Data Sampling using Congressional sampling

by Juhani Heliö

Overview

1. Introduction
2. Data sampling as a concept
3. Uniform random sampling
4. Congressional sampling
5. Results of Congressional sampling
6. Summary

Explanations for the presentation

In this presentation I will use following words with following meanings:

- Data set: The entire data being sampled.
- Data point/point of data: Any single tuple or entry or similar that is found in the data set.

1. Motivation

Motivation

- Storing big data can be difficult task on its own right but how does one actually use the stored data? Because of the large volume of the data, using the data can be difficult.

Example: Company has enormous amounts of data stored and wants to construct an average sales record to use in its decision making. How could this be done as fast and efficient?

2. Data sampling as a concept

What is data sampling?

- Main idea is to take a statistically significant sample of data and then analyse this sample rather than having to use the whole original data set.
- This way analysing huge amounts of data can be done faster and more efficiently.

3. Uniform random sampling

1. What is Uniform Random Sampling
2. Why Uniform random sampling is good...
3. ...and why it’s not very good after all
4. Example 1: US Census database
5. Example 2: Very sparse data

What is Uniform Random Sampling

Uniform random sampling is a simple and old sampling method

Key concept:

- Select points of data at random from the whole data set to the sample.
- Selection is done so that all the points of data have the same chance to be chosen to the sample.

Pros

1. Works well with simple queries like trying to find average of the whole data set.
2. Uniform random sampling is fast, O(n)

Cons:

Uniform random sampling has despite its good sides a critical flaw that can lead to inaccuracies in the result sample.

Key problem lies with grouped data. If the size difference between groups is too large, uniform random sampling can cause problems. In the following examples we examine 2 similar scenarios and see that uniform random sampling struggles with largely sparse data and large differences between sizes of groups.

Example: US Census database

-US Census database contains data of all the citizens in the nation. In this example an analyst wants to make a query to the Census DB asking for average income of each state. Because of the large volume of data in the DB a sample will be made. If uniform random sampling would be used in this instance inaccuracies would occur possibly rendering the sample unusable.

-This occurs because of the differences in the populations of each state. States with low population will not have many points of data selected. If too few data points are produced into the sample the resulting average calculation will not be accurate enough to be used. This problem could be fixed by creating an larger sample but this would reduce the effectiveness of sampling.

Problem summary

Given large number of groups from which large majority are small, uniform random sampling needs to either consume nearly the entire data set to satisfy the error bound or give inaccurate answer which probably will be useless to the user.

This leads to having less benefit from the sampling or even negative benefit due to the sampling overhead.

Solutions

There are many solutions for the problem and in this presentation will focus on Congressional sampling developed to enhance the Aqua system.

4. Congressional sampling

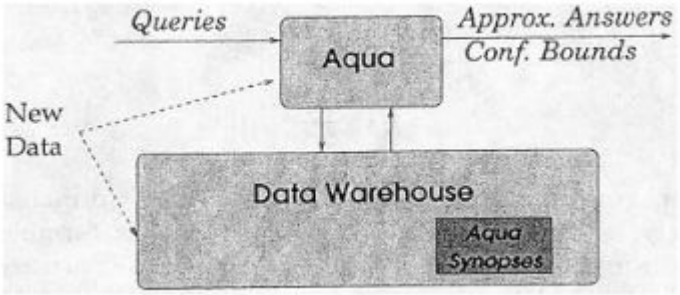
1. Introduction
2. Aqua
3. House
4. Senate
5. Basic congress
6. Congress

Introduction

Congressional sampling is a biased sampling method developed to enhance the Aqua system. This sampling method is actually four methods of sampling. Congressional sampling has taken its inspiration from the US political system, hence the name.

Aqua

- A system designed to sit between traditional DBMS and the users of the database.
- Aqua provides approximate query answering.
- Enhancing this system has been the main motivation for developing this sampling method.



House

- Do an uniform random sample over the entire data set.
- This will favor the large subgroups of the data set as per with uniform random sampling.

- This also means that House in itself is bad at sampling groupBys

Senate

- Take an equal sized sample from all subgroups of the sample
  - This division is done by dividing the sample size by the number of subgroups.
- This method heavily favors small subgroups of the data set.
  - Because the groups are even sized the small groups get disproportionately large amounts of points of data in their samples compared to large groups.
- The Senate thus will perform worse than the House with data containing only a few small groups.

Basic congress

- The basic congress is a combination of the house and senate samples.
- This method of sampling would be fair to both large and small groups
- However, this would also mean that the sample created would be twice as big
- This is mitigated by the following strategy:
  - For all subgroups g in the samples made with House h_g and Senate s_g do:
    - Take the larger of h_g and s_g into the basic Congress sample
    - Then the sample sizes are uniformly scaled down so that the overall sample size is the same as house or senate would have.

