产品相似度计算---蒋来

# 旅游产品特征向量整理

目标表:

TB\_REC\_CH\_TOURISM\_TAG 列簇:info1 取描述

TB\_REC\_CH\_TOURISM\_NEW\_TAG 列簇:info3,info4 取目的地

其他价格,团信息等

整合进t\_bigdata\_travel\_product\_eigen\_vector两个列族

逗留时间分段：1~2 3~5 5~7 8~

价格区间：c(0,500,1500,8000,15000,30000,Inf)

：

金棕榈的线路信息：酒店星级等 要看哦 机票（直飞还是转机）

自由行 跟团游

## pearson相似度计算

根据旅游产品的特征向量，利用pearson相似度的计算公式，计算每个产品与其他产品的相似度，并排序，取前n个（n可配置）存放在hbase里，按照相似度降序排列，存储的内容包括旅游产品的id和相似度。

**Mahout**

类名：PearsonCorrelationSimilarity

原理：用来反映两个变量线性相关程度的统计量

范围：[-1,1]，绝对值越大，说明相关性越强，负相关对于推荐的意义小。

说明：1、 不考虑重叠的数量；2、 如果只有一项重叠，无法计算相似性（计算过程被除数有n-1）；3、 如果重叠的值都相等，也无法计算相似性（标准差为0，做除数）。

    该相似度并不是最好的选择，也不是最坏的选择，只是因为其容易理解，在早期研究中经常被提起。使用Pearson线性相关系数必须假设数据是成对地从正态 分布中取得的，并且数据至少在逻辑范畴内必须是等间距的数据。Mahout中，为皮尔森相关计算提供了一个扩展，通过增加一个枚举类型 （Weighting）的参数来使得重叠数也成为计算相似度的影响因子。

## Jaccard相似度

