Chapter VII
Continuous Post-Mining of Association Rules in a Data Stream Management System

Hetal Thakkar
University of California, Los Angeles, USA

Barzan Mozafari
University of California, Los Angeles, USA

Carlo Zaniolo
University of California, Los Angeles, USA

ABSTRACT

The real-time (or just-on-time) requirement associated with online association rule mining implies the need to expedite the analysis and validation of the many candidate rules, which are typically created from the discovered frequent patterns. Moreover, the mining process, from data cleaning to post-mining, can no longer be structured as a sequence of steps performed by the analyst, but must be streamlined into a workflow supported by an efficient system providing quality of service guarantees that are expected from modern Data Stream Management Systems (DSMSs). This chapter describes the architecture and techniques used to achieve this advanced functionality in the Stream Mill Miner (SMM) prototype, an SQL-based DSMS designed to support continuous mining queries.

INTRODUCTION

Driven by the need to support a variety of applications, such as click stream analysis, intrusion detection, and web-purchase recommendation systems, much of recent research work has focused on the difficult problem of mining data streams for association rules. The paramount concern in previous works was how to devise frequent itemset algorithms that are fast and light enough
Continuous Post-Mining of Association Rules in a Data Stream Management System

for mining massive data streams continuously with real-time or quasi real-time response (Jiang, 2006). The problem of post-mining the association rules, so derived from the data streams, has so far received much less attention, although it is rich in practical importance and research challenges. Indeed, the challenge of validating the large number of generated rules is even harder in the time-constrained environment of on-line data mining than it is in the traditional off-line environment. On the other hand, data stream mining is by nature a continuous and incremental process, which makes it possible to apply application-specific knowledge and meta-knowledge acquired in the past, to accelerate the search for new rules. Therefore, previous post-mining results can be used to prune and expedite both (i) the current search for new frequent patterns and (ii) the post-processing of the candidate rules thus derived. These considerations have motivated the introduction of efficient and tightly-coupled primitives for mining and post-mining association rules in Stream Mill Miner (SMM), a DSMS designed for mining applications (Thakkar, 2008). SMM is the first of its kind and thus must address a full gamut of interrelated challenges pertaining to (i) functionality, (ii) performance, and (iii) usability. Toward that goal, SMM supports

- A rich library of mining methods and operators that are fast and light enough to be used for online mining of massive and often bursty data streams,
- The management of the complete DM process as a workflow, which (i) begins with the preprocessing of data (e.g., cleaning and normalization), (ii) continues with the core mining task (e.g., frequent pattern extraction), and (iii) completes with post-mining tasks for rule extraction, validation, and historical preservation.
- Usability based on high-level, user-friendly interfaces, but also customizability and extensibility to meet the specific demands of different classes of users.

Performing these tasks efficiently on data streams has proven difficult for all mining methods, but particularly so, for association rule min-

![Figure 1. Post-mining flow](image-url)
Related Content

Computing Join Aggregates Over Private Tables
[www.igi-global.com/article/computing-join-aggregates-over-private/1816?camid=4v1a](www.igi-global.com/article/computing-join-aggregates-over-private/1816?camid=4v1a)

User-Centric Similarity and Proximity Measures for Spatial Personalization
[www.igi-global.com/article/user-centric-similarity-proximity-measures/42152?camid=4v1a](www.igi-global.com/article/user-centric-similarity-proximity-measures/42152?camid=4v1a)

Weak Ties and Value of a Network in the New Internet Economy
Davide Di Fatta, Roberto Musotto, Vittorio D’Aleo, Walter Vesperì, Giacomo Morabito and Salvatore Lo Bue (2017). *Social Media Data Extraction and Content Analysis* (pp. 66-84).

Improved Data Partitioning for Building Large ROLAP Data Cubes in Parallel
[www.igi-global.com/article/improved-data-partitioning-building-large/1761?camid=4v1a](www.igi-global.com/article/improved-data-partitioning-building-large/1761?camid=4v1a)