Paper Shredding 3: Deep DB

Summary:

The authors take a different route and propose a new data-driven approach for learned DBMS components which directly supports changes of the workload and data without the need of retraining. The results of the empirical evaluation demonstrate that data-driven approach not only provides better accuracy than state-of-the-art learned components but also generalizes better to unseen queries.

The authors indicated RSPN Rational sum-product network) based on traditional sum-product networks. It has three main pros compared with SPN: 1. While SPNs support only single tables and simple queries (i.e., no joins and no aggregation functions), RSPNs can be built on arbitrary schemata and support complex queries with multiway joins and different aggregations. 2. RSPNs support direct updates. 3. RSPNs also include a set of database-specific extensions such as NULLvalue handling and support for functional dependencies. They also constructs base ensemble for a given database scheme which can calculate the RDC to judge the independency between two tables and build RSPN if necessary.The authors indicate extensions on basic schemes about confidence intervals and ensemble optimization. On the evaluation part, the authors used different methods including perfect selectivities to calculate the cardinality and generalization of the DeepDB to show its performance compared with the traditional training process.

Pros:

1. The result of data-driven model can be used in different DBMS tasks such as query answering and cardinality estimation, whereas the combination use of data-driven and workload-driven models can have a better performance.
2. A model in DeepDB augments a database similar to indexes to speed-up queries and to provide additional query capabilities while we can still run standard SQL queries over the original database, which supports generalization for different database tasks.
3. RSPN on AQP: RSPNs can support joins of multiple tables. And DeepDB supports ad-hoc queries and is thus not limited to the query types covered by the training set.
4. DeepDB performs better not only on cardinality but also on training time and storage cost, which is a great part in the traditional training process on databases.

Cons:

1. The confidence interval needs two simplifying assumptions: (i) the estimates for the

expectations and probabilities are independent, and (ii) the resulting estimate is normally distributed, whereas the normally distributed may not be true in many circumstances.

2. Using the relative cost to select the RSPN which has highest mean RDC value and the lowest relative creation cost which is a simplification assumption, especially on complex relative tables.
3. The probabilistic database lacks sensible complexity calculation methods for query for DeepDB, which needs more data support and further research.