Simultaneous Equation Systems for Query Processing on Continuous-Time Data Streams

Ying Zheng and Steve Gu
Instructor: Jun Yang
Feb 12, 2008

Outline

- Motivation
- Overview of Pulse System
- Query Processing
- Validation
- Experiments
- Conclusion

Motivation

- Many physical processes fundamentally continuous
  - E.g. temperature, trajectory of cars
- Continuous model to represent discrete data streams
  - Random access to data at arbitrary points in time
  - Compact representation of data as model parameters

Overview of Pulse System

- Piecewise polynomial model
  - Segmentation of stream data
  - N-th degree polynomial for an attribute "a" in i-th segment

$$a(t) = \sum_{i=0}^{n} c_{a,i} t^i$$

- Model parameters as inputs to continuous query processing
  - Filters, joins, aggregates

Overview of Pulse System

- Pre-compute the results given the predicted model
- Ignore tuples if we have predicted them
- Can we ignore more?
  - ensure that we do not miss any mispredicted tuples

Query Processing

- Selective operators
  - Basic idea: transform predicates into equations
    - Comparison operator R (i.e. <, >, ≤, ≥, =, !=)
    - X R Y -> X-Y R 0
- Filters

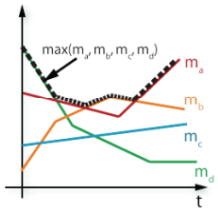
$$x(t) - c_0 = \sum_{i=0}^{n} c_{x,i} t^i - c_0 = \sum_{i=1}^{n} c_{x,i} t^i + (c_{x,0} - c_0)$$

- Joins
  - Implicit Constraint: X, Y should happen at the same time

$$x(t) - y(t) = \sum_{i=0}^{n} c_{x,i} t^i - \sum_{i=0}^{n} c_{y,i} t^i = \sum_{i=1}^{n} (c_{x,i} - c_{y,i}) t^i$$

Query Processing

- Min/Max aggregates
  - Single stream: compute derivatives on polynomial segments
  - Multiple streams: maintain a lower/upper envelope



Query Processing

- Sum aggregates
  - Sum of values in a segment

$$wf_{sum} = \int_{t-w}^{t} \sum_{i=0}^{n} c_{x,i} t^i dt = \sum_{i=1}^{n+1} \frac{c_{x,i-1}}{i} t^i \Big|_{t-w}^{t}$$

  - Sum of values spanning multiple segments
- Average aggregates

$$wf_{avg} = \frac{wf_{sum}}{w}$$

Query Processing

- Output: segments
- Produce discrete tuple answers by sampling
  - User-defined sampling rate
- Consequence
  - False-positive
  - False-negative

Validation

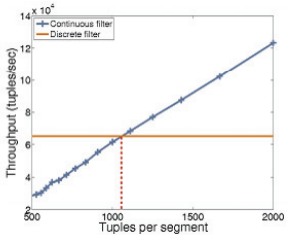
- Slack – for selective operators

$$\begin{aligned} &\mathbf{slack} = \min_t \|\mathbf{Dt}\|_\infty \\ &\text{s.t } t \in \bigcap [t^l, t^u)_i \qquad \forall i. [t^l, t^u)_{update} \cap [t^l, t^u)_i \neq \emptyset \end{aligned}$$

- Error bound inversion
  - invert error bounds at outputs to bound at inputs
  - using heuristics to apportion bounds across input

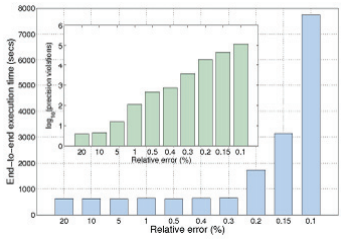
Experiment

- Synthetic data – filter query



Experiment

- NYSE dataset – stock trade prices



Conclusion

- Model data by user-specified continuous polynomial model
  - Compact representation of data
  - Capture data continuity
  - Prediction & pre-compute the answers
- Query Processing
  - Can do: Filter, join (equi-time constraint), aggregates (max/min, sum/average)
  - Cannot do: general join, aggregates involving frequency, mixed queries

Conclusion

- False positive & False negative
  - Tuple-based output
  - Due to continuous model
- Data validation to tolerate error
  - Ignore some inputs and use predicted answers for efficiency