Global Journal of Computer Science and Technology
Volume 11 Issue 22 Version 1.0 December 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Extended Apriori for association rule mining: Diminution based utility weightage measuring approach

By P. Laxmi, A. Poongodai, D. Sujatha
Department of CSE

Abstract - The field of Association rule mining is a dynamic area for innovation of knowledge through which uncountable procedures have been expounded. Recently, by including significant components viz. value (utility), volume of items (weight) etc, the researchers have enhanced the quality of association rule mining for industry by bringing out the association designs. In this note, a proficient methodology has been put forward based on weight factor and utility for effective digging out of important association rules. At the very beginning, a traditional Apriori algorithm has been utilized that make use of the anti-monotone property which states that if n items are recurring continuously then n-1 items should also recur by which the scores of weightage(W-Gain), utility(U-Gain) and diminution(D-sum), are derived at. Eventually, we derive a subset of important association rules through which EUW-Score is generated. The tentative outcome demonstrates the effectiveness of the methodology in generating high utility association rules that is profitably used for the business improvement.

Keywords : *Association Rule Mining (ARM), Recurrent item set, Utility, Weightage, Apriori, Utility Gain (U-Gain), Weighted gain (W-Gain), Diminution sum (D-sum), Exact Utility Weighted Score (EUW-score).*

GJCST Classification : *H.2.8*

Strictly as per the compliance and regulations of:

Extended Apriori for association rule mining: Diminution based utility weightage measuring approach

P. Laxmi [α], A. Poongodai [Ω], D. Sujatha [β]

Abstract - The field of Association rule mining is a dynamic area for innovation of knowledge through which uncountable procedures have been expounded. Recently, by including significant components viz. value (utility), volume of items (weight) etc, the researchers have enhanced the quality of association rule mining for industry by bringing out the association designs. In this note, a proficient methodology has been put forward based on weight factor and utility for effective digging out of important association rules. At the very beginning, a traditional Apriori algorithm has been utilized that make use of the anti-monotone property which states that if n items are recurring continuously then n-1 items should also recur by which the scores of weightage(W-Gain), utility(U-Gain) and diminution(D-sum), are derived at. Eventually, we derive a subset of important association rules through which EUW-Score is generated. The tentative outcome demonstrates the effectiveness of the methodology in generating high utility association rules that is profitably used for the business improvement.

Keywords : Association Rule Mining (ARM), Recurrent item set, Utility, Weightage, Apriori, Utility Gain (U-Gain), Weighted gain (W-Gain), Diminution sum (D-sum), Exact Utility Weighted Score (EUW-score).

I. Introduction

Since accessibility of huge amounts of data and knowledge is on the rise, data mining has occupied a significant place in the field of information industry [1]. Data mining is a vital part of the process of Knowledge Discovery in Databases (KDD) [22] which is a non-trivial excavation of hidden, implied, never revealed data and comparatively has a large usage [6].

In broad, data mining methods are categorized into two ways:

1. Descriptive mining

It is an account of putting forward a group of data and its characteristics in a succinct and summarized way. One of the most significant of the descriptive kind mining is Association Rule Mining (ARM) which was introduced by Agarwal et.al. [2].

Author α : M.Tech, Department of CSE, ATRI, Parvathapur, Uppal, Hyderabad, India. E-mail : laxmi.p16@gmail.com
Author Ω : Assoc.Prof, Department of CSE, ATRI, Parvathapur, Uppal, Hyderabad, India. E-mail : a_poongodai@yahoo.co.in
Author β : Assoc.Prof and HOD Department of CSE, ATRI, Parvathapur, Uppal, Hyderabad, India.
E-mail : sujatha.dandu@gmail.com

2. Predictive mining

This method makes to surmise the outline of the information to give some assumptions [13].

Progressively many network side scholars and researchers of computer science, particularly those who are dedicatedly working in the area of Knowledge Discovery in Data (KDD), mainly concentrate and accentuate on Association Rule Mining (ARM) [14]. Association rule mining [2, 3] has been extensively utilized to detect and unravel data mining complicated issues which involve financial troubles, and business dealings [4].

The difficulties of using the mining association rules are divided into two steps.

