International Journal of the Physical Sciences Vol. 6(2), pp. 301-307, 18 January, 2011
Available online at http://www.academicjournals.org/IJPS
DOI: 10.5897/IJPS09.303
ISSN 1992 - 1950 ©2011 Academic Journals

Full Length Research Paper

Mining utility-oriented association rules: An efficient approach based on profit and quantity

Parvinder S. Sandhu[1]*, Dalvinder S. Dhaliwal[2] and S. N. Panda[3]

[1]Department of Computer Science and Engineering, Rayat and Bahra Institute of Engineering and Bio-Technology, Sahauran, Distt. Mohali (Punjab)-140104 India.
[2] Department of Computer Science and Engineering, RIMIT Institute of Engineering and Technology, Mandi Gobindgarh (Punjab)- India.
[3]Regional Institute of Management and Technology, Mandi Gobindgarh (Punjab)- India.

Accepted 20 December, 2010

Association rule mining has been an area of active research in the field of knowledge discovery and numerous algorithms have been developed to this end. Of late, data mining researchers have improved upon the quality of association rule mining for business development by incorporating the influential factors like value (utility), quantity of items sold (weight) and more, for the mining of association patterns. In this paper, we propose an efficient approach based on weight factor and utility for effectual mining of significant association rules. Initially, the proposed approach makes use of the traditional *Apriori* algorithm to generate a set of association rules from a database. The proposed approach exploits the anti-monotone property of the *Apriori* algorithm, which states that for a k-itemset to be frequent all (k-1) subsets of this itemset also have to be frequent. Subsequently, the set of association rules mined are subjected to weightage (W-gain) and utility (U-gain) constraints, and for every association rule mined, a combined utility weighted score (UW-Score) is computed. Ultimately, we determine a subset of valuable association rules based on the UW-Score computed. The experimental results demonstrate the effectiveness of the proposed approach in generating high utility association rules that can be lucratively applied for business development.

Key words: Association rule mining (ARM), frequent itemset, utility, weightage, *apriori*, utility gain (U-gain), weighted gain (W-gain), utility factor (U-factor), utility weighted score (UW-score).

INTRODUCTION

Due to the wide availability of huge amounts of data and imminent need for turning such data into useful information and knowledge, data mining has attracted a great deal of attention in the information industry in recent years (Ali et al., 2009). Data mining is an important part of the process of knowledge discovery in databases (KDD) (Oded and Lior, 2005). The knowledge discovery in databases (KDD) is defined as, "The non-trivial extraction of implicit, previously unknown, and potentially useful information from data" (Frawley et al., 1992).

In general, data mining tasks can be classified into two categories: Descriptive mining and predictive mining. Descriptive data mining is the description of a set of data in a concise and summarized manner and the presentation of the general properties of the data. Predictive data mining is the process of inferring patterns from data to make predictions (Han and Fu, 1996). One important descriptive data mining technique, association rule mining (ARM) introduced by Agarwal et al. (1993), have received considerable attention in data mining research and applications. More and more computer science scholars and researchers, especially those who specialize in the field of knowledge discovery in data (KDD), focus and emphasize on association rule mining (ARM) (Bing et al., 1999).

Association rule mining (Agrawal et al., 1993; Agrawal and Srikant, 1994) has been widely used to solve data

*Corresponding author. E-mail: parvinder.sandhu@gmail.com. Tel: +91-98555-32004.

mining problems in numerous applications, including financial analysis, the retail industry, and business decision-making (Chen et al., 1996). The problem with mining association rules can be distilled into two steps. The first step involves finding all frequent itemsets (or say large itemsets) in databases. The next step is the generation of association rules. Once the frequent itemsets are found, generating association rules is straightforward and can be accomplished in linear time (Chun-Jung et al., 2008). The most important task of traditional association rule mining (ARM) is to identify frequent itemsets. Traditional ARM algorithms treat all the items equally by assuming that the weightage of each item is always 1 (item is present) or 0 (item is absent). Obviously, it is unrealistic and will lead to loss of some useful patterns (Guangzhu et al., 2008). In order to overcome the weakness of the traditional association rules mining, utility mining model (Zubair and Alasubram, 2009; Hong et al., 2004) and weighted association rule mining (Ke Sun and Fengshan, 2008) have been proposed.

