

# K-频繁项集挖掘并行化算法

## 1. 设计思路

本题目的要求是在规定的标准数据集上进行频繁项集的挖掘，因而自然而然地想到了两种著名的算法：Apriori算法和FP-Growth算法。由于测试数据集约含1800万条事务记录，预赛时考虑到Apriori算法在计算过程中需要重复扫描数据集（如题要求1-频繁项集到8-频繁项集的挖掘，则需要扫描8遍数据集）以及产生大量的候选集，而FP-Growth算法只需进行2遍数据集的扫描，以将事务记录压缩成树来推导频繁项集的方式来代替候选集的生成，效率更高。因而最终决定采用并行化的FP-Growth算法来解决这一问题。

然而在决赛中，作为对比，尝试实现了Spark上的Apriori算法，发现其运行速度要远远超过并行化的FP-Growth算法。经查找原因，发现是由于数据集的不稀疏性，经过FP-Growth的Map过程，原数据集并不能被划分为一个个大小均等的小数据集，往往被划分后的最后一个子数据集的大小与原数据集基本相同；同时在对生成的FP-Tree迭代挖掘的过程中很难做到挖掘频繁项集的并行化。而Apriori算法在可并行化的程度上要大大超过FP-Growth算法，而且在不同稀疏性的数据集下的运行效率基本一致。因而在决赛上提交的程序最终选定为并行化的Apriori算法实现。

## 2. 实现方案

本算法的实现主要分为以下两个步骤：

### 2.1 挖掘1-频繁项集

利用简单的类单词计数算法求出不同项的出现频数，并根据支持度进行过滤，将过滤后的1-频繁项集存储到哈希表oneItemHS中。根据oneItemHS哈希表对原始数据中的每一条事务记录进行过滤，仅保留包含在该哈希表中的项，最终形成filteredTransactions。

### 2.2 递归用K-频繁项集生成(K+1)-频繁项集

1. 将之前生成的K-频繁项集lastItemSet两两union，distinct之后选择长度为(K+1)的候选项并作剪枝生成候选项集（将候选项集中的每条候选项转换成由逗号分隔的字符串）。

注：剪枝的依据为若生成的(K+1)长度的候选项的某子集不曾出现在K-频繁项集lastItemSet中，则该候选项被剪枝。

2. 根据上一步生成的候选项集，利用类似2.1的单词计数算法求出每条候选项在filteredTransactions中的出现频数，并根据支持度进行过滤，将过滤后的(K+1)-频繁项集存储到lastItemSet中。

3. 递归执行上述两步。

## 3. 测试结果

## 3.1 jar包使用说明

spark-submit

—class edu.seu.cloud.jn3

—master <master-url>

jn3.jar

input-path

output-path

## 3.2 任务提交

```
[bookcold@namenode mathpanda]$ hadoop dfs -rmr /user/mathpanda/jn3
Picked up _JAVA_OPTIONS: -Xms30g -Xmx80g
Picked up _JAVA_OPTIONS: -Xms30g -Xmx80g
Deleted hdfs://namenode:9000/user/mathpanda/jn3
[bookcold@namenode mathpanda]$ nohup ./spark-1.0.1/bin/spark-submit --class edu
.seu.cloud.jn3.Main --master spark://192.168.1.18:7077 ./jn3.jar /user/mathpand
a/apriori_data.bat /user/mathpanda/jn3 &
[1] 15603
[bookcold@namenode mathpanda]$ nohup: 忽略输入并把输出追加到"nohup.out"
[bookcold@namenode mathpanda]$
```

## 3.3 运行情况

### Completed Applications

ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
app-20150324221812-0006	JN3	192	40.0 GB	2015/03/24 22:18:12	bookcold	FINISHED	15 min

minPartitions = 192时，Duration = 15min

## 3.4 执行结果

```
MP - bookcold@namenode: /mathpanda/jn3/result-2 - ssh - 80x24
[bookcold@namenode mathpanda]$ hadoop dfs -get /user/mathpanda/jn3 jn3
Picked up _JAVA_OPTIONS: -Xms30g -Xmx80g
Picked up _JAVA_OPTIONS: -Xms30g -Xmx80g
[bookcold@namenode mathpanda]$ cd jn3
[bookcold@namenode jn3]$ ls
result-1 result-2 result-3 result-4 result-5 result-6 result-7 result-8
[bookcold@namenode jn3]$ cd result-2
[bookcold@namenode result-2]$ ls
part-00000 part-00001 part-00002 part-00003 part-00004 part-00005
part-00006 part-00007 part-00008 part-00009 part-00010 part-00011
part-00012 part-00013 part-00014 part-00015 part-00016 part-00017
part-00018 part-00019 part-00020 part-00021 part-00022 part-00023
part-00024 part-00025 part-00026 part-00027 part-00028 part-00029
part-00030 part-00031 part-00032 part-00033 part-00034 part-00035
part-00036 part-00037 part-00038 part-00039 part-00040 part-00041
part-00042 part-00043 part-00044 part-00045 part-00046 part-00047
part-00048 part-00049 part-00050 part-00051 part-00052 part-00053
part-00054 part-00055 part-00056 part-00057 part-00058 part-00059
part-00060 part-00061 part-00062 part-00063 part-00064 part-00065
part-00066 part-00067 part-00068 part-00069 part-00070 part-00071
part-00072 part-00073 part-00074 part-00075 part-00076 part-00077
part-00078 part-00079 part-00080 part-00081 part-00082 part-00083
part-00084 part-00085 part-00086 part-00087 part-00088 part-00089
part-00090 part-00091 part-00092 part-00093 part-00094 part-00095
part-00096 part-00097 part-00098 part-00099 part-00100 part-00101
part-00102 part-00103 part-00104 part-00105 part-00106 part-00107
part-00108 part-00109 part-00110 part-00111 part-00112 part-00113
part-00114 part-00115 part-00116 part-00117 part-00118 part-00119
part-00120 part-00121 part-00122 part-00123 part-00124 part-00125
part-00126 part-00127 part-00128 part-00129 part-00130 part-00131
part-00132 part-00133 part-00134 part-00135 part-00136 part-00137
part-00138 part-00139 part-00140 part-00141 part-00142 part-00143
part-00144 part-00145 part-00146 part-00147 part-00148 part-00149
part-00150 part-00151 part-00152 part-00153 part-00154 part-00155
part-00156 part-00157 part-00158 part-00159 part-00160 part-00161
part-00162 part-00163 part-00164 part-00165 part-00166 part-00167
part-00168 part-00169 part-00170 part-00171 part-00172 part-00173
part-00174 part-00175 part-00176 part-00177 part-00178 part-00179
part-00180 part-00181 part-00182 part-00183 part-00184 part-00185
part-00186 part-00187 part-00188 part-00189 part-00190 part-00191
part-00192 part-00193 part-00194 part-00195 part-00196 part-00197
part-00198 part-00199
```

```
MP - bookcold@namenode: /mathpanda/jn3/result-2 - ssh - 80x24
[bookcold@namenode result-2]$ head part-00198
58,54:8,958077405482433
58,41:8,8658194543458433
58,50:8,958055352881333
[bookcold@namenode result-2]$
```