International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:11, 2007

SUPAR: System for User-Centric Profiling of Association Rules in Streaming Data

Sarabjeet Kaur Kochhar

***Abstract*—** With a surge of stream processing applications novel techniques are required for generation and analysis of association rules in streams. The traditional rule mining solutions cannot handle streams because they generally require multiple passes over the data and do not guarantee the results in a predictable, small time. Though researchers have been proposing algorithms for generation of rules from streams, there has not been much focus on their analysis.

We propose *Association rule profiling*, a user centric process for analyzing association rules and attaching suitable profiles to them depending on their changing frequency behavior over a previous snapshot of time in a data stream.

Association rule profiles provide insights into the changing nature of associations and can be used to characterize the associations. We discuss importance of characteristics such as predictability of linkages present in the data and propose metric to quantify it. We also show how association rule profiles can aid in generation of user specific, more understandable and actionable rules.

The framework is implemented as SUPAR: System for User-centric Profiling of Association Rules in streaming data. The proposed system offers following capabilities:

i) Continuous monitoring of frequency of streaming item-sets and detection of significant changes therein for association rule profiling.

ii) Computation of metrics for quantifying predictability of associations present in the data.

iii) User-centric control of the characterization process: user can control the framework through a) constraint specification and b) non-interesting rule elimination.

***Keywords*—**Data Streams, User subjectivity, Change detection, Association rule profiles, Predictability.

I. Introduction

STREAM databases depict characteristics such as online processing, continuity, rapid growth rates, infrequent access, and possibly non-persistent data. With a surge of stream processing applications in the areas of network traffic management, financial data, web-server logs, click streams, data feeds from sensor networks, and telecom call records, novel techniques are required for generation and analysis of association rules in streams.

Association rule profiling is a user centric process for analyzing association rules and attaching suitable profiles to them depending on their changing frequency behavior over a previous snapshot of time in a data stream. A *profile* is a mapping of the changes in support of an association over a user specified time, into a linguistic tag that represents an assessment about the past support behavior of the associations. Such an assessment gives insights into the changing nature of associations and can be used to characterize the associations. We discuss importance of characteristics such as predictability of linkages present in the data and propose metric to quantify it.

Another problem recognized as a challenge by the data mining community is that the traditional rule mining solutions produce a very large number of rules, most of which may be uninteresting for the user. Such an output is also difficult for the user to comprehend. The user must sift the output possibly using some post-processing filters for associations meaningful to him.

A solution to this problem is to lay emphasis on the generation of user specific rules. We show that association rule profiles can aid in meeting this goal. The rule profiles present a sketch of the nature of an association that is reflective of the changes in its popularity over a user specified period of time and increases understandability and action-ability of a rule. For example, knowledge of new and recurrent association rules is important for a retail chain to keep abreast of the latest trends while fading association rules may indicate declining trends in the market.

To meet the aforesaid goals, we implement our framework as SUPAR: System for User-centric Profiling of Association Rules in streaming data, to support characterization of association rules. The proposed system offers following capabilities:

i) Continuous monitoring of frequency of streaming item-sets for association rule profiling.

ii) Computation of metrics for quantifying predictability of associations present in the data.

iii) User-centric control of the characterization process: user can control the framework through a) constraint specification and b) non-interesting rule elimination.

SUPAR mines association rules using an extension of FP tree algorithm [12] for data streams. The user is involved during the pre-mining and mining phases of rule generation. He is allowed to specify his mining requirements through a set of constraints before mining. During the mining phase flexibility is provided to the user for elimination of non-

Manuscript received January 25, 2006. Sarabjeet Kaur Kochhar is a research scholar at Department of Computer Science, University of Delhi, Delhi, India. (phone: +91-9811444650; e-mail: skochhar@cs.du.ac.in).

interesting rules. The mined association rules are assigned profiles that are also used to compute metrics to characterize the evolving data.

A. Related Work

The problem of generating understandable rules has been identified as a serious problem and many solutions have been proposed in the past [2], [9]. Previous works include user subjectivity for rule pruning [8], constrained mining [7], [3] and specification of rule templates [4]. Interestingness measures such as minimum rule cover, usability, action-ability and unexpectedness [5], [6] have also been studied.

Our approach, instead of performing post rule analysis as has been seen in most of the above works, involves user during the pre-mining and mining processes to specify mining requirements through constraints and uses association rule profiles to eliminate entire families of rules that are not interesting for the user.

To the best of our knowledge, no work has yet been reported on gauging the predictability of evolving data. However, some important related work can be found in [10], [11].

