



# 基于完全前缀展开的无序日志修复研究

1/23

硕士开题答辩

肖永博

指导老师：闻立杰 副教授



# 主要内容

1

选题背景与意义

2

国内外研究现状

3

研究内容与方案

4

预期成果及可能的创新点

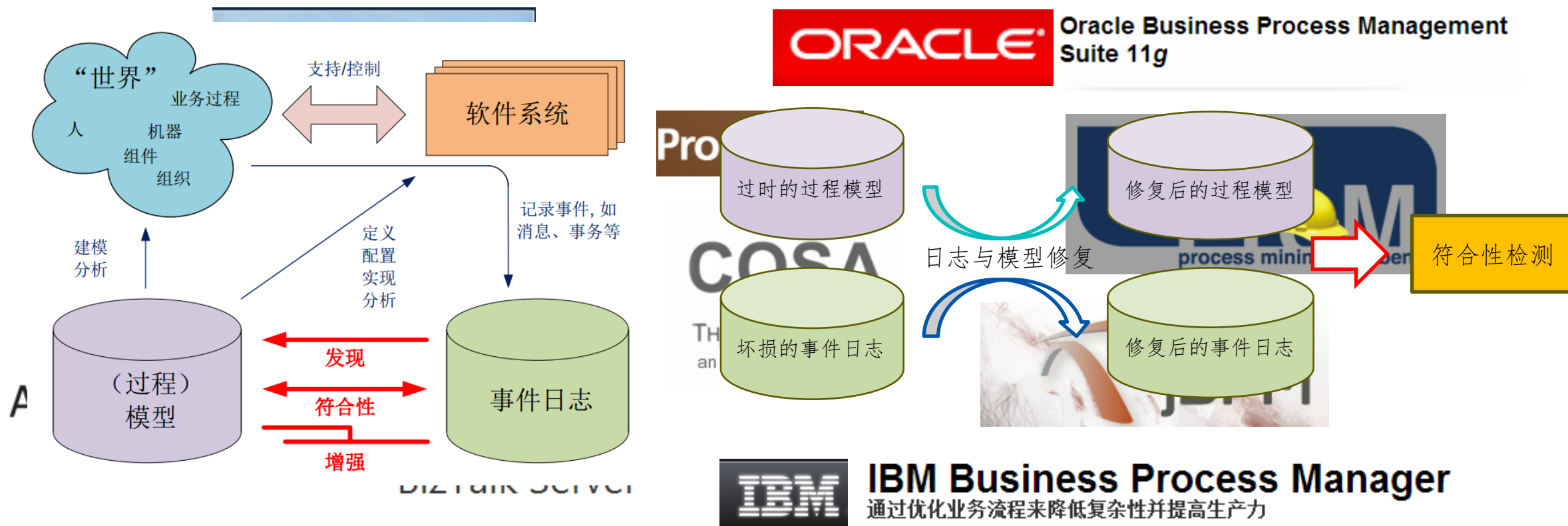
5

研究计划



# 选题背景与意义

## 国内外业务过程管理产品



大量的过程感知信息系统已经被广泛的应用于各行业，为企业提供业务流程建模和业务执行监测等服务。



# 选题背景与意义

4/23



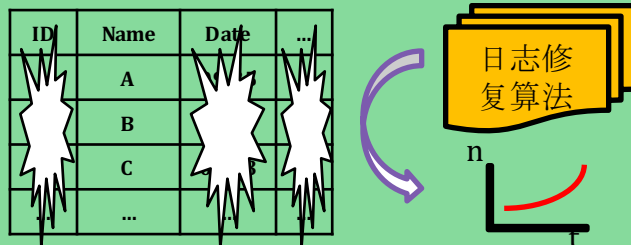
分布式计算等应用场景下  
很容易产生执行顺序缺失  
的日志

ID	Name	Date	...
1	A	09.25	...
2	B	09.27	...
3	C	09.23	...
...	...	...	...

