CONGRESSIONAL SAMPLES FOR APPROXIMATE ANSWERING OF GROUP BY QUERIES

Swaroop Acharya, Philip B Gibbons, Vishwanath Poosala

By Agasthya Padisala
Anusha Reddy Rachapalli Muni
OUTLINE:
1. INTRODUCTION
2. AQUA SYSTEM
3. REQUIREMENTS ON GROUP BY ANSWERS
4. SOLUTIONS
5. REWRITING
6. EXPERIMENTS
7. EXTENSIONS
8. RELATED WORK
9. CONCLUSION
LIMITATIONS OF UNIFORM SAMPLING:

- Not suitable for group by query.

- For Example, Group by query on the U.S.census database could be used to determine the per capita income per state.

- There can be large discrepancy in size of groups. The size of one state can be more than the size of other state, e.g., the state of California has nearly 70 times the population of Wyoming.

- Uniform random sample of the relation will contain disproportionately fewer tuples from the smaller groups (states), which leads to poor accuracy for reliable answers of a group.
BIASED SAMPLING FOR GROUP BY QUERIES:

- In order to get an unbiased answer for groupby queries we use biased sample.
- Briefly, our techniques involve taking group-sizes into consideration while sampling, in order to provide highly-accurate answers.
- The techniques in this paper are tailored to precomputed or materialized samples, such as used in Aqua.
AQUA SYSTEM:

- Aqua maintains smaller-sized statistical summaries of the data called *synopses*, and uses them to answer queries.

- A key feature of Aqua is that the system provides probabilistic error/confidence bounds on the answer, based on the Hoeffding and Chebyshev formulas.
Figure 1: The Aqua architecture.
select l_returnflag, l_linenstatus, sum(l_quantity)
from lineitem
where l_shipdate <= '01-SEP-98'
group by l_returnflag, l_linenstatus;

(a) Original query

select l_returnflag, l_linenstatus, 100*sum(l_quantity),
       sum_error(l_quantity) as error1
from bs_lineitem
where l_shipdate <= '01-SEP-98'
group by l_returnflag, l_linenstatus;

(b) Rewritten query

Figure 2: Query rewriting in Aqua.
<table>
<thead>
<tr>
<th>l_returnflag</th>
<th>l_linenstatus</th>
<th>sum(l_quantity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>F</td>
<td>3773034</td>
</tr>
<tr>
<td>N</td>
<td>F</td>
<td>100245</td>
</tr>
<tr>
<td>N</td>
<td>O</td>
<td>7459912</td>
</tr>
<tr>
<td>R</td>
<td>F</td>
<td>3779140</td>
</tr>
</tbody>
</table>

Figure 3: Exact answer.

<table>
<thead>
<tr>
<th>l_returnflag</th>
<th>l_linenstatus</th>
<th>sum(l_quantity)</th>
<th>error1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>F</td>
<td>3.778e+06</td>
<td>1.4e+04</td>
</tr>
<tr>
<td>N</td>
<td>F</td>
<td>1.194e+05</td>
<td>2.6e+04</td>
</tr>
<tr>
<td>N</td>
<td>O</td>
<td>7.457e+06</td>
<td>1.9e+04</td>
</tr>
<tr>
<td>R</td>
<td>F</td>
<td>3.782e+06</td>
<td>1.4e+04</td>
</tr>
</tbody>
</table>

Figure 4: Approximate answer.
The user has two requirements on the approximate answer to a group-by query.

1. The approximate answer should contain all the groups that appear in the exact answer.

2. The estimated answer for every group should be close to the exact answer for that group.
SOLUTIONS:

- Congressional samples are hybrid union of uniform and biased samples.

- The strategy adopted is to divide the available sample space $X$ equally among the $g$ groups, and take a uniform random sample within each group.
Consider US Congress which is hybrid of House and Senate.

- **House** has representative from each state in proportion to its population. So, it represents a uniform random sampling for entire relation.

- **Senate** has equal number of representative from each state. So, it represents a sample having an equal number of tuples for each group.
Let $X$ be the sample size then, the final sample size allocated to group $g$ is given by:

$$c_g = X \frac{\max \left( \frac{n_g}{|R|}, \frac{1}{m_T} \right)}{\sum_{j \in g} \max \left( \frac{n_j}{|R|}, \frac{1}{m_T} \right)}$$

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>House $s_{q,0}$</th>
<th>Senate $s_{q,AB}$</th>
<th>Basic Congress (before scaling)</th>
<th>Basic Congress</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>$b_1$</td>
<td>30</td>
<td>25</td>
<td>30</td>
<td>27.3</td>
</tr>
<tr>
<td>$a_1$</td>
<td>$b_2$</td>
<td>30</td>
<td>25</td>
<td>30</td>
<td>27.3</td>
</tr>
<tr>
<td>$a_1$</td>
<td>$b_3$</td>
<td>15</td>
<td>25</td>
<td>25</td>
<td>22.7</td>
</tr>
<tr>
<td>$a_2$</td>
<td>$b_3$</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>22.7</td>
</tr>
</tbody>
</table>
Let $X$ be the sample size then, the final sample size allocated to group $g$ is given by:

$$\text{SampleSize}(g) = X \frac{\max_{T \subseteq G} s_{g,T}}{\sum_{j \in G} \max_{T \subseteq G} s_{j,T}}$$

