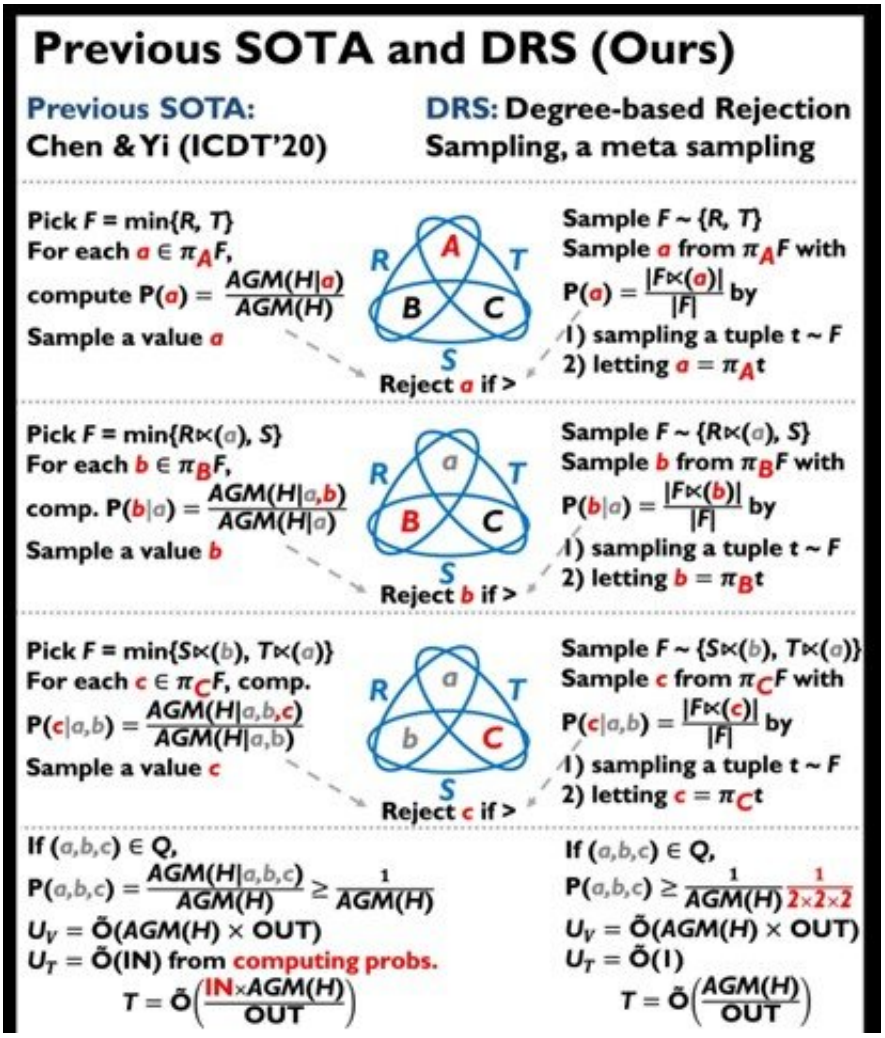
Advancing toward optimal sampling: The zenith of data analysis

July 10 2023



The graphical representation delineates the difference between the conventional methodologies and DRS. Unlike traditional methods that compute intricate probability distributions at each stage - A, B, and C - the DRS permits swift sampling by first establishing a sample space with straightforward probabilities for A, B, and C, and subsequently selecting values. This eliminates the need for

elaborate probability distribution calculations. Credit: POSTECH

The world witnessed a monumental face-off between human intelligence and artificial intelligence in March 2016. The computer program AlphaGo honed its skills from a substantial database and emerged victorious against a human opponent in Go, a game renowned for its complexity in calculating countless possible moves.

The importance of quality data for AI's continuous evolution is undeniable. AI has seamlessly integrated into such sectors as healthcare, finance, and education, while its advancements rely heavily on the availability of robust data for learning.

Data is typically stored in distributed groups known as tables. For an AI to glean insights from these table-stored data, a "join" process is deployed to amalgamate these disparate tables into one comprehensive table. The sheer scale of this resultant table presents challenges in terms of storage, while the join process itself can be quite time-consuming. Even now, developing techniques for swift and uniform data sampling from tables remains a complex puzzle yet to be solved in data science.

In a significant breakthrough, a POSTECH research team led by Professor Wook-Shin Han (Graduate School of Artificial Intelligence) along with Ph.D. candidate Kyoungmin Kim (Department of Convergence IT Engineering) proposed a novel method for optimal sampling of data stored across various tables. This new technique managed to generate results rapidly.

The research was published as part of the *Proceedings of the 42nd ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems* (PODS 2023). This marked a momentous occasion as it was the first

instance of a paper from a Korean research team being presented at this symposium in its 42-year history.

The researchers pioneered a method named degree-based rejection sampling (DRS), which falls under the umbrella of meta-sampling. Conventional methods necessitated pre-calculating probabilities for every value in the sample space before any value could be extracted directly. By contrast, the DRS method proposed by the team initiates with the extraction of a sample space with a simple probability distribution based on the degree of specific values, subsequently drawing values from this sample space.

The team convincingly demonstrated that at least one sample space affords a greater probability than the elaborate probabilities computed via traditional methodologies for any random value that can be selected. This implies that values can be obtained with similar probabilities as traditional methods via rejection sampling. In this way, only the probability of extracting a sample space is merely multiplied as a constant value to the probability of sampling a value, avoiding complex probability calculations, and allowing for rapid data sampling.

Moreover, the team employed a technique known as generalized hypertree decompositions (GHDs) to extend the method, which involves analyzing a query in a tree format during the join procedure of integrating tables. If an entire query is processed using a singular join algorithm, it can lead to high time complexity, particularly when the query contains multiple join relations.

Using GHDs allows for conducting join operations on smaller sub-queries instead of the entire query, and subsequently combining the results, thereby reducing time complexity. The research team integrated GHDs with DRS to augment the latter, guaranteeing a lower complexity than the original DRS in certain instances.

Heading the research, Professor Wook-Shin Han expressed high hopes for the innovative method, stating, "This technique can be universally applied to all queries, regardless of whether the data structures form a tree, exhibiting hierarchical relationships, or a cycle, depicting circular relationships. It promises to significantly improve both speed and accuracy in the data sampling process for machine learning."

More information: Kyoungmin Kim et al, Guaranteeing the Õ(AGM/OUT) Runtime for Uniform Sampling and Size Estimation over Joins, *Proceedings of the 42nd ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems* (2023). DOI: 10.1145/3584372.3588676

Provided by Pohang University of Science and Technology