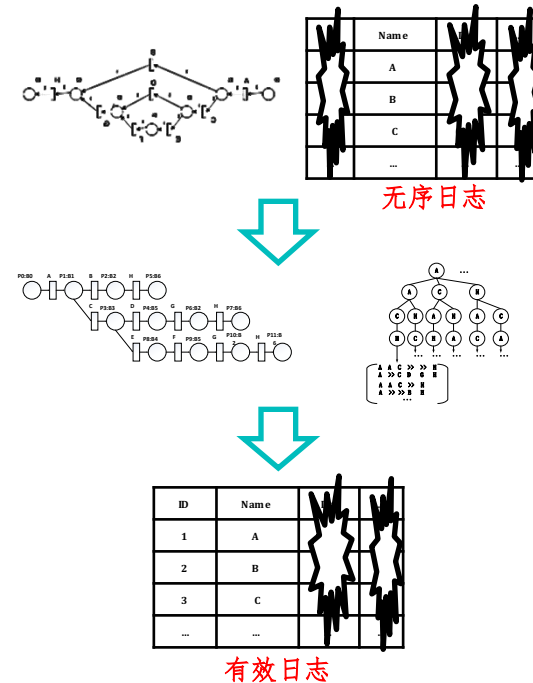
ID	Name	Date	...
1	A		
2	B		
3	C		
...	...		