1. First step is to evaluate the item sets that often occur in the databases.
2. The second one is to produce the association rules. Just the once if the some item sets are found occurring often, then the production of the association rules is uncomplicated and can be accomplished in a specified time [5].

Conventional ARM algorithms judge all the data items in the similar passion by assuming that their weight-age as always 1 if they are identified or 0 if they are not identified which intangibly drives to miss some of the very functional outlines of the data [7]. In order to trounce these drawbacks of the traditional way of mining, utility mining method [9] [10] and weighted association rule mining [16] has come into existence.

Utility data mining is a latest field of development entranced in every type of utility factors in data mining procedures and accentuated to assimilate the services also called as utilities in the data mining methods [15]. Utility of any particular data or item is a reliant on the individuals and is measured in terms of aesthetic values or other expressions of individual inclination [7]. When an action in the database and its related minimum utility threshold value and a utility chart are monitored then the objective of the utility mining is to identify and determine each and every high utility item set [12]. In general pros and cons of the item set in the business is not possible to be derived by the utilization of the values that shore up the rules. So this rationally proves that utility mining can be more advantageous than the conventional association rule mining [12].

But while considering the weighted association rule mining (WARM) [16], there has been a modification of not counting the item sets that occurred in an action of the database which made a compulsion to acclimatize the conventional help to the weighted one [23]. This method also segments the consumers on the basis of their reliable nature and impending count of procurements [8]. For an illustration, one consumer may buy 15 items and some other may have 5 items at a time but the conventional association rule method considers these couple of actions in a similar way. Hence the procedure of considering the actions in a same conduct in the conventional association rule method mislays some of the vital information [8]. As a result, Weighted ARM deals mainly with the magnitude of individual substances in a database [17, 18, 19]. For a case in point, goods that has more advantages or which are under the process of endorsement are given prior able constraints in comparison to rest [20].

Incorporating the mutual characteristics (weightage and utility) for excavation of rules is treated as an addition to the weighted association rule mining which means that the data weights are most important in a particular set of actions besides it also deals with the number of possible appearances of the data in those actions. This has made a concern in categorizing the data appearances and their weights and also in detecting the prior able data which put in more to the benefits of the business [21]. By considering this Sandhu et al[28] proposed a model that identifying association rules based on UW-Score, which calculated based on characteristics weightage and utility. In their proposal they were not considering diminution occurred in contextual factors

Here, we put forward an efficient method that makes the utilization of the traditional Apriori algorithm to engender a group of association rules from a database out of which a pooled Exact Utility Weighted Score (EUW-Score) is calculated. In due course, sub values of the priority given weighted, utility and diminution considered constraints are derived on the source of the EUW-score and the tentative outcome exhibit the effectiveness.

The later part of the paper is structured as given: The section 2 explains the methods involved. The proposed methodology based on usage of mining utility-oriented association rules is explained in section 3. Section 4 concludes the paper.

II. Methods Involved

Apriori, a notable algorithm for ARM, is one of the frequently used processes of discovering an assortment of data properties which functionally is associated to bring about the data and are chiefly based on range of occurrences. But the extraction through the number of occurrences does not bring in the attention of the scholars, and to overcome this some more measures are included in the Apriori algorithm for an efficient mining of association rules. They are:

a) Weightage

Unlike the general transaction database which projects the total amount of characteristics by some number, the traditional algorithms like Apriori mine association rules utilizes a binary mapped database that depicts the occurrence of the data or an item in one course of an action, thus allowing to gather and verify some good number of information related to the characteristics of the data, that results in recurrent but few number of weight-age rules. Even in an ordinary user transaction, some times the data that has a good weight-age occurs rarely, but it must also be involved in the recurrent item set. This procedure is followed in our approach, for mining a subset that has high significance.

b) Utility

The individual utility (Gain) of the characteristics is the subsequent measure involved in the approach to give a good standard to the ARM.

c) Diminution

The individual Diminution that occurs when item failed to raise the utility (Gain), which balance the utility and provide actual gain of the characteristics is the subsequent measure involved to the approach to give a good standard to the ARM. Some of the service standards in the business would be neglected in a process of mining. As these rules, when mined without these service standards will lead to a plausible loss. Those standards are attained by this method through utility measure (U-gain). Weight-age and utility measures are individually incorporated in copious researches [21, 24, 25, 26] so as to make their methods more efficient but those procedures need high capability. These procedures are effectively utilized in this methodology to extract the association rules from a database.