Problem of Basic congress

The Basic congress method is still somewhat flawed: Consider a data set with 4 groups of tuples with sizes respectively: $\{a_1, b_1\}$ 3000, $\{a_1,b_2\}$ 3000, $\{a_1,b_3\}$ 1500 and $\{a_2,b_1\}$ 2500. We take samples with sample size X = 100.

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

Figure 5: Expected sample sizes for various techniques, for $X = 100$.

In the table we can see the different samples done with house and senate and also with Basic congress and Congress.

The problem in Basic congress is that it focuses on the extremes.

- In the case we would like to make a sample with the values of A, Basic congress will allocate 77.3 and 22.7 units of space in these groups. This could lead to inaccuracies in the a_2 group.

This problem is addressed in the Congress method of sampling

Congress

Basic concept of the congress is to use stratified biased sampling to construct a sample.

Unlike the Basic congress,the Congress method considers all the possible groupings in the data and constructs the sample out of those.

In the case with the figure above, possible groups would be {A, B}.

The sample would then be taken using these groups and then combining them using the same method used in Basic congress.

Optimization is then done to ensure the sample size stays the same.

5. Results of Congressional sampling

To test the validity of this method three tests were conducted with different groupings: No groupBy columns, two groupBys and three groupBys. Results:

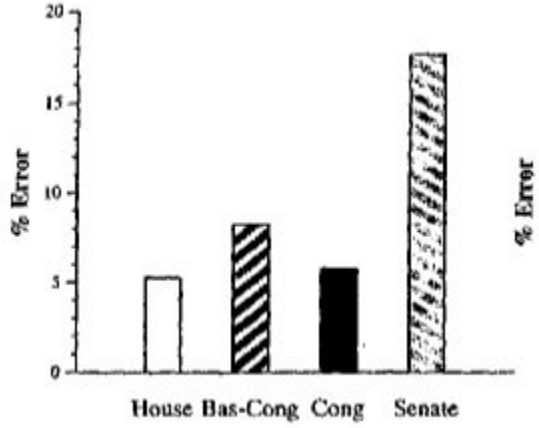


Figure 13: Query Q_{g0} Error

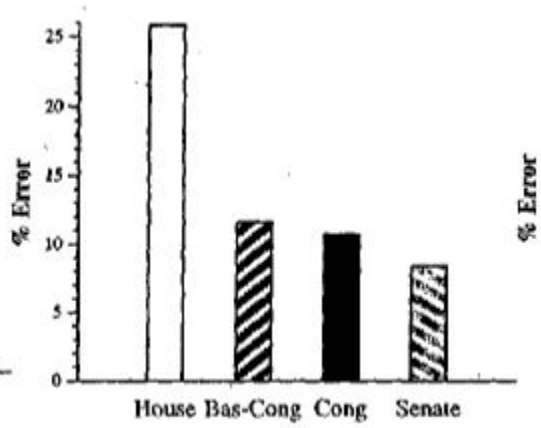


Figure 14: Query Q_{g3} Error

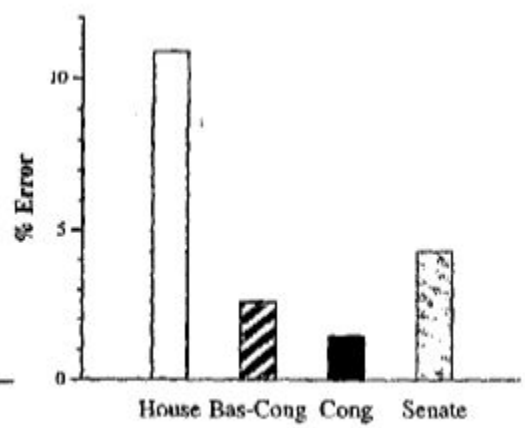


Figure 15: Query Q_{g2} Error

- The house performs poorly with any groupBys but when no groupBys were made house was the most accurate. This is due to the focus of House on the allocation of space to the large subgroups.
- Senate on the other hand focuses on the allocation of space to the small groups and thus performs poorly with no groupBys.
- Basic congress performs slightly poorly than Congress as it gives more focus on the extreme groups but still tries to balance them out.
- Congress performs the best or nearly the best in all of the cases and is the most consistent. As the other methods try to focus on one aspect of sampling, congress does not focus on any particular aspect and thus performs the best.

6. Summary

In this presentation we have explored different sampling methods:

-Uniform random sampling despite its appealing simple and fast nature was found to be lacking with more complex queries

-Congressional sampling, a biased sampling method, was found to be a good alternative with an effective solution to this problem.

Q&A