A Framework for Incremental Mining of Interesting Temporal Association Rules

Ahmed Sultan Al-Hegami
Associate Professor of
IntelligentInformation Systems
,University of Sana'a, Yemen.

ABSTRACT
Association rules are an important problem in data mining. Massively increasing volume of data with temporal dependencies in real life databases has motivated researchers to design novel and incremental algorithms for temporal association rules mining.

In this paper, an incremental association rules mining algorithm is proposed that integrates interestingness criterion during the process of building the model called SUMA. One of the main features of the proposed framework is to capture the user background knowledge, which is monotonically augmented. The incremental model that reflects the changing data over the time and the user beliefs is attractive in order to make the over all KDD process more effective and efficient. The proposed framework is implemented and experiment it with some public datasets and found the results quite promising.

General Terms
Knowledge discovery in databases (KDD), Data mining, Incremental Association rules, Temporal association rule, Domain knowledge, Interestingness, Novelty measure.

Keywords
Knowledge discovery in databases (KDD), Data mining, Incremental Association rules, Temporal association rule, Domain knowledge, Interestingness, Novelty measure.

1. INTRODUCTION
Association rule mining is one of the most important techniques that aims to extract interesting correlations, frequent patterns, associations among sets of items in the transaction databases. The task of association rules mining usually performed in two steps process. The first step aims at finding all *frequent* itemsets that satisfy the minimum support constraint. The second step involves generating association rules that satisfy the minimum confidence constraint from the frequent itemsets.

One of the main drawbacks with the classical association rules algorithms is that they do not consider the time in which the data arrive. In practice, data is acquired in small batches over the time. In such scenario a combination of old and new data is used to build a new model from scratch.

As time advances, some old transactions may become obsolete and thus are discarded from the database. Consequently, some previously discovered knowledge (PDK) becomes invalid while some other new rules may show up. Researchers therefore have been strongly motivated to propose techniques that update the association rule model as new data arrives, rather than running the algorithms from scratch [1,2,3], resulting in incremental models.

Incremental algorithms build and refine the model as new data arrive at different points in time, in contrast to the traditional algorithms where they perform model building in batch manner [1,3]. The incremental association rules algorithms that reflect the changing data trends and the user beliefs are attractive in order to make the over all KDD process more effective and efficient.

Temporal association rule adds the time constraint to association rules. The database storing temporal information would be named as temporal database. The traditional association rule mining ignores the time characters of data; however, the application areas are always changing with time. The task of temporal association rule discovery, both association rules and temporal features are expected to be extracted from the database. Practically, it is too expensive to find all possibly hidden temporal association rules from large databases without any restrictions.

Although objective measures are commonly used techniques to discover interesting patterns in most KDD endeavours, they are partially effective unless combined with subjective measures of interestingness. Capturing the user subjectivity in dynamic environment is a challenging task to KDD community with respect to the time and user. Thus, at two different points in time and/or for two different end users, *interestingness* of the discovered knowledge may vary. The in which user background knowledge is monotonically augmented to reflect the changing data over the time, have addressed the issue of user subjectivity to help the user focus on the interesting knowledge [4,5,6,8].

In this paper, an incremental algorithm, SUMA, is proposed. It is based on the premise that unless the underlying data generation process has changed dramatically, it is expected that the rules discovered from one set are likely to be similar (in varying degrees) to those discovered from another set [2,8,10]. Interestingness measures can be used as an effective way to filter the rule set discovered from the target data set thereby, reducing the volume of the output. This paper extends the approaches presented in [8,10,11] and integrates it into temporal association rule algorithm in an incremental manner. It pushes the novelty measure during temporal association rule mining to form a constraint to the algorithm in order to discover only novel rules. The proposed approach is a self-upgrading filter that keeps *known knowledge* (previously discovered knowledge (*PDK*) and the user domain knowledge (*DK*) rule base updated as new rules discovered.

The SUMA algorithm operates on the incremental dataset and builds a model. During the frequent itemsets generation, the algorithm computes the novelty measure of interestingness against the *known knowledge* and prunes the items that do not meet the user interest. A detailed description of computation of novelty measure can be found in [10]. Further iteration of

the algorithm is performed only for the frequent itemsets that have novelty degree higher than a user specified threshold.

The incremental nature of the proposed algorithm makes it advantageous to discover interesting patterns at current time with respect to the previously discovered patterns (rules), rather than exhaustively discovering all patterns.

