Summarizing A 3 Way Relational Data Stream

Baptiste Csernel[1][2], Fabrice Clerot[2] and Georges Hébrail[1]

[1] École Nationale Supérieure des Télécommunications, 46 rue Barrault, 75013 Paris
Département Informatique et Réseaux.
[2] France Télécom R&D, 2 avenue P. Marzin, 22307 Lannion

Abstract. In this paper, we present a novel method to summarize a group of three data streams sharing a relational link between each other. That is, using the relational data base model as a reference, two entity streams and one stream of relationships. Each entity stream contains a stream of entities, each one of them referenced by a key, much in the same way as a primary key references all objects in a regular database table. The stream of relationships contains key couples identifying links between objects in the two entity streams as well as attributes characterising this particular link.

The algorithm presented here produces not only a summary of both entity streams considered independently of each other, but also gives a summary of the relations existing between the two entity streams as well as information on the join obtained joining the two entity streams through the relationship stream.

Keywords: Relational Data Mining, Data Stream Mining, Clustering, Data Stream Summaries, Data Stream Joins.

1 Introduction

The last decades have seen a huge inflation in the amount of information generated by most commercial processes and the rates at which it is produced. This has led to the development of a new field in the data analysis community devoted to the study of infinite streams of data, arriving at a rhythm so fast, and so large that they can't fit in storage and thus have to be treated in one pass. This new field, called data stream analysis has been the subject of a growing attention by different communities including but not limited to those of databases, data mining or machine learning.

This work studies the design of summaries built from such data streams. Much work has already been done to design algorithms capable of producing summaries of any given data streams. However, most real world data does not stand on its own but includes references to different data produced by different streams. That's why we have chosen to interest ourselves in the summary not of a single data stream, but of several data streams joined by relationships. To simplify the problem, we will only consider here a small example of three data streams, one relationship stream, and two entity streams.

2 Related Work

This work is related to much previous work done on data streams in the past few years [1] [2], however, it shares particular relations with two specific problems. The first one is the summary of a single data stream, and the second, the problem associated with the join of two data streams. Both of these problems have been studied separately. However, while the problem treated here might seem like the conjunction of the two previously cited, it is a new and different problem. The goal here is not to first join two streams and then summarize the resulting stream but to make a summary of the streams without having to process the join but while still taking into account the relationship information.

This particular problem has not been much considered yet to our knowledge, and that is why we have decided to propose a solution for it.