B. Organization of the paper

The outline of the paper is as follows: Section II presents a 3-stage strategy for capturing user focus, Section III describes our approach to rule profiling and its applications. Section IV presents the overall architecture of SUPAR and Section V presents experimental study. Section VI concludes the paper.

II. User Focus

Given the high volume and unbounded characteristics associated with data streams, user focusing is very important for processing streams. Focusing allows user to communicate his interests and mining requirements to the system thus restricting the monitoring space. This leads to the generation of a small number of comprehend able and interesting rules that satisfy user requirements. We capture user interest through the following 3-stage filter which allows quick, fine grained control of system by the user and also yields benefits of space and time.

A. Stage I: Item Constraints.

The first stage of filter captures the user's current interest and communicates it to the rest of system. It allows the user to constrain the set of items that he wishes to study [7]. The user can dynamically select items of perceived importance or de-select items that do not currently seem interesting to him [1].

Definition 1: An item i ϵ I is a Base Item if it is selected for monitoring.

Let $\boldsymbol{B}$ be the set of base items, then $\boldsymbol{M^S}$, the maximal set of monitored item-sets is $\{Pow(\boldsymbol{B}) - \emptyset\}$.

.

B. Stage II: Support and Confidence Constraints.

The second stage of the filter allows the user to specify support and confidence thresholds. In real world applications, it may be interesting to study items in different support ranges. This requires providing user with a flexibility of setting multiple support monitoring ranges.

	Consistent	Novel	Reappearing	Soaring	Crumbling
Consistent	Consistent	Novel	Reappearing	Soaring	Crumbling
Novel	Novel	Novel	Novel	Novel	Novel
Reappearing	Reappearing	Novel	Reappearing	Reappearing	Reappearing
Soaring	Soaring	Novel	Reappearing	Soaring	Crumbling
Crumbling	Crumbling	Novel	Reappearing	Crumbling	Crumbling

Fig. 1 Profiling Matrix

A *monitoring range* is a closed interval bounded by two user specified support levels that consists of a support vector. Only the base items belonging to the monitoring range are selected for current study. The user specifies multiple support levels S^i_h and S^i_l that establish boundaries for multiple monitoring ranges R^i. The definition of $\boldsymbol{M^S}$ can now be refined as: $\boldsymbol{M^S} = \{ x | x \epsilon \boldsymbol{B} \Lambda (S^i_l \leq Supp(x) \leq S^i_h) \}$

C. Stage III: Rule Constraints.

A Specification of rule constraints is oriented towards generation of specific, useful knowledge. Rule constraints enable user to restrict the type of rules to be generated. Using rule constrains, output of the system can be restricted to Novel, Reappearing, Consistent, Soaring or Crumbling association rules (Formal definitions in Section III).

III. Rule Profiling

The mined knowledge at any given instant represents knowledge at the lowest level of abstraction. A higher level of abstraction of mined knowledge can give important insights into the overall nature of linkages in data. The mined support of base items can be analyzed to derive association rule profiles that represent knowledge at a higher level of abstraction. The concept of deriving knowledge by abstracting knowledge from lower levels of abstractions is called *knowledge differentiation* [1].

An association rule profile presents an overall sketch of the nature of an association rule of the form $\{x \rightarrow y | x, y \epsilon \boldsymbol{B}\}$, w.r.t. changes in its popularity over a user specified period of time. The profiles present an assessment about the long term support behavior of the associations based on the consolidation and analysis of the history of changes in frequency of the corresponding base items in the antecedent and consequent of the rule.

Rule profiles are also used to honor rule constraints and for computation of some useful metrics that can help characterize the underlying data generation process. The association rules are pro¬ led as a two-step process. Base items present in the antecedent and consequent of rules are first categorized on the basis of support behavior and then assigned a profile. These two steps are described in the following two subsections.

