需求背景：

itembased主要是两个步骤：

1 item相似度的计算

2根据user所评分过的item，以及item之间的相似度，预测未知item的分数

mahout的itembased现有的问题：

mahout集成的itembased算法，

里面的每个步骤耦合度太强，难以分割。

我们希望上面两个步骤能分开进行，

一来是步骤1和步骤2结果的更新频率不一定是相同的，

二来是我们可能考虑其他特征来计算item的相似度。

因此，我对itembased进行了一定的改造，将相似度计算和预测评分两个步骤拆分开，两者互相独立。

itembased算法的主程序是org.apache.mahout.cf.taste.hadoop.item.RecommenderJob

改造后，主程序有两个

org.apache.mahout.cf.taste.hadoop.item.RecommenderFirstJob

org.apache.mahout.cf.taste.hadoop.item.RecommenderSecondJob

前者的输出是item之间的相似度，

后者需要前者的输出作为输入，预测user对未知item的评分。

好处是显然易见的，

1预测评分的更新比item相似性计算更加频繁

2 我们融入其他特征或者算法优化item相似度计算，输入给RecommenderSecondJob，

以此优化预测评分的效果。

使用说明：

hadoop jar mahout-examples-0.9-job.jar org.apache.mahout.cf.taste.hadoop.item.RecommenderFirstJob -Dmapred.output.compress=false

-Dmapreduce.user.classpath.first=true

-i $hdfs\_input\_path

-o $hdfs\_result\_path

--tempDir $hdfs\_tmp\_path

--booleanData true

--minPrefsPerUser 1

--similarityClassname org.apache.mahout.math.hadoop.similarity.cooccurrence.measures.CooccurrenceCountSimilarity --outputPathForSimilarityMatrix $hdfs\_similarity\_path;

hadoop jar mahout-examples-0.9-job.jar org.apache.mahout.cf.taste.hadoop.item.RecommenderSecondJob -Dmapred.output.compress=false

-Dmapreduce.user.classpath.first=true

-i $hdfs\_input\_path

-o $hdfs\_result\_path

--numRecommendations 100

--tempDir $hdfs\_tmp\_path

--booleanData true

--minPrefsPerUser 1

--similarityClassname org.apache.mahout.math.hadoop.similarity.cooccurrence.measures.CooccurrenceCountSimilarity --outputPathForSimilarityMatrix $hdfs\_similarity\_path;

hadoop jar mahout-examples-0.9-job.jar org.apache.mahout.cf.taste.hadoop.item.RecommenderJob

-Dmapred.output.compress=false

-Dmapreduce.user.classpath.first=true

-i $hdfs\_input\_path

-o $hdfs\_result\_path

--numRecommendations 100

--tempDir $hdfs\_tmp\_path

--booleanData true

--minPrefsPerUser 1

--similarityClassname org.apache.mahout.math.hadoop.similarity.cooccurrence.measures.CooccurrenceCountSimilarity --outputPathForSimilarityMatrix $hdfs\_similarity\_path;

代码改动说明：

主要是RecommenderSecondJob从hdfs读取item相似度，

将item原始ID映射回内部索引，

将相似度读进一个矩阵中。

需要注意的是，相似度数据的存储格式，

输入时，数据格式，每一行是这样子的：item1,item2,similarity

输入后，每一行大概是这样子的：item1 vector[(item2,similarity),(item3,similarity)]

<https://gitlab.bigdata.leshiren.com/liulian/mahout-0.9-custom/tree/master>