# 不确定数据中的效用模式挖掘方法研究

学号：SY1306225
姓名：王　洋
导师：兰雨晴

# 目录

- 论文工作计划
- 已经完成的工作
- 关键技术或难点
- 下一阶段工作计划

# 目录

- 论文工作计划
- 已经完成的工作
- 关键技术或难点
- 下一阶段工作计划

# 研究目标

- 已有的数据挖掘研究（确定数据中频繁模式挖掘、高效用模式挖掘，不确定数据中频繁模式挖掘）为销售商、互联网企业、医疗机构、政府等各行各业的信息和知识的挖掘，进而进行相应的决策提供了支持，但是已有的研究成果在面临更新、更复杂的应用场景时，不能很好的满足需求。为了解决这一难题，本文首先定义不确定数据中高效用模式挖掘的概念，进而分别用基于期望和基于概率的模型设计挖掘高效用模式的方法，并探索可用的剪枝策略和优化方法以提高挖掘的执行效率，最大化满足用户的需求。

# 研究内容

- 对数据挖掘领域已有的问题和成果的研究
  - 确定数据中的频繁模式挖掘（数据挖掘经典问题）
  - 确定数据中的效用模式挖掘
  - 不确定数据中的频繁模式挖掘
- 不确定数据中效用模式挖掘问题定义及方法的设计与实现
  - 潜在的高效用项集挖掘
  - 基于期望语义模型的高效用项集挖掘
  - 基于概率语义模型的高效用项集挖掘

# 论文工作计划

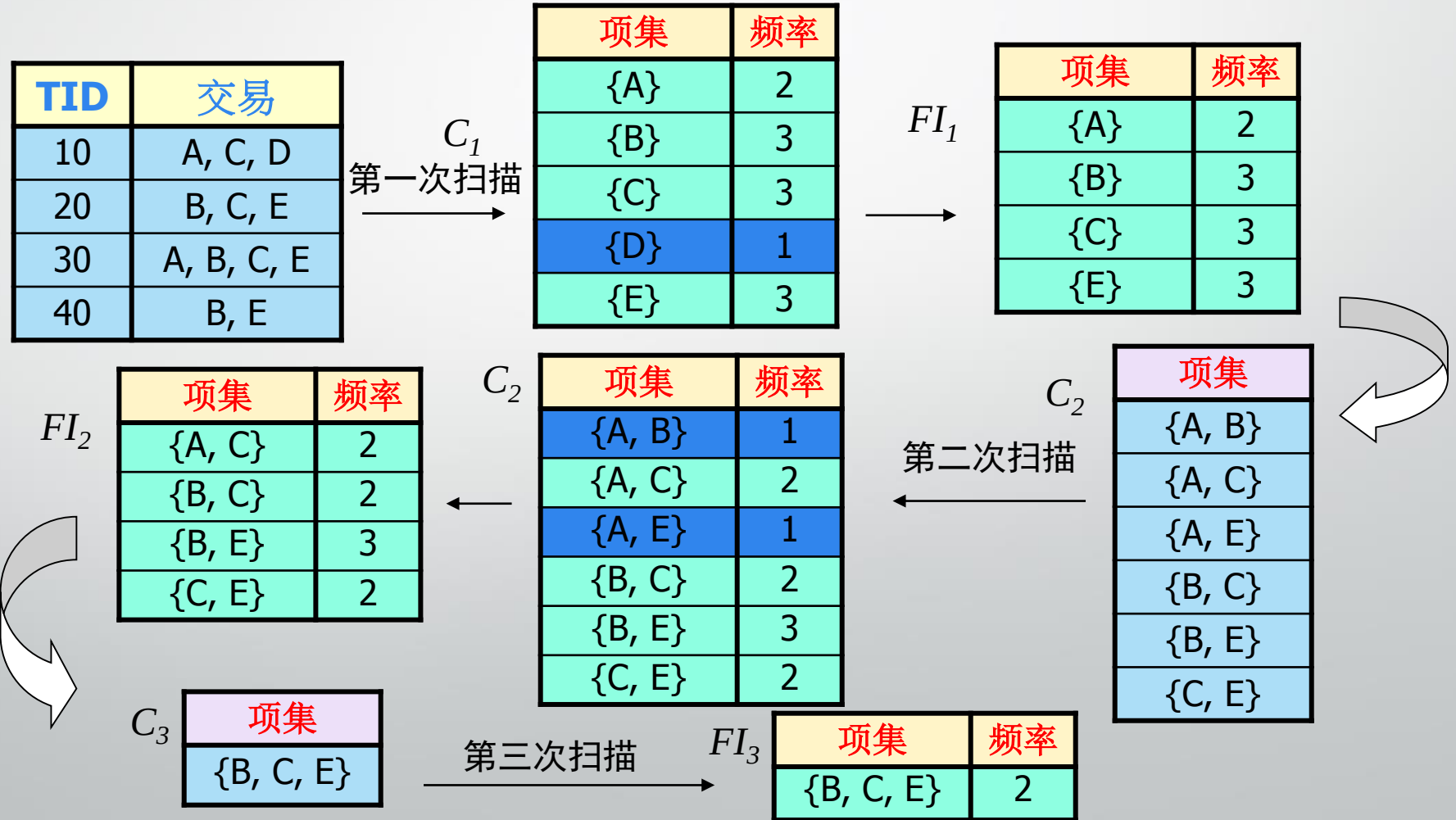
| 序号 | 时间段 | 计划完成的工作 |
| --- | --- | --- |
| 1 | 2014.12~2015.01 | 确定数据中频繁模式挖掘、高效用模式挖掘方法研究，不确定数据中频繁模式挖掘方法研究 |
| 2 | 2015.02~2015.03 | 不确定数据中潜在高效用项集挖掘问题的定义及方法的设计与实现 |
| 3 | 2015.04~2015.05 | 不确定数据中基于期望的高效用项集挖掘问题的定义及方法的设计与实现 |
| 4 | 2015.06~2014.08 | 不确定数据中基于概率的高效用项集挖掘问题的定义及方法的设计与实现 |
| 5 | 2015.09~2015.10 | 收集真实应用场景下的数据，完成不确定数据中高效用项集挖掘方法的测试与bug修复 |
| 6 | 2015.11~2015.11 | 毕业论文撰写 |

