



Finding Aggregates from Streaming Data in Single Pass

Medha Atre

Course Project for CS631 (Autumn 2002) under Prof. Krithi Ramamritham
(IITB).



Overview

- The need
- Type of solutions
- Choice of solution
- Problems addressed
- How does *wavelet transform* work ?
- Implementation
- Results



The need ...

- Huge data streams encountered at routers, telephone switches, stock exchanges etc.
- Necessity to analyze this data for trend-related analysis, and fraud detection.
- Analysis to be done as fast as possible for mission-critical tasks as detecting fraud, security breaches etc.
- What are the possible ways of analysis ?



Solutions ...

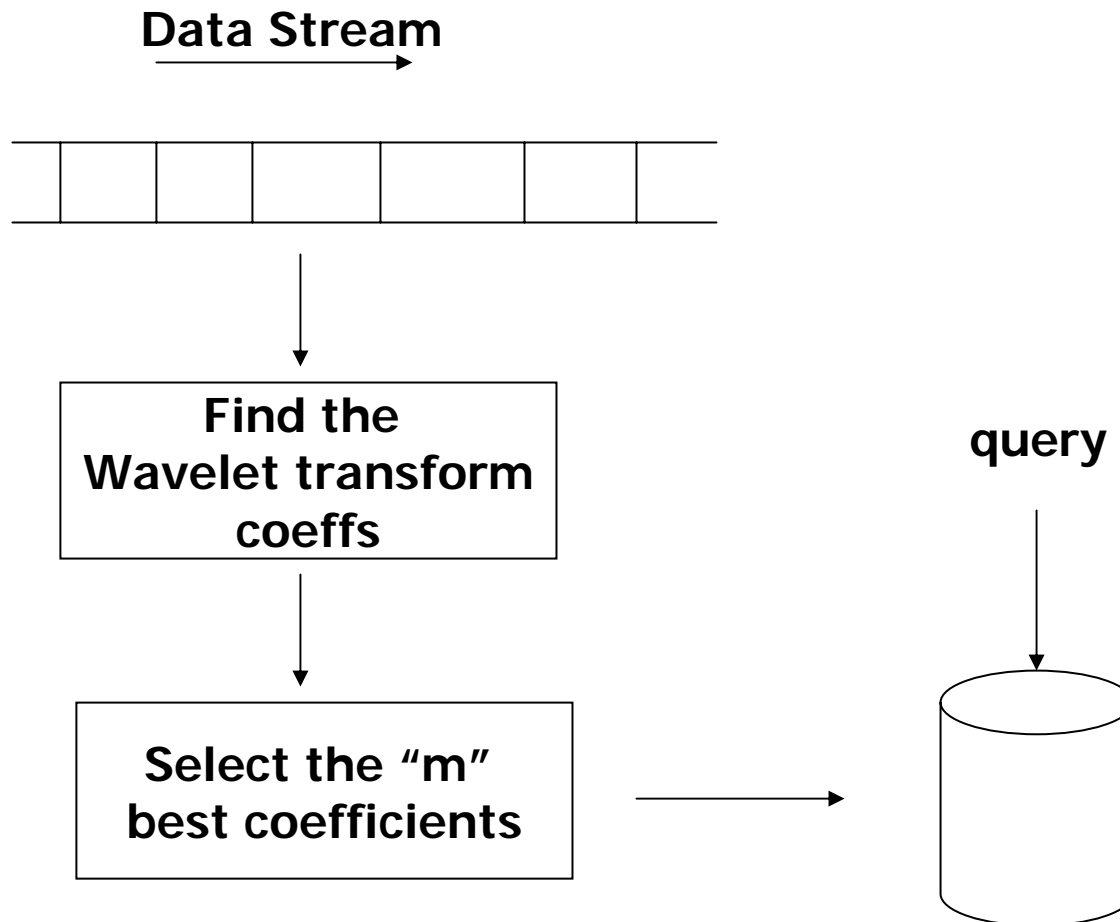
- Offline processing –
Archive whole data in real-time, and analyze it offline. (Slower w.r.t. basic motives of analysis i.e. fraud-detection and performance.)
- Real-time processing –
Analyze the data as it arrives ...
 - In Multiple passes – easiest method .. But slower and inefficient w.r.t. load of system.
 - In *Single pass* – Requires special implementation techniques .. But faster and efficient.