III. Proposed Methodology

Assuming D as a database having n number of transactions T and m number of attributes I= [i1,i2 ,.....,im] with positive real number weights Wi. Ui specifies the profit associated with the i attribute of utility table U with m count of utility values.

The methodology based on weight-age and utility involves some key steps like:

Step1: Extraction of the association rules from D by utilizing Apriori.
Step 2: Generation of W-gain value.
Step 3: Generation of U-gain value.
Step 4: Generation of D-sum value.
Step 5: Generation of DUW-score through W-gain and U-gain.
Step 6: Deriving the vital association rules by taking UW-score into consideration.

a) Extraction of the Association Rules through Apriori

To begin with, association rules are excavated from a transaction database D with n transactions. We represent the database D as:

$$D = \begin{bmatrix} T_1 \\ T_2 \\ . \\ . \\ T_n \end{bmatrix} \quad (1)$$

Each transaction T in D encompasses with 'm' number of attributes I= [i1,i2 ,.....,im] related to it and each attribute i is symbolized by weights Wi.

To extract the rules, a typical Apriori algorithm is used in our methodology. A binary mapped database BT is applied for extracting the association rules in conventional Apriori process by which the initial database D is converted to binary mapped database BT such that it comprises of binary 0 and 1 denoting the non-existence and existence of attributes. By the succeeding equation the weights Wi are mapped onto the binary values.

$$B_r = \begin{cases} 0 \ if \ W_i = 0 \ \forall T_k \\ 1 \ if \ W_i \neq 0 \ \forall T_k \end{cases} \quad (2)$$

Consequently, an input to the Apriori algorithm [2] is produced by the binary mapped database BT for extraction of association rules which are processed in two steps of Apriori as follows:

- ***Recurrent Item set Generation:*** Produces min-support which signifies that each and every feasible set of attributes that comprise support value higher than a predefined threshold.
- ***Association Rule Generation:*** Produces min-confidence which signifies that association rules from the item sets that comprise confidence higher than a predefined threshold.

The composition of a typical association constraint is: $A \rightarrow B$, where A symbolizes the antecedent and B symbolizes the consequent and these are subset of the items in the binary mapped database, such that $A \subset I$, $B \subset I$ and $A \cap B = \phi$ and is construed as B co-existing if A exists. The support S and confidence C clasps the constraint $A \rightarrow B$ in the transaction database D, if item sets A and B are contained in S% and C % of the transactions. Therefore:

$$\text{Support} \quad (A \rightarrow B) = P(A \cup B) \quad (3)$$

$$\text{Confidence}\ (A \rightarrow B) = P(B|A) = \text{support}\ (A \cup B) / \text{support}\ (A) \quad (4)$$

The pseudo code for the Apriori algorithm is:

$I_1 = \{large\ 1-itemsets\};$

$for (k = 2; I_{k-1} \neq 0; k++)\ do\ begin$

$C_k = apriori - gen(I_{k-1});\ //\ New\ candidates$

$for\ all\ transactions\ T \in D\ do\ begin\ C_r = subset(C_k, T);$

$for\ all\ candidates\ c \in C_T\ do$

$c.count++;$

end

end

$I_k = \{c \in C_k \mid c.count \geq \min sup\}$

end

$Answer = \bigcup_k I_k;$

A k amount of association rules R=[R1,R2 ,.....,Rk] are produced with apriori algorithm and is sent as input to the successive part in the methodology for weight-age and utility calculation. Each attribute of k association rules of R the is determined with some measures, that is for an association rule R i of the form, [A, B] => C, where, A, B and C are considered as the attributes, the derivations U-gain, W-gain and UW-score are evaluated for each attribute A, B and C independently.

At first, a decremented arrangement is done for the produced k association rules considering their respected confidence level. The listing of rearranged association rules is specified by

S = { R1', R2 ',.., Rk ' }, $S \in R$, where conf (R1 ') $\geq$ conf (R2 ') $\geq$ conf (R3 ')….. $\geq$ conf (Rk ').

b) Generating the value of W- gain

At the start the initial rule R1' is chosen from the rearranged list S and the independent attributes of R1' are derived followed by the computation of W-gain.

