Stream Summarization Framework

CS641 Project Proposal (Fall 2002)

Mohamed G. Elfeky   Ihab F. Ilyas   Mohamed F. Mokbel
Department of Computer Sciences, Purdue University
West Lafayette, IN 47907-1398
{mgelfeky, ilyas, mokbel}@cs.purdue.edu

October 24, 2002

1 Problem Statement

Summarizing data streams is an essential task in streams handling and processing [2]. As the number of streams increases and the rate at which data is pushed into the system increases, it is infeasible to store every single data item or even to process it in time. Building summaries and aggregates on incoming data becomes increasingly important in practically answering user queries [3]. Changing the data processing modules to process the compued aggregates on input data streams becomes an efficient approach to answer user queries on large number of data streams [1].

In this project we try to address the following problem: How to efficiently build a stream summarization framework to efficiently represent the incoming streams in a manageable form that can be used to answer user queries. Several challenges are there in building such framework.

- The framework should be flexible and extensible. Different ways of summarization can be plugged to best suit the application. Summaries include histograms [6, 7], wavelet transform [5], sampling [4] and others.
- How to schedule incoming streams and what are the different parameters that can be introduced to prioritize the processing of each stream.
- What are the different summarization parameters that can be introduced to tune the performance of answering the user queries. Example parameters are the sampling rate, bin sizes of histogram, the truncation rate, etc.

2 Our Approach

The main purpose of this study is to build a prototype of a stream summarization system that can receive many data streams as input, schedule these streams for processing and builds summaries of the incoming data to be used in query answering. The goal is to use the system to explore the different research issues in streams summarization and perform a limited sets of experiments to show the effect of the different system parameters. A strawman architecture of the proposed prototype is depicted in Figure 1.

References

[1] Alin Dobra, Minos N. Garofalakis, Johannes Gehrke, and Rajeev Rastogi. Processing complex aggregate queries over data streams. In *Proc. of the intl. conf. on Management of data, SIGMOD*, Madison, WI, May 2002.

[2] Minos N. Garofalakis and Johannes Gehrke. Querying and mining data streams: You only get one look. In *Proc. of the 28th Intl. Conf. on Very Large Data Bases, VLDB*, Hong Kong, August 2002.

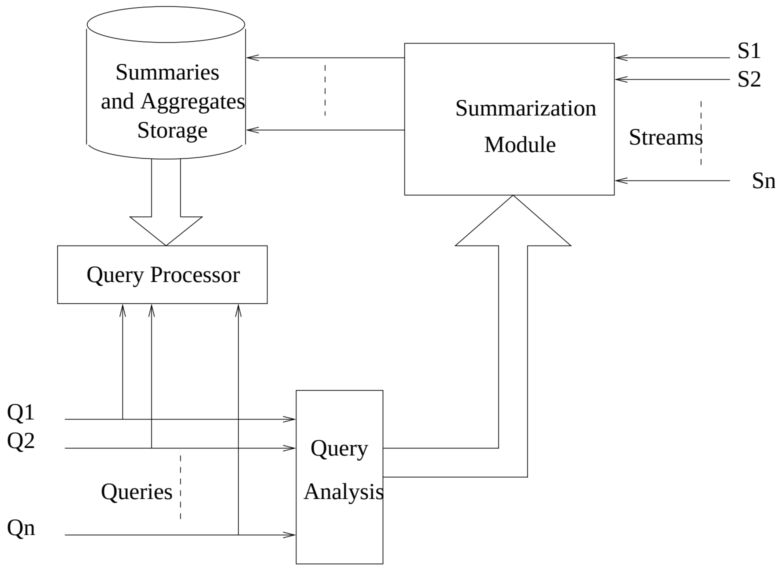

Figure 1: The proposed streams summarization architecture.

[3] Johannes Gehrke, Flip Korn, and Divesh Srivastava. On computing correlated aggregates over continual data streams. In *Proc. of the intl. conf. on Management of data, SIGMOD*, Santa Barbara, CA, May 2001.

[4] Phillip B. Gibbons and Yossi Matias. New sampling-based summary statistics for improving approximate query answers. In *Proc. of the intl. conf. on Management of data, SIGMOD*, Seattle, WA, June 1998.

[5] Anna C. Gilbert, Yannis Kotidis, S. Muthukrishnan, and Martin Strauss. Surfing wavelets on streams: One-pass summaries for approximate aggregate queries. In *Proc. of the 27th Intl. Conf. on Very Large Data Bases, VLDB*, Rome, Italy, September 2001.

[6] Gurmeet Singh Manku and Rajeev Motwani. Approximate frequency counts over data streams. In *Proc. of the 28th Intl. Conf. on Very Large Data Bases, VLDB*, Hong Kong, August 2002.

[7] Viswanath Poosala Yannis E. Ioannidis. Histogram-based approximation of set-valued query-answers. In *Proc. of the 25th Intl. Conf. on Very Large Data Bases, VLDB*, Edinburgh, Scotland, September 1999.