Real time processing of Data in Single pass

- Methods used – Wavelet Transform, Sampling techniques, MaxDiff algorithm.
- Why *Wavelet Transform* ? –
 - Storing *fairly approximate "sketch"* of data in smaller space.
 - Answering simple point and range queries with *quite good* approximation from stored "sketch".
 - Known to perform better than other techniques and easier for implementation. *Note: Comparative analysis of these techniques is outside the scope of this project.*

Block diagram of implementation technique

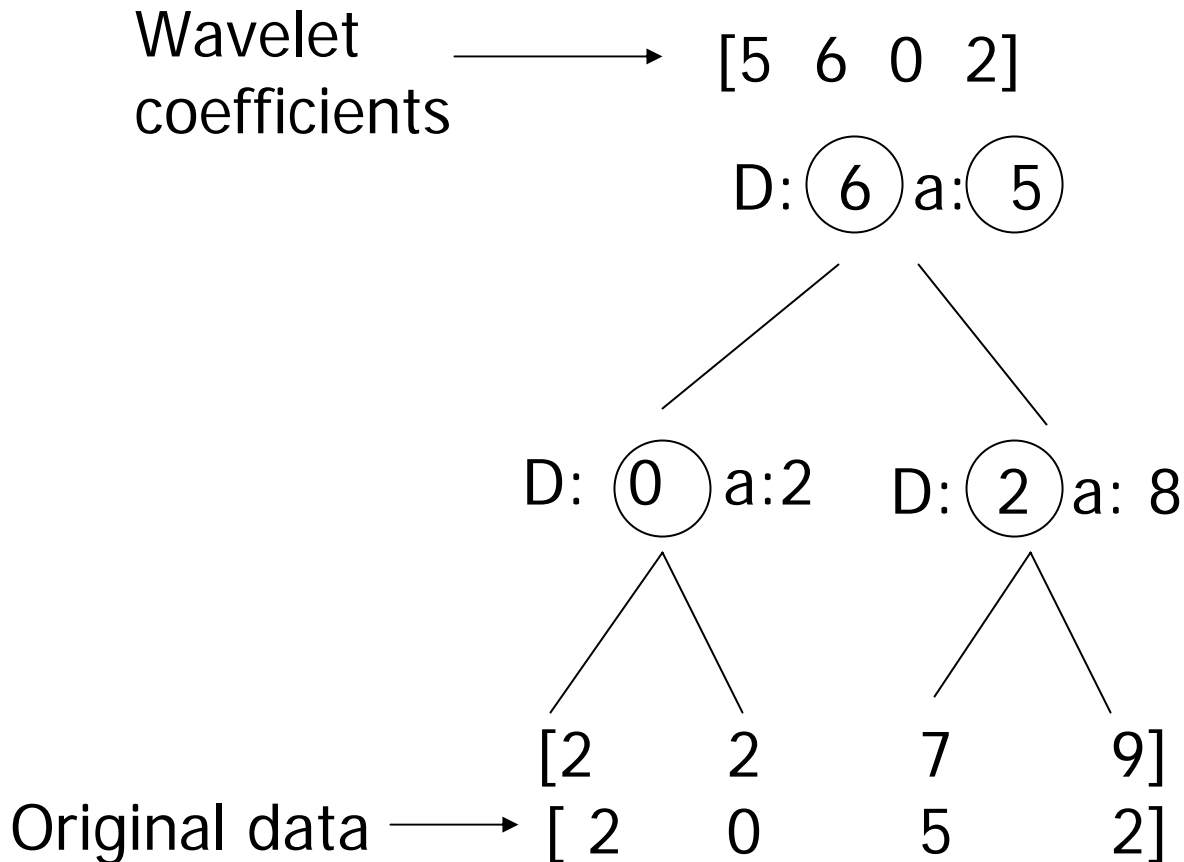


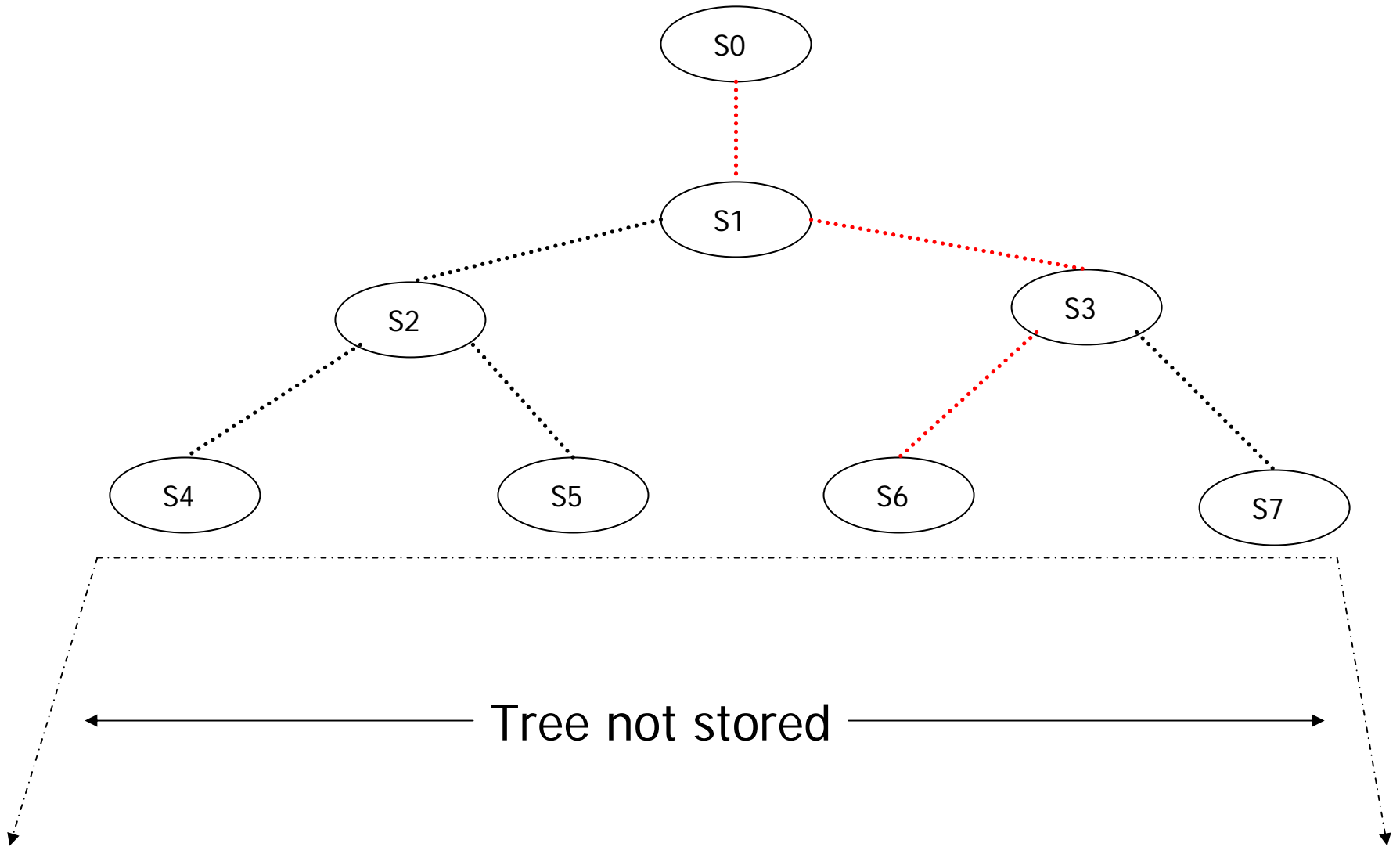


Key aspects of implementation ..

