as running time of the results is measured based on the minimum utility threshold value in the dataset . Next, show the results of the efficiency of the proposed and existing FP and FP++ .The sample utility mining results of the FP++ and the proposed STA-ABC based tree mining algorithm transaction frequent itemset results are illustrated in the following screen such as Figure 1,Figure 2 for existing and proposed methods .


Efficient Algorithms for Mining High Utility Itemsets
Process
Graph View
Show Result
Exit
Data Iteration : Process 10
Transaction Utility
Close

A	B	C	D	E	F	G	TWU
	B,2,4	C,1,1		E,1,3			:20
		C,1,1					:3
	B,4,8	C,3,3		E,1,3			:20

Item	Profit
E	40
D	43
C	43
B	40

Figure 1:Mined High Utility itemset results for FP growth


Efficient Algorithms for Mining High Utility Itemsets
Process
Graph View
Show Result
Exit
Data Iteration : Process 10
Transaction Utility
Close

A	B	C	D	E	F	G	TWU
	B,2,4	C,1,1		E,1,3			:20
		C,1,1					:3
	B,4,8	C,3,3		E,1,3			:20

Item	Profit
E	40
D	43
C	43
B	40

Figure 2: Mined High Utility itemset results for STA-ABC growth

The running time comparison results of the proposed STA-ABC and the existing FP,FP++ growth mining algorithm results are shown in the figure 3,it shows that the proposed STA-ABC have less running time when compare to existing methods, candidate generation results of the proposed and existing method are shown in Figure 4 ,it shows that proposed STA-ABC have less candidate generation results .

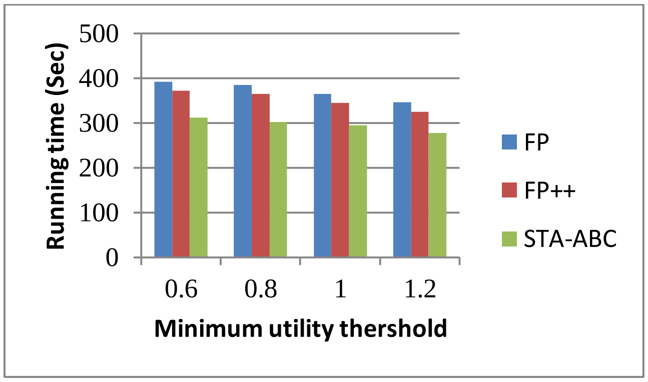


Figure 3: Running time comparison results of the methods

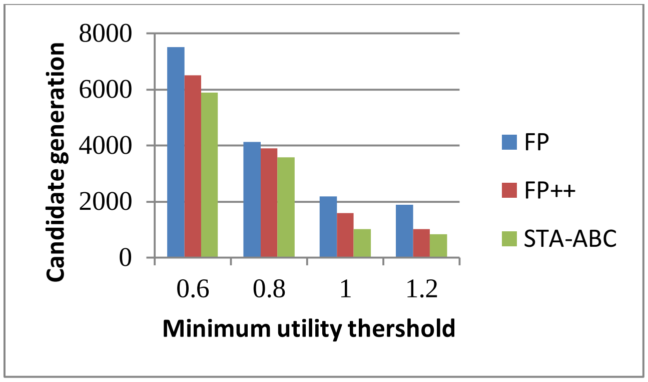


Figure 4: Candidate generation comparison results of the methods

5. CONCLUSION

In this paper presents a novel efficient frequent itemset mining based on the creation of systolic tree with automatic weight generation .The automatic weight generation process use an frequency and count values of the items in the transaction, weight values are automatically generated for each items using artificial bee colony (ABC) optimization algorithm for mining the high utility of the itemset for each number of the items in the transaction database. The tree structure the items values of each information are stored in the data structure for each frequent itemset ,best weight are determined using ABC and thus improves the itemset mining results in the larger transactional database. The proposed system overcomes the problem of the overestimation and underestimation utility problem for each utility mining. Results show that the proposed STA-ABC have higher utility mining results than the existing FP growth and UP growth algorithm in terms of the execution time and memory space through reducing the number of irrelevant candidate frequent itemset patterns in the database. Wide-ranging performance investigation shows with the intention of our tree structures are extremely well-organized designed for interactive high utility pattern mining and they go one better than the existing algorithms.

References

1. R. Agrawal, T. Imielinski, and A. Swami, **"Mining Association Rules Between Sets of Items in Large Databases,"** *Proc. 12th ACM SIGMOD*, pp. 207-216, 1993.
2. R. Agrawal and R. Srikant, **"Fast Algorithms for Mining Association Rules,"** *Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94),* pp. 487-499, 1994.
3. J. Han, J. Pei, Y. Yin, and R. Mao**, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach,"** *Data Mining and Knowledge Discovery*, Vol. 8, pp. 53-87, 2004.