2. PROBLEM STATEMENT

Given a dataset D collected over the time $[t_0,t_1,t_2,\ldots t_n]$. At t_0, D_0 represents an empty database. At time instance t_i, an incremental dataset D_i , $i \in \{1,\ldots,n\}$, is collected such that $D = D_1 \cup D_2 \cup \ldots D_i$. Let T_i and T_{i+1} be two models discovered at time instances t_i and t_{i+1} from datasets $\bigcup_{j=1}^{i} D_j$ and $\bigcup_{j=1}^{i+1} D_j$ respectively. The major volume of discovered rules in T_i and T_{i+1} would be similar to some extent. A small set of rules, which are either present or absent in T_{i+1} represents change in data characteristics. The objective is to update T_i to T_{i+1} using D_{i+1} and T_i. T_i — the model discovered at time t_i now represents PDK. T_{i+1} is the up-to-date model obtained by adding interesting rules discovered from D_{i+1}. This is achieved by constructing a model $\check{T}_{i+1}$ from D_{i+1} such that association rules in $\check{T}_{i+1}$ have user specified degree of shocking interestingness with respect to the rules in T_i. Subsequently, $\check{T}_{i+1}$ is used to update T_i to T_{i+1}.

3. RELATED WORK

There are many approaches that discovered temporal association rules [11, 12, 14, 13]. In [11], authors limit the total transactions to the ones belonging to the items' lifetime. Those associations would be now discovered, as they would count on enough support. Another difficulty is the large number of rules that could be generated, for which many solutions have been proposed. Using age as an obsolescence factor for rules helps reduce the number of rules to be presented to the user. They expand the notion of association rules to incorporating time to the frequent itemsets discovered. The concept of temporal support is introduced and the known algorithm Apriori is modified to incorporate the temporal notions.

[12] presents an algorithm "T-Apriori", which based on time constraint, is designed and implemented on the basis of analyzing the related definitions and general steps of temporal association rule mining.

In [14], a SPFA algorithm (Standing for Segmented Progressive Filter Algorithm) is proposed .The basic idea behind SPFA is to first segment the database into sub-databases in such a way that item in each sub-database will have either the common starting time or the common ending time. Then, for each sub-database, SPFA progressively filters candidate 2-itemsets with cumulative filtering thresholds either forward or backward in time. This feature allows SPFA to adopt the scan reduction technique by generating all candidate k-itemsets from candidate 2-itemsets directly.

In [13], a framework is proposed to derive temporal rules from time series. The approach is based on episode rule mining that discovers temporal rules from time series in the frequency domain using the discrete cosine transform. The rules are then translated to temporal relations between time series patterns of arbitrary length.

Several approaches have been proposed for developing incremental algorithms of association rules mining [15,16,17,18,19,20]. The main assumption of these approaches is to update the discovered model when new data arrive. In [15], DEMON algorithm is proposed that works effectively and efficiently with evolving data over the time. [16] proposed an incremental algorithm that uses statistical methods for updating process called DELI algorithm. DELI algorithm applies a sampling technique to estimate the support counts using an approximate upper/lower bounds on the amount of changes in the set of newly introduced association rules. A low bound would mean that changes in association rules are small and there should be no maintenance. In addition, these algorithms are incremental in nature as they use and reuse the previously discovered knowledge and integrated when new data occur. In [17], a parallel approach is designed that reduces the computational requirement in the algorithm.

In [19,20], the FUP (Fast Update) algorithm is proposed for the maintenance of discovered association rules in large databases. FUP handles incremental database by scanning the database to check whether there are large itemsets or not. FUP algorithm is introduced for quantifying the large itemsets in the updated database. The goal of this algorithm is to solve the efficient update problem of association rule in updated database.

The novelty measure of interestingness of discovered patterns is studied in many proposals [10,21, 22,23, 25].

In this research, the novelty measure of the discovered rules presented in [10,21,22,23,25,26] is used and framework is proposed that utilizes this measure in temporal and incremental association rule mining in order to find not only accurate knowledge but also comprehensible and novel knowledge.

4. THE SUMA FRAMEWORK

In this paper ,the SUMA framework is proposed. It uses the novelty measure of interestingness presented in [10,21, 22,23,25] as a constraint to discover temporal association rules that are novel and hence interesting in an incremental manner.