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>House $s_{g,\ell}$</th>
<th>Senate $s_{g,\ell} AB$</th>
<th>Basic Congress (before scaling)</th>
<th>Basic Congress (sc)</th>
<th>$s_{g,A}$</th>
<th>$s_{g,B}$</th>
<th>Congress (before scaling)</th>
<th>Congress (sc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>$b_1$</td>
<td>30</td>
<td>25</td>
<td>30</td>
<td>27.3</td>
<td>20 (of 50)</td>
<td>33.3</td>
<td>33.3</td>
<td>23.5</td>
</tr>
<tr>
<td>$a_1$</td>
<td>$b_2$</td>
<td>30</td>
<td>25</td>
<td>30</td>
<td>27.3</td>
<td>20 (of 50)</td>
<td>33.3</td>
<td>33.3</td>
<td>23.5</td>
</tr>
<tr>
<td>$a_1$</td>
<td>$b_3$</td>
<td>15</td>
<td>25</td>
<td>25</td>
<td>22.7</td>
<td>10 (of 50)</td>
<td>12.5 (of 33.3)</td>
<td>25</td>
<td>17.7</td>
</tr>
<tr>
<td>$a_2$</td>
<td>$b_3$</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>22.7</td>
<td>50</td>
<td>20.8 (of 33.3)</td>
<td>50</td>
<td>35.3</td>
</tr>
</tbody>
</table>
Query rewriting involves two key steps:
1) scaling up the aggregate expressions and
2) deriving error bounds on the estimate.

For each tuple, let its scale factor Scale Factor be the inverse sampling rate for its strata.

There are two approaches to doing this:
1) store the Scale Factor(SF) with each tuple in sample relation- Integrated
2) use a separate table to store the Scale Factors for the groups- Normalized, Key-normalized, Nested-integrated
<table>
<thead>
<tr>
<th>Key</th>
<th>Grouping Columns</th>
<th>Aggregate Column</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>k₁</td>
<td>a₁</td>
<td>b₁</td>
</tr>
<tr>
<td>k₂</td>
<td>a₁</td>
<td>b₁</td>
</tr>
</tbody>
</table>

**Figure 6:** Relation Rel with two example tuples

```
select A, B, sum(Q)
from Rel
group by A, B;
```

**Figure 7:** User Query $Q_2$

```
select A, B, C, Q
from SampRel
group by A, B;
```

**Figure 8:** Integrated Rewriting

```
select SR.A, SR.B, sum(Q*SF)
from SampRel SR, AuxRel AR
where SR.A = AR.A and SR.B = AR.B
     and SR.C = AR.C
group by SR.A, SR.B;
```

**Figure 9:** Normalized Rewriting
(a) SampRel schema

(b) AuxRel schema

(c) Rewritten Query Q₂

Figure 10: Key-normalized Rewriting

Figure 11: Nested-integrated Rewriting
EXPERIMENTS:

- Testbed: On Aqua with Oracle(v7) as the backend DBMS

- Accuracy of Sample allocation strategies: Performance of different query sets (queries with no Group-bys, Three Group-bys, Two Group-bys) are given below:
Figure 14: Query $Q_{q0}$ Error  Figure 15: Query $Q_{q3}$ Error  Figure 16: Query $Q_{q2}$ Error
Effect of Sample Size:
- Error is inversely proportional to sample size. Congress – error drops rapidly with increasing sample size and provide high accuracy even for arbitrary group-bys.

Figure 17: Sample Size vs. Accuracy (Query $Q_{g2}$)
Performance of Rewriting strategies:

<table>
<thead>
<tr>
<th>Technique</th>
<th>Sample Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1%</td>
</tr>
<tr>
<td>Integrated</td>
<td>1.3</td>
</tr>
<tr>
<td>Nested-integrated</td>
<td>1.2</td>
</tr>
<tr>
<td>Normalized</td>
<td>1.7</td>
</tr>
<tr>
<td>Key-normalized</td>
<td>1.8</td>
</tr>
</tbody>
</table>

Table 3: Times Taken for Different Sample Percentages (actual query time = 40 sec)
EXTENSIONS:

- Generalization to multiple criteria. The congressional samples framework can be even extended to support grouping attributes with weight vectors.

- Generalization to Other Queries
  This can be achieved by replacing the values in grouping column.
RELATED WORK:

- Online Aggregation scheme
- Histograms
- Wavelets
- Stratified Sampling
CONCLUSION:

- Congressional samples will minimize errors over queries on a set of possible group-by relations.
- New strategies were validated to produce accurate estimates to group-by queries and has good execution efficiency.
REFERENCES

- http://portal.acm.org/citation.cfm?id=335191.335450&coll=ACM&dl=ACM&CFID=3197914&CFTOKEN=98178154

- http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.32.7042
QUESTIONS???