2022 VLDBSS 汇报

华东师范大学
数据科学与工程学院

翁思扬

CONTENT

1 小组介绍

2 实操结果

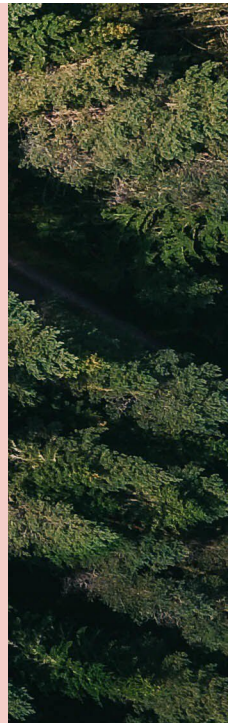
3 方法亮点

4 实践收获

1

PART ONE

小组介绍



2

PART TWO

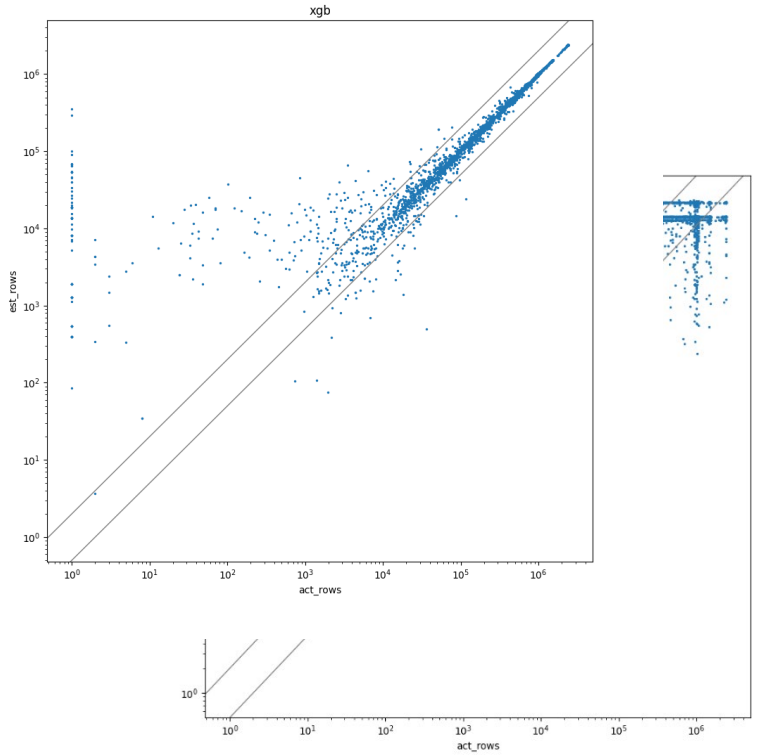
实操结果

实操结果

LAB 1 基数预估

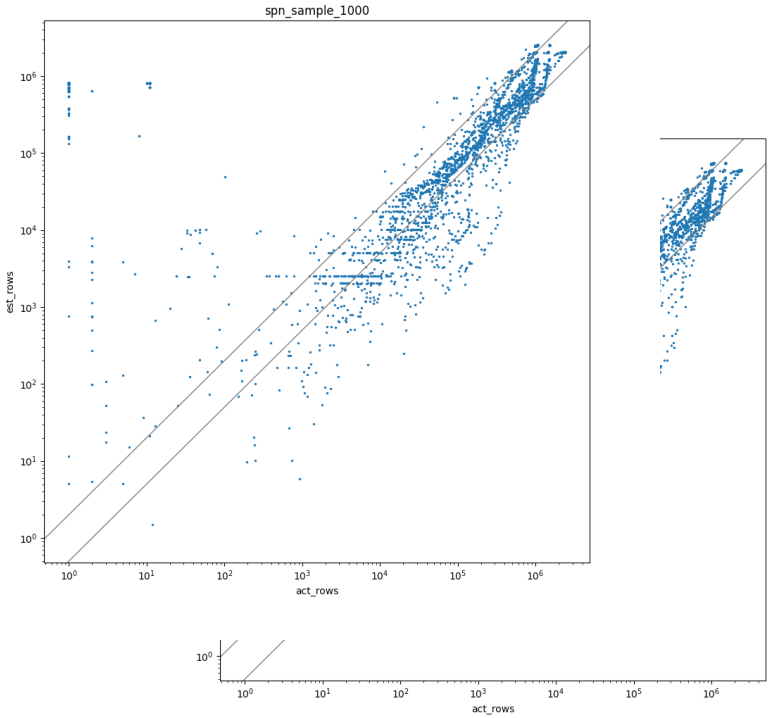
基于NN的查询驱动方法

提取查询特征，训练神经网络