One of the main features of the proposed approach is to capture the user background knowledge, which is monotonically augmented. The temporal rules that reflects the changing data and the user beliefs is attractive in order to make the overall KDD process more effective and efficient.

At each level of generating temporal frequent itemsets, partial temporal rules[1] are generated from the frequent itemsets at that level. These partial temporal rules are evaluated using confidence measure and prune the partial temporal rules that do not satisfy this measure resulting in a set of strong partial temporal association rules. The strong temporal rules are subjected to the novelty criterion [10,21, 22,23,25] in order to decide either these rules are interesting or not. The novelty measure (NM) of a partial temporal rule (P) is computed to the closest rule (R) in the existing model (M). If the partial temporal rule is found not interesting, it computes the Relevance Factor (RF) [25] of the partial temporal rule (P) to decide if the partial temporal rule will continue and will be used in the second stage or not. Computation of Relevance factor of P indicates the relevance of rule R with respect to current itemset in database (D). If the Relevance factor is acceptable, the partial temporal rule will not continue to be

1 We use partial rule because the fully temporal association rule has not been generated so far.

used in the second stage and hence will be pruned. The general architecture of the of proposed framework is described in Figure 1.

At time T, database D is pre-processed and subjected to the algorithm. the algorithm takes into account the existing model M, the known temporal association rules. The algorithm expands only those temporal itemset that are likely to lead to novel rules. For each temporal frequent itemset, a rule is extracted and used to update the model M.

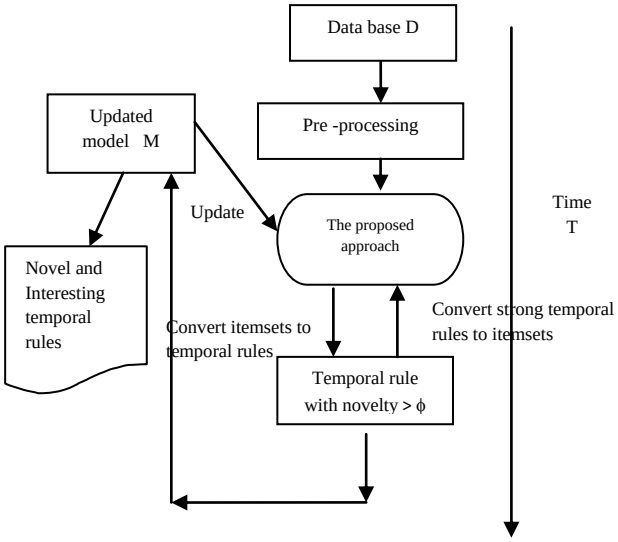

Fig 1: SUMA framework

4.1 Generation of Frequent Itemsets

The proposed algorithm is based on the T-Apriori algorithm [11]. We compute the frequent itemset as in equation (1) [11, 12]:

$$S(X,I_x,D) \geq \sigma \text{ and } |I_x| \geq \tau \quad (1)$$

Where,

$S(X,I_x,D)$=The support of X in D over lifespan I_x.

S= set of transaction in D.

X=Item in S.

I_x=The period or lifespan of item in D.

D=The transaction database.

σ=The threshold of support $\sigma \in [0, 1]$

τ= The threshold of temporal support.

For every temporal frequent itemsets, the following steps are performed:

1. Extracting the strong partial temporal rules which have confidence higher than the confidence threshold value.
2. Computation of Novelty measure (NM) of the partial temporal rules (P) with respect to the existing model M,
3. Computation of Relevance factor of the partial temporal rule (P) with with $RF_P^R = 0$.

The process of candidate generation is continued as the traditional T- Apriori algorithm, taking into account the only novel frequent items.

4.2 Generation of Temporal Association Rules

To generate temporal rules for the partial or complete temporal rule, it is necessary to find frequent itemset and then compute a confidence for rules through the time. The strong (partial or complete) temporal rules have confidence higher than the confidence threshold value. We compute the confidence of itemset as in equation (2) [11, 12]:

$$conf(X \Rightarrow Y, [t1,t2], D) = S(X \cup Y, IX \cup Y, D) / S(X, Ix, D) \quad (2)$$

Where,

$I_{X \cup Y} = [t_1, t_2]$

$S(X \cup Y, I_{X \cup Y}, D)$ =The transaction in D that contain $X \cup Y$ over lifespan $I_{X \cup Y}$.

S(X,Ix,D) = The transaction in D that contain X over lifespan I_X.

