Stratified Sampling over Streaming Sample Join for Approximate Aggregation

Abe Wits
CWI
wits@cwi.nl

Lefteris Sidirourgos
CWI
e.sidirourgos@cwi.nl

Hannes Mühleisen
CWI
hannes@cwi.nl

ABSTRACT

Approximations of aggregations are sufficient for many purposes. However it is hard to estimate an aggregation over a join efficiently. With a stream sample algorithm, it is possible to obtain a uniform sample of the join result without calculating the join explicitly. You can then obtain an unbiased estimate of the aggregate using this sample. However, the time complexity of stream sample join is linear in the input size. We propose a faster alternative that is linear in the output size by using non-uniform samples with known distribution. The non-uniformity introduces a bias in the estimate. To provide an unbiased estimate, we modify the aggregation step. We improve on the precision of the aggregation further by combining and extending sample join techniques and (stratified) sampling techniques.

We introduce three new stream sample join algorithms. The first is a generalized version of the stream sample join algorithm, called the weighted stream sample join algorithm (WS-join). The second is the approximate weighted sampling join algorithm (AWS-join), a fast heuristic variation of this algorithm. The third is the uniform sampling join algorithm (US-join), an even faster exact alternative that provides samples according to one fixed weight distribution.

We combine these stream sample join techniques with several aggregation techniques, and analyse which combination of join and aggregation algorithm and what choice of weights is optimal in different settings.

We develop approximate weighted sampling (the core of AWS-join). This is the first algorithm (to our knowledge) to take approximate weighted samples. It has a better asymptotic complexity than all known exact weighted sampling algorithms. AWS is analysed theoretically to provide heuristics to estimate the error in the selection probability. These heuristics are tested experimentally.

In Figure 1, the different types of weight distributions are shown (the lower part), and a high level overview of our aggregation of join algorithm is shown. Different aggregation techniques exist for each type of weight distribution

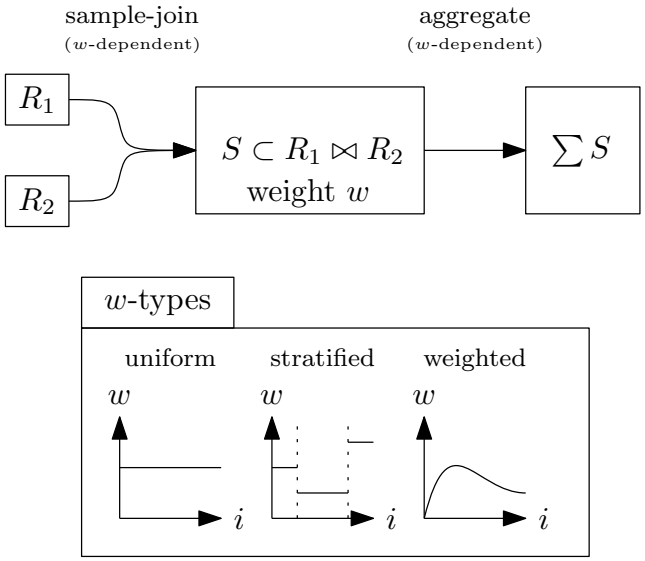

Figure 1: Global overview of our approximate aggregation through joins.

(w-type), all with different variance properties. To obtain samples of the different types of weight distribution, different sample-join algorithms have to be used. The choice of weights influences both the time complexity of the sample-join step and the precision of the aggregation step.