# 目录

- 论文工作计划
- 已经完成的工作
- 关键技术或难点
- 下一阶段工作计划

# 国内外相关技术、原理、方法的研究

- 频繁模式挖掘经典方法（频繁模式挖掘的基础是频繁项集挖掘）
  - Apriori算法
  - FP-Tree
  - H-Mine



# 国内外相关技术、原理、方法的研究

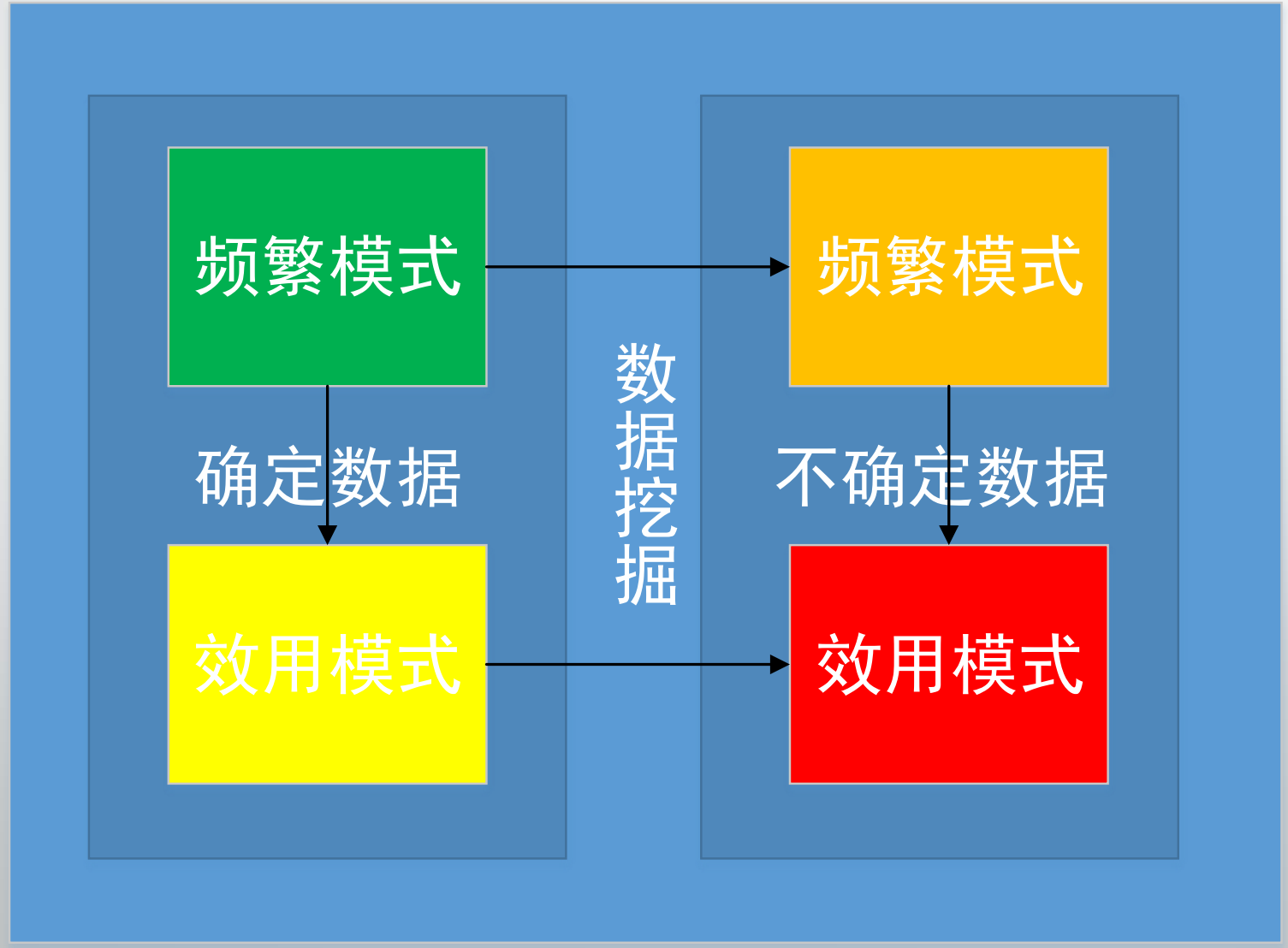
## 效用模式挖掘的研究

- Umining算法
- IHUP算法
- UP-Growth算法

## 不确定数据中的频繁模式挖掘的研究

- 基于期望语义的频繁模式挖掘方法
- U-Apriori、UFP-Growth、UH-Mine
- 基于概率语义的频繁模式挖掘方法
- I-PFIM、TODIS

# 相关研究与本次研究的关系



# 潜在高效用项集挖掘的定义与方法的实现

- 在不确定数据中的高效用项集挖掘中，每条记录（transaction）中的一项（item），都包含三个信息：数量q（quantity），单位效益p（profit），出现概率pr（probability）。其中，每条记录中每一项的效用等于它是数量和单位效益的乘积，即效用U（utility）=q*p。

  效用阈值 min_util　　概率阈值 min_prob

  若U（X）≥min_util 且 pr（X）≥min_prob
  则X为潜在的高效用项集

- 挖掘方法——确定数据下效用模式挖掘+不确定数据下频繁模式挖掘

  如“UP-Growth”+“TODIS”执行结果取交集

# 基于期望语义模型的高效用项集挖掘的定义与方法实现

- 不确定数据库中每条记录（transaction）中的一项（item），都包含三个信息：数量q（quantity），单位效益p（profit）,出现概率pr（probability）。

  期望效用EU=q*p*pr。我们规定：期望高效用阈值min_eutil

  若EU（X）≥min_eutil，则X是这个数据库中的高效用项集

- 方法

  Apriori算法进行扩充可以解决这一问题，但由于在效用模式挖掘中downward closure property不成立，无法对挖掘过程剪枝和优化

  为此我们提出了一种新的结构（U-list）以及算法很好的解决该问题

# 与研究相关的小论文撰写及发表情况

- Mining High Utility Itemsets over Uncertain Database
  IEEE CyberC 2015（EI会议）（已接收）
- Mining Probabilistic High Utility Itemsets over Uncertain Database
  WISE 2015（CCF C类会议）（评审中）
- Mining Potential High Utility Itemsets over Uncertain Database
  （正在撰写，计划投会议）