Utility based data mining is a new research area entranced in all types of utility factors in data mining processes and focused at integrating utility considerations in data mining tasks (Shankar et al., 2009). Utility of an item is a subjective term dependent on users and applications; it could be measured in terms of profit, cost, risk, aesthetic value or other expressions of user preference (Guangzhu et al., 2008). Given a transaction database, a minimum utility threshold, and a utility table, the goal of utility mining is to discover all high utility itemsets (Yu-Chiang et al., 2008). The profitability of an itemset or the total cost of stocking an itemset cannot be determined using the support value alone. Thus, in practical terms, utility mining can be more useful than traditional association rule mining (Yu-Chiang et al., 2008).

In weighted association rule mining (WARM) (Ke Sun and Fengshan, 2008), itemsets are no longer simply counted as they appear in a transaction. This change of counting mechanism makes it necessary to adapt traditional support to weighted support (Feng et al., 2003). Weighted association rules cannot only improve the confidence in the rules, but also provide a mechanism to do more effective target marketing by identifying or segmenting customers based on their potential degree of loyalty or volume of purchases (Zubair and Alasubram, 2009). For example, a customer may purchase 13 bottles of coke and 6 bags of snacks and another may purchase 4 bottles of coke and 1 bag of snacks at a time. The conventional association rule approach treats the aforementioned transactions in the same manner, which could lead to the loss of some vital information (Zubair and Alasubram, 2009). So, weighted ARM deals with the importance of individual items in a database (Cai et al., 1998; Wang et al., 2000; Zubair and Alasubram, 2009). For example, some products are more profitable or may be under promotion, therefore more interesting as compared to others, and hence rules concerning them are of greater value (Sulaiman et al., 2009).

Recently, researchers are interested at incorporating both the attributes (weightage and utility) for mining of valuable association rules. The incorporation, weighted utility association rule mining (WUARM) can be considered as the extension of weighted association rule mining in the sense that it considers items weights as their significance in the dataset and also deals with the frequency of occurrences of items in transactions. Thus, weighted utility association rule mining is concerned with both the frequency and significance of itemsets and is also helpful in identifying the most valuable and high selling items which contribute more to the company's profit (Sulaiman et al., 2008). Here, we propose an efficient approach based on weight factor and utility for effectual mining of high utility association rules. Initially, the proposed approach makes use of the traditional *Apriori* algorithm to generate a set of association rules from a database. Subsequently, the set of association rules mined are subjected to (1) weightage (W-gain) and (2) utility (U-gain) constraints, and for every association rule mined, a combined utility weighted score (UW-Score) is computed. Ultimately, we determine a subset of valuable weighted and utility based association rules on the basis of the UW-score computed. The experimental results demonstrate the effectiveness of the proposed approach in mining high weighted and utility based association rules that can be lucratively applied for business development.

PROPOSED METHODOLGY BASED ON PROFIT AND QUANTITY FOR MINING UTILITY-ORIENTED ASSOCIATION RULES

Measures employed

Association rule mining (ARM), one of most commonly used descriptive data mining task, is the process of discovering a collection of data attributes that are statistically related in the underlying data. *Apriori* has been recognized as the most renowned algorithm for ARM. Most algorithms for ARM including *apriori*, mine potential data patterns chiefly based on frequency. The data patterns thus extracted based only on frequency would not be of the highest value for decision makers in business development. Hence in recent times, the incorporation of interestingness, utility and item weightage into the standard association rule mining algorithms have attracted voluminous research. In the proposed approach, we have incorporated two of the aforesaid measures together with *Apriori* for effectual mining of association rules. The two attribute measures chosen in the proposed research are, weightage and utility.

