读书报告内容：（复制下面部分到一个新文档，名称命名为自己的名字，填写完交给组长，没做第四部分可以不复制第四部分）

1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：P30如果一个规则有百分之百的置信度，那么往这个规则里加入更多前件，置信度仍为百分之百，如何理解？

讨论后的理解：一个规则具有百分之百的置信度，说明只要包含了这个规则，就有确定的对应结果，那么无论前件加了什么，不会有任何改变。

1. 提出的问题2：序列模式挖掘出频繁序列假设为<A,B>，这表明A发生B，但这个间隔时间是多久呢？如何给序列添加时间约束？

讨论后的理解：有两种约束方法，最大跨度约束及最小间隔和最大间隔约1.最大跨度约束，最大跨度约束指定整个序列中所允许的事件的最晚和最早发生时间的最大时间差。一般地，最大时间跨度maxspan越长，在数据序列中检测到模式的可能性就比较大。然而较长的maxspan也可能捕获不真实的模式，因为增加这两个不相关的事件成为时间相关事件的可能性，此外，模式也可能涉及陈旧事件。2.最小间隔和最大间隔约束时限约束也可以通过限制序列中两个相继元素之间的时间差来指定，使用最大间隔约束的一个旁效就是可能违反先验原理。可能会出现序列中事件数增加时，支持度也增加，这就违背了先验原理。

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）：

3. 问题3：为什么只有当s1的第一个项目或者s2的最后一个项目是各自序列中唯一具有最小MIS值的项目时才会发生过度剪枝问题？

自己的理解：因为之前我们合并的规则是，两串候选序列分别去掉首项目和尾项目的串相等。如果s1的第一个项目或者s2的最后一个项目是各自序列中唯一的MIS值最小的项目时，可能会被剪枝，导致下一步无法生成。

1. 问题4：为什么需要两串候选序列分别去掉首项目和尾项目的串相等？

自己的理解：如果s1去掉第一项的余串和s2去掉最后一项的余串相等，就把s2的最后一个项目加入s1最后。这样产生的候选序列集合，是有可能成为频繁序列的，即子集有可能是Fk中的频繁项目。

1. （必填）读书计划

1、本周完成的内容章节：如2.5-2.7

2、下周计划：2.8-2.9+3.5

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）

2.6序列模式的基本概念

给定一个由不同序列组成的集合，其中，每个序列由不同的元素按顺序有序排列，每个元素(交易)由不同项目组成，同时给定一个用户指定的最小支持度阈值，序列模式挖掘就是找出所有的频繁子序列，即该子序列在序列集中的出现频率不低于用户指定的最小支持度阈值。

2.7基于GSP挖掘序列模式

GPS算法是类Apriori算法。用于从候选项集中发现具有时序先后性的频繁项集。两个步骤：进行自连接、进行剪枝。缺点：每次计算支持度，都需要扫描全部数据集；对序列模式很长的情况，由于其对应的短的序列模式规模太大，算法很难处理。可以使用SPADE算法把扫描的结果存入列表。