- Mining High Utility Itemsets from Uncertain Data
  （正在撰写，计划投期刊）

# 目录

- 论文工作计划
- 已经完成的工作
- 关键技术或难点
- 下一阶段工作计划

# 关键技术及难点

- 效用模式挖掘中向下封闭性(downward closure property)不成立

在频繁模式挖掘中，存在这样一条性质：

若sup（AB）≥min_sup，则sup（A）与sup（B） 均≥min_sup
同理，若sup（A）< min_sup，则任何sup（AXX） < min_sup。
即向下封闭性(downward closure property)。

但在效用模式挖掘中，这一性质不再适用。比如，A共在数据库中出现20次，A和B共同出现10次，若A、B单位效用分别为2和1，则A效用（40）大于AB效用（30），如A、B单位效用分别为1和2，则A效用（20）小于AB效用（30）。

这一情况给，提高挖掘的效率增加了难度。

# 关键技术及难点

- 不确定数据中的效用模式挖掘向下不封闭将带来更大困难
- 就基于期望语义模型而言

EU=q*p*pr 比确定数据中 U=q*p复杂度高一个数量级

- 就基于概率语义模型而言

用二项分布筛选效用不小于某值的概率，由于变量由常数变为X方，很难找到有效的优化和剪枝策略

# 目录

- 论文工作计划
- 已经完成的工作
- 关键技术或难点
- 下一阶段工作计划

# 下一阶段工作计划

- 不确定数据中基于概率的高效用项集挖掘问题的定义及方法的设计与实现
- 收集真实应用场景下的数据，完成不确定数据中高效用项集挖掘方法的测试与bug修复
- 毕业论文撰写

# 核心参考文献


- Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." VLDB 1994.
- Han, Jiawei, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation." SIGMOD 2000.
- Chui, Chun-Kit, Ben Kao, and Edward Hung. "Mining frequent itemsets from uncertain data." PAKDD 2007.
- Bernecker, Thomas, et al. "Probabilistic frequent itemset mining in uncertain databases." SIGKDD 2009.
- Tseng, Vincent S., et al. "UP-Growth: an efficient algorithm for high utility itemset mining." SIGKDD 2010.
- Tong, Yongxin, et al. "Mining frequent itemsets over uncertain databases." VLDB 2012.

谢谢各位老师！