

频繁项集挖掘算法

说明

关联规则挖掘算法的原理和算法步骤在正文已有叙述，这里主要讲一下编译连接运行时候的注意事项

代码结构

FreqItemMain.java	主程序入口
FreqItemRun.java	算法运行的主类，包括mapper，combiner和reducer的描述
FindSubset.java	查找子集算法

输入

输入为文本文件， 文件格式：每一行代表一个集合，集合元素之间以逗号分隔。如：

```
1,2,3,4
1,2,3
5,6,7
...
```

输出

输出的结果是文本文件（默认是part-r-XXXXX）， 每一行代表一个满足要求的项集，项集元素间以逗号分隔，最后空格，之后是项集出现次数。如：

```
1,2,3 5
```

说明{1,2,3}这个项集满足结果，该项集出现了5次。

编译运行

编译

```
java *.java
```

运行

```
hadoop jar FreqItemSet.jar FreqItemSetMain <dfs_path> <input> <k> <spt_dg> <output>
```

示例

```
hadoop jar FreqItemSet.jar FreqItemSetMain /user/data/freqItemSet/ data 2 2 out
```