序列模式

GIven a set of sequences, find the complete set of frequent subsequences

项目集(Itemset)

各种项目组成的集合

序列(Sequence)

不同项目集(ItemSet)的有序排列，序列 s可以表示为$s =< s_1 s_2 \dots s_l >$，$s_j(1 <= j <= l)$为项目集(Itemset)，也称为序列s的元素

序列的长度

一个序列所包含的项目集(ItemSet)的个数。

子序列

设$\alpha =< a_1 a_2 \dots a_n >$，$\beta =< b_1 b_2 \dots b_m >$，如果存在整数 $1 <= j_1 < j_2 < \dots < j_n <= m$，使得$a_1 \subseteq b_{j1}$，$a_2 \subseteq b_{j2}$，$\dots$，$a_n \subseteq b_{jn}$，则称序列α为序列β的子序列，又称 序列β包含序列α，记为α ⊆ β

Maximal Sequence

给定一个序列集合，如果序列s不包含于任何一个其它的序列中，则称s是最大的(Maximal Sequence)

litemset (Large itemset)

A itemset with minimum support

Example (Customer-sequence)

All the transactions of a customer, ordered by increasing transaction-time, corresponds to a sequence

---

GSP 算法

`Step`

1. Sort phase
2. Large itemset phase
3. Transformation phase
4. Sequence phase (Apriori)
5. Maximal phase

Example (Apriori Candidate Generation)

`Customer Sequences`

> 2 frequent pattern -b.ppt

FP-Growth

1. 第一次扫描数据库:
   类似于Apriori算法，找出频繁的1-itemset和他们的计数值， 将频繁项目按频度降序排列
2. 第二次扫描数据库
   - 构造fp-tree (频度从大到小)
   - 挖掘该树(频度从小到大)

> PS: 数据库大时，fp-tree可能在内存中装不下，需要 采取partition方法。