Approximate Answers to OLAP Queries on Streaming Data Warehouses

Michel DE ROUGEMONT, Phuong Thao CAO
Paris II Univ., South Paris Univ.
November 2, 2012
Outline

1. **Context: OLAP Queries**
   - Approximate answers
   - Streaming data

2. **Data exchange**
   Approximate answers with:
   - Sampling algorithm on the Sources
   - Use of Statistical dependencies

3. **Implementation**
1. Context

- **OLAP Schema**

```
RecordID | SensorID | Date     | Sun (hours) | Rain (hours)
---------|----------|----------|-------------|-------------
1001     | 8        | 11/02/12 | 8           | 2           
...      | ...      | ...      | ...         | ...         
```

- **Different streams feed the Fact table**

- **OLAP queries**
  (Sum of Measure)

  Measure=Hours of Sun  
  Analysis by Country  
  Analysis by Country/Manuf.
Approximation

- **Distance $L_1$**
  100% error for the blue area

- **Sampling:**
  - classical technology to approximate
  - streaming: It is hard to approximate (Cormode et al. 2003)

- **Data Exchange**
2. OLAP Data Exchange

- Different sources

- Different streams: hard to approximate in the worst case (Cormode et al. 2012)

How can we approximate queries in some special case? **statistical dependencies**
2.1 Streams with Different Rates

- Data warehouse
  - Union of different Sources
  - Rate of tuples of each Source is different
  (rate: relative number of tuples per unit of time)
Uniform samples on the Streams

• Approximate Algorithm
  – Step 1: sampling on each Source with uniform distribution. \#samples % to the rate of the Source.
  – Step 2: combine all samples according to the rates
  – Step 3: approximation on the union of samples

**Theorem:** On a window of size $T$, OLAP queries are $\varepsilon$-approximated with $N$ samples (which depend on $T$ and $\varepsilon$) with high probability.
2.2 Special Case: Statistical Dependencies

- Some attributes imply a distribution $\mu$ on the measure: $A.B.C \triangleleft M$
  - $(a,b,c)$ determines a fixed distribution on $M$
  - Generalization of functional dependencies

- City $\triangleleft$ Sun ($\mu$ distribution)

  Marseille:

  London:
# Distribution of pairs

## City.Country

<table>
<thead>
<tr>
<th>City</th>
<th>Country</th>
<th>Density of tuples</th>
</tr>
</thead>
<tbody>
<tr>
<td>London</td>
<td>U.K</td>
<td>1/12</td>
</tr>
<tr>
<td>Berlin</td>
<td>Germany</td>
<td>1/12</td>
</tr>
<tr>
<td>Paris</td>
<td>France</td>
<td>1/6</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

## Manufacturer.City (δ)

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>City</th>
<th>Density of tuples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thomson</td>
<td>London</td>
<td>1/12</td>
</tr>
<tr>
<td>Thomson</td>
<td>Berlin</td>
<td>1/12</td>
</tr>
<tr>
<td>Siemens</td>
<td>Paris</td>
<td>1/12</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country</th>
<th>Distribution of Sun</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.K.</td>
<td>0.64</td>
</tr>
<tr>
<td>Germany</td>
<td>0.21</td>
</tr>
<tr>
<td>France</td>
<td>0.15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>Distribution of Sun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Siemens</td>
<td>0.39</td>
</tr>
<tr>
<td>Thomson</td>
<td>0.61</td>
</tr>
</tbody>
</table>
Use of Statistical Hypothesis:
Distributed Algorithm

- **Each Source** \( i \), we sample by uniform distribution and:
  - Learn the \( \mu_i \)
  - Estimate the distribution on pairs \( \delta_i \)
  - Estimate its rate: \( r_i \)

- **Data Warehouse**:
  - Combine rates \( r_i \), \( \delta_i \) and \( \mu_i \) to approximate the OLAP query on A (Manufacturer)

\[
Q^M_{C=Siemens} = (r_1 \cdot Q^M_{C=Siemens})^1 + (r_2 \cdot Q^M_{C=Siemens})^2
\]

\[
= \frac{2}{3} \left[ \sum_{City} \delta(Siemens, City) \cdot Avg(\mu_{City}) \right] + \frac{1}{3} \left[ \sum_{City} \delta(Siemens, City) \cdot Avg(\mu_{City}) \right]
\]

\[
= 0.39
\]
Statistical Model

Advantages:
- Statistical dependencies: more intuitive
- Sources send only statistical dependencies
  (constant size of information on finite domains)
- Sources do not send samples
Our contribution