缺少针对无序事件日志的  
高效修复算法

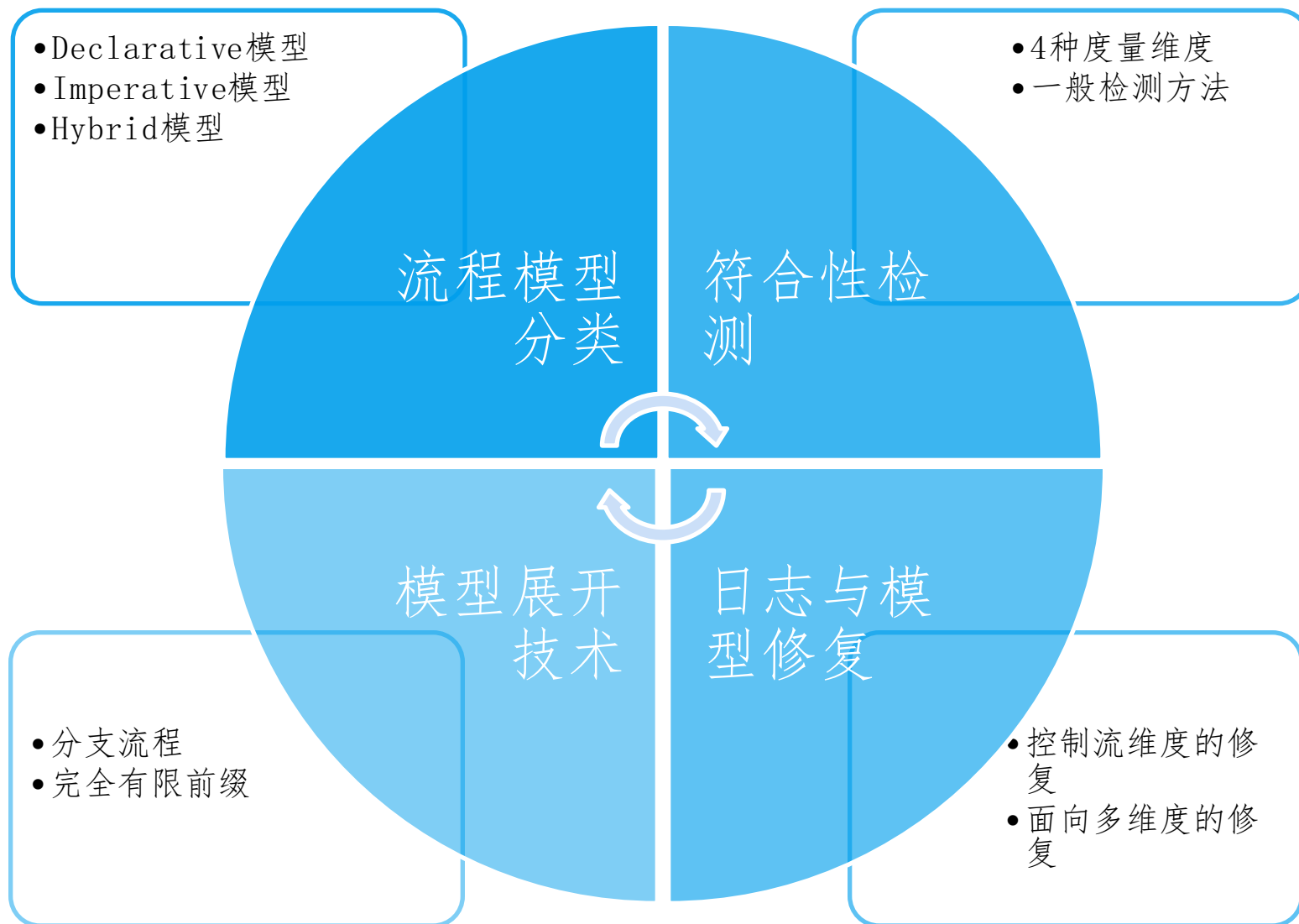


设计一种新的  
日志修复算法





# 国内外研究现状



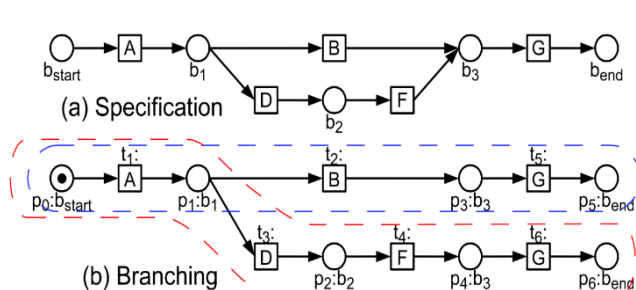


# 国内外研究现状

## ➤ 流程模型展开技术

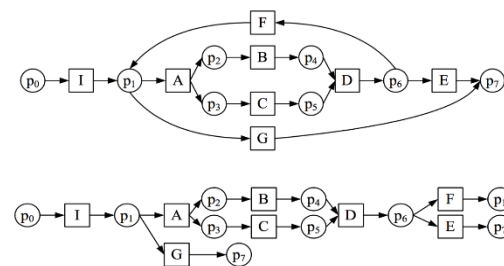
### 分支流程

- ❑ 一种基于偏序的Petri网结构的展开形式
- ❑ 简化对模型的分析难度



### 完全有限前缀

- ❑ Petri网展开形成的出现网的前缀部分
- ❑ 包含所有状态
- ❑ 在状态重复的位置产生截断



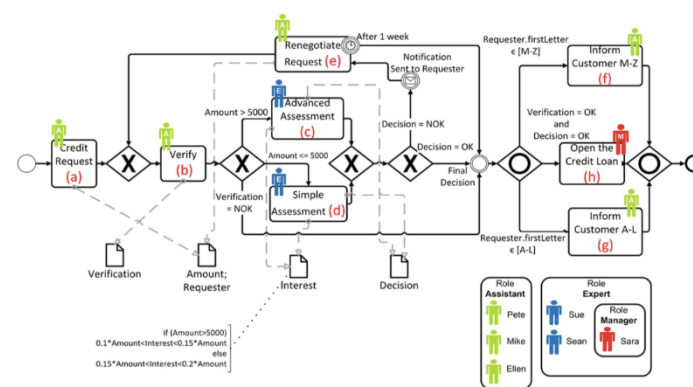
