# Mahout谱聚类

该模块主要为在Mahout中进行Spectral Clustering准备数据。

主要实现：对用户和宾馆ID进行映射；构建<hid, uid, rating>三元组；填补三元组形成方阵；利用Mahout的Spectral Clustering进行谱聚类。

涉及到的数据文件有：

* Original matrix: originalMatrix.txt
* Parsed matrix: parsedMatrix.txt

类说明如下：

1. Preprocessing/BuildMatrix.java
   1. 用途：从原始文件中提取user与hotel的信息，构建<hid, uid, rating >三元组，即original matrix
   2. 步骤：
      1. 从HotelUserInfo中将数据读入内存
      2. 与blacklist中的黑名单用户比较，去掉黑名单用户
      3. 提取UserID, HotelID和Overall Rating，写入文件
   3. 输入：HotelUserInfo
   4. 输出：originalMatrix.txt
   5. 输出示例： 46704行

|  |
| --- |
| 80802,gourmettwo,4  80802,LNTraveller,4  80802,Bonvoyage79,3  80802,Northcutts,4  80802,TunaSlayer,5  80802,wish71,5  80802,travelbee07,2  80802,AllieL,4  80802,BillCarl,5  80802,q4quality,3 |

1. Preprocessing/ReflectMatrix.java
   1. 用途：对用户和宾馆ID映射，形成ParsedMatrix，再补全矩阵形成SquareMatrix
   2. 步骤：
      1. 将OriginalMatrix读入内存
      2. 利用HashMap生成User和Hotel的映射表userReflectionTable和hotelReflectionTable，并对OriginalMatrix进行映射替换，生成ParsedMatrix
      3. 构建邻接矩阵，填充的规则为：

newHid = uid + # of hotels

newUid = hid + # of users

* + 1. 将生成的邻接矩阵写入SquareMatrix
  1. 输入：originalMatrix.txt
  2. 输出：parsedMatrix.txt, squarematrix.txt, userReflectionTable.txt, hotelReflectionTable.txt
  3. 输出示例：
     1. parsedMatrix.txt (46704)

|  |
| --- |
| 0,0,4  0,1,4  0,2,4  0,3,5  0,4,2 |

* + 1. squareMatrix.txt (93408)

|  |
| --- |
| 0,0,4  392,396  33,40,1,4  393,39633,4  0,2,4  394,39633,4  0,3,5  395,39633,5  0,4,2  396,39633,2 |

* + 1. userReflectionTable.txt (39633)

|  |
| --- |
| 28664 janisbanis  10902 gasLosAngeles  3323 Bridgetpianos  27818 California2  36717 BigPod  37627 roomfinder  30793 California3 28638 MaryStarbucks |

* + 1. hotelReflectionTable.txt (392)

|  |
| --- |
| 50 614244  255 613399  216 81511  82 99392  308 81514  147 80788  290 124765  105 80786  251 100554 |

1. analysis/SpectralClustering.java
   1. 用途：用Java代码调用mahout进行谱聚类
   2. 步骤：
      1. 上传squareMatrix到HDFS

|  |
| --- |
| hadoop fs -put /home/claire/IdeaProjects/GroupRec/DataResource/squareMatrix /claire/mahout\_input |

* + 1. 对hdfs上的mahout\_input文件进行谱聚类，保留控制台的输出

|  |
| --- |
| mahout spectralkmeans -i hdfs://localhost:9000/claire/mahout\_input/squareMatrix -o /claire/mahout\_input/clusterResultTest1000\_20 -d 40025 -k 20 -x 100 >> /tmp/thousand20.txt 2>&1 & |

* 1. 输入： parsedMatrix.txt
  2. 输出：clusteringResult.txt
  3. 输出示例：

|  |
| --- |
| 1140:16  1141:16  1142:16  1143:16  1144:16  1145:16  1146:16  1147:6  1148:16  1149:16  1150:16 |

# 距离计算

1. Preprocessing/DurationComputing
   1. 用途：利用GoogleMap API进行两点之间到达时间计算
   2. 步骤：
      1. 创建Person对象，填入其经纬度
      2. 设置destination经纬度
      3. 创建Group对象，加入Person的list与destination
      4. GoogleMap Call得到每个Person需要的时长
   3. 输入：List<Person> + Destination
   4. 输出：durationList