TABLE 1
EXPERIMENTAL RESULTS

Time Instant	Average Previous Support T $\sum PS^i_j / T$ j=0	Support Differential T $\Delta^i_s = CS^i - (\sum PS^i_j / T)$ j=0	Base Item Category	Generated Rules	Rule Summary
To	-	-	-	-	-
T1	-	-	-	-	-
T2	-	-	-	Profile: Novel a->b a->c a->d b->c b->d c->d	Total No. Of Rules Generated: 6 No. Of Soaring Rules: 0 No. Of Consistent Rules: 0 No. Of Novel Rules: 6
T3	a = 0.43 b = 0.4 c = 0.3 d = 0.43 e - f -	0.7 -0 .43 = 0.27 0.7 - 0.3 = 0.4 0.4-0.3 = -0.1 0.4-0.43 = -0.03 - -	Soaring Soaring Consistent Consistent - -	Profile: Soaring a->b a->c a->d b->c b->d Profile: Consistent c->d Profile: Novel a->e b->e c->e d->e	Total No. Of Rules Generated: 10 No. Of Soaring Rules: 5 No. Of Consistent Rules: 1 No. Of Novel Rules: 4
T4	a = 0.53 b = 0.56 c = 0.4 d = 0.43 e = 0.33 f -	0.8 -0 .53 = 0.27 0.6 - 0.56 = 0.04 0.5-0.4 = -0.1 0.5-0.43 = 0.07 0.6 – 0.33 = 0.27 -	Soaring Consistent Consistent Consistent Soaring -	Profile: Soaring a->b a->c a->d a->e b->e c->e d->e Profile: Consistent b->c b->d c->d	Total No. Of Rules Generated: 10 No. Of Soaring Rules: 7 No. Of Consistent Rules: 3 No. Of Novel Rules: 0
T5	a = 0.63 b = 0.6 c = 0.43 d = 0.43 e = 0.46 f = 0.03	0.7 -0 .63 = 0.07 0.3 - 0.6 = -0.3 0.5 - 0.43 = 0.07 0.4 - 0.43 = -0.03 0.7 – 0.46 = 0.24 0.3 – 0.03 = 0.27	Consistent Crumbling Consistent Consistent Soaring Novel	Profile: Soaring a->e c->e d->e Profile: Consistent a->c a->d c->d Profile: Novel a->f b->f c->f d->f e->f	Total No. Of Rules Generated: 11 No. Of Soaring Rules: 3 No. Of Consistent Rules: 3 No. Of Novel Rules: 5

Predictability is thus an important characteristic of an evolving database. It indicates whether the given current state of data is likely to persist in future. It may be noted that this is different from estimation of the future values of data.

Predictability may be very important in certain application domains. For example, in a stock market scenario, let portfolio of a customer consist of the stocks he is interested in (Please note that this portfolio maps to the set of Base items in our framework).

The user may be interested in monitoring the sales volume of the selected stocks or any associations between them. But in a stock market, the trends in the sales frequency of stocks as well as the associations observed are always likely to change. Predictability of stock data implies knowing how stable and therefore predictable the data is i.e. whether the trends observed today are likely to be stay or not, whether the associations observed between the different stocks are likely to change tomorrow or not. Needless to say such an information can provide valuable inputs to a prospective investor.

Intuitively more stable the data, more its predictability [10]. Based on the rule profiles, we propose Predictability Quotient to quantify the predictability as a function of consistency of the underlying data.

Predictability Quotient Þ at time t can be given as:

$$\text{Þ} = \frac{\text{Count (Consistent Associations)}}{\text{Count (Total Associations).}}$$

. The association rules with consistent profile indicate the consistency of the linkages present in the data and hence can be used for quantification of predictability. A fraction of associations that are consistent give an indication of the stability and hence predictive capability of underlying data.

IV. SUPAR

The SUPAR system, consists of three distinct components. These components viz. focusing unit, mining unit and the rule-generation unit, collaborate and cooperate on a continuous basis to facilitate rule generation and profiling.

A. Focusing Unit

The focusing unit of SUPAR system comprises of a 3-stage Filter. The focusing unit captures and communicates user interest to both mining unit and the rule-generation unit. It allows the user to specify constraints on the SUPAR as per his requirements and/or interests. This unit shares data structures with the other two units, in order to guide them in their respective tasks.

B. Mining Unit

The mining unit continuously mines the data stream for current support of item-sets and saves the mining results in a synopsis data structure. It maintains the support count of all singleton items in transactional database D and all item-sets of base items. We use an extension of the FP tree algorithm to mine supports of the item-sets in streaming data [12].

C. Rule-generation Unit

The rule-generation unit analyzes the mined knowledge from the mining unit to generate association rules and categorize them as Novel, Fading, Deviating, Consistent or Re-appearing association rules. Working in accordance with the rule constraints communicated by the focusing unit, this unit produces only specified type of rules as constrained by the user.

V. Experimental Study

We developed multi-threaded C++ program to implement the proposal. The program was compiled using gcc compiler and executed under Red Hat Linux 7.3 operating system. The hardware environment consisted of 2.3 GHz AMD Athlon XP processor and 256 MB DDR RAM. The program was run in a stand alone environment, with no other user process. The program consists of two threads. The first thread simulates both Focusing unit and the Mining unit (since these program units must run in mutually exclusive mode). The second thread simulates Rule generation unit and computes predictability statistics.