Weightage

Generally in a transaction database, attributes comprise of numerical assets that gives the actual quantity of the attribute (count of the items) involved in the transaction. But traditional algorithms like *Apriori* mine association rules from a binary mapped database that only depicts the presence of the item in a transaction or not. So, standard ARM algorithms possibly overlook the quantitative information associated with an attribute, leading to frequent but less-weightage rules. In most cases of a customer transaction, some of the attributes may be actually more weighted in one transaction but it may not have occurred frequently in the database. However, these set of attributes would be of significant value in the business point of view and should have been included in the frequent itemset. Hence, the proposed research will consider the weightage measure (W-gain) of the individual items in every transaction, for mining a subset of most significant rules from the set of frequent association rules mined.

Utility

The second measure employed in the proposed research to improve the quality of ARM is the individual utility (Gain) of the attributes. In general, a supermarket is likely to consist of attributes (items) that will yield different margins of profit. Hence, the rules mined without considering those utility values (profit margin) will lead to a probable loss of profitable rules. So, to attain a subset of high utility rules from the *Apriori* mined rules, the proposed approach makes use of a utility measure (U-gain).

A lot of researches (Sulaiman et al., 2008; Hong and Howard, 2006; Chun-Jung et al., 2009; Unil, 2007) have incorporated the weightage and utility measures individually into ARM so as to achieve effective rules from large databases. But, possibly the incorporation of both the measures together into ARM will enable more potential utility-oriented association rules. In this research, we incorporate the two measures weightage (W-gain) and utility (U-gain) to mine the association rules from a transaction database.

Proposed methodology

Let D be a database consisting of n number of transactions T and m number of attributes $I = [i_1, i_2, \ldots\ldots, i_m]$ with positive real number weights W_i. A utility table U comprising of m number of utility values U_i, where U_i denotes the profit associated with the i attribute. The major steps involved in the proposed approach for association rule mining based on weightage and utility are:

1. Mining of association rules from D using *Apriori.*
2. Computation of the measure W-gain.
3. Computation of the measure U-gain.
4. Computation of UW-score from W-gain and U-gain.
5. Determination of significant association rules based on UW-score.

Association rule mining using *Apriori*

Initially, association rules are mined from a transaction database D with n transactions. The database D is denoted as

$$D = \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_n \end{bmatrix}$$

Every transaction T in D comprises of ‘m’ number of attributes $I = [i_1, i_2, \ldots\ldots, i_m]$ associated with it. Every attribute i is represented by weights W_i.

Apriori, a standard ARM algorithm is used in the proposed approach to mine the association rules. Classical *Apriori* generally processes on a binary mapped database B_T for mining association rules. Hence, the input database D is transformed to a binary mapped database B_T that consists of binary values 0 and 1 denoting the non-existence and existence of attributes in the transactions respectively. The weights W_i associated with the individual attributes in the database D is mapped onto the binary values using the following equation.

$$B_T = \begin{cases} 0 \text{ if } \; \mathrm{W_i} = 0 \;\; \forall \, \mathrm{T_k} \\ 1 \text{ if } \; \mathrm{W_i} \neq 0 \;\; \forall \; \mathrm{T_k} \end{cases}$$

Subsequently, the binary mapped database B_T is given as an input to the *Apriori* algorithm (Agrawal et al., 1993) for mining of association rules. Generally, the process of association rule mining using *Apriori* is composed of two steps namely,

1. Frequent itemset generation: Generate all possible sets of attributes that have support value greater than a predefined threshold, called minsupport.
2. Association rule generation**:** Generate association rules from the generated frequent itemsets that have confidence greater than a predefined threshold called minconfidence.

