Spatial Hierarchy & OLAP-Favored Search in Spatial Data Warehouse

Fangyan Rao
IBM China Research Lab
Outline

- Motivation
- Spatial hierarchy
- OLAP-favored search
- Heuristic OLAP-favored search
- Test Result
- Conclusions
Motivation - Pervasive Spatial Data

80% of the data stored in computers has a spatial aspect

Spatial data warehouse - provide insight into business data from geospatial point of view
Queries against Spatial Data Warehouse and OLAP

- What is the total number of calls initiated along the light-railway? [Telecommunication]
- What is the average income of customers with a home insurance policy living within 1000 yards of the creek who DO NOT have flood insurance option in 2002. [Insurance]
- What is the average traffic volume of the roads within 1km of each gym at rush hour during last month? [Traffic Control]
Spatial Data and Spatial Data Warehouse

- Spatial data: multi-dimensional spatial objects described in coordinates
- Spatial data warehouse
  Dimensions and measures can contain spatial data
Star Schema of a Spatial Data Warehouse

<table>
<thead>
<tr>
<th>Station_ID</th>
<th>Station_Name</th>
<th>Address</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>77</td>
<td>Beiti Gas Station</td>
<td>23rd Street, Shangdi, Beijing</td>
<td>(116.37,39.51)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gas_ID</th>
<th>Unit_Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>3.25</td>
</tr>
<tr>
<td>93</td>
<td>3.60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trans_No.</th>
<th>Station_ID</th>
<th>Gas_ID</th>
<th>Vehicle_Group_ID</th>
<th>Time_ID</th>
<th>Quantity</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>5478912</td>
<td>77</td>
<td>90</td>
<td>2</td>
<td>3455</td>
<td>40</td>
<td>130.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vehicle_Group_ID</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Private car</td>
</tr>
<tr>
<td>2</td>
<td>Taxi</td>
</tr>
<tr>
<td>3</td>
<td>Bus</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time_ID</th>
<th>Time</th>
<th>Day</th>
<th>Month</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>3455</td>
<td>7:20</td>
<td>12</td>
<td>Feb</td>
<td>2002</td>
</tr>
<tr>
<td>3456</td>
<td>7:25</td>
<td>13</td>
<td>Feb</td>
<td>2002</td>
</tr>
</tbody>
</table>

- Problem - limited prior knowledge of the concept hierarchy on spatial dimension
Spatial Index Tree as Hierarchy on Spatial Dimension

Hierarchy

- Adult: 18-120
  - Young: 18-39
  - Middle-aged: 40-59
  - Senior: 60-120
- Teenager: 13-17
- Child: 0-12

R-tree

From J. Han and M. Kamber. Data Mining: Concepts and Techniques
Building Spatial Data Cube

Post order traversal:

(((S1, S2, S7) R4), (...) R5) R2, (...) R3) R1

Summary Table

<table>
<thead>
<tr>
<th>NID</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td>8000.00</td>
</tr>
<tr>
<td>S2</td>
<td>3000.00</td>
</tr>
<tr>
<td>S7</td>
<td>4000.05</td>
</tr>
<tr>
<td>R4</td>
<td>15000.00</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>R5</td>
<td>20000.00</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>35000.00</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>R3</td>
<td>45000.00</td>
</tr>
<tr>
<td>R1</td>
<td>75000.00</td>
</tr>
</tbody>
</table>
OLAP-favored Search

What’s the average sales of gas-stations within the given query region last year?