- Single pass over data (obviously!!) ☺
- At any point while processing data only $O(N)$ memory is used where N is the number of data items being considered.
- Selecting “ m ” best data coefficients out of N data items .. such that they give minimum error in retrieval of original value of data-key.
- Storing these “ m ” coefficients instead of all N data items ($m \ll N$).

How does Wavelet Transform work ..







How queries are answered ...

- Point queries – Find the number of calls made by telephone number 2422 5074
Find value of key 5 ..
$$Value(24225074) = S(0) + \frac{1}{2} * S(1) + \frac{1}{2} * S(3) - \frac{1}{2} * S(6)$$
- Range queries –
Find number of calls made from exchange 2422 ..
Answer to this query is the root of tree having numbers from exchange 2422 as leaves. i.e. S1, S2 etc.



Brief about implementation ..

- Data input from a file
- Reading file sequentially to simulate single pass over data-stream, and not accessing previous data of file.
- Forming the coefficient tree in the form of linked list.
- Storing “m” best coefficients.
- A program to calculate point and range queries from coefficient and answer back to user.



Few points to note ...

- Very basic implementation ... cannot handle data fed in any arbitrary format.
- Assumptions –
 - Assumes incoming data in key-value pair (e.g. key is tele-number 2422 5074, value is number of calls made from it in last 1 hr. "24225074" => 6
 - Incoming data stream is in ordered-aggregate form.
- Selection of "m" best coefficients changes according to data-stream types.



contd ...

- Currently we take highest first “m” coefficient by sorting them ... not the best approach.
- Multi-dimensional data-streams not considered (discussed in research papers referred for implementation).



Our results ...

N = 4096

VALUE OF $m = N$

RESULT OF THE QUERY: SELECT VALUE WHERE KEY = 1011

IS 85631.0

RESULT OF THE QUERY: SELECT VALUE WHERE KEY = 3067

IS 7505

THE RESULT OF THE QUERY: SELECT VALUE WHERE KEY = 2015

IS 1480.0

RESULT OF RANGE QUERY VALUE FROM 1016 TO 1021 IS : 20562.0

VALUE OF $m = 50\%$ of N

RESULT OF THE QUERY: SELECT VALUE WHERE KEY = 1011

IS 85630.0

THE RESULT OF THE QUERY: SELECT VALUE WHERE KEY = 3067

IS 8296.0

THE RESULT OF THE QUERY: SELECT VALUE WHERE KEY = 2015

IS 6323.0

RESULT OF QUERY VALUE FROM 1016 TO 1021 IS : 22416.0



contd ...

VALUE OF m = 25% of N

THE RESULT OF THE QUERY: SELECT VALUE WHERE KEY = 1011
IS 85630.0

THE RESULT OF THE QUERY: SELECT VALUE WHERE KEY = 3067
IS 8297.0

THE RESULT OF THE QUERY: SELECT VALUE WHERE KEY = 2015
IS 8338.0

RESULT OF QUERY VALUE FROM 1016 TO 1021 IS : 22415.938



References

- *Surfing Wavelets on Streams: One-Pass Summaries for Approximate Aggregate Queries.*
S. Muthukrishnan, A.C. Gilbert, Y. Kotidis, M. Strauss, 2001.
- *Wavelet-based histograms for selectivity estimation.*
J. Vitter, Y. Matias, M. Wang, 1998.



Thank you