Definition 1: Item weight (Wi): Item weight value Wi, is a nonnegative integer which is termed as the total magnitude evaluation of the attribute present in the transaction database D.

Definition 2: Weighted Gain (W-gain): W-gain is termed as the summation of weights of each item W i of an attribute that is involved in each and every transaction of the database D as referred in the given equation:

$$W - gain = \sum_{i=1}^{|T|} W_i \quad (5)$$

Here we term, w_i as the weight of item in an attribute and $|T|$ as the amount of transactions in the database D.

c) Generating the value of U-gain

Correspondingly the initial rule R1' is preferred from the rearranged list S and the independent attributes of R1' are derived. By considering the U-factor and the utility value Ui ,the value of U-gain for each character attribute is determined.

Definition 3: Item Utility (Ui): In general every character has a precincts of the gain related to that particular character or attribute and is delineated as the Item utility Ui.

Definition 4: Utility table U: A quantity of 'm' utility values Ui are encompassed in the utility table U with the attributes related in the transaction database D. We represent the utility table as:

$$U = \begin{bmatrix} U_1 \\ U_2 \\ \\ . \\ U_m \end{bmatrix} \quad (6)$$

Definition 5: Utility factor (U-factor): The constant value of utility factor (U-factor) is derived by the addition of every utility items (Ui) of the utility table U .We define it as:

$$U - factor = \frac{1}{\sum_{i=1}^{m} U_i} \quad (7)$$

Consider, m is the amount of attributes involved in the transaction database.

Definition 6: Utility Gain (U-gain): The calculation of an attribute's authentic utility by considering its U-factor is referred as the Utility Gain and we define it as follows:

$$U - gain = U_i \quad u - factor \quad (8)$$

For every attribute in the association rule R1' the value of U-gain is calculated.

Definition 7: Diminution table: A quantity of 'm' diminution values DMi are encompassed in the diminution table DM with the attributes related in the transaction database D. We represent the diminution table as:

$$DM = \begin{bmatrix} DM_1 \\ DM_2 \\ \\ . \\ DM_m \end{bmatrix} \quad (9)$$

Definition 8: Diminution factor (D-factor): The constant value of diminution factor (D-factor) is derived by the addition of every diminution items (DM_i) of the diminution table DM. We define it as:

$$D - factor = \frac{1}{\sum_{i=1}^{m} DM_i} \quad (10)$$

Consider, m is the no of attributes involved in the transaction database.

Definition 9: Diminution Sum (D-sum): The calculation of an attribute's authentic utility by considering its U-factor is referred as the Utility Gain and we define it as follows:

$$D - sum = DM_i * D - factor \quad (11)$$

For every attribute in the association rule R1' the value of D-sum is calculated.

d) Generation of EUW-score through W-gain, U-gain and D-sum

From the values derived by calculating W-gain, U-gain and D-sum for the each attribute, they are merged together into one value named as UW-score for every independent association rule.

Definition 10: Exact Utility Weighted Score (EUW-score): EUW-score is derived by computing the proportion between the addition of products of W-gain, U-gain and D-sum for each attribute in the association rule to the total amount of attributes present in the rule.

$$EUW - score = \frac{\sum_{i=1}^{|R|} (W - gain)_i * ((U - gain) - (D - sum))}{|R|} \quad (12)$$

Here, | R | denotes the amount of attributes present in the association rule.

The equations (5),(8) and (11) and (12) aimed to determine the W-gain, U-gain, D-sum and EUW-score These equations are looped for the remaining association rules R2 ' to R k' involved in the rearranged list S. And for total 'k' number of association rules in the rearranged list S will be calculated with a EUW-Score related to it and the association rules in the rearranged list S are consequently rearranged by taking EUW-score into consideration to get

$S' = \{R_1'', R_2'', ..., R_k''\}$, where

$EUW- Score(R_1'') \geq EUW- Score(R_2'') \geq EUW- Score(R_3'').....\geq EUW- Score(R_k'')$.

e) Deriving the vital association rules by considering EUW-score

As a final point, we choose certain rules from a group of important weighted utility association rules R_{EUW} in the rearranged list S', whose EUW-Score is greater than the predefined threshold. The

consequential values of the weighted and utility related association rules is given by

$R_{EUW} = \{R_{EUW1}, R_{EUW2}, R_{EUW3,.....,R_{EUWl}}\}$, where $k \geq l$ and $R_{EUW} \subseteq S'$.