4.3 Dynamic Pruning Based On Novelty Criterion

The characteristic feature of the proposed approach is its ability to facilitate pruning based on novelty [8,10]. The objective is to reduce the size of the frequent itemsets generation with assurance that the resulting rules does not compromise in terms of accuracy and provides the user with novel temporal association rules.

The algorithm computes novelty measure (NM) at every stage of frequent itemsets generation to determine whether an itemset is likely to lead to an interesting rule, or not. An itemset becomes a candidate for next stage frequent itemsets generation if its novelty measure (NM) value is 1 or the relevance factor of the closest rule in M is less than the relevance factor threshold value. An interestingness value of 1, of the partial temporal rule indicates that this rule is unlikely to expand to any existing temporal association rule. A novelty measure value (NM) of 0 of the partial temporal rule indicates that the partial rule is likely to expand to some existing temporal association rule.

We compute the novelty measure (NM) denoted by as in equation (3) [10]:

$$NM = \frac{\{|S_1| + |S_2| - 2*k\} + \sum_{i=1}^{k} \delta(c_1^i, c_2^i)}{|S_1| + |S_2|} \quad (3)$$

Where,

S_1 and S_2 be two conjunct sets with cardinalities $|S_1|$ and $|S_2|$ respectively.

K= the pairs of compatible conjuncts between S_1 and S_2.

(c^i_1, c^i_2) is the i^{th} pair of compatible conjuncts.

For more details of computation of novelty measure, reader can refer to our work presented in [10,21, 22,23, 25].

4.4 Relevance Factor of a Partial Temporal Rule

The algorithm computes the relevance factor (RF) at each partial temporal rule [25] P with $NM^{R}_{P} = 0$, to judge relevance of the rule in the current database D. The computation of relevance factor (RF) is required in order to be assured that the expected expansion of the current partial temporal rule is relevance in the database. A higher relevance factor of the expected expansion of a partial temporal rule with respect to the database indicates that the rule is still relevance at current time t. A smaller relevance factor, on the other hand, indicates that the expected expansion of the partial temporal rule is now obsolete and is not valid in this dataset. As a result, the itemset must be further expanded and may lead to an interesting rule.

Having computed the relevance factor of the partial temporal rule, the algorithm expands the itemset if the relevance factor is lower than the specified relevance factor threshold value and stops expanding otherwise.

We compute the relevance factor as in equation (4,5)[25]:

The relevance factor (RF) of $A \rightarrow R$ $[t_1,t_2]$is given as follows:

$$RF \quad (A \rightarrow R) = \frac{|\Gamma(A \cap R)|}{|\Gamma A|} \qquad (4)$$

where

$|\Gamma (A \cap R)|$ = the number of tuples that contain both A and R,

$|\Gamma A|$ =the number of tuples that contain antecedent A.

Let R^P be the partial rule and R^S $(A \rightarrow R)$ be the closest rule in T such that $NM(R^P,R^S) = 0$. Then $(R^P,R^S) = 0$. Then

$$RF \quad (R^P) = \frac{|\Gamma(A \cap R)|}{|\Gamma A|} \qquad (5)$$

5. EXPERIMENTAL STUDY

The framework is implemented and tested using real-life datasets. The system is built using c programming language.
We will explain our framework through the following experiments.

5.1 Experiment 1

This experiment compares the proposed approach and T-Aprior algorithm with the same dataset example in [11], with σ= 0.5, τ=3, θ=0.8 and ϕ=0.8.The dataset contains 6 transactions and 9 items each item has lifespan in the database defined by a time interval $[t_1,t_2]$. The main features of our proposed approach is to reduce the number of the discovered rules. The proposed approach is applied with different value of minimum support σ, minimum temporal τ , minimum confidence θ and novelty threshold ϕ as shown in table 1.

Table 1: Comparing our algorithm with T-Aprior algorithm with various thresholds

σ	τ	θ	ϕ	Our approach	T- Aprior
				Rule discovery	Rule discovery
0.4	3	0.7	0.7	3	2
0.5	3	0.8	0.8	1	3
0.6	4	0.6	0.8	2	4

5.2 Experiment 2

The second experiment was performed using Sick, Zoo and Monk datasets available at http://kdd.ics.uci.edu/. These datasets are considered as evolving with time, and partitioned into three increments: D_1, D_2 and D_3 assumed to have arrived at times t_1, t_2 and t_3 respectively. The proposed approach is against the three partitions of each datasets to compute the overall accuracy of the models. It has been shown that the accuracy is increased as the number of training instances increases. Figure 2(a,b,c) shows the change in the accuracy of the generated rules using Sick, Zoo and Monk datasets respectively.