## ➤ 日志与模型修复技术

### 控制流维度的日志模型修复技术

- ❑ 模型修复
- ❑ 日志修复
  - 基于统计学的方法
  - A\*方法
  - 针对事件缺失修复的加速方法
- ❑ 标签修复

### 面向多维度的日志模型修复技术

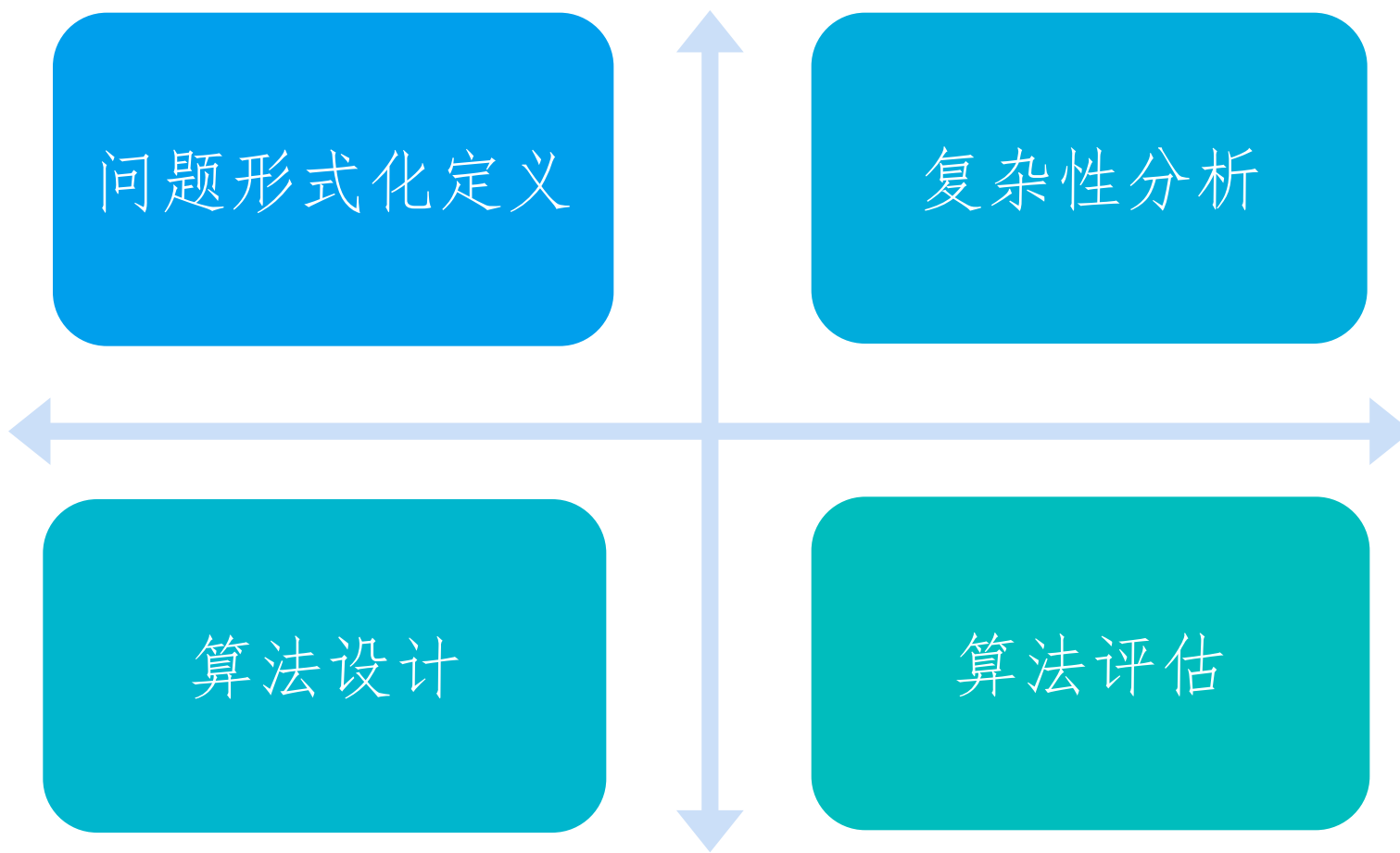
- ❑ 控制流
- ❑ 数据
- ❑ 资源





# 研究内容与方案

## ► 研究内容





# 研究内容与方案

## ► 问题的形式化定义

### ► 问题描述

给定一个流程模型和一个活动多集，判定该活动多集是否能构成符合该流程模型的一条轨迹。若可以，其轨迹是什么；若不可以，多余或者欠缺的事件是哪些，其中最小代价的修复方案是什么？