The significant improvement in minimizing number of rules can be observable in following graphs.

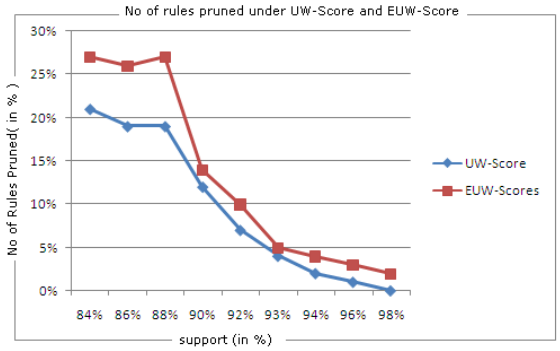

Fig 1 : % of Rules pruned by UW-Score and EUW-Score-line chart representation

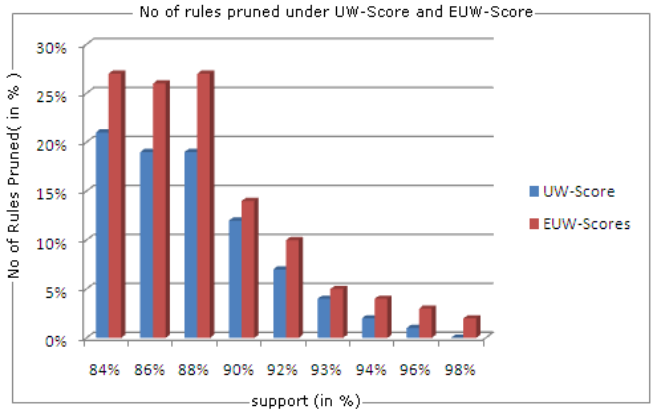

Fig 2 : % of Rules pruned by UW-Score and EUW-Score-bar chart representation

From fig 1 and fig 2, we can observe that number of rules minimized by EUW-Score is significantly better than UW-Score.

IV. Conclusion

By considering the weight factor, utility and diminution, the methodology used by us has given a chance to provide a proficient high utility association rules. At the outset, the planned methodology has enabled to make utilization of the conventional Apriori algorithm to create a group of association rules from a database. Depending on weightage (W-gain), utility (U-gain) and diminution (D-sum) complications a joint Exact Utility Weighted Score (EUW-Score) is generated for each association rule extracted. Considering the EUW-Score generated, eventually a subset of notable association rules are derived at.

References Références Referencias

1. Ali Rajabzadeh Ghatari, Nasibeh Mohamadi, Aida Honarmand, Parviz Ahmadi and Nasibeh Mohamadi," Recognizing & rioritizing 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, 2009.
2. R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large databases", In proceedings of the international Conference on Management of Data, ACM SIGMOD, pp. 207–216, Washington, DC, May 1993.
3. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", In Proceedings of 20th International Conference on Very Large Data Bases, Santiago, Chile, pp. 487–499, September 1994.
4. [4] M. S. Chen, J. Han, P.S. Yu, "Data mining: an overview from a database perspective", IEEE Transactions on Knowledge and Data Engineering, vol. 8, no.6, pp. 866–883, 1996.
5. Chun-Jung Chu, Vincent S. Tseng and Tyne Liang, "Mining temporal rare utility Itemsets in large databases using relative utility thresholds", International Journal of Innovative Computing, Information and Control, Vol. 4, no. 11, November 2008.
6. Frawley, W., Piatetsky-Shapiro, G., Matheus, C., "Knowledge Discovery in Databases: An Overview", AI Magazine, fall 1992, pp.213-228, 1992.
7. Guangzhu Yu, Shihuang Shao and Xianhui Zeng, "Mining Long High Utility Itemsets in Transaction Databases", WSEAS Transactions On Information Science & Applications, vol.5, no. 2, pp.202-210, February 2008.
8. A.M.J. Md. Zubair Rahman and P. Balasubram, "Weighted Support Association Rule Mining using Closed Itemset Lattices in Parallel", International Journal of Computer Science and Network security, Vol.9 No. 3 pp. 247-253, March 2009.
9. Hong Yao, Howard J. Hamilton, and Cory J. Butz, "A Foundational Approach to Mining Itemset Utilities from Databases", In Proceedings of the Third SIAM International Conference on Data Mining, pp. 482-486, Orlando, Florida, 2004.
10. Jing Wang, Ying Liu, Lin Zhou, Yong Shi, and Xingquan Zhu, "Pushing Frequency Constraint to Utility Mining Model", In Proceedings of the 7th international conference on Computational Science, pp.685 - 692, Beijing, China, 2007.
11. Jianying Hu and Aleksandra Mojsilovic, "High-utility pattern mining: A method for discovery of high-utility item sets", Pattern Recognition, vol. 40, no. 11, pp.3317-3324, November 2007.
12. Yu-Chiang Li , Jieh-Shan Yeh and Chin-Chen Chang, "Isolated items discarding strategy for discovering high utility itemsets", Data & Knowledge

