Group Skyline算法实现

2018214131 王彬彬 软硕181

目录

[Group Skyline算法实现 1](#_Toc531160317)

[1 实验环境 2](#_Toc531160318)

[2 实现内容 2](#_Toc531160319)

[3 实验结果 3](#_Toc531160320)

[4 实现亮点 3](#_Toc531160321)

[5 文件说明 4](#_Toc531160322)

# 1 实验环境

* Windows 10
* C++ 11
* g++

# 2 实现内容

本次实验实现了论文*Finding Pareto Optimal Groups: Group-based Skyline*中找出G-Skyline groups的Point-Wise Algorithm和Unit Group-Wise Algorithm，如论文中所示，分别简记为PWise和UWise+。论文中给出的算法是一种广度优先搜索算法，当数据规模较高或维度较高时，空间复杂度太高，因此实现这两个算法的深度优先搜索版本：PWiseDFS和UWise+DFS。在实验中发现，在有些高维数据集上，例如corr-8d，G-Skyline groups数量过多，如果将结果一直保存在内存中，则会出现内存分配错误，无法跑出结果，因此，当找到的groups数量达到一定阈值时，将结果写入磁盘文件，记为Disk版本，实验中这一阈值设为1000,000。同时，实现了BaseLine算法，枚举所有可能的group，从中挑选出合法的group。

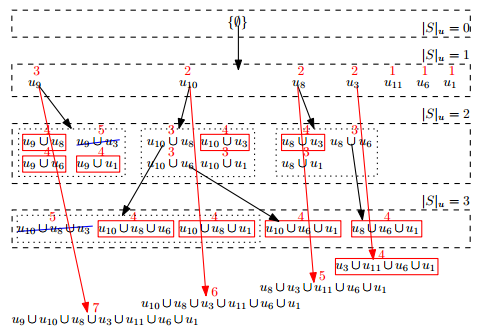
本次实验实现的算法总结如下表。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 实现算法 | 结果数量达到阈值时写入磁盘 | | 结果一直保留在内存 | |
| BFS | PWiseDisk | UWise+Disk | PWise | UWise+ |
| DFS | PWiseDFSDisk | UWise+DFSDisk | PWiseDFS | UWise+DFS |
|  | BaseLineDisk | | BaseLine | |

# 3 实验结果

# 4 实现亮点

1、对于算法UWise+，对subset pruning过程进行了微小的优化，如下图所示：



论文中给出的算法从到进行遍历，每次需要计算集合的大小，而实际上可以从到遍历，利用计算，避免重复计算，同时计算到即可停止subset pruning过程。

2、实现了论文中两个算法的深度优先搜索版本，有效地降低了空间复杂度。

3、实现了算法的Disk版本，能够在高维数据集上得到结果。

# 5 文件说明

|  |  |
| --- | --- |
| 文件名 | 说明 |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |