CONGRESSIONAL SAMPLES FOR APPROXIMATE ANSWERING OF GROUP BY QUERIES

Swaroop Acharya,Philip B Gibbons,
VishwanathPoosala

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

INTRODUCTION:

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 BYQUERIES:

- 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.

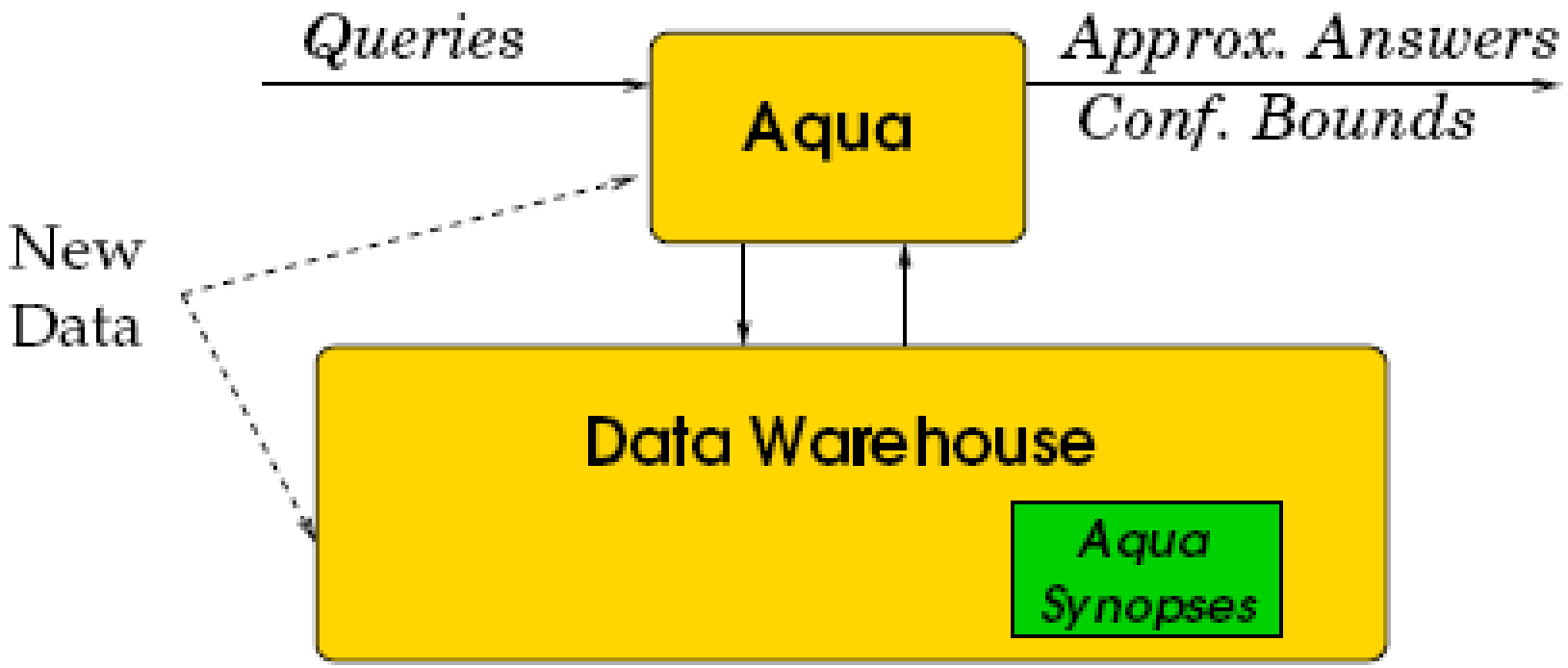


Figure 1: The Aqua architecture.

```
select l_returnflag, l_linestatus, sum(l_quantity)
from lineitem
where l_shipdate <= '01-SEP-98'
group by l_returnflag, l_linestatus;
```

(a) Original query

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

(b) Rewritten query

Figure 2: Query rewriting in Aqua.

l_returnflag	l_linestatus	sum(l_quantity)
A	F	3773034
N	F	100245
N	O	7459912
R	F	3779140

Figure 3: Exact answer.

l_returnflag	l_linestatus	sum(l_quantity)	error1
A	F	3.778e+06	1.4e+04
N	F	1.194e+05	2.6e+04
N	O	7.457e+06	1.9e+04
R	F	3.782e+06	1.4e+04

Figure 4: Approximate answer.