基于SPN的数据驱动方法

切分数据，独立估计每个子集再合并

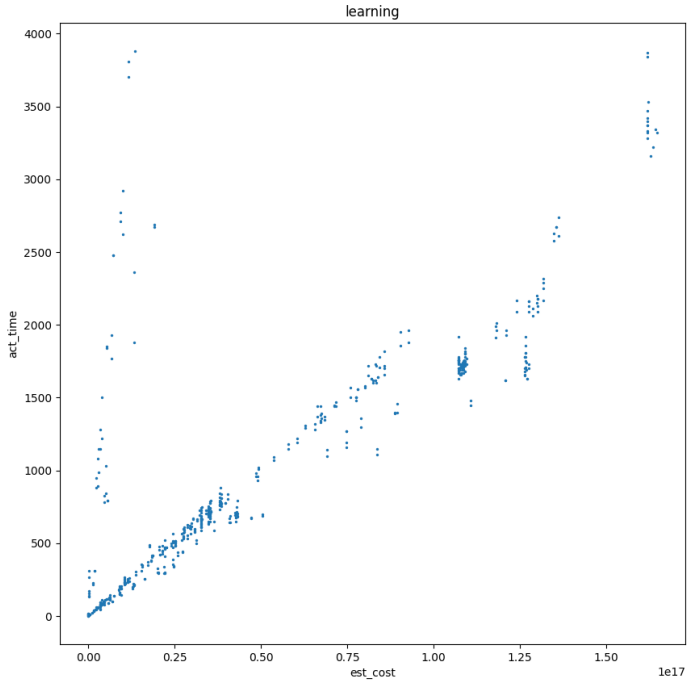


实操结果

LAB 2 代价估算

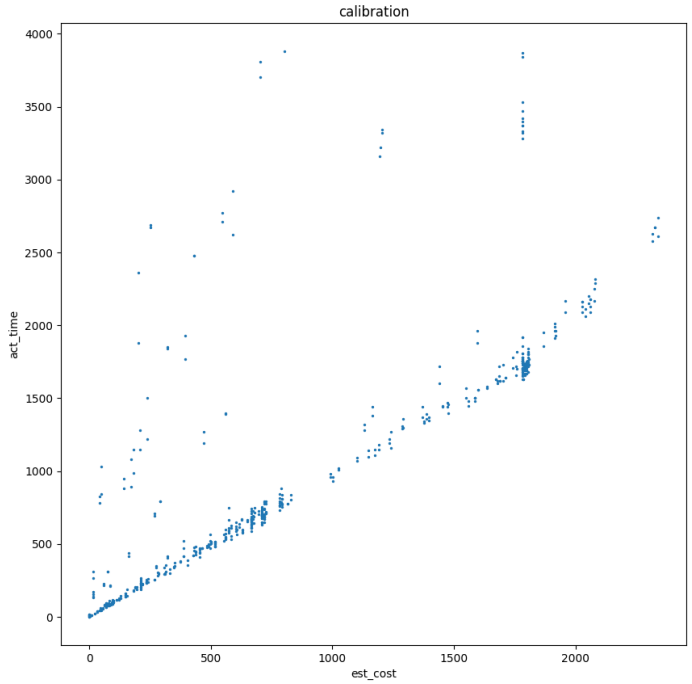
基于计划特征的NN预估

提取计划特征，训练神经网络



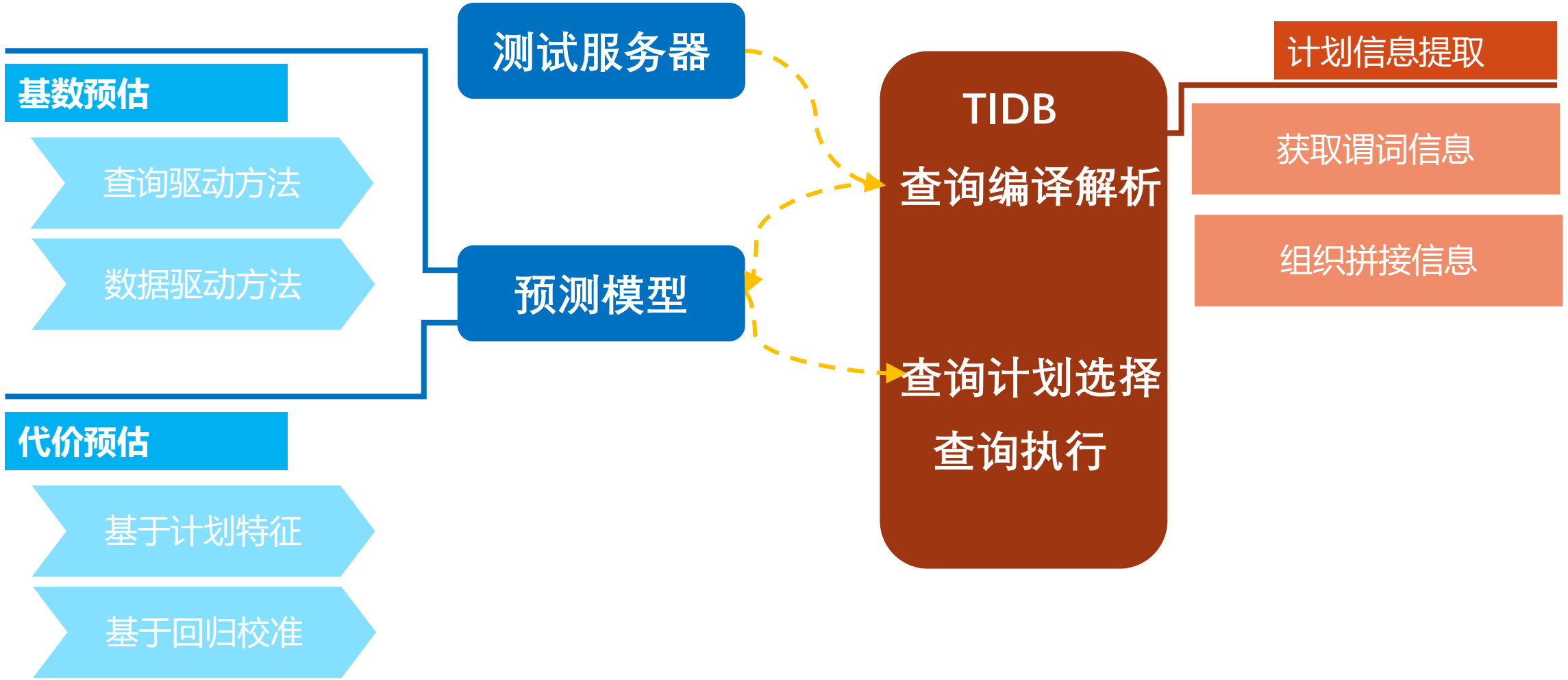
基于回归校准的代价计算

计算代价公式，回归校准权重



实操结果

LAB 3 模型结合数据库



实操结果

LAB 4 端到端的代价估计

计划特征提取

遍历计划算子，逐一提取特征

