Fast Stratified Sampling for Approximate Queries with Indexes

Yunnan Yu

Department of Computer Science and Engineering
University at Buffalo
yunnanyu@buffalo.edu

1 Introduction

In typical business intelligence and reporting applications, ad-hoc queries are very common [3]. Users often execute range aggregation queries over large and frequently updated datasets and require low latency. Sampling-based Approximate Query Processing (S-AQP) is a popular technique that draws samples from the query range and computes approximate answers with a form of probabilistic error guarantee as confidence interval. With the help of sampling indexes, the total cost of an approximate query is usually sublinear to database size.

However, the query efficiency of such system, in terms of the total running time to achieve a given confidence interval, may vary heavily depending on the data distribution and query selectivity. It is known that very selective predicates or heavy data skewness can cause poor estimations with a given sample size, which makes them not quite reliable in practice. The root cause is that the system may have to draw an excessive number of samples, which involves excessive number of random accesses in the sampling indexes, to achieve the desired confidence interval if the inherent data variance is high.

Stratified sampling with Neyman allocation is a well-known technique for reducing the estimator variance. Consider an ad-hoc range query on the US airline on-time performance dataset [1] to count the number of cancelled flights in some ad-hoc date range for analysis. Most date ranges have consistently low cancellations, allowing accurate `COUNT` estimates with few samples. However, ranges containing spikes (e.g., due to 9/11) lead to high variance, requiring more samples and increasing query latency. However, existing stratified sampling algorithms in AQP systems [5, 4, 2] require apriori data statistics or online table scans, and thus cannot leverage sampling indexes to deliver very low latencies.

2 Proposed System

Motivated by that, we propose a two-phase index-assisted stratified sampling framework for online aggregation and implement it in PostgreSQL based on a state-of-the-art index-assisted S-AQP system [6]. When an ad-hoc approximate query arrives along with a pre-specified confidence interval, the system will draw samples with user-specified size from the entire query range in the initial phase to collect the data statistics, and then perform a re-partitioning optimization using dynamic programming and compute the optimal sample size allocation for each partition using modified Neyman allocation. In the second phase, the samples will be drawn independently from each disjoint query range. Experiments show that the system can significantly improve the approximate query cost for query ranges with high variances compared to existing S-AQP systems and exact query processing by up to 76% and 99.97%. We will also discuss some possible future extensions for more complex queries.

References

[1] Data Expo 2009: Airline on time data, 2008. URL: `https://doi.org/10.7910/DVN/HG7NV7`.

[2] Surajit Chaudhuri, Gautam Das, and Vivek Narasayya. Optimized stratified sampling for approximate query processing. *ACM Trans. Database Syst.*, 32(2):9–es, June 2007.

[3] Bolin Ding, Silu Huang, Surajit Chaudhuri, Kaushik Chakrabarti, and Chi Wang. Sample + seek: Approximating aggregates with distribution precision guarantee. In *SIGMOD '16*, page 679–694, 2016.

[4] Srikanth Kandula, Anil Shanbhag, Aleksandar Vitorovic, Matthaios Olma, Robert Grandl, Surajit Chaudhuri, and Bolin Ding. Quickr: Lazily approximating complex adhoc queries in bigdata clusters. In *SIGMOD '16*, pages 631–646, 2016.

[5] Yongjoo Park, Barzan Mozafari, Joseph Sorenson, and Junhao Wang. Verdictdb: Universalizing approximate query processing. In *SIGMOD '18*, page 1461–1476, 2018.

[6] Congying Wang, Nithin Sastry Tellapuri, Sphoorthi Keshannagari, Dylan Zinsley, Zhuoyue Zhao, and Dong Xie. Approximate queries over concurrent updates. *Proc. VLDB Endow.*, 16(12):3986–3989, August 2023.