A standard association rule is of the structure: A→B, where A (the antecedent) and B (the consequent) is subset of items in the binary mapped database, such that, A$\subset$ I, B $\subset$ I and A $\cap$B=ϕ. The rule A→B is interpreted as “if A exists then it is likely that B also co-exists”. The rule A $\longrightarrow$ B holds in the transaction database D with a support S and confidence C, if, S% of the transactions in D contains the itemsets A and B, and C% of the transactions that contain A also contain B.

Support (A $\longrightarrow$ B) = P (AUB)

Confidence (A $\longrightarrow$ B) = P (B|A) = support (AUB)/ support (A)

The pseudo code for the *Apriori* algorithm is,

```
$I_1 = \{l\arg e \;\; 1 - itemsets\};$
$for\ (\mathrm{k}=2;\ \mathrm{I_{k\text{-}1}} \neq 0;\ \mathrm{k}{+}{+})$ do begin
$C_k = apriori - gen(I_{k-1});$    // New candidates
$forall$ transactions $\mathrm{T} \in \mathrm{D}$ do begin
$C_T = subset(C_k, T);$ // Candidates contained in T
$forall$ candidates $\mathrm{c} \in \mathrm{C_T}$ do
$c.count++;$ )
$end$
$end$
```

$$I_k = \{c \in C_k \mid c.count \geq \min\sup\}$$
$$end$$
$$Answer = \bigcup_k I_k;$$

The *Apriori* algorithm generates a k number of association rules $R = \{R_1, R_2, ..., R_k\}$. The set of association rules R is fed as input to the next phase of the proposed research, weightage and utility computation. The measures W-gain (weightage) and U-gain (utility) are calculated for every attribute present in the k association rules of R. For example, say an association rule R_i of the form, (A, B) $\Longrightarrow$ C, where, A, B and C are the attributes in the rule R_i, the measures U-gain, W-gain and UW-score are calculated for every attribute A, B and C individually.

Initially, the k association rules generated is first sorted in descending order based on their confidence level. The sorted list of association rules is given by $S = \{R'_1, R_2', ..., R_k'\}$, $S \in R$, where conf (R_1') $\geq$ conf (R_2') $\geq$ conf (R_3')..... $\geq$ conf (R_k').

Computation of W- gain

From the sorted list S, the first rule R_1' is selected and the individual attributes of R_1' are determined. Subsequently, the measure W-gain is calculated for every attribute in the rule R_1'.

Definition 1: Item weight (Wi): Item weight is the quantitative measure of the attribute contained in the transaction database D. Item weight value W_i is a non-negative integer.

Definition 2*:* Weighted gain (W-gain): W-gain is defined as the sum of item weights W_i of an attribute contained in every transaction of the database D.

$$\text{W-gain} = \sum_{i=1}^{|T|} W_i$$

where, W_i is the item weight of an attribute and $|T|$ is the number of transactions in the database D.

Computation of U-gain

Similarly, for U-gain computation, the first rule R_1' from the sorted list S is selected and the individual attributes of R_1' are determined. Subsequently, the U-gain measure is calculated for every individual attribute present in the rule R_1', based on the U-factor and the utility value U_i of the attribute.

Definition 3: Item Utility (U_i): The Item utility is generally defined as the margins of profit associated with that particular attribute. It is denoted as U_i.

Definition 4: Utility table U : The utility table U comprises of 'm' utility values U_i associated with the attributes present in the transaction database D. The utility table is represented by,

$$U = \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ \mathrm{U_m} \end{bmatrix}$$

Definition 5: Utility factor (U-factor): The utility factor (U-factor) is a constant that is determined by the sum of the all items utility (U_i) contained in the utility table U. It is defined as,

$$\text{U-factor} = \frac{1}{\sum_{i=1}^{m} U_i}$$

where, m is the number of attributes present in the transaction database.

Definition 6: Utility gain (U-gain): Utility gain refers to the measure of an attribute's actual utility based on the U-factor."

