Wavelet-Based OLAP Approaches

Motivation: New Multidimensional Data Intensive Applications

- Multidimensional data sets: (with dimension & measure)
  - Remote sensory data (from JPL):
    - <latitude, longitude, altitude, time, temperature>
  - Sensor readings from GPS ground stations (from NASA):
    - <lat, long, t, velocity>
  - Petroleum sales (from Digital-Government research center):
    - <location, product, year, month, volume>
  - ACOUSTIC data (from UCLA sensor-network project):
    - <IPAQ-id, volume-id, event#, time, value>
  - Market data (from NCR): <store-location, product-id, date, price, sale>

- Large size, e.g., current (toy!) NASA/JPL data set:
  - Past 10 years, sampling twice a day, at a lat-long-alt grid of 64 * 128 * 16, recording 8 bytes of temperature & 16 bytes of dimensions
  - This is 6 MB of data per day; a total of 21 GB for 10 years
  - Increase: twice an hour sampling, 1024 * 4096 * 128 grid, …
Motivation: Multidimensional Applications

- I/O and computationally complex queries
  - Range-aggregate queries (with aggregate function)
    - Average temperature, given an area and time interval
    - Average velocity of upward movement of the station
    - Total petroleum sales volume of a given product in a given location and year
    - Number of jackets sold in Seattle in Sep. 2001
  - Tougther queries:
    - Covariance of temperature and altitude (correlation)
    - Variance of sale of petroleum in 2002 in CA

- Quick response-time (interactive):
  - The results can be approximate and/or progressively become exact

Recap!

- Multidimensional data
- Large data
- Aggregate queries
- Approximate answers
- Progressive answers
- Multi-resolution compression
- Wavelets!
Data Cube Approximation and Histograms via Wavelets


Outline

- Motivation
- Technique
- Complexity Analysis
Motivation

- There are a number of scenarios in which a user may prefer an approximate answer in a few seconds over an exact answer that requires tens of minutes or more to compute.
- Another consideration is that the data cube may be remote and currently unavailable, so that finding an exact answer is not an option, until the data again become available.

Technique

1. In a preprocessing step, they form the partial sum data cube $P$ from the (raw) data cube $A$. (In their method, they further process $P$ by replacing each cell value by its natural logarithm.)
2. They compute the wavelet decomposition of $P$, obtaining a set of $N$ coefficients, where $N$ is the size of array $A$.
3. They keep only the $C$ most significant wavelet coefficients, for some $C$ that corresponds to the desired storage usage and accuracy. The choice of which $C$ coefficients to keep depends upon the particular thresholding method they use.
1. Computing the Partial Sum Datacube

- Why wavelet decompose P and not A?
  - P is monotone nondecreasing, and a compact data cube built on P seems to give a better approximation than one built directly on A.
  - To answer a range-sum query using the compact data cube built on P, all they need to do in the online phase is to reconstruct the values corresponding to the boundaries of the ranges (instead of reconstructing all the values covered by the query, as in an extended data cube).

2. Wavelet Decomposition of the Partial Sum Data Cube P

- Wavelet basis function: Haar wavelets:
  - $h = \left[ \frac{1}{2}, \frac{1}{2} \right]$
  - $g = \left[ -\frac{1}{2}, \frac{1}{2} \right]$
- One dimensional “signal”: [2, 2, 7, 11]

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Averages</th>
<th>Detail Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>[2, 2, 7, 11]</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>[2, 9]</td>
<td>[0, 2]</td>
</tr>
<tr>
<td>1</td>
<td>[5½]</td>
<td>[3½]</td>
</tr>
</tbody>
</table>

- Wavelet transform of the original signal is the overall average of the original signal followed by the detail coefficients in the order of increasing resolutions: [5½, 3½, 0, 2]
2. Wavelet Decomposition of the Partial Sum Data Cube $P$

- Loss less: the original signal can be reconstructed
- If signal correlated, lots of zeros in details
- General: other filters, convolution, down-sampling
- Multidimensional transform:
  - Perform a series of one-dimensional decompositions.
  - For example, in the two-dimensional case, we first apply the one-dimensional wavelet transform to each row of the data. Next, we treat these transformed rows as if they were themselves the original data, and we apply the one-dimensional transform to each column.

