1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：P37页，为什么具有最小MIS的项目（设为x），不是s1中第一个项目，则s2中必然含有这个x? （例23？）

讨论后的理解：因为对于书中例23提到的问题，只有当序列s1的第一个项目或者序列s2的最后一个项目是各自序列中的唯一的MIS值最小的项目时才有可能发生，所以具有最小MIS的项目（设为x），不是s1中第一个项目，则s2中必然含有这个x。

1. 提出的问题2：PrefixSpan算法需要构建映射数据库，如果序列多且每个序列都需要构建映射数据库那么开销会很大，有没有什么办法可以优化？

讨论后的理解：算法主要开销在于投影数据库的构造，如果序列多且每个序列建立一个投影数据库，那么开销就比较大,可通过伪投影技术(Pseudo-projection)或者Bi-level projection减少投影数据库的数量和大小。

1. （必填）别人提出的问题的理解（选择几个问题罗列，并给出理解）：
2. 问题3：prefixspan算法的核心思想是什么？

自己的理解：先找出所有长度为1的前缀和对应的投影数据库；对长度为1的前缀进行计数，将支持度低于阈值α的前缀对应的项从数据集S删除，同时得到所有的频繁1项序列，i=1；对于每个长度为i满足支持度要求的前缀进行递归挖掘：找出前缀所对应的投影数据库，如果投影数据库为空，则递归返回；统计对应投影数据库中各项的支持度计数。如果所有项的支持度计数都低于阈值α，则递归返回；将满足支持度计数的各个单项和当前的前缀进行合并，得到若干新的前缀； 令i=i+1，前缀为合并单项后的各个前缀，分别递归执行第3步。

1. 问题4: 序列模式挖掘出频繁序列假设为 <A,B> ,这表明A发生后隔一段时间可能发生B，但隔一段时间到底是隔多长时间呢？换句话说，如何给序列挖掘添加时限约束。

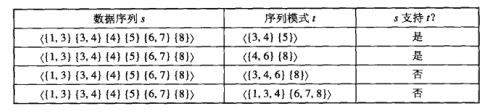
自己的理解：序列s是序列 w=<e1e2⋯ek> 的邻接子序列必须满足下面条件之一:

1. s是从w删除e1或ek中一个事件后得到
2. s是从w删除至少包含两个事件的ei中删除一个事件后得到
3. s是t的邻接子序列，而t是w的邻接子序列



用邻接子序列概念，可以用如果一个k-序列是频繁的，则它的所有邻接k-1序列也一定是频繁的，来处理最大间隔约束。

元素sj中的事件不必同时出现，定义一个窗口大小阈值（ws）来指定序列模式的任意元素中最晚和最早出现之间的最大时间差，窗口为0表示统一元素中的所有事件必须同时出现。



1. 问题5: GSP算法每次计算支持度的时候都要重新遍历数据，有什么办法可以优化

自己的理解：Spade算法是对GSP算法的优化，在扫描的时候不是扫描整个数据库，而是只扫描ID\_LIST，通过便利一次数据库得到的经验知识来降低多次对数据库的扫描。

1. （必填）读书计划

1、本周完成的内容章节：2.6-2.9

2、下周计划：5.1

四、（选做）读书摘要及理解或伪代码的具体实现（读书摘要、伪代码的具体实现代码等可以写到这个部分）

1、读书摘要及理解（选做）