### ► 问题输入

- 一个流程模型（Imperative模型、Declarative模型或者Hybrid模型）
- 一个活动多集

### ► 问题输出

- 一条有效的、修复代价最小的轨迹

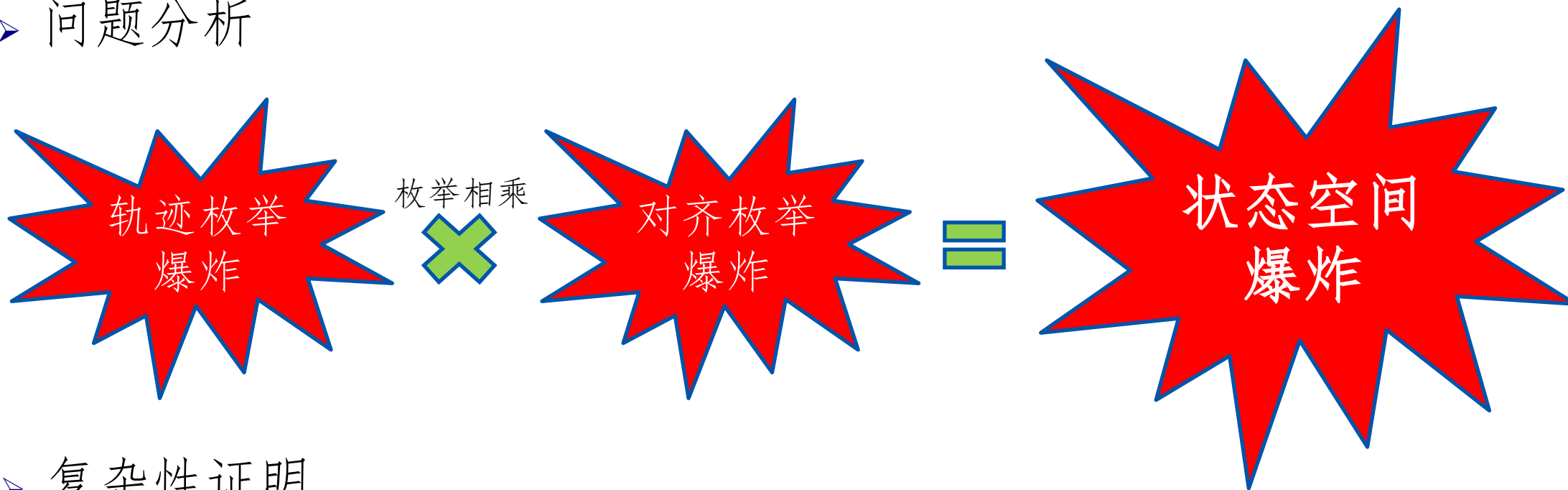




# 研究内容与方案

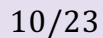
## 问题的复杂性分析

### 问题分析



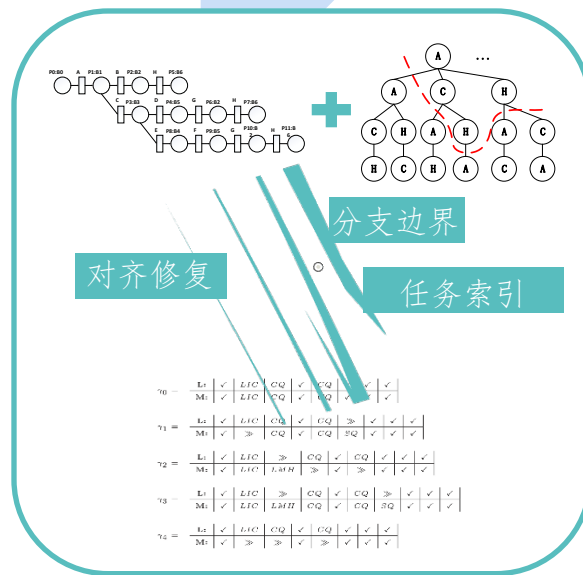
### 复杂性证明

- 验证解决该问题是否为NP-Hard问题、是否有多项式解、是否有启发式解
- 尝试使用归约方法将问题转换为经典NP-Hard问题



- ▶ 算法设计
  - ▶ 基本思路

