Data Stream Synopsis Using SaintEtiQ*

Quang-Khai Pham, Noureddine Mouaddib, and Guillaume Raschia

LINA - Polytech'Nantes
ATLAS-GRIM Group
2, rue de la Houssinire, BP 92208
44 322 Nantes cedex 03, France
{Quang-Khai.Pham, Noureddine.Mouaddib, Guillaume.Raschia}@univ-nantes.fr

Abstract. In this paper, a novel approach for building synopses is proposed by using a service and message-oriented architecture. The SaintEtiQ summarization system initially designed for very large stored databases, by its intrinsic features, is capable of dealing with the requirements inherent to the data stream environment. Its incremental maintenance of the output summaries and its scalability allows it to be a serious challenger to existing techniques. The resulting summaries present on the one hand the incoming data in a less precise form but is still on the other hand very informative on the actual content. We expose a novel way of exploiting this semantically rich information for query answering with an approach mid-way between blunt query answering and mid-way between data mining.

1 Introduction

Emerging applications are generating and exploiting data in the form of *data streams*. Such information, as opposed to the traditional way of managing information, is by essence on-line, potentially unlimited and unbounded. Interesting domains of application include network traffic surveillance and administration, financial analysis, sensor data feeds or web applications. The constraints of these application are related to their need for timely answers for decision making purposes. For example, when a broker has to decide whether to buy or sell stocks according to the evolution of different indexes, he needs a final answer within seconds or tens of seconds. When considering the streams, each one corresponding to the real-time evolution of a stock market index, generated by all the stock markets, it is virtually impossible to store the transiting data. Thus, we pinpoint the need for adapted structures for managing this versatile data.

While the data stream domain is relatively new, as opposed to the traditional database paradigm, synopses remain relatively unformalized. Gibbons and Matias define in [6] those, for a class of queries Q, as a function$f(n)$ that provides *(exact or approximate) answers to queries from* Q *that uses* $O(f(n))$ *space for a data set of size* n*, where* $f(n) = O(n^{\epsilon})$ *for some constant* $\epsilon < 1$. The evaluation criteria proposed are thus (i)the coverage of $f(n)$, (ii)the answer quality, (iii)the

* This work supported by the ACI APMD and SemWeb.

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 530–540, 2006.

footprint, (iv)the query time and (v)the computation/update time (the reader is invited to refer to [6] for further details). We retain criteria (i), (ii), (iii) and (v) as a basis to evaluate and position our system comparatively to existing approaches. The challenge that then rises is to find a compromise optimizing these opposing criteria.

Motivating example. A certain number of present applications and needs motivate our research in this direction.

As an example, it is well-known that the only way for car industries to make profit is to reduce their costs. An option is to introduce more and more on-board electronics to replace previously hydraulic and pneumatic systems. Thus, Bosch is working on *motronic* (motorised + electronic) brakes that would be completely independent and would communicate via a wireless connection to a central on-board system. These are monitored in real-time through the sensor feeds they send to the server. Considering the limits of these on-board electronics, it is impracticable to store the data and necessary to provide timely answers.

Formulation of the problem. Recent applications are providing and using more and more input feeds in the form of data streams either structured or not. Due to the host system physical limitations(hard drive access times, computing capabilities, etc...) and the timely answers required, it is not realistic to try to deal with these new inputs with the existing techniques that have been developed for the traditional stored databases and/or data warehouses. On the other hand, even though timely answers are required by such applications, 100% exact answers may not be necessary.

Thus, the main problem for managing and exploiting data streams can partially be answered by providing data structures, called *synopses*, and algorithms to construct and maintain them, in a time and space cost-efficient fashion. These are part of the major criteria retained but these solutions also need to propose answers adapted to the requirements of the application considered, such as the class of queries addressed, and guarantee to a certain extent the quality of the answers provided.

Roadmap. The rest of the paper is organized as follows. First, an overview of the existing techniques and of their performances in designing synopses for data streams is presented in section 2. Then, we will discuss the SaintEtiQ approach for building summaries and the novel decision making-oriented paradigm that we propose for exploiting SaintEtiQ summaries in a data stream environment through section 3. Our perspectives and short & long term work will be introduced in section 4. Finally we will conclude this paper in section 5.

2 Related Work

Recent works on data streams have mostly focused on designing synopses, algorithms and/or (adapting) techniques from the traditional database domain for estimating one dimensional data values.