Random Sampling over Data Streams for Sequential Pattern Mining

Chedy Raïssi
LIRMM, EMA-LGI2P/Site EERIE
161 rue Ada
34392 Montpellier Cedex 5, France
France
raissi@lirmm.fr

Pascal Poncelet
EMA-LGI2P/Site EERIE
Parc Scientifique Georges Besse
30035 Nîmes Cedex, France
Pascal.Poncelet@ema.fr

February 16, 2007

Abstract

In recent years the emergence of new real-world applications such as network traffic monitoring, intrusion detection systems, sensor network data analysis, click stream mining and dynamic tracing of financial transactions, calls for studying a new kind of model. Named *data stream*, this model is in fact a continuous and potentially infinite flow of information as opposed to finite and statically stored data sets. We study the problem of sequential pattern mining in data streams. This problem has been extensively studied for the conventional case of disk resident data sets. In the case of data streams, this problem becomes more challenging as the volume of data is usually too huge to be stored on permanent devices, main memory or to be scanned thoroughly more than once. In this case, it may be acceptable to generate approximable solutions for our mining problem. In this paper we introduce a new approach based on biased reservoir sampling to achieve a more efficient mining of sequential patterns. Furthermore, we theoretically prove that our biased reservoir size is always bounded whatever the size of the stream is. This property often allows us to keep the entire relevant reservoir in main memory. We also show a simple algorithm to build the biased reservoir for the special case of sequential pattern mining. Experimental evaluation supports the claim that sequential pattern mining based on biased reservoir sampling needs small memory requirements. Besides, we also propose an adapted approach to handle the case of mining sequential patterns in a sliding window model. The experiment show that the results are accurate.

1 Introduction

Recently, the data mining community has focused on a new challenging model where data arrives sequentially in the form of continuous rapid streams. It is often referred to as data streams or streaming data. Since data streams are continuous, high-speed and unbounded flow of informations, it is often impossible to mine patterns with classical algorithms that require multiple scans. As a consequence new approaches were proposed to mine itemsets [5, 3, 2, 4, 8] using different approaches based on the *landmark*, *sliding windows* or *time-fading* models. However, few researches focused on sequential patterns extraction over data streams. In this paper we consider that transactions are ordered into the streams and grouped under different identifiers. We