Chapter 7.2
Approximate Range Querying over Sliding Windows

Francesco Buccafurri
University "Mediterranea" of Reggio Calabria, Italy

Gianluca Caminiti
University "Mediterranea" of Reggio Calabria, Italy

Gianluca Lax
University "Mediterranea" of Reggio Calabria, Italy

ABSTRACT

In the context of Knowledge Discovery in Databases, data reduction is a pre-processing step delivering succinct yet meaningful data to sequent stages. If the target of mining are data streams, then it is crucial to suitably reduce them, since often analyses on such data require multiple scans. In this chapter, we propose a histogram-based approach to reducing sliding windows supporting approximate arbitrary (i.e., non biased) range-sum queries. The histogram is based on a hierarchical structure (as opposed to the flat structure of traditional ones) and it results suitable to directly support hierarchical queries, such as drill-down and roll-up operations. In particular, both sliding window shifting and quick query answering operations are logarithmic in the sliding window size. Experimental analysis shows the superiority of our method in terms of accuracy w.r.t. the state-of-the-art approaches in the context of histogram-based sliding window reduction techniques.

INTRODUCTION

It is well known that data pre-processing techniques, when applied prior to mining, may significantly improve the overall data mining results. This is particularly true in the context of data stream mining, where data comes continuously and mining may be done on the basis of sliding windows including only the most recent data (Babcock & Babu, 2002; Cohen, 2006; Lin, 2005). In order to operate on meaningful

data, an important issue is keeping the sliding windows size as large as possible. As a consequence, any technique capable of both reducing (i.e., compressing) sliding windows, maintaining a good approximate representation of data distribution inside them, and smoothing possible outliers, is certainly notable in the field of data stream mining, because provides a number of advantages. On the one hand, reducing sliding windows allows to simultaneously keep more than just one approximate sliding window, in order to implement *similarity queries* or *change mining queries* (Bulut, 2005; Dong, 2003), which are useful to perform trend analysis of the data stream. On the other hand, since in a typical streaming environment only limited memory resources are available (Garofalakis, 2002; Li, 2005), reduction is a key factor enabling the processing of queries which require multiple scans on data.

Thus, several properties emerge that a sliding window reduction technique has to satisfy:

1. The reduced sliding window should maintain in a certain measure the *semantic nature* of original data, in such a way that meaningful queries for mining activities can be submitted to reduced data in place of original ones.
2. For a given kind of query, accuracy of the reduced structure should be enough independent of the position where the query is applied. Indeed, mining needs the possibility of freely querying data.
3. The reduction technique should not limit too much the capability of drilling-down and rolling-up data.

In this chapter, we propose a histogram-based technique to reduce sliding windows. Our approach supports approximate arbitrary range-sum queries satisfying all the above properties. Observe that range-sum queries represent a class of queries very frequent in the field of data stream mining. Our histogram, differently from traditional ones, is based on a hierarchical temporal structure, referenced to as *c-tree*, which is a binary tree whose nodes contain, in a hierarchical fashion, pre-computed range-sum queries, that are stored by approximate (via bit-saving) encoding. Thus, range-sum queries are either embedded in the histogram or derivable from such embedded queries by means of linear interpolation. As a consequence, the c-tree structure directly supports the estimation of arbitrary range-sum queries.

Concerning data reduction, it results both from data aggregation implemented by leaves of the tree (discretization), and from the saving of bits which is obtained by representing range queries with less than 32 bits (assumed enough for an exact representation). In particular, the number of bits used to represent range queries decreases as the level of the tree increases.

The c-tree structure is designed as dynamic. Each operation updating the c-tree to the current sliding window can be applied in logarithmic time w.r.t. the window size, in the worst case. Moreover, answering to a range query requires at most logarithmic time too.

Observe that the c-tree hierarchical structure directly supports querying at different abstraction levels, thus allowing drill-down and roll-up operations.

Bucket summarization smoothes each data value by consulting the "neighborhood" or values around it, thus enforcing data noise reduction.

Finally, the main feature we have to remark for our histogram concerns its accuracy. Indeed, in order the reduction technique to have significance, error should be either guaranteed or heuristically shown to be low (and this is our case), compared with that of the state-of-the-art techniques.

BACKGROUND

Data streams is an emergent issue that in the last years has captured the interest of many scientific communities. The crucial problem, arising in sev-