# 对齐修复





# 研究内容与方案

## ► 算法设计

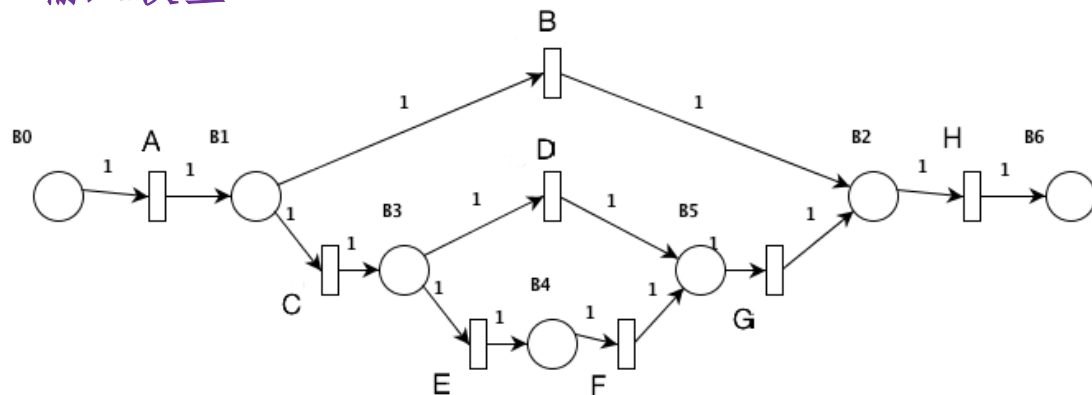
### ► A\*算法

#### ► 状态空间（两阶段枚举相乘）

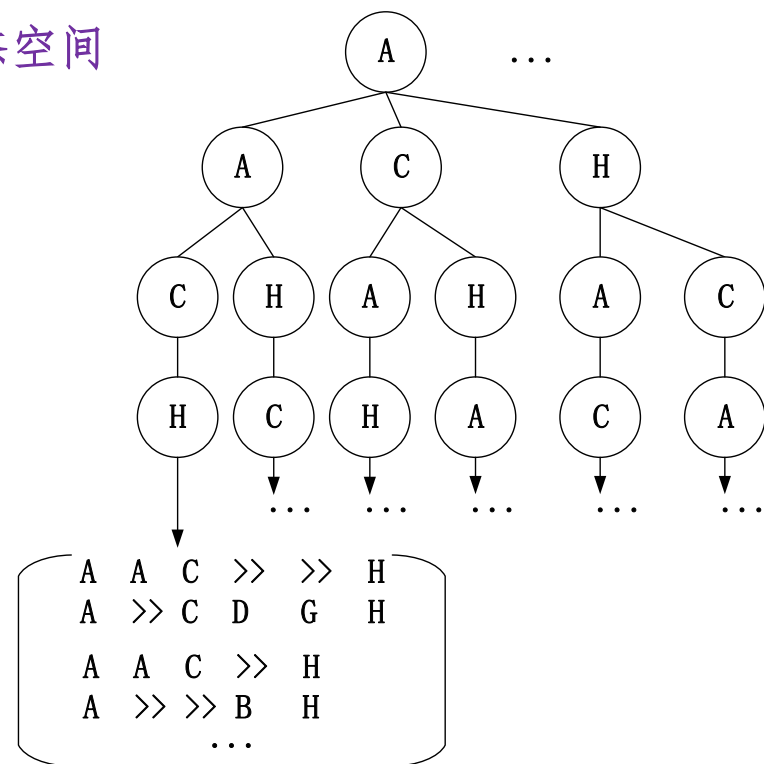
输入活动多集

{A, A, H}

输入模型



状态空间



轨迹集合

对齐集合



# 研究内容与方案

## ➤ 算法设计

### ➤ A\*算法

#### ➤ 实际代价函数 $g(n)$

- 记录当前状态下的实际代价
- 总结、分析各种操作的代价，为其定义相应的代价函数
  - ▣ 轨迹枚举阶段：从活动多集中选择一个活动
  - ▣ 对齐修复阶段：出现模型中存在但在活动多集中不存在的活动
  - ▣ 对齐修复阶段：出现在活动多集中存在但在模型中不存在的活动

