Abstract

Join is one of the central problems in database theory, which has been extensively studied in both theory and practice. In this thesis, we focus on estimating join size and sampling from joins.

We present three results in this area. The first result is a new algorithm for estimating the size of cycle join of k binary relations in the property-testing model. The algorithm gives an unbiased estimator for the join size and it achieves running time of $O\left(\frac{\mathrm{IN}^{k/2}}{\mathrm{OUT}}\right)$ in expectation to obtain a constant-factor approximation with constant probability, where IN is total number of tuples in the input, and OUT is the join size. Moreover, for cycle join of k binary relations, $\mathrm{IN}^{k/2}$ is a tight upper bound for OUT.

The second result is an algorithm for generating uniform samples from the join results without computing the full join in a model that allows a linear-time preprocessing step. For the class of sequenceable queries, a sample can be drawn uniformly at random from the join results in $O(\mathrm{IN}^{\rho}/\mathrm{OUT})$ time in expectation where ρ is fractional edge cover number of the hypergraph. For non-sequenceable queries, the running time for drawing a sample is $O(\mathrm{IN}^{\rho+1}/\mathrm{OUT})$. These results hold for both full join queries and join-project queries. Prior to this work, the only solution to this problem with formal guarantees is to either precompute the full join results, which has $O(\mathrm{IN}^{w})$ preprocessing and storage cost where w is the *fractional hypertree width* of the query and $O(1)$ sampling cost, or compute the full join at sampling time, which has no preprocessing cost but $O(\mathrm{IN}^{w})$ sampling cost. We also adapt our algorithm to solve the join size estimation problem. We show that after drawing a constant number of samples, a constant-factor approximation to the join size can be obtained with constant probability.

The previous two results for join size estimation all collect samples after the query is given. In the third work, we present an algorithm for join size estimation using synopses collected separately for each relation before the query is given. A major challenge here is that join queries often come with ad hoc selection predicates that are only known at query time, thus there is no way to take the predicates into account when the statistics are collected. Our new sampling method, called *two-level sampling*, combines the advantages of Bernoulli sampling, Correlated sampling and End-biased sampling while making further

improvements. Two-level sampling gives unbiased estimator for join size and provides error guarantees on its estimates. It is also very easy to implement, requiring just one pass over the data. It only relies on some basic statistical information about the data, such as the ℓ_k-norms and the heavy hitters. The algorithm can also be extended to support estimating other aggregates like `SUM` and `AVG`. Both analytical and empirical comparisons show that the new algorithm outperforms all the previous algorithms on a variety of joins, including primary key-foreign key joins, many-to-many joins, and multi-table joins.