构造扫描bitmap

计划特征重编码

编码谓词条件

编码计划特征

模型训练

采用Tree-LSTM模型

采用q-error作为损失函数: $\min(\frac{act}{pred}, \frac{pred}{act})$

3

PART THREE

方法亮点

方法亮点

LAB 1 基数预估

基于NN的查询驱动方法

轻量化模型

- 简单模型：快速训练，便于加载
- 复杂模型：训练缓慢，加载困难

基于SPN的数据驱动方法

节点分割策略

- 优先提高各子节点的独立性
- 优先减少节点数量

LAB 2 代价估算

基于计划特征的NN预估

特征充分抽取

- 包含算子内部信息和计划结构信息
- 简单统计不同算子数量

基于回归校准的代价计算

轻量化模型

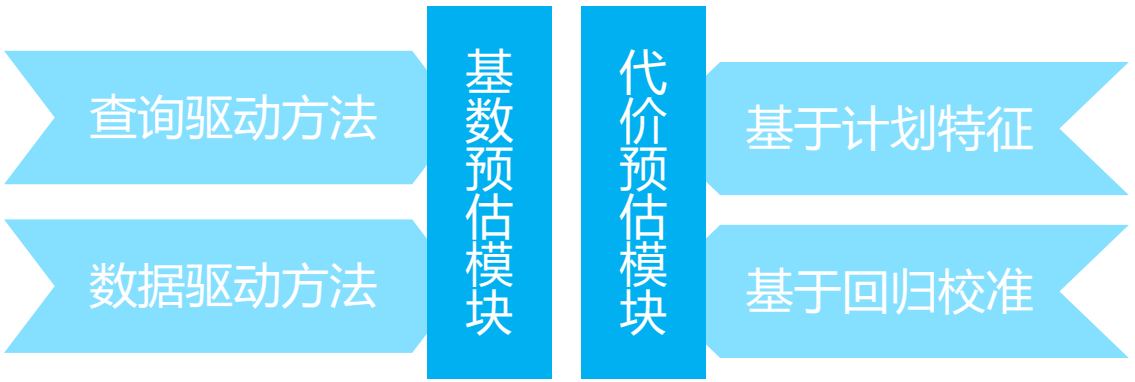
- 简单模型：快速训练，便于加载
- 复杂模型：训练缓慢，加载困难

方法亮点

LAB 3 模型结合数据库

模型加载解耦化

实现解耦接口，便于不同方法插拔



LAB 4 端到端的代价估计

全流程打通

不需要其他模型提供信息

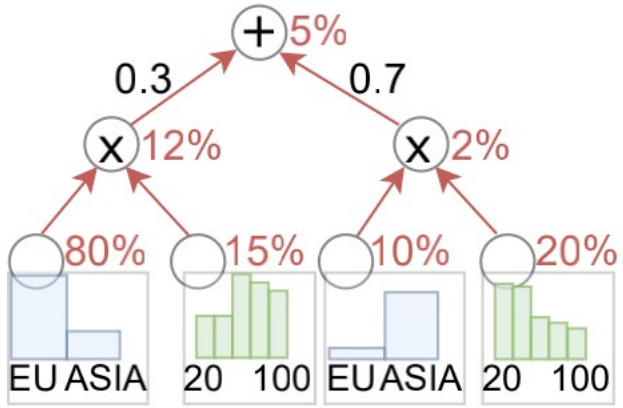
需要额外基数估计模型提供数据

4

PART FOUR

实践收获