3. Thresholding and Error Measures

- Keeping only $C \ll N$ coefficients
- Question: Which are the “best” $C$ coefficients to keep, so as to minimize the error of approximation
- It is well-known that thresholding by choosing the $C$ largest (in absolute value) wavelet coefficients after normalization is provably optimal in minimizing the 2-norm (Euclidian distance) of the absolute errors, among all possible choices of $C$ nonzero coefficients, assuming that the wavelet basis functions are orthonormal.
- The $C$ wavelet coefficients together with their $C$ indices (in the one-dimensional order of cells), form the compact data cube.
Complexity Analysis

- Storage: 2C
- Off-line Transformation of P:
  \[ O(\frac{N}{B} \log_{\frac{M}{B}} \frac{N}{B}) \] where B is the disk block size, N is the total cube size, and M is the internal memory size
- Partial sum value for a given cell can be computed using \( O(dC) \) space in
  \[ O(\Sigma_{1 \leq i \leq d} \min\{C, \log|D_i|\}) \] where \( |D_i| \) is the size of dimension i of the cube (out of d dimensions)

ProPolyne: A Fast Wavelet-based Algorithm for Progressive Evaluation of Polynomial Range-Sum Queries

Rolfe Schmidt and Cyrus Shahabi
University of Southern California
Dept. of Computer Science
Los Angeles, CA 90089-0781

shahabi@usc.edu
http://infolab.usc.edu
Approach:
Enabling Data Manipulation, Query & Analysis in the WAVELET Domain

- Everybody else’s idea: let’s compress data
  - Reason: save space? No not really!
  - Implicit reason: queries deal with smaller data sets and hence faster (not always true!)
  - More problems: not only query results can never be 100% accurate anymore, but also different queries can have very different error rates given their areas of interest
  - Why? At the data population time, we don’t know which coefficients are more/less important to our queries! (also observed by [Garofalakis & Gibbons, SIGMOD’02], but they proposed other ways to drop coefficients assuming a uniform workload)
    - Different than the signal-processing objective to reconstruct the entire signal as good as possible

- Our idea/distinction: storage is cheap and queries are ad-hoc; let’s keep all the wavelet coefficients! (no data compression)

- Opportunity: At the query time, however, we have the knowledge of what is important to the pending query

- ProPolyne: Progressive Evaluation of Polynomial Range-Aggregate

(C. Shahabi)
Outline

- ProPolyne: Overview and Features
- ProPolyne: Details
- Comparison Table
- Performance Results
- Conclusion
- How to Evaluate Multiple Range-Sum Queries Progressively

Overview of ProPolyne

- Define range-sum query as dot product of query vector and data vector
- Offline: Multidimensional wavelet transform of data
- At the query time: “lazy” wavelet transform of query vector (very fast)
- Dot product of query and data vectors in the transformed domain \( \Rightarrow \) exact result in \( O(2 \log N)^d \)
- Choose high-energy query coefficients only \( \Rightarrow \) fast approximate result (90% accuracy by retrieving < 10% of data)
- Choose query coefficients in order of energy \( \Rightarrow \) progressive result
ProPolyne Features

- “Measure” can be any polynomial on any combination of attributes
  - Can support COUNT, SUM, AVERAGE
  - Also supports Covariance, Kurtosis, etc.
  - All using one set of pre-computed aggregates
- Independent from how well the data set can be compressed/approximated by wavelets
  - Because: We show “range-sum queries” can always be approximated well by wavelets (not always HAAR though!)
- Low update cost: $O(\log^d N)$
- Can be used for exact, approximate and progressive range-sum query evaluation

Outline

- ProPolyne: Overview and Features
- ProPolyne: Details
  - Polynomial Range-Sum Queries as Vector Queries
  - Naive Evaluation of Vector Queries
  - Fast Evaluation of Vector Queries
  - Progressive/Approximate Evaluation of Vector Queries
- Comparison Table
- Performance Results
- Conclusion
- How to Evaluate Multiple Range-Sum Queries Progressively
Polynomial Range-Sum Queries

- Polynomial range-sum queries: \(Q(R,f,I)\)
  - \(I\) is a finite instance of schema \(F\)
  - \(R\) is a subset of \(\text{Dom}(F)\), is the range
  - \(f: \text{Dom}(F) \rightarrow R\) is a polynomial of degree \(\delta\)

\[
Q(R, f, I) = \sum_{x \in I} f(x)
\]