U-gain= U_i * U-factor

The measure U-gain is computed for every attribute in the association rule R_1'.

Computation of UW-score from W-gain and U-gain

Based on the calculated W-gain and U-gain measures for the individual attributes of an association rule, a single consolidated value termed UW-score is computed for every individual association rule.

Definition 7: Utility weighted score (UW-score): UW-score is defined as the ratio between the sum of products of W-gain and U-gain for every attribute in the association rule to the number of attributes present in the rule.

$$\text{UW-score} = \frac{\sum_{i=1}^{|R|} (W-gain)_i * (U-gain)}{|R|}$$

where, $|R|$ represents the numbers of attributes in the association rule.

The aforesaid processes of W-gain, U-gain and UW-score computation are repeated for the rest of the association rules R_2' to R_k' present in the sorted list S. Now, all 'k' number of

Table 1. Test dataset description.

Datdaset	Size	No. of transactions	No. of items
T10I4D100K	3.93MB	100,000	870

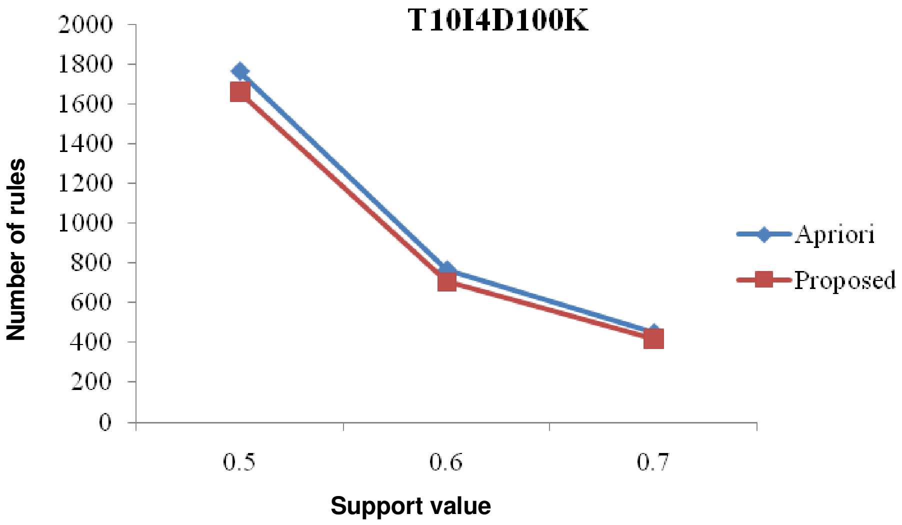

Figure 1. No. of rules generated using various support threshold with UW-Score=0.3.

association rules in the sorted list S possess a UW-Score associated with it. Subsequently, the association rules in the sorted list S are sorted based on the UW-score to get $S^{'}$ = $\{R_1", R_2", \ldots\ldots, R_k"\}$ where UW-score ($R_1"$) ≥ UW-score ($R_2"$) ≥ UW-score ($R_3"$)...... ≥ UW-score ($R_k"$).

Determination of significant association rules based on UW-score

Finally, from the sorted list $S^{'}$, a set of significant weighted utility association rules R_{WU} whose the UW-Score is above a predefined threshold are selected. The resultant weighted and utility based association rules is given by $R_{WU} = \{R_{WU1}, R_{WU2}, \ldots\ldots, R_{WUl}\}$, where, $k \geq l$ and $R_{WU} \subseteq S^{'}$.

EXPERIMENTAL RESULTS

The data employed in our experiments are either real-world data obtained from different fields or widely-accepted synthetic data generated using existing tools that are used in scientific and statistical simulations. We have evaluated our approaches in two different datasets, namely T10I4D100K (Sulaiman et al., 2008). The synthetic data, T10I4D100K is attained from the IBM dataset generator.