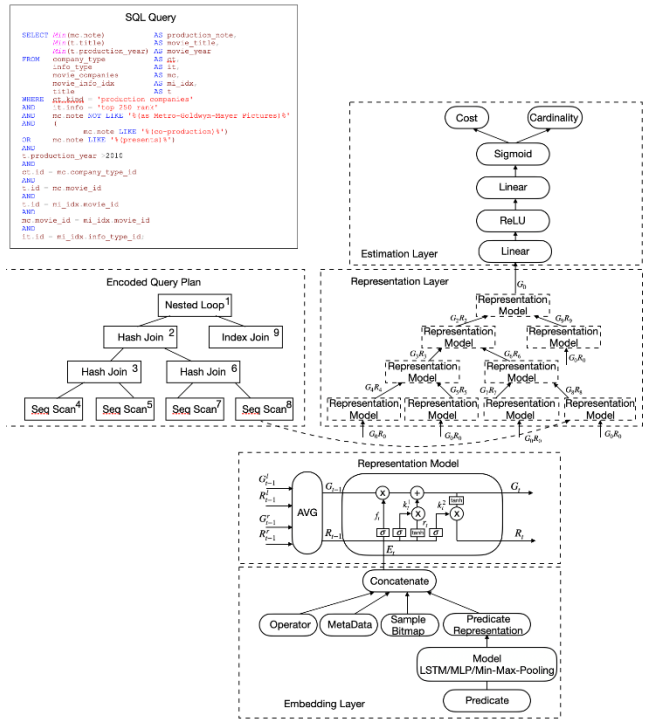
实践收获



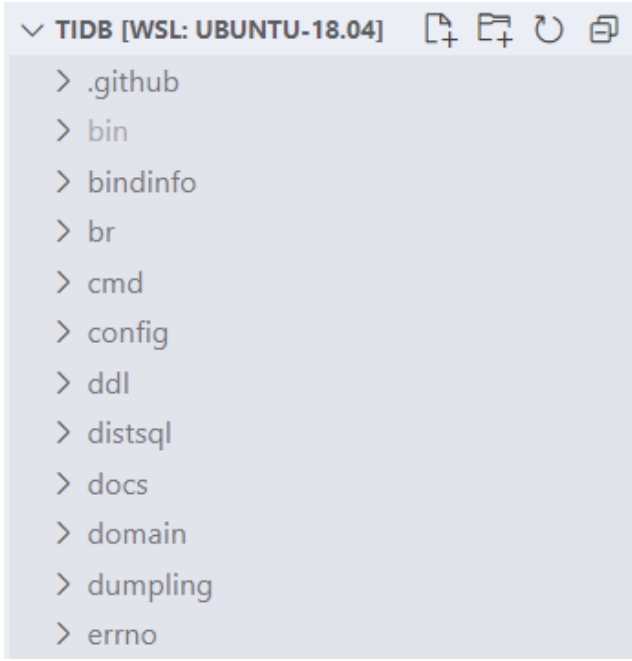

```
class SPN:
    """
    SPN represents a sum-product network.
    """

    def __init__(self, root):
        self.root = root
```

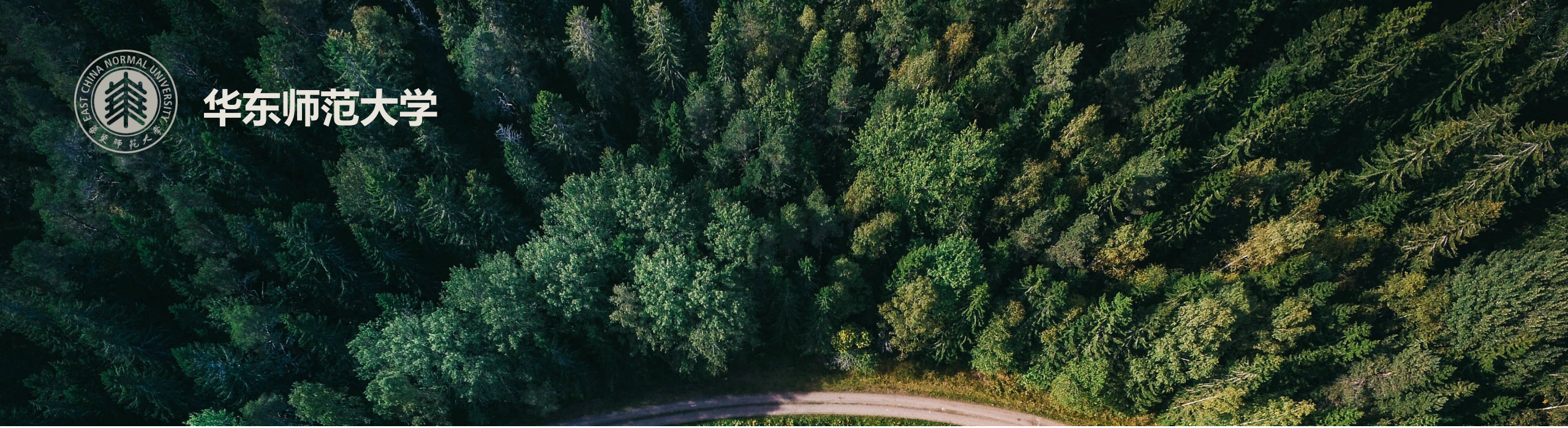
印证讲座知识



加深算法理解



接触数据库内核



THANKS