**Example:** \(F = (\text{Age}, \text{Salary})\)
- \(R\): \((25 < \text{age} < 40) \land (55k < \text{salary} < 150k)\)

**COUNT** : \(f(\overline{x}) = 1(\overline{x}) = 1\)

\[
Q(R, 1, I) = \sum_{x \in I} f(x) = 1(28,55k) + 1(30,58k) = 2
\]

**SUM** : \(f(\overline{x}) = \text{salary}(\overline{x})\)

\[
Q(R, \text{salary}, I) = \sum_{x \in I} f(x) = \text{salary}(28,55k) + \text{salary}(30,58k) = 113k
\]

\[
Q(R, \text{salary} \times \text{age}, I) = \sum_{x \in I} \text{salary}(\overline{x}) \times \text{age}(\overline{x}) = f(28,55k) + f(30,58k) = 3280 M
\]

\[
\text{Cov}(\text{age}, \text{salary}) = \frac{Q(R, \text{salary} \times \text{age}, I) - Q(R, \text{age}, I)Q(R, \text{salary}, I)}{Q(R, I)^2}
\]

Polynomial Range-Sum Queries as “Vector Queries”

- The data frequency distribution of \(I\) is the function \(\Delta: \text{Dom}(F) \rightarrow \mathbb{Z}\) that maps a point \(x\) to the number of times it occurs in \(I\)
- To emphasize the fact that a query is an operator on the data frequency distribution, we write

\[
Q(R, f, I) = Q(R, f, \Delta)
\]

**Example:** \(\Delta(25,50) = \Delta(28,55) = \ldots = \Delta(57,120) = 1\) and \(\Delta(x) = 0\) otherwise.

Hence:

\[
Q(R, f, \Delta) = \sum_{x \in \text{Dom}(F)} f(x) \chi(x) \Delta(x)
\]

where:

\[
\chi(x) = \begin{cases} 
1 & \text{if } x \in R \\
0 & \text{if } x \notin R
\end{cases}
\]

Or:

\[
\langle f \chi(x), \Delta \rangle
\]
Overview of Wavelets

- H operator: computes a local average of array \( a \) at every other point to produce an array of summary coefficients: \( H[a] \)
- G operator: measures how much values in the array \( a \) vary inside each of the summarized blocks to compute an array of detail coefficients: \( G[a] \)
- \( \hat{a} \): aka wavelet coefficients of \( a \)

Naive Evaluation of Vector Queries Using Wavelets

- Hence, vector queries can be computed in the wavelet-transformed space as:

\[
Q(R, f, \Delta) = \langle f\hat{\Delta}, \hat{\Delta} \rangle = \sum_{\eta_0, ..., \eta_{d-1}} f_{\eta_0, ..., \eta_{d-1}}(\hat{\Delta}(\eta_0, ..., \eta_{d-1}))
\]

- Algorithm:
  - Off-line transformation of data vector (or “data distribution function”, i.e., \( \Delta \), to be exact)
    - \( O(|\mathbf{I}|d\log^2 N) \) for sparse data, \( O(|\mathbf{I}|) = N^d \) for dense data
  - Transform the query vector at submission
    - \( O(N^d) \)
  - Sum-up the products of the corresponding elements of data and query vectors
    - Retrieving elements of data vector: \( O(N^d) \)
Fast Evaluation of Vector Queries Using Wavelets

Main intuitions:
- “query vector” can be transformed quickly because most of the coefficients are known in advance
- “Transformed query vector” has a large number of negligible (e.g., zero) values (independent on how well data can be approximated by wavelet)
- Example: Haar filter & COUNT function on $R=[5,12]$ on the domain of integers from 0 to 15:

$$\chi_q = \{0,0,0,0,1,1,1,1,1,0,0,0\}$$

$$\tilde{\chi}_q = \{2, \frac{1}{2}, \frac{3}{2\sqrt{2}}, -\frac{3}{2\sqrt{2}}, \frac{0}{2}, \frac{1}{2}, -\frac{1}{2}, \frac{0}{\sqrt{2}}, -\frac{1}{2}, \frac{0}{\sqrt{2}}, 0\}$$

The Lazy Wavelet Transform

Computing Summary Coefficients (Haar Filter, COUNT function)

At boundary of range, summary coeff is $\frac{1}{2} * 0 + \frac{1}{2} * 1 = \frac{1}{2}$

Inside range, summary coeffs are $\frac{1}{2} * 1 + \frac{1}{2} * 1 = 1$

All summary coefficients computed in CONSTANT time!