T10I4D100K

This dataset contains 100,000 transactions and 870 distinct items. T10I4D100K denotes the average size of the transactions (T), average size of the maximal potentially large itemsets (I) and the number of transactions (D) (Table 1).

Experimentations on both datasets are undertaken using two different association rule mining algorithms. The algorithms utilized for our results analysis are standard *Apriori* algorithm and an efficient proposed approach based on profit and quantity. Here, two thresholds, namely (1) support values (min_sup and min_conf) and (2) utility weighted score (UW-Score) are used for obtaining the utility based association rules. By changing these thresholds, different results are obtained and the results are analyzed for fixing the accurate threshold.

The experimental results are taken by varying the support and UW-Score on T10I4D100K dataset. The obtained results are plotted in the graphs as shown in Figures 1 to 4, which show the performance of the approaches on T10I4D100K dataset in mining the utility based association rules. By analyzing the plotted graphs, the behavior of the two approaches (Standard *Apriori* and the proposed approach) are almost same, when UW-Score=0.3. So, in this threshold, the performance of the proposed approach fails to produce better results. Then, we increased the UW-Score value as 0.4 and the number

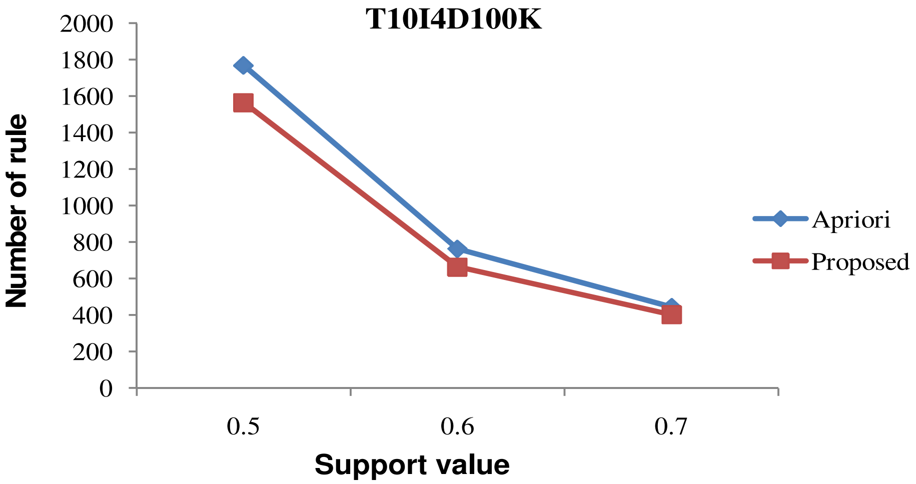

Figure 2. Number of rules generated using various support threshold with UW-Score=0.4.

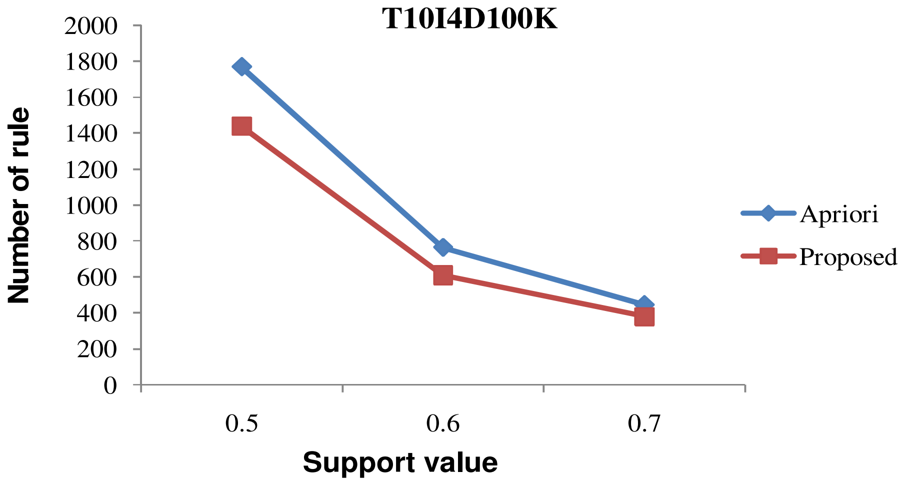

Figure 3. Number of rules generated using various support threshold with UW-Score=0.5.

