Mining Sequential Patterns from Data Streams: a Centroid Approach

Alice Marascu Florent Masseglia

INRIA Sophia Antipolis
2004 route des Lucioles - BP 93
06902 Sophia Antipolis, France
E-mail: {Alice.Marascu,Florent.Masseglia}@sophia.inria.fr

Abstract

In recent years, emerging applications introduced new constraints for data mining methods. These constraints are typical of a new kind of data: the *data streams*. In data stream processing, memory usage is restricted, new elements are generated continuously and have to be considered as fast as possible, no blocking operator can be performed and the data can be examined only once. At this time only a few methods has been proposed for mining sequential patterns in data streams. We argue that the main reason is the combinatory phenomenon related to sequential pattern mining. In this paper, we propose an algorithm based on sequences alignment for mining approximate sequential patterns in Web usage data streams. To meet the constraint of one scan, a greedy clustering algorithm associated to an alignment method is proposed. We will show that our proposal is able to extract relevant sequences with very low thresholds.

Keywords: data streams, sequential patterns, web usage mining, clustering, sequences alignment.

1 Introduction

The problem of mining sequential patterns from a large static database has been widely addressed [2, 11, 14, 18, 10]. The extracted relationship is known to be useful for various applications such as decision analysis, marketing, usage analysis, etc. In recent years, emerging applications such as network traffic analysis, intrusion and fraud detection, web clickstream mining or analysis of sensor data (to name a few), introduced new constraints for data mining methods. These constraints are typical of a new kind of data: the *data streams*. A data stream processing has to satisfy the following constraints: memory usage is restricted, new elements are generated continuously and have to be considered in a linear time, no blocking operator can be performed and the data can be examined only once. Hence, many methods have been proposed for mining items or patterns from data streams [6, 3, 5]. At first, the main