The experimental study was performed using a synthetically generated dataset. A customized data generating program was required to control the frequency behavior of item-sets selected for study. Some pre-configured variations were introduced in the support of base items to demonstrate the capability of SUPAR system to detect them faithfully. The program allows the user to control the size, cardinality and average transaction length of the data-set. The generated data-set contained 1000 items and 10 million transactions which were continually streamed in for performing the experiments.

TABLE II
SUPPORT OF BASE ITEMS

	a	b	c	d	e	f
T0	0.4	0.5	0.4	0.4	0.2	0.1
T1	0.5	0.5	0.4	0.5	0.2	0.1
T2	0.4	0.5	0.4	0.4	0.3	0.0
T3	0.7	0.7	0.4	0.4	0.5	0.0
T4	0.8	0.6	0.5	0.5	0.6	0.1
T5	0.7	0.3	0.5	0.4	0.7	0.3

Following parameters were supplied by the user through a parameter file:

1. Set of Base items, B = {a, b, c, d, .e, f.} (The set was chosen to be small so that the results could be illustrated.) The support for all the item-sets except for the members of set B was random. The support of Base items was varied as shown in Table II.
2. $S^l = 0.3$, $S^h = 0.9$
3. Rule Constraints: Soaring, Consistent and Novel Profiles.
4. $\delta = 0.2$

5. PS-window size = 3

Table I presents the results of the experimental study. The results are generated at each time instant Ti after an initial gestation time of 3 time units from T0 – T2 (since PS-window size = 3).

Experiment1: First Experiment was performed to show the efficacy of SUPAR's rule interestingness and elimination techniques. As illustrated in the results, a very small number of rules are generated in accordance with the rule constraints. This is in contrast with the usual rule mining algorithms which produce a voluminous, uncomprehendable and unclassified output.

Experiment2: The second experiment was performed to show that the system is able to capture the temporal nature of rules with the evolving data. At time T2 all the rules generated were Novel (table I) which kept on changing their profiles with the changes in the support of underlying base items (as shown in table II).

VI. Conclusion

In this paper we introduced Association rule profiles to characterize association rules based on their changing frequency behavior in a data stream and demonstrated their use for quantifying the predictability of evolving data and for generating user specific, more understandable and actionable rules.

The SUPAR framework captures changes in the discovered support trends and analyzes them to assign profiles to association rules. It also recognizes the importance of user involvement in this process and uses a 3-stage constraint specification strategy to capture user subjectivity. The Experimental study shows that the system is able to achieve its stated objectives.

References

[1] V. Bhatnagar and S. K. Kochhar. User subjectivity in change modeling of streaming itemsets. In ADMA, pages 812ñ823, 2005.
[2] G. Dong and J. Li. Interestingness of discovered association rules in terms of neighborhood based unexpectedness. In PAKDD, pages 72ñ86, 1998.
[3] B. G. Helsinki. Interactive constrained association rule mining.
[4] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo. Finding interesting rules from large sets of discovered association rules. In N. R. Adam, B. K. Bhargava, and Y. Yesha, editors, Third International Conference on Information and Knowledge Management (CIKM'94), pages 401ñ407. ACM Press, 1994.
[5] B. Liu, W. Hsu, and Y. Ma. Identifying non-actionable association rules. In KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pages 329ñ334, New York, NY, USA, 2001. ACM Press.
[6] B. Liu, W. Hsu, L.-F. Mun, and H.-Y. Lee. Finding interesting patterns using user expectations, Knowledge and Data Engineering, 11(6):817ñ832, 1999.
[7] R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints. In D. Heckerman, H. Mannila, D. Pregibon, and R. Uthurusamy, editors, Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining, KDD, pages 67ñ73. AAAI Press, 14ñ17 1997.
[8] H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hatonen, and H. Mannila. Pruning and grouping of discovered association rules, 1995.
[9] P. S. M. Tsai and C.-M. Chen. Mining interesting association rules from customer databases and transaction databases. Inf. Syst., 29(8):685ñ696, 2004.
[10] D. Kifer, S.B.David and J.Gehrke. Detecting Change in Data Streams, in proc. Of VLDB 2004
[11] A. Tsymbal. The Problem of Concept Drift: Definitions and Related Work, 2004
[12] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. Technical Report TR-99-12, Computing Science Technical Report, Simon Fraser University, October 1999.