Learning-Based Frequency Estimation Algorithms

Chen-Yu Hsu, Piotr Indyk, Dina Katabi, Ali Vakilian

This paper introduces learning-based frequency estimation algorithms that leverage patterns in input data to improve accuracy while still maintaining their formal error guarantees. Traditional streaming algorithms previously discussed in this course such as CM Sketch, the authors address, rely on hash-based approximations and make no assumptions about input distributions. However, the authors highlight how these methods suffer from hash collisions, inducing estimation errors, especially for high-frequency elements which they refer to as heavy hitters. The authors propose augmenting these classical methods with machine learning by training a model to predict properties of heavy hitters based on a subset of the data stream. The learned model helps assign high-frequency elements to unique buckets, preventing them from colliding with low-frequency elements, thereby reducing overall estimation errors. The paper provides theoretical guarantees that CM Sketch with learning achieves an error bound logarithmically smaller than pure CM Sketch and that it is asymptotically optimal. The authors then perform experiments on two real-world datasets: internet traffic estimation and search query frequency. They demonstrate that performance improves but across a wide range (18-71%) when compared to traditional streaming algorithms. The authors make the assumption that the learned model can generalize from a small training set to unseen data and that frequency distributions exhibit predictable patterns, such as Zipfian distributions in natural language and network traffic. However, their method may be less effective in cases where the data does not exhibit predictable patterns or where frequent elements shift dynamically or where they simply do not have enough input data to effectively learn. Possible extensions include federated learning to reduce the compute overhead that learning has and to explore more complex models instead of simple neural networks.

SNARF: A Learning-Enhanced Range Filter

Kapil Vaidya, Subarna Chatterjee, Eric Knorr, Michael Mitzenmacher, Stratos Idreos, Tim Kraska

This paper presents a novel learned range filter that efficiently supports range queries for numerical data. Traditional range filters optimize for different workload characteristics, but there is no flagship filter that consistently outperforms others across all query distributions. The authors introduce SNARF as a new approach by using a simple learned model to approximate the empirical CDF of a dataset. This model maps keys into a sparse bit array which is then compressed for efficient storage. When a query is received, the model estimates the key's position within the compressed bit array, significantly reducing false positives compared to existing range filters both in theory and in practice. The authors performed experiments with results that show that SNARF achieves a false positive rate up to 50 times better than SuRF and 100 times better than Rosetta under the same space constraints. The authors also show that integrating SNARF into RocksDB improves query execution times by up to 10 times for read-heavy workloads. The authors make the assumption that numerical datasets follow predictable distributions that can be effectively modeled with a CDF approximation and that learned models can generalize well across different datasets which isn't always true in practice. It is possible that SNARF struggles with irregular distributions or distributions where learned approximations fail. Additionally, its performance depends on the chosen compression scheme

which adds another hyperparameter to the fine-tuning of the algorithm. Potential extensions include dynamic update support. Overall, the observed performance boost was extremely impressive and is a very exciting insight in the field, suggesting that algorithms with learning can achieve very good results in practice.