Summary coefficient array looks almost exactly like original array.

The only “interesting” activity happens on the boundary.
The Lazy Wavelet Transform

Computing Detail Coefficients (Haar Filter, COUNT function)

Outside the range, detail coeffs are
\( \frac{1}{2} \times 0 - \frac{1}{2} \times 0 = 0 \).

At lower boundary of range, detail coeff is
\( \frac{1}{2} \times 0 - \frac{1}{2} \times 1 = -\frac{1}{2} \).

At upper boundary of range, detail coeff is
\( \frac{1}{2} \times 1 - \frac{1}{2} \times 0 = \frac{1}{2} \).

All detail coefficients computed in CONSTANT time!

All but 2 detail coefficients at each level are equal to zero!

The only “interesting” activity happens on the boundary.

Fast Evaluation of Vector Queries Using Wavelets ...

- Technical Requirements:
  - Wavelets should have small support (i.e., the shorter the filter, the better)
  - Wavelets must satisfy a “moment condition”
  - Supports any Polynomial Range-Sum up to a degree determined by the choice of wavelets
    - E.g., Haar can only support degree 0 (e.g., COUNT), while db4 can support up to degree 1 (e.g., SUM), and db6 for degree 2 (e.g., VARIANCE)

- Standard DWT: \( O(N) \)
- Our lazy wavelet transform: \( O(l \log N) \),
  where \( l \) is the length of the filter
Exact Evaluation of Vector Queries

**Query:**
SUM(salary) when
(25 < age < 40) &
(55k < salary < 150k)

# of Nonzero Coordinates: 4380
# of Wavelet Coefficients: 837

Approximate Evaluation of Vector Queries

With 150 coefficients:
It is as if this query is being Submitted or evaluated
Progressive Evaluation of Vector Queries

<table>
<thead>
<tr>
<th>Name of Technology</th>
<th>Research Group</th>
<th>Year</th>
<th>Query Cost</th>
<th>Update Cost</th>
<th>Storage Cost</th>
<th>Aggregate Function Support</th>
<th>Query Evaluation Support</th>
<th>Known at Population?</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROPOLYNE</td>
<td>USC Schmidt &amp; Shahabi</td>
<td>2001</td>
<td>$O(d N \lg N)$</td>
<td>$O(d N \lg N)$</td>
<td>$O(d N \lg N)$</td>
<td>Exact, Approximate, Progressive</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>PROPOLYNE-FM</td>
<td>USC Schmidt &amp; Shahabi</td>
<td>2001</td>
<td>$O(d N \lg N)$</td>
<td>$O(d N \lg N)$</td>
<td>$O(d N \lg N)$</td>
<td>Exact, Approximate, Progressive</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Space Efficient Dynamic Data Cube 2000</td>
<td>UCSB El Abbadi &amp; Agrawal et. al</td>
<td>2000</td>
<td>$O(d N)$</td>
<td>$O(d N)$</td>
<td>$O(d N)$</td>
<td>Exact, Approximate, Progressive</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Relative Prefix-Sum 1999</td>
<td>UCSB</td>
<td>1999</td>
<td>$O(d N)$</td>
<td>$O(d N)$</td>
<td>$O(d N)$</td>
<td>Exact, Approximate, Progressive</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Prefix-Sum 1997</td>
<td>IBM Agrawal et. al</td>
<td>1997</td>
<td>$O(d N)$</td>
<td>$O(d N)$</td>
<td>$O(d N)$</td>
<td>Exact, Approximate, Progressive</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>pCube/MAETree 2000</td>
<td>UCSB and UC Irvine (Mehrota et. al)</td>
<td>2000</td>
<td>$O(d N)$</td>
<td>$O(d N)$</td>
<td>$O(d N)$</td>
<td>Exact, Approximate, Progressive</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Compact Data Cube 1990-2001</td>
<td>Duke and IBM (Yin &amp; al.)</td>
<td>1990-2001</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>Exact, Approximate, Progressive</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Compact Data Cube 1990-2001</td>
<td>Duke and IBM (Yin &amp; al.)</td>
<td>1990-2001</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>Exact, Approximate, Progressive</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Optimal Histograms 2001</td>
<td>Microsoft (Fayyad et. al)</td>
<td>2001</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>Exact, Approximate, Progressive</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Kernel Density Estimation 2001</td>
<td>Microsoft (Fayyad et. al)</td>
<td>2001</td>
<td>small</td>
<td>small</td>
<td>small</td>
<td>Exact, Approximate, Progressive</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Conclusion

