1. 数据预处理
   1. 从原始Web日志记录中提取有用的日志数据，将数据转换为模式挖掘中所需的类型
2. 模式挖掘
   1. 根据挖掘目的，如分类、聚类、关联规则发现或者序列模式分析等决定选择何种挖掘算法
3. 模式分析
   1. 将模式挖掘发现出来的模式进行分析，剔除冗余和无关模式
4. 将原始的Web日志数据导入数据库中，建立原始Web访问日志数据表（id, ip, identd, user, time, url, status, size, refer, agent组成 ）
5. 数据预处理
   1. 数据清理：剔除无用数据
   2. 用户识别：针对不同IP建立个人库，通过基于IP启发式规则来识别
   3. 会话识别：时间阈值、最大向前引用
   4. 路径填充：使用站点拓扑结构、用户访问页面的参引页面
   5. 事务识别：引用长度法、最大向前引用法、时间窗口
6. 模式挖掘模块（采用 Apriori 算法）：基于逐级搜索的思想。它采用多轮搜索的方法，每一轮搜索扫描一遍整个数据集，并最终生成所有的频繁项目集。Apriori算法常常会产生数量巨大的项集，随着序列模式的长度增加，候选序列会成指数级增长，虽然通过Apriori的向下封闭属性缩小了候选项集的大小，但是算法的时间复杂度仍然达不到比较理想的程度。另外，该算法需要多次的扫描日志数据库，只要候选序列的长度增加，就必须扫描一遍数据库，这样对整个算法的执行效率有很大的影响。

改进：缩小候选集大小；缩小事务数据库大小

有共同userID属性的候选集才可能是频繁项目集.

* 1. 路径分析：判定最频繁访问路径
  2. 关联规则：寻找同一事件中出现不同项的相关性
  3. 序列模式：挖掘数据库中高频率出现有序列
  4. 分类分析：识别一个特殊群体的公共属性的描述，决策树方法或神经元网络
  5. 聚类分析：聚类分析是把具有相似特征的用户或数据项归类,在网站管理中通过聚类具有相似浏览行为的用户。基于模糊理论的Web页面聚类算法与客户群体聚类算法的模糊聚类定义相同。 K—means、DBSCAN
  6. 统计：从Web 站点中抽取知识的最常用方法, 它通过分析会话文件, 对浏览时间、浏览路径等进行频度、平均值等统计分析
  7. 协同过滤：采用最近邻技术，利用客户的历史、喜好信息计算用户之间的距离，目标客户对特点商品的喜好程度由最近邻居对商品的评价的加权平均值来计算

1. 分析目标：
   1. 对被频繁访问的资源预警
   2. 网站上不存在资源的请求（恶意传参等）
   3. 观察搜索引擎蜘蛛的来访情况
   4. 观察访客行为
   5. 作用： 对访问时间进行统计，可以得到服务器在某些时间段的访问情况； 对IP进行统计，可以得到用户的分布情况； 对请求URL的统计，可以得到网站页面关注情况； 对错误请求的统计，可以更正有问题的页面。
2. 的