Engineering, vol.64, no.1, pp.198-217, January 2008.
13. J. Han and Y. Fu, "Attribute-Oriented Induction in Data Mining," Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press, pp. 399- 421, 1996.
14. Bing Liu, Wynne Hsu, Ke Wang, and Shu Chen, "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, 1999.
15. Shankar.S, Babu Nishanth, T. Purusothaman and Jayanthi. S., "A Fast Algorithm for Mining High Utility Itemsets", In proceedings of IEEE International Advance Computing Conference, Patiala,India, 2009.
16. Ke Sun and Fengshan Bai, "Mining Weighted Association Rules without Preassigned Weights", IEEE Transactions on Knowledge and Data Engineering, Vol. 20, no. 4, April 2008.
17. Cai. C.H., Fu. A.W.C., Cheng. C. H and Kwong. W.W, "Mining Association Rules with Weighted Items", In Proceedings of the International Symposium on Database Engineering and Applications, pp. 68-77, Cardiff, Wales, UK, July 1998.
18. Wang. W., Yang. J and Yu. P. S, "Efficient Mining of Weighted Association Rules (WAR)", In Proceedings of the KDD, pp. 270-274, Boston, MA, August 2000.
19. Lu, S., Hu, H., Li, F, "Mining Weighted Association Rules", Intelligent Data Analysis, vol.5 , no. 3, pp.211 - 225, August 2001.
20. M. Sulaiman Khan, Maybin Muyeba and Frans Coenen, "Fuzzy Weighted Association Rule Mining with Weighted Support and Confidence Framework", International Workshops on New Frontiers in Applied Data Mining, Osaka, Japan, pp. 49 - 61, May 20-23, 2009.
21. M. Sulaiman Khan, Maybin Muyeba and Frans Coenen, "A Weighted Utility Framework for Mining Association Rules", In proceedings of European Symposium on Computer Modeling and Simulation,pp.87- 92,Liverpool, September 2008.
22. Oded Z. Maimon, Lior Rokach, "Decomposition Methodology for Knowledge Discovery and Data Mining: Theory and Applications", World Scientific Publishing Company, ISBN-13: 9789812560797, May 2005.
23. Feng Tao, Fionn Murtagh and Mohsen Farid, "Weighted Association Rule Mining using Weighted Support and Significance Framework", In Proceedings of the International Conference on Knowledge Discovery and Data Mining, pp.661 - 666, Washington, 2003.
24. Hong Yaoa and Howard J. Hamilton, "Mining itemset utilities from transaction databases", Data & Knowledge Engineering, vol. 59, no.3, pp. 603-626, December 2006.
25. Chun-Jung Chua, Vincent S. Tsengb and Tyne Liang, "An efficient algorithm for mining high utility itemsets with negative item values in large databases", Applied Mathematics and Computation, vol. 215, no.2, pp. 767-778, 2009.
26. Unil Yuna, "Efficient mining of weighted interesting patterns with a strong weight and/or support affinity", Information Sciences, vol.177, no. 17, pp. 3477-3499, September 2007.
27. Sandhu, P.S.; Dhaliwal, D.S.; Panda, S.N.; Bisht, A.; , "An Improvement in Apriori Algorithm Using Profit and Quantity," Computer and Network Technology (ICCNT), 2010 Second International Conference on , vol., no., pp.3-7, 23-25 April 2010.