4. G. Grahne and J. Zhu, "**Fast Algorithms for Frequent Itemset Mining Using FP-Trees,"** *IEEE Trans. Knowledge and Data Eng.*, Vol. 17, No. 10, pp. 1347-1362, Oct. 2005.
5. J. Dong and M. Han, **"BitTableFI: An Efficient Mining Frequent Itemsets Algorithm,"** *Knowledge-Based Systems,* Vol. 20, pp. 329-335, 2007.
6. M. Song and S. Rajasekaran, **"A Transaction Mapping Algorithm for Frequent Itemsets Mining,"** *IEEE Trans. Knowledge and Data Eng.*, Vol. 18, No. 4, pp. 472-481, Apr. 2006.
7. Yao, H., Hamilton, H.J., Buzz, C.J.**," A Foundational Approach to Mining Itemset Utilities from Databases",** *In: 4th SIAM International Conference on Data Mining*. Florida USA (2004)
8. Yao, H., Hamilton, H.J**.," Mining itemset utilities from transaction databases",** *Data & Knowledge Engineering* Vol.59,No.3, 603–626 (2006)
9. Liu, Y., Liao, W.K., Choudhary, A**.," A Fast High Utility Itemsets Mining Algorithm",** *In: 1st Workshop on Utility-Based Data Mining*. Chicago Illinois (2005).
10. H. F. Li, H. Y. Huang, Y. C. Chen, Y. J. Liu and S. Y. Lee, **"Fast and Memory Efficient Mining of High Utility Itemsets in Data Streams,"** *in Proc. of the 8th IEEE Int'l Conf. on Data Mining*, pp. 881-886, 2008.
11. B.-E. Shie, H.-F. Hsiao, V. S. Tseng and P. S. Yu, **"Mining high utility mobile sequential patterns in mobile commerce environments,"** *in Proc. of the 16th Intl. Conf. on DAtabase Systems for Advanced Applications (DASFAA 2011) and Lecture Notes in Computer Science (LNCS),* Vol. 6587/2011, pp. 224-238, 2011.
12. R. Chan, Q. Yang and Y. Shen. **"Mining high utility itemsets,"** *in Proc. of Third IEEE Int'l Conf.* on Data Mining, pp. 19-26, Nov., 2003.
13. H. Xiong, P.-N. Tan, and V. Kumar, "**Hyperclique Pattern Discovery,"** *Data Mining and Knowledge Discovery*, Vol. 13, pp. 219-242, 2006.
14. J. Wang, J. Han, Y. Lu, and P. Tzvetkov**, "TFP: An Efficient Algorithm for Mining Top-K Frequent Closed Itemsets,"** *IEEE Trans. Knowledge and Data Eng.*, Vol. 17, no. 5, pp. 652-664, May 2005.
15. C. Lucchese, S. Orlando, and R. Perego, **"Fast and MemoryEfficient Mining of Frequent Closed Itemsets,"** *IEEE Trans. Knowledge and Data Eng*., Vol. 18, No. 1, pp. 21-36, Jan. 2006.
16. B. Barber and H.J. Hamilton, **"Extracting Share Frequent Itemsets with Infrequent Subsets,"** *Data Mining and Knowledge Discovery*, Vol. 7, pp. 153-185, 2003.
17. H. Yao, H.J. Hamilton, and C.J. Butz, **"A Foundational Approach to Mining Itemset Utilities from Databases,"** *Proc. Fourth SIAM Int'l Conf. Data Mining (SDM '04),* pp. 482-486, 2004.
18. H. Yao and H.J. Hamilton, **"Mining Itemset Utilities from Transaction Databases,"** *Data and Knowledge Eng.,* Vol. 59, pp. 603-626, 2006.
19. A. Erwin, R.P. Gopalan, and N.R. Achuthan, **"CTU-Mine: An Efficient High Utility Itemset Mining Algorithm Using the Pattern Growth Approach,"** *Proc. Seventh IEEE Int'l Conf. Computer and Information Technology (CIT '07),* pp. 71-76, 2007.
20. Y.-C. Li, J.-S. Yeh, and C.-C. Chang, "**Isolated Items Discarding Strategy for Discovering High Utility Itemsets,"** *Data and Knowledge Eng*., vol. 64, pp. 198-217, 2008.
21. D. Karaboga, "**An Idea Based On Honey Bee Swarm For Numerical Optimization"**, *Technical Report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department*, 2005.
22. B. Basturk, D.Karaboga, "**An Artificial Bee Colony (ABC) Algorithm for Numeric function Optimization"**, *IEEE Swarm Intelligence Symposium 2006,* May 12- 14, 2006, Indianapolis, Indiana, USA.