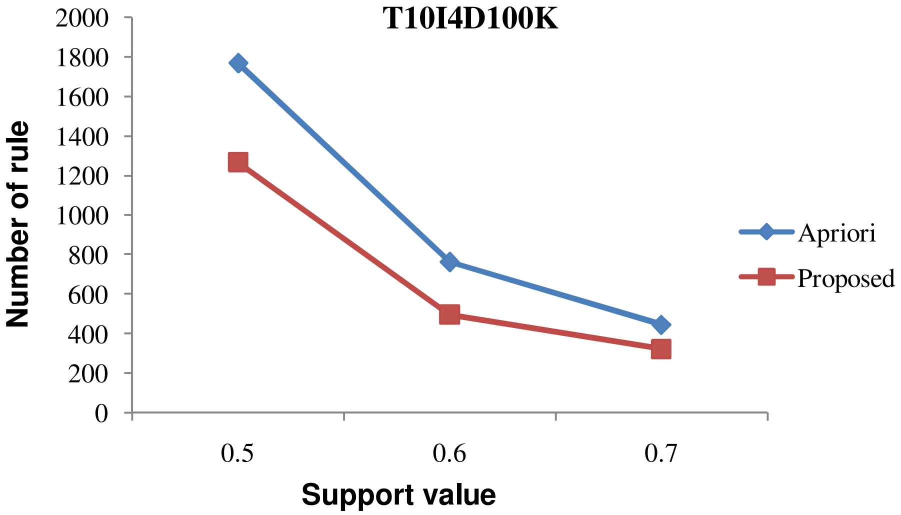

Figure 4. Number of rules generated using various support threshold with UW-Score=0.6.

of association rules generated by the proposed approach is somewhat constraint from the *Apriori* algorithm. The proposed approach provides better results, while we increase the UW-Score value as 0.5 and 0.6. With these values, the proposed approach generated a set of constraint utility based association rules compared with *Apriori* algorithm.

Conclusion

We have proposed an efficient approach based on weight factor and utility for effectual mining of high utility association rules. Initially, the proposed approach has made use of the traditional *Apriori* algorithm to generate a set of association rules from a database. For every association rule mined, a combined utility weighted score (UW-Score) is computed based on weightage (W-gain) and utility (U-gain) constraints. Ultimately, we have determined a subset of significant association rules based on the UW-Score computed. The experimental results have demonstrated the effectiveness of the proposed approach in generating high utility association rules that can be lucratively applied for business development.

REFERENCES

Agrawal R, Imielinski T, Swami A (1993). “Mining association rules between sets of items in large databases”. In proceedings of the International Conference on Management of Data, ACM SIGMOD, Washington, DC, 5: 207–216.

Agrawal R, Srikant R (1994). “Fast algorithms for mining association rules”, In Proceedings of 20th International Conference on Very Large Data Bases, Santiago, Chile, 9: 487–499.

Ali RG, Nasibeh M, Aida H, Parviz A, Nasibeh M (2009). "Recognizing and Prioritizing Of Critical Success Factors (CSFs] On Data Mining Algorithm's Implementation. In Banking Industry: Evidence From Banking Business System", In Proceedings of EABR & TLC, Prague, Czech Republic.

Bing L, Wynne H, Ke W, Shu C (1999). “Visually Aided Exploration of Interesting Association Rules”, In Proceedings of the Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining, pp. 380–389.