• Special situation: model of statistical dependencies on streaming data

• Approximation algorithms:
  – Sampling: each Source samples and we combine all the samples
  – Statistical model: combine statistical dependencies and distributions of pairs

• Worst case is not approximable
3. Implementation

- **Program**
  - Mondrian OLAP engine
  - Jpivot interface
- **Data warehouse**
  - $10^6$ tuples
Approximate answer on sources:

- **Data warehouse**
  - 12 sensors: 6 in France, 3 in Germany, 3 U.K.
  - 2 manufacturers: Siemens, Thomson
  - 9 cities
  - $1 \leq \text{Sun, Rain} \leq 10$
- Statistical dependencies:
  - City $\triangleleft$ Sun
  - Distribution of pairs
    - City.Country
    - Manufacturer.City
Example 1: Analysis by country

Source from U.K. $m_1$ samples

Source from Germany $m_2$ samples

Source from France $m_3$ samples

Data warehouse $m$ samples = $m_1$, $m_2$, $m_3$
Approximate answer on sources:
Analysis by country

• Learn distributions $\mu_i, \delta_i$ from samples

<table>
<thead>
<tr>
<th>City</th>
<th>Average value of Sun : $\text{Avg}_\mu(a_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>London</td>
<td>3.5</td>
</tr>
<tr>
<td>Berlin</td>
<td>5</td>
</tr>
<tr>
<td>Paris</td>
<td>7.5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>City</th>
<th>Country</th>
<th>Density of tuples : $\delta(a_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>London</td>
<td>U.K.</td>
<td>1/12</td>
</tr>
<tr>
<td>Berlin</td>
<td>Germany</td>
<td>1/12</td>
</tr>
<tr>
<td>Paris</td>
<td>France</td>
<td>1/6</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

\[ m_i = m \times \frac{\delta(a_i) \times \text{Avg}_\mu(a_i)}{\sum_i \delta(a_i) \times \text{Avg}_\mu(a_i)} \]
Approximate answer on sources: Analysis by country

\[ m_i = m \times \frac{\delta(a_i) \times \text{Avg}_{\mu}(a_i)}{\sum_i \delta(a_i) \times \text{Avg}_{\mu}(a_i)} \]
Example 2: Analysis by Manufacturer

<table>
<thead>
<tr>
<th>City</th>
<th>Avg value of Sunlight : $\text{Avg}_\mu(a_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>London</td>
<td>3.5</td>
</tr>
<tr>
<td>Berlin</td>
<td>5</td>
</tr>
<tr>
<td>Paris</td>
<td>7.5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>City</th>
<th>Density of tuples : $\delta(a_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thomson</td>
<td>London</td>
<td>1/12</td>
</tr>
<tr>
<td>Thomson</td>
<td>Berlin</td>
<td>1/12</td>
</tr>
<tr>
<td>Siemens</td>
<td>Paris</td>
<td>1/12</td>
</tr>
<tr>
<td>Thomson</td>
<td>Paris</td>
<td>1/12</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Manufacturer Distribution of Sun

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>Distribution of Sun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Siemens</td>
<td>0.39</td>
</tr>
<tr>
<td>Thomson</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Approximate answer: Analysis by Manuf.
Analysis of errors

- All algorithms: rate of errors < 4%
- Statistical model is better than uniform sampling
- Statistical model is better than Measure based sampling

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>Uniform sampling</th>
<th>Measure-based sampling</th>
<th>Linear estimation by the data exchange</th>
<th>Exact answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Siemens</td>
<td>0.3851</td>
<td>0.4100</td>
<td>0.3890</td>
<td>0.3911</td>
</tr>
<tr>
<td>Thomson</td>
<td>0.6149</td>
<td>0.5900</td>
<td>0.6110</td>
<td>0.6089</td>
</tr>
<tr>
<td>TOTAL ERROR</td>
<td>0.0120</td>
<td>0.0378</td>
<td>0.0042</td>
<td></td>
</tr>
</tbody>
</table>
Conclusion and Perspective

• Conclusion
  – In the case of statistical dependencies, the algorithm keeps a good approximation to OLAP queries
  – Constant information exchanged on finite domains
  – Required memory in the worst case: $\Omega(N)$

• Perspective:
  – Application to RSS
  – Decision tree for the statistical model: discover the statistical dependencies
Thank you!

Questions & Answers?