REQUIREMENTS ON GROUP BY ANSWERS

- 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.**

- BASIC CONGRESS:

Let X be the sample size then, the final sample size allocated to group g is given by :

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

A	B	*House* $s_{g,\emptyset}$	*Senate* $s_{g,AB}$	*Basic Congress* (before scaling)	*Basic Congress*
a_1	b_1	30	25	30	27.3
a_1	b_2	30	25	30	27.3
a_1	b_3	15	25	25	22.7
a_2	b_3	25	25	25	22.7

- CONGRESS:

Let X be the sample size then, the final sample size allocated to group g is given by :

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

A	B	*House* $s_{g,\emptyset}$	*Senate* $s_{g,AB}$	*Basic Congress* (before scaling)	*Basic Congress*	$s_{g,A}$	$s_{g,B}$	*Congress* (before scaling)	*Congress*
a_1	b_1	30	25	30	27.3	20 (of 50)	33.3	33.3	23.5
a_1	b_2	30	25	30	27.3	20 (of 50)	33.3	33.3	23.5
a_1	b_3	15	25	25	22.7	10 (of 50)	12.5 (of 33.3)	25	17.7
a_2	b_3	25	25	25	22.7	50	20.8 (of 33.3)	50	35.3

REWRITING :

- Query rewriting involves two key steps:
  a) scaling up the aggregate expressions and
  b) 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:
  a) store the Scale Factor(SF) with each tuple in sample relation- Integrated
  b) use a separate table to store the Scale Factors for the groups- Normalized, Key-normalized, Nested-integrated

Key	Grouping Columns			Aggregate Column
K	A	B	C	Q
k_1	a_1	b_1	c_1	q_1
k_2	a_1	b_1	c_2	q_2

Figure 6: Relation `Rel` with two example tuples

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

Figure 7: User Query $\mathcal{Q}_2$

K	A	B	C	Q	SF

(a) `SampRel` schema

```
select A,B, sum(Q*SF)
from SampRel
group by A,B;
```

(b) Rewritten Query $\mathcal{Q}_2$

Figure 8: *Integrated* Rewriting

K	A	B	C	Q

(a) `SampRel` schema

A	B	C	SF

(b) `AuxRel` schema

```
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;
```

(c) Rewritten query $\mathcal{Q}_2$

Figure 9: *Normalized* Rewriting

(a) SampRel schema

(b) AuxRel schema

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

(c) Rewritten Query Q_2

Figure 10: *Key-normalized* Rewriting

(a) SampRel schema

```
select A,B, sum(SQ*SF)
from (select A, B, SF, sum(Q) as SQ
          from SampRel
          group by A, B, SF)
group by A,B;
```

