2024春季大规模数据处理技术 第2次作业

2024.04.28

一、实验简介

本次实验旨在让同学们了解并掌握PageRank算法及其优化与实现。我们以Stanford web graph作为原始数据集，在此基础上选择了其中一部分完全连通的子图作为我们的实验数据集。同学们需要手动实现PageRank算法及其相关优化算法，并在我们提供的数据集上进行排序，提交源代码、Top1000排序结果以及实验报告。

二、数据集结构

数据集为以‘,’分隔的 `web_links.csv` 表格文件。该表包含2列2,300,000行，其中第一行为表头，之后的每一行为两个整数表示一条边的两个节点。

- FromNodeId：表示该边出发节点对应的Id
- ToNodeId：表示该边终止节点对应的Id

三、实验要求

1. 载入数据 `web_links.csv` ，在其上实现PageRank算法并对其进行排序
2. 实现一种优化方式，以达到更小的内存开销以及更短的排序时间
3. 将Top1000排序结果保存至 `test_prediction.csv` 中，要求包含两列，第一列为NodeId，第二列为PageRank值。要求表的第一行为表头，表头名称为NodeId、PageRank_Value，因此最终表的大小应为 `(1001,2)`
4. 撰写实验报告，包括方法介绍、实现细节、机器配置、内存/时间开销与对比、排序结果等内容

四、提交内容

1. 源代码
2. 报告，保存为pdf格式
3. 输出结果文件 `test_prediction.csv`

以上内容**打包为一个.zip压缩文件**，重命名为“<学号>_<姓名>.zip”（如“123033919999_张三.zip”）提交至canvas平台，**截止日期为5月12日23:59**，迟交将扣分

五、评分标准

本次实验的评分会根据报告内容、在测试集上的预测评估、算法开销以及代码内容等方面进行综合评分