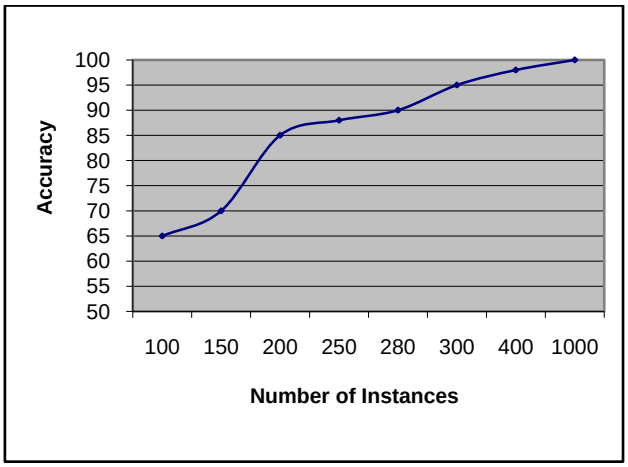

Fig. 2(a). Change in AR accuracy on the Sick dataset.

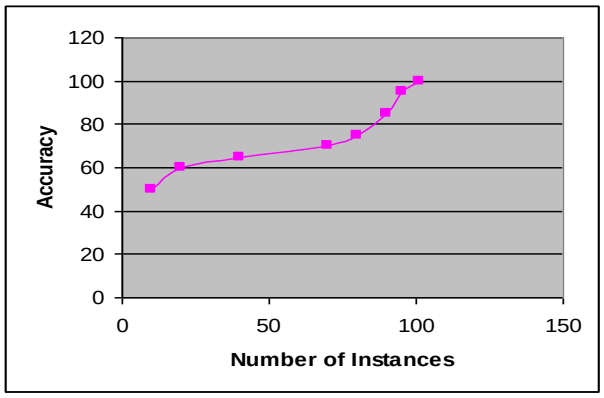

Fig. 2(b). Change in AR accuracy on the Zoo dataset.

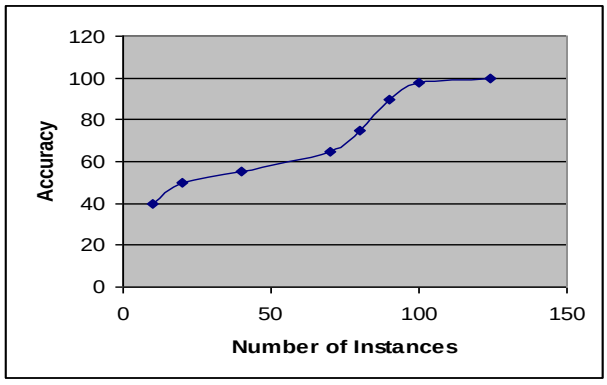

Fig. 2(c). Change in AR accuracy on the Monk dataset.

6. CONCLUSION

In this research a SUMA framework is proposed for mining temporal and incremental association rules based on novelty measure. The novelty criterion is pushed into a temporal association rules to reduce the volume of discovered rules.

The SUMA framework makes use of interestingness measure as the basis of extracting interesting patterns. This important feature of the proposed algorithm is attractive and desirable in many real life applications as the volume of data keeps on growing and changing over the time and therefore the user background knowledge is monotonically augmented. This changing environment updates the user understandability and comprehensibility about the domain.

7. REFERENCES

[1] Han, J. and Kamber, M.: Data Mining: Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers, (2001).

[2] Dunham M. H.: Data Mining: Introductory and Advanced Topics. 1st Edition Pearson Education (Singapore) Pte. Ltd. (2003).

[3] Hand, D., Mannila, H. and Smyth, P.: Principles of Data Mining, Prentice-Hall of India Private Limited, India, (2001).

[4] Liu, B. and Hsu, W. : Post Analysis of Learned Rules. In Proceedings of the 13th National Conference on AI (AAAI'96), (1996).

[5] Liu, B. and Hsu, W., Lee, H-Y. And Mun, L-F.: Tuple-Level Analysis for Identification of Interesting Rules. *Technical Report TRA5/95, SoC*. National University of Singapore, Singapore, (1996).