Cai CH, Fu AWC, Cheng CH, Kwong WW (1998). “Mining Association Rules with Weighted Items”. In Proceedings of the International Symposium on Database Engineering and Applications, Cardiff, Wales, UK, July, pp. 68-77.

Chen MS, Han J, Yu PS (1996). “Data mining: An overview from a database perspective”. IEEE Trans. Knowl. Data Eng., 8(6): 866–883.

Chun-Jung C, Vincent ST, Tyne L (2008). "Mining temporal rare utility Itemsets in large databases using relative utility thresholds". Int. J. innov. Comput. Inf. Cont., 4: 11.

Chun-Jung C, Vincent ST, Tyne L (2009). "An efficient algorithm for mining high utility itemsets with negative item values in large databases", Appl. Math. Comput., 215(2): 767-778.

Feng T, Fionn M, Mohsen F (2003). "Weighted Association Rule Mining using Weighted Support and Significance Framework". In Proceedings of the International Conference on Knowledge Discovery and Data Mining, Washington, pp. 661–666.

Frawley W, Piatetsky-Shapiro G, Matheus C (1992). “Knowledge Discovery in Databases: An Overview”. AI Mag. Fall, pp. 213-228.

Guangzhu Y, Shihuang S, Xianhui Z (2008). "Mining Long High Utility Itemsets in Transaction Databases". WSEAS Trans. Inf. Sci. Appl., 5(2): 202-210.

Han J, Fu Y (1996). “Attribute-Oriented Induction in Data Mining,” Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press, pp. 399-421.

Hong Y, Howard JH (2006). “Mining itemset utilities from transaction databases". Data Knowl. Eng., 59(3): 603-626.

Hong Y, Howard JH, Cory JB (2004). “A Foundational Approach to Mining Itemset Utilities from Databases”. In Proceedings of the Third SIAM International Conference on Data Mining, Orlando, Florida, pp. 482-486.

Ke S, Fengshan B (2008). "Mining Weighted Association Rules without Preassigned Weights". IEEE Trans. Knowl. Data Eng., 20(4).

Md. Zubair Rahman AMJ, Balasubram P (2009). "Weighted Support Association Rule Mining using Closed Itemset Lattices in Parallel". Int. J. Comput. Sci. Network Security, 9(3): 247-253.

Oded ZM, Lior R (2005). "Decomposition Methodology for Knowledge Discovery and Data Mining: Theory and Applications". World Scientific Publishing Company May, ISBN-13: 9789812560797.

Sandhu PS, Dhaliwal DS, Panda SN, Bisht A (2010). "An Improvement in *Apriori* Algorithm Using Profit and Quantity". Second Int. Conf. Comput. Network Technol., (ICCNT], April 23-25, 2010, Bangkok, pp. 3-7.

Shankar S, Babu N, Purusothaman T, Jayanthi S (2009). "A Fast Algorithm for Mining High Utility Itemsets", In proceedings of IEEE International Advance Computing Conference, Patiala, India.

Sulaiman KM, Maybin M, Frans C (2008). "A Weighted Utility Framework for Mining Association Rules". In Proceedings of European Symposium on Computer Modeling and Simulation, Liverpool, September, pp. 87-92.

Sulaiman MK, Maybin M, Frans C (2009). "Fuzzy Weighted Association Rule Mining with Weighted Support and Confidence Framework". Int. Workshops New Frontiers Appl. Data Mining, Osaka, Japan, 49–61: 20-23.

Unil Y (2007). "Efficient mining of weighted interesting patterns with a strong weight and/or support affinity". Inf. Sci., 177(17): 3477-3499.

Wang W, Yang J, Yu PS (2000). “Efficient Mining of Weighted Association Rules (WAR]”, In Proceedings of the KDD, Boston, MA, August, 270: 27.

Yu-Chiang L, Jieh-Shan Y, Chin-Chen C (2008). "Isolated items discarding strategy for discovering high utility itemsets". Data Knowl. Eng., 64(1): 198-217.