(a) "Within" Query

(b) Traditional Search

(c) OLAP-favored Search

Number of Nodes Returned: 4
Edge Visited: 9

Number of Nodes Returned: 2
Edge Visited: 6

What's the average sales of gas-stations within the given query region last year?
Implementation – Based on GiST

- Extendibility - Consistent() and search()
  - Introduce a third state (partial true) of Consistant()

\[
\begin{array}{c}
A \\
\text{Within (A,B) = Partial True}
\end{array}
\]

- Invoke search on the subtree rooted from a node when Consistent returns partial true
Consistent() in OLAP-favored Search

<table>
<thead>
<tr>
<th>Consistent(E,q)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong></td>
</tr>
<tr>
<td><strong>Output:</strong></td>
</tr>
</tbody>
</table>
OLAP-favored Search in GiST

---

Search(R, q)

Input: GiST rooted at node R, search predicate q
Output: all nodes/tuples that satisfies q
Sketch: Recursively descend all paths in tree whose keys are partially consistent with q
Procedure: List ret = null

//1. [Search subtrees]
IF R is not a leaf THEN
    FOR Each entry E on R
        BEGIN
            IF Consistent(E, q) is true THEN
                add E to ret
            ELSE
                IF Consistent(E, q) is partial true THEN
                    // invoke Search on the subtree whose root node is referenced by E.ptr
                    List subret = Search(E.ptr, q)
                    add subret to ret
            END
        END
    END
ELSE // R is a leaf
    FOR Each entry E on R
        BEGIN
            IF Consistent(E, q) is true THEN
                add E to ret
        END
    END
RETURN ret
Heuristic OLAP-favored Search

- **AF (Approximate Factor)** to quantify the Partial True state
  
  \[ AF = \frac{\text{Approximate query result of a tree node}}{\text{exact aggregation result of its subtree}} \]
  
  \( 0 < AF < 1 \), higher AF indicates less error in approximate value

  Approximate value = AF * Aggregated value on the tree node

- **MinAF** is set for the tree traversal threshold

- **Example evaluation of AF**
  
  \[ AF = \frac{\text{Area of Intersecting}}{\text{Area of node's MBR}} \]
Implementation – Based on GiST

- Consistant() return a number ranging from 0 to 1

  ![Diagram](image)

  Within (A,B) = 0.4

- Invoke search on the subtree rooted from a node when Consistent returns a value bigger than MinAF
Consistent() in Heuristic OLAP-favored Search

\[
\text{Consistent}(E,q)
\]

**Input:** Index entry \( E = (p; \text{ptr}) \), query predicate \( q \)
where \( p \) is the predicate of the index entry,
\( \text{ptr} \) is the pointer to the child node in the
index entry.

**Output:**
returns 0 if for each descendant entry \( \text{DE} \) of \( E \),
\( \text{DE}.p \land q \) can be guaranteed unsatisfiable,
1 if for each descendant entry \( \text{DE} \) of \( E \),
\( \text{DE}.p \land q \) can be guaranteed satisfiable,
\( \text{AF}(0 < \text{AF} < 1) \) otherwise.
Testing Data

- 3 spatial data sets: uniform, normal, real data from TIGER
- Sales data generated using uniform distribution
- 9 groups of queries are generated with 100 query regions in each group. Query regions in each group are same in size while the group region size grows from 10% to 90% of the whole data area.
Measurements

- Number of records returned (NRR)
- Aggregation ratio (AR) = NRR of OLAP-favored search / NRR of traditional search
- Number of edge visited (NEV)
- Error in heuristic search method

- Measures the group average value.
NRR of OLAP-favored Search

Query Window Size (%) vs. Number of Records Returned for different Fanout values (30, 50, 100) and traditional search.
NRR of OLAP-favored Search

Real Data Set

Query Window Size(%) vs Number of RecordsReturned

- Fanout=30
- Fanout=50
- Fanout=100
- Trad. Search
AR of OLAP-favored Search

Graphs showing the relationship between query window size and aggregation ratio for different fanout values.
NEVR of OLAP-favored Search

![Graphs showing the ratio of number of edges visited against query window size for different fanout values (30, 50, 100).]
NRR of Heuristic OLAP-favored Search when Fanout=100
NRR of Heuristic OLAP-favored Search when Fanout=100

