Technical Perspective: Relative Error Streaming Quantiles

Rasmus Pagh
BARC
University of Copenhagen, Denmark
pagh@di.ku.dk

The paper *Relative Error Streaming Quantiles*, by Graham Cormode, Zohar Karnin, Edo Liberty, Justin Thaler and Pavel Veselý studies a fundamental question in data stream processing, namely how to maintain information about the distribution of data in the form of *quantiles*. More precisely, given a stream S of elements from some ordered universe $\mathcal{U}$ we wish to maintain a compact summary data structure that allows us to estimate the number of elements in the stream that are smaller than a given query element $y \in \mathcal{U}$, i.e., estimate the *rank* of y. Solutions to this problem have numerous applications in large-scale data analysis and can potentially be used for range query selectivity estimation in database engines.

The challenge is to make the size s of the summary data structure (also known as a *sketch*) as small as possible while bounding the error on rank queries. If data is available in sorted order and has a known size n, an optimal solution is to store s *quantiles*, that is, elements in S with specific ranks. If S contains n elements we can store the elements in S that have rank $n/(s+1)$, $2n/(s+1), \ldots, sn/(s+1)$, which would allow us to approximate the rank of any query element $y \in \mathcal{U}$ up to an additive error of $n/(s+1)$. (Here we ignore rounding issues for simplicity.) In other words, to achieve additive error εn it suffices to use space $1/\varepsilon$.

Surprisingly, it is possible to achieve almost the same space efficiency in the setting where elements are presented in arbitrary order, and no bound on n is known. This was established in a series of papers culminating in the work of Karnin, Lang, and Liberty [4]. The simplest version of their data structure, now often referred to as the "KLL sketch" has become widely used in data processing pipelines. The popularity may be partly due to another property of the KLL sketch: Sketches for datasets S_1 and S_2 can be combined to form a sketch for the multiset union of the data sets. This property, referred to as *mergeability* means that computation of KLL sketches can be efficiently carried out in parallel and distributed settings with little communication.

In *Relative Error Streaming Quantiles*, which appeared in PODS '21 [2], the authors consider the situation where more precise information about the *tails* of the distribution (i.e., the largest and smallest elements) must be maintained. This setting is important because many data distributions have few elements in the tails, and thus the KLL sketch may have little or no information about the distribution of the data in the tails. A natural way of asking for more precise information on elements on the lower tail is to ask for a rank error guarantee that is *multiplicative*, i.e., a rank estimate that is within some factor $1+\varepsilon$ from the true rank. In contrast, the KLL sketch offers an *additive* error guarantee, which is much weaker for elements with rank $\ll n$. When a multiplicative rank guarantee is possible, by symmetry, improved guarantees can also be achieved for query elements with rank close to n.

Though multiplicative error is a desirable property, past solutions were significantly less efficient than the KLL sketch. A gap in performance is unavoidable for a broad class of algorithms, as shown in a lower bound of Cormode and Veselý [3]. Still, past solutions required larger overheads in the dependence on the stream length n or the approximation parameter ε. To improve on this, the authors go back to the KLL sketch and describe a simple but clever modification that improves past results and in fact nearly matches the lower bound. Though the algorithm, called *ReqSketch*, is relatively simple, analyzing it is not and requires ideas that go significantly beyond those needed for the KLL sketch.

The new, mergeable sketch, is not only a great theoretical contribution — it is already available along with the KLL sketch as part of the Apache Datasketches library of streaming algorithms [1] and is ready to impact computing practice!

1. REFERENCES

[1] Apache datasketches. `https://datasketches.apache.org/`. Accessed: 2022-03-10.

[2] G. Cormode, Z. S. Karnin, E. Liberty, J. Thaler, and P. Veselý. Relative error streaming quantiles. In *Proceedings of the Symposium on Principles of Database Systems (PODS)*, pages 96–108. ACM, 2021.

[3] G. Cormode and P. Veselý. A tight lower bound for comparison-based quantile summaries. In *Proceedings of Symposium on Principles of Database Systems (PODS)*, pages 81–93. ACM, 2020.

[4] Z. S. Karnin, K. J. Lang, and E. Liberty. Optimal quantile approximation in streams. In *Proceedings of Symposium on Foundations of Computer Science (FOCS)*, pages 71–78. IEEE Computer Society, 2016.