(b) Rewritten Query Q_2

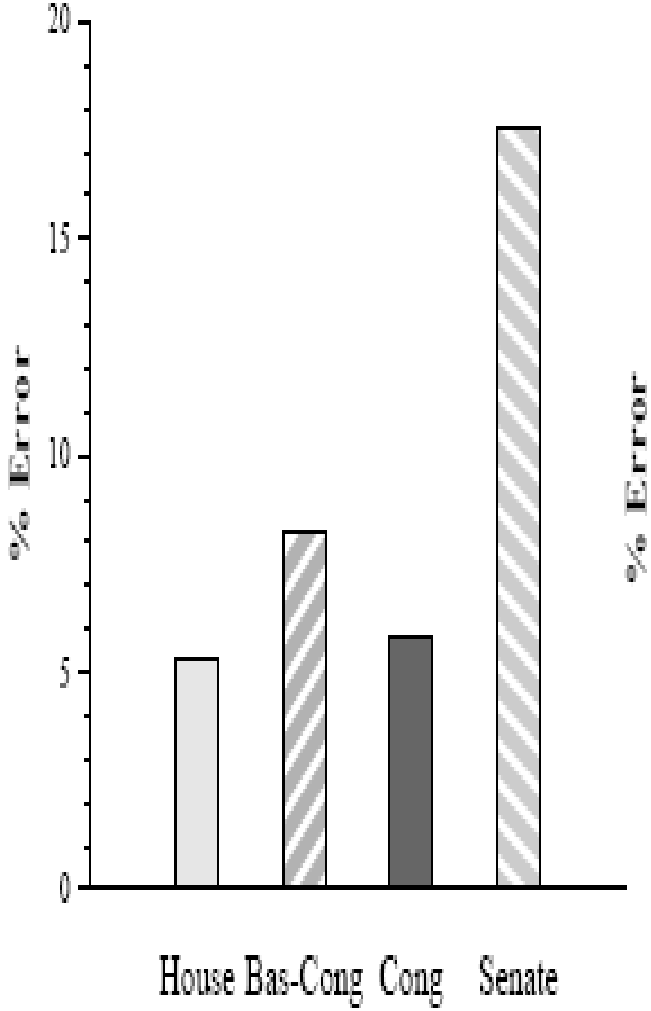
Figure 11: *Nested-integrated* Rewriting

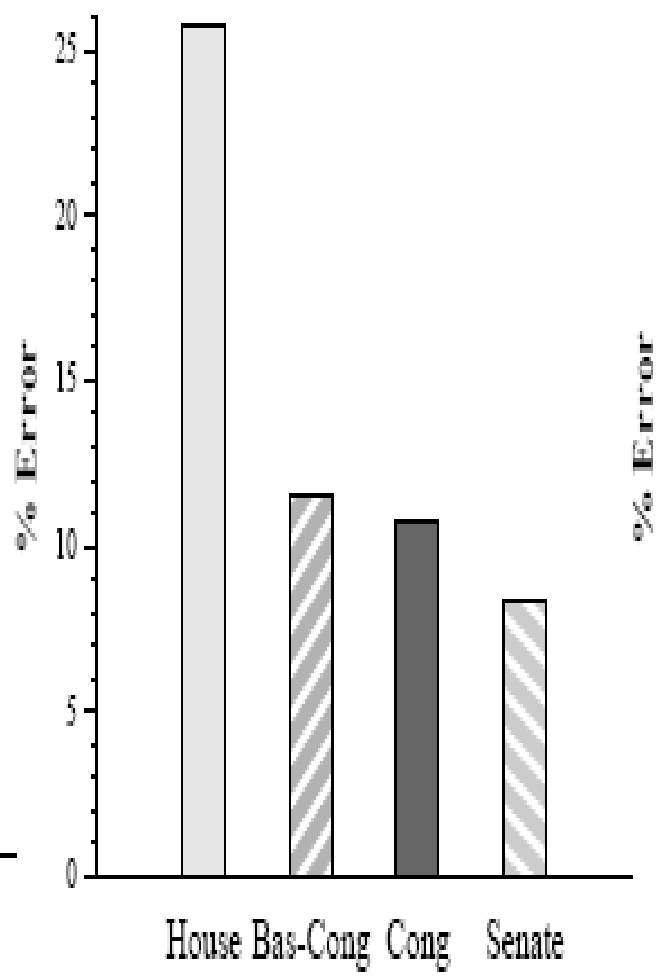
EXPERIMENTS:

- 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:





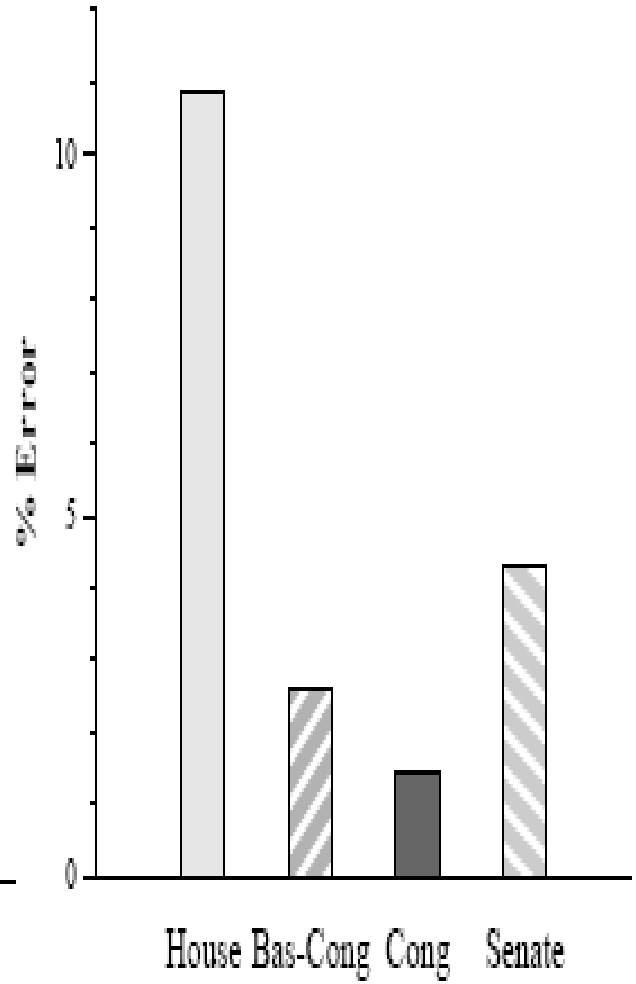


Figure 14: Query Q_{g0} Error Figure 15: Query Q_{g3} Error Figure 16: Query Q_{g2} Error

- 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

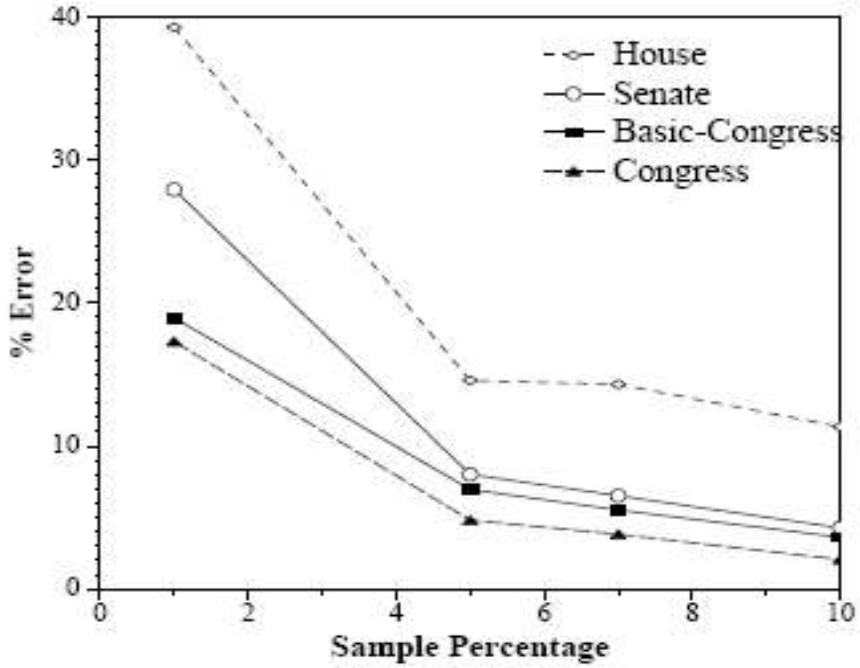


Figure 17: Sample Size vs. Accuracy (Query Q_{g2})

Performance of Rewriting strategies:

Technique	Sample Percentage		
	1%	5%	10%
Integrated	1.3	3.8	6.8
Nested-integrated	1.2	3.3	6.0
Normalized	1.7	14.0	27.3
Key-normalized	1.8	14.3	28.4

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

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???