Real Data Set

- ▲ Trad. search
- ◊ OLAP-favored search
- → Heuristic w/ MinAF=0.1
- ✫ Heuristic w/ MinAF=0.3
- ☻ Heuristic w/ MinAF=0.5
- ■ Heuristic w/ MinAF=0.7
- ○ Heuristic w/ MinAF=0.9
Error in Heuristic OLAP-favored Search when Fanout = 100
Error in Heuristic OLAP-favored Search when Fanout = 100

Real Data Set

Query Window Size

- Heuristic w/ MinAF=0.1
- Heuristic w/ MinAF=0.3
- Heuristic w/ MinAF=0.5
- Heuristic w/ MinAF=0.7
- Heuristic w/ MinAF=0.9
Related Work

- Spatial data warehouse firstly proposed by Han et al[13]
  - Non-spatial
  - Spatial to non-spatial
  - Spatial to Spatial
- Store aggregation result in index tree, Papadias et al[17][18]
- Static spatial OLAP query, Li et al[16]
- Statistical trees, SIAM[9], TBSAM[22]
Conclusions

- Extend hierarchy into multi-dimensional data space, and use index tree as the hierarchy on spatial dimension
- Developed an OLAP-favored search algorithm
- Heuristic query method
- Empirical experiments
Thanks
Can current data warehouse handle locations?

Implicit hierarchy of address: (street, district, city, province, nation)
Spatial data: location string and coordinates

<table>
<thead>
<tr>
<th>Hotel</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hotel InterContinental New Orleans</td>
<td>444 St Charles Avenue New Orleans Louisiana 70130 USA</td>
</tr>
</tbody>
</table>

Geocoding

<table>
<thead>
<tr>
<th>Hotel</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hotel InterContinental New Orleans</td>
<td>30.00, -90.00</td>
</tr>
</tbody>
</table>

What is the total amount of sales of restaurants less than 1KM from the hotel in 2002?
### Fact Table

<table>
<thead>
<tr>
<th>NID</th>
<th>Vehicle</th>
<th>Gas</th>
<th>Year</th>
<th>Month</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>R4</td>
<td>1</td>
<td>90</td>
<td>2002</td>
<td>Jan.</td>
<td>27865.00</td>
</tr>
<tr>
<td>R4</td>
<td>1</td>
<td>90</td>
<td>2002</td>
<td>Feb.</td>
<td>24500.00</td>
</tr>
<tr>
<td>R4</td>
<td>1</td>
<td>90</td>
<td>2002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R4</td>
<td>1</td>
<td>90</td>
<td>2002</td>
<td></td>
<td>190234.00</td>
</tr>
<tr>
<td>R4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>348976.00</td>
</tr>
<tr>
<td>R4</td>
<td>1</td>
<td>93</td>
<td>2002</td>
<td>Jan.</td>
<td>3456.00</td>
</tr>
<tr>
<td>R4</td>
<td>1</td>
<td>93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R4</td>
<td>1</td>
<td>93</td>
<td></td>
<td></td>
<td>1098768.00</td>
</tr>
<tr>
<td>R4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R4</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td>2310087.00</td>
</tr>
<tr>
<td>R4</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R4</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td>45619876.00</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>R2</td>
<td>1</td>
<td>90</td>
<td>2002</td>
<td>Nov.</td>
<td>3017.00</td>
</tr>
<tr>
<td>R2</td>
<td>1</td>
<td>90</td>
<td>2002</td>
<td></td>
<td>455231.00</td>
</tr>
<tr>
<td>R2</td>
<td>1</td>
<td>null</td>
<td></td>
<td></td>
<td>1765234.00</td>
</tr>
<tr>
<td>R2</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td>28754395.00</td>
</tr>
<tr>
<td>R2</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td>63549857.00</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Spatial OLAP Server

Developer
Maintenance
(Building, reorganization)

Spatial OLAP construction request

Spatial OLAP Query GUI
Analyzer
analysis predict decision making

Spatial OLAP Server

Spatial OLAP Cube Builder

Spatial Index Engine

Spatial OLAP Query Processor

ROLAP Engine

Spatial Index

Spatial OLAP Cube