Name: Bharghav Srikhakollu

CS549 – ST/Scalable Querying & ML – Spring 2022

Topic:

Scalable and Usable Relational Learning with Automatic Language Bias

https://web.engr.oregonstate.edu/~termehca/ready/papers/sig.mod-rdm365.pdf

What is the problem discussed in the paper?

From the list of possible relational models over the relational databases and knowledge bases, to understand the most effective relational model over the possible ones, users of the current learning systems must restrict the structure of these candidate models using "Language Bias". To develop these language bias, Machine Learning experts usually spend lot of time for inspection of data and to finalize the model they must perform multiple rounds of trial and error for finding the effective language bias. For scaling relational learning to large data there are primarily two challenges. One, the space of possible hypothesis that the relational learning algorithm should explore depend on the Datalog programs. These are dependent on the schema of the database where there is huge data on the database. Second, for the defined hypothesis in the space, the algorithm should evaluate the quality of these which is time consuming over the large data.

Why is it important?

At present, the language bias for relational learning is something written manually by the expert user, and these should be rewritten for each of the learning task and even when the new data evolves over the database. Most importantly it requires many lines of code to build the language bias. To handle these situations, it is important to understand a better model.

What are the main ideas of the proposed solution for the problem?

For scaling relational learning to large data, a novel and usable approach is proposed called "Auto Bias". Without much user intervention, this novel method automatically generates language bias and techniques were proposed to explore the hypothesis space.

To induce language bias automatically, it focuses on the database constraints available in the schema of the database. It uses the exact and approximate database constraints. To understand the bias over the large data, for this method they proposed random sampling techniques so that the learning algorithm explore large hypothesis space effectively. Stratified sampling method is used for gaining diverse hypothesis. Also, these sampling techniques helps to evaluate the quality of each hypothesis efficiently. The proposed methods were evaluated over large databases to understand the efficiency of the model.