- A novel pre-aggregation strategy
- Supports conventional aggregates: COUNT, SUM and beyond: multivariate statistics
- First pre-aggregation technique that does not require measures be specified a priori
  - Measures treated as functions of the attributes at the time
- Provides a data independent progressive and approximate query answering technique
- With provably poly-logarithmic worst-case query and update costs
- And storage cost comparable or better than other pre-aggregation methods
Future Work

- Almost Complete
  - Batch queries
  - Efficient layout on disk
- In Progress ….
  - I/O efficient wavelet transformation and update
  - Hybrid ordering of data and query coefficients
  - Error forecasting
- Longer Term
  - Best basis per dimension
  - ProPolyne on GRID using p2p
  - ProPolyne on relation algebra operators

The remainder is FYI….
**Experimental Setup**

**PETROL Data Set:**
- Petroleum sales volume
- 56,504 tuples
- Five dimensions:
  - <location, product, year, month, volume>
- Sparseness: 0.16%
  - Traditional data approximation works well

- 250 range queries generated
  - Randomly from all possible ranges with the uniform dist.
  - Ranges which select fewer than 100 tuples were discarded
  - Median Relative error

**GPS Data Set:**
- Sensor readings from GPS ground stations in CA
- 3,358 tuples
- Four dimensions:
  - <lat, long, t, velocity>
  - Velocity of upward movement of the station
- Sparseness: 0.01%
  - Data approximation works poorly

\[
\text{Relative Error} = \frac{|\text{est}(q) - \text{propolyne}(q)|}{\text{Max}(1, \text{propolyne}(q))}
\]

**Performance Results**

- Compact Data Cube (as a representative of data approximation techniques): under 10% error after using less than 10% of wavelet coefficients (wavelet coefficients sorted in the order of energy)
- CDC needs 5 times as many coefficients as there were tuples in the original table before providing a median relative error of 10% (because data cannot be compressed well)

- Temperature Data Set
  - Model data from JPL
  - Temperature values at:
    - 128 Longitudes
    - 64 Latitudes
    - 16 altitudes
    - Each 12 hours over 4 days in March 2001
  - Total ~ 1.05 million records
  - Temperature discretized to 8 bits
  - Domain size: $2^{28}$
Batches of Related Queries and Structural Error [PODS'02]

- **Typical scientific data exploration scenario:**
  - Users specify a coarse partition of the domain of a data set, and request aggregate query results from each cell of this partition.
  - This provides a data synopsis which is used to identify interesting regions of the data domain.
  - Users then drill-down into the interesting regions by partitioning them further and requesting aggregate query results on the new cells.
  - User is interested in both the individual cell values that are returned and the way the query results change between neighboring cells.

- **Observations:**
  - [Scientific] OLAP queries are submitted in batches
  - The structure of error across queries is more important than error of each individual query
Our PODS’02 Contributions

- Batch-Biggest-B, an extension of PROPOLYNE that exploits I/O sharing to provide fast exact query results for a batch of queries.
- The introduction of structural error for batch queries, and the definition of structural error penalty functions.
  - This generalizes common error measures such as sum of square errors (SSE) and $L^p$ norms.
- A progressive version of Batch-Biggest-B that can accept any penalty function specified at query time, and will minimize the worst case and average penalty at each step of the computation.

Experimental Setup

- Dataset: Temperature Model
  - Temperature at different Latitudes, Longitudes and Altitudes every 12 hours from March 1 to May 1, 2001
  - 15.7 million records
  - 5 dimensions: Lat, Long, Alt, Time, Temperature
- Queries
  - Batches of 512 range sum queries
  - Ranges partition data domain
Progressive Relative Error

Less than 1% error after retrieving only less than 10 coefficients!

Different Penalty Functions: Optimizing for SSE
Different Penalty Functions: Optimizing for Weighted SSE