[6] Liu, B. and Hsu, W., Mun, L-F, and Lee, H-Y.: Finding Interesting Patterns Using User Expectations:. *Technical Report:TRA7/96*, Department of Information Systems and Computer Science, National University of Singapore, (1996).

[7] Kaur H., Wasan. S. K, Al-Hegami A. S., Bhatnagar, V.: A Unified Approach for Discovery of Interesting Association Rules. In Proceedings of Industrial Conference on Data Mining (ICDM), (2006).

[8] Al-Hegami, A. S.: Pushing Novelty Criterion into Incremental Mining Algorithm, International Journal of Computer Science and Network Security,Korea, VOL.7 No.12, December (2007).

[9] Yafi, E., Al-Hegami, A. S, Alam, M. A., and Biswas, R.: Incremental Mining of Shocking Association Patterns. In Proceedings of World Academy of Science, Engineering and Technology, Volume 37, Dubai, UAE, (2009).

[10] Bhatnagar, V., Al-Hegami A. S., and N. Kumar: A hybrid approach for Quantification of Novelty in Rule Discovery. In Proceedings of International Conference on Artificial Learning and Data Mining (ALDM'05), Turkey, Feb. 25-27, pp 39-42 (2005).

[11] Ale, J. M. and Rossi, G. H. : An approach to discovering temporal association rules. In Proc. of the 2000 ACM Symposium on Applied Computing, pages 294–300, (2000).

[12] Lianga, Z. , Xinming, T., Lin, L. and Wenliang, J.: Temporal Asocation Rule Mining Based On T-Aprior Algorthim And Its Typical Application. ,Chine, (2008).

[13] Dafas, P. A and d'Avila Garcez, A. S.: Applied temporal rule mining to time series, (2005).

[14] Hang, J. and Wei, W.: Efficient Algorithm for Mining Temporal Association Rule. International Journal of Computer Science and Network Security, VOL.7 No.4, (2007).

[15] Ganti, V., Gehrke, J. and Ramakrishnan, R.: DEMON: Mining and Monitoring evolving data. In Proceeding of the 16th International Conference on Data Engineering, San Diego, USA. (2000).

[16] Lee, S., and Cheung, D.: Maintenance of discovered association rules. When to update? In Research Issues on Data Mining and Knowledge Discovery. (1997).

[17] Zaki, M. and Hsiao, C.: Charm: An efficient algorithm for closed itemset mining. In Proceeding of the 2nd SIAM International Conference on Data Mining, Arlington, USA. (2002).

[18] Cheung, D. W., Han, J., Ng, V.T., Wong, C.Y.: Maintenance of discovered Association Rules in Large Databases: An Incremental Updating Technique, Proc. the International Conference On Data Engineering, (1996) 106-114.

[19] Cheung, D. W., Ng, V.T., Tam, B.W.: Maintenance of Discovered Knowledge: A case in Multi-level Association Rules, Proc. 2nd International Conference on Knowledge Discovery and Data Mining, (1996) 307-310.

[20] Cheung, D. W., Lee, S.D., Kao, B.: A general Incremental Technique for Mining Discovered Association Rules, Proc. International Conference on

Database System for Advanced Applications, (1997) 185-194.

[21] Bhatnagar, V., Al-Hegami A. S., and N. Kumar: Novelty as a Measure of Interestingness in Knowledge Discovery. In International Journal of Information Technology, Volume 2, Number 1, (2005).

[22] Al-Hegami, A. S.: Pushing Novelty Criterion into Incremental Mining Algorithm. International Journal of Computer Science and Network Security, Korea, VOL.7 No.12, (2007).

[23] Al-Hegami, A. S. : Subjective Measures and their Role in Data Mining Process. In Proceedings of the 6th International Conference on Cognitive Systems, New Delhi, India, (2004).

[24] Yafi, E., Al-Hegami, A. S, Alam, M. A., and Biswas, R.: YAMI: Incremental Mining of Interesting Association Patterns. In International Arab Journal of Information T, Vol. 9, No. 6, (2012).

[25] Al-Hegami, A. S.: On Quantification of Novelty in Knowledge Discovery Systems. Ph.D. Dissertation. Department of Computer Science, University of Delhi, India, (2006).

[26] Al-Hegami, A. S and Al-Ariki, S.: Constraint Based Mining of Interesting Temporal Association Rules. Submitted for publication in international conference of machine learning and applications (ICMLA), 2010, USA.