链滴

Spark mllib API- fpm

作者：Zing

原文链接：https://ld246.com/article/1461208831259

来源网站：链滴

spark 在该模块中提供了两种发现频繁项的算法：FPGrowth 和 PrefixSpan。

##FPGrowth

spark并行化的FPGrowth算法，用于挖掘频繁项。FPGrowth算法基于Apriori，采用FP树减少扫描数集的次数。更多和FPGrowth、Apriori算法相关可看我的另一篇博文：数据挖掘算法初窥门庭--关联

- 类：FreqItemset

该类用于表示频繁项，数据结构为 (items, freq) 元组。

- 类： pyspark.mllib.fpm.FPGrowth
  - 方法： train(data, minSupport=0.3, numPartitions=-1)
    - data：数据集
    - minSupport：支持度，默认0.3
    - numPartitions：用于FPGrowth算法计算的分区数，默认和输入数据的分区数一致。
- 类：pyspark.mllib.fpm.FPGrowthModel
  - 方法： 'freqItemsets()'

返回该模型的FreqItemset集合

##PrefixSpan

spark并行化的PrefixSpan算法，用于挖掘频繁序列模式。PrefixSpan算法是韩家炜老师在2004年提的序列模式算法。

prefixspan算法的核心是产生前缀和对应的后缀，每次递归都将合适的后缀变为前缀。

- 类：FreqSequence

该类用于表示频繁序列，数据结构为(sequence, freq) 元组

- 类：pyspark.mllib.fpm.PrefixSpan
  - 方法： train(data, minSupport=0.1, maxPatternLength=10, maxLocalProjDBSize=32000000)
    - data：输入数据集，每个样本代表一个序列
    - minSupport：最小支持度，任何出现次数大于 minSupport*size-of-the-dataset 的模式会被输出，默认为0.1
    - maxPatternLength：序列的最大长度，默认为10
    - maxLocalProjDBSize：本地处理前，数据库允许的最大样本数量，若超过此数量，会执行一个分布式prefix growth迭代。默认32000000
- 类： pyspark.mllib.fpm.PrefixSpanModel
  - 方法： freqSequences()

返回该模型的频繁序列集