#### ➤ 估计代价函数 $h(n)$

- 估计当前状态到目标状态的代价
- 如何设计有效的估计代价函数，使算法求解过程向目标前进

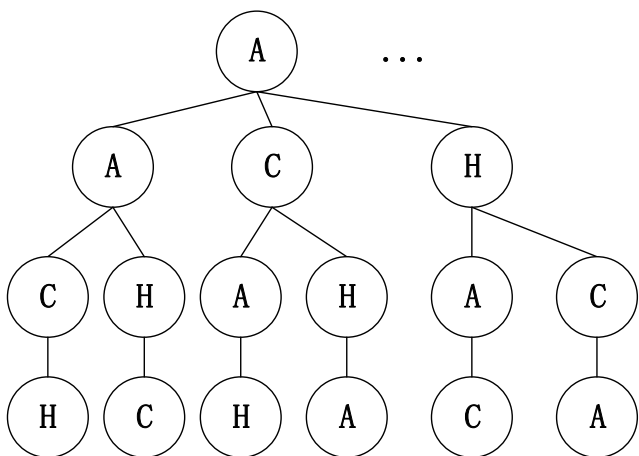


# 研究内容与方案

## ▶ 算法设计

- ▶ 加速算法（基于A\*算法的状态空间+辅助索引）
  - ▶ 针对轨迹枚举使用可达索引

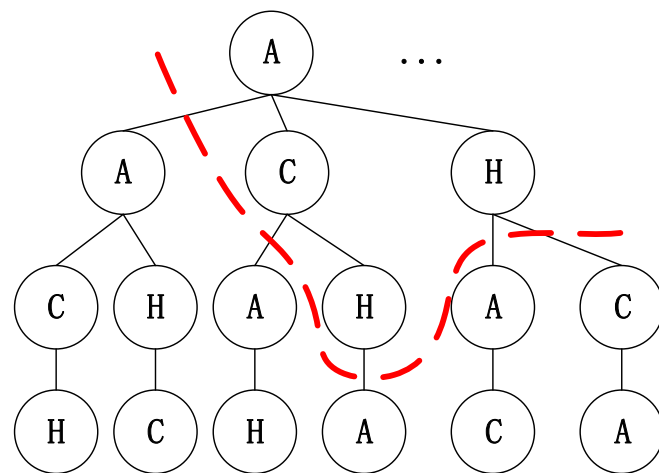
空间状态

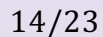


可达索引

	...	A	C	H	...
...	...	...	...	...	...
A	...	↗	→	→	...
C	...	↗	↗	→	...
H	...	↗	↗	↗	...
...	...	...	...	...	...

空间剪枝

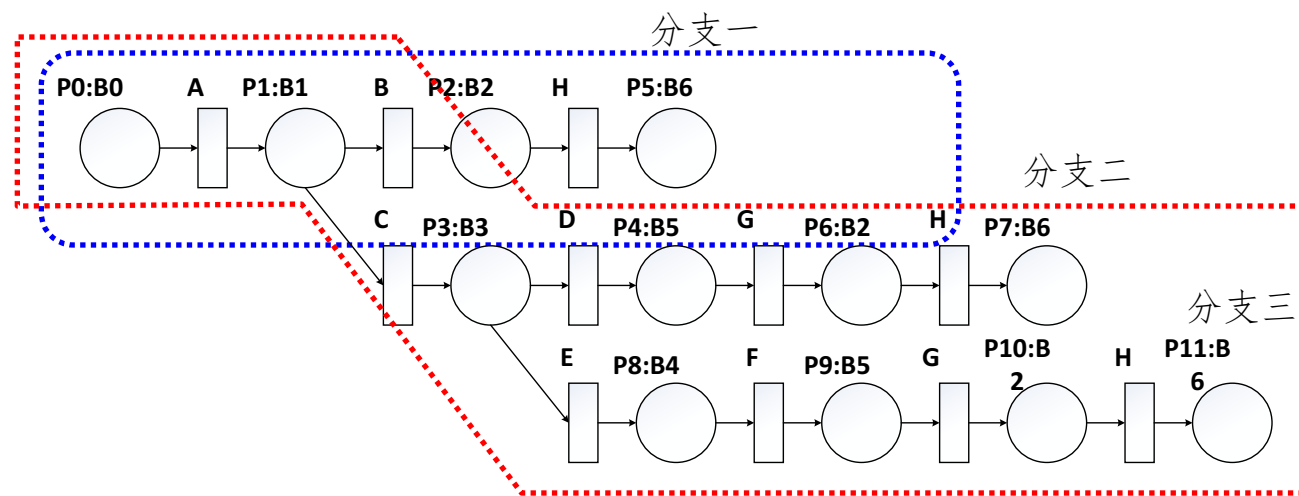




## ► 算法设计

## ➤ 加速算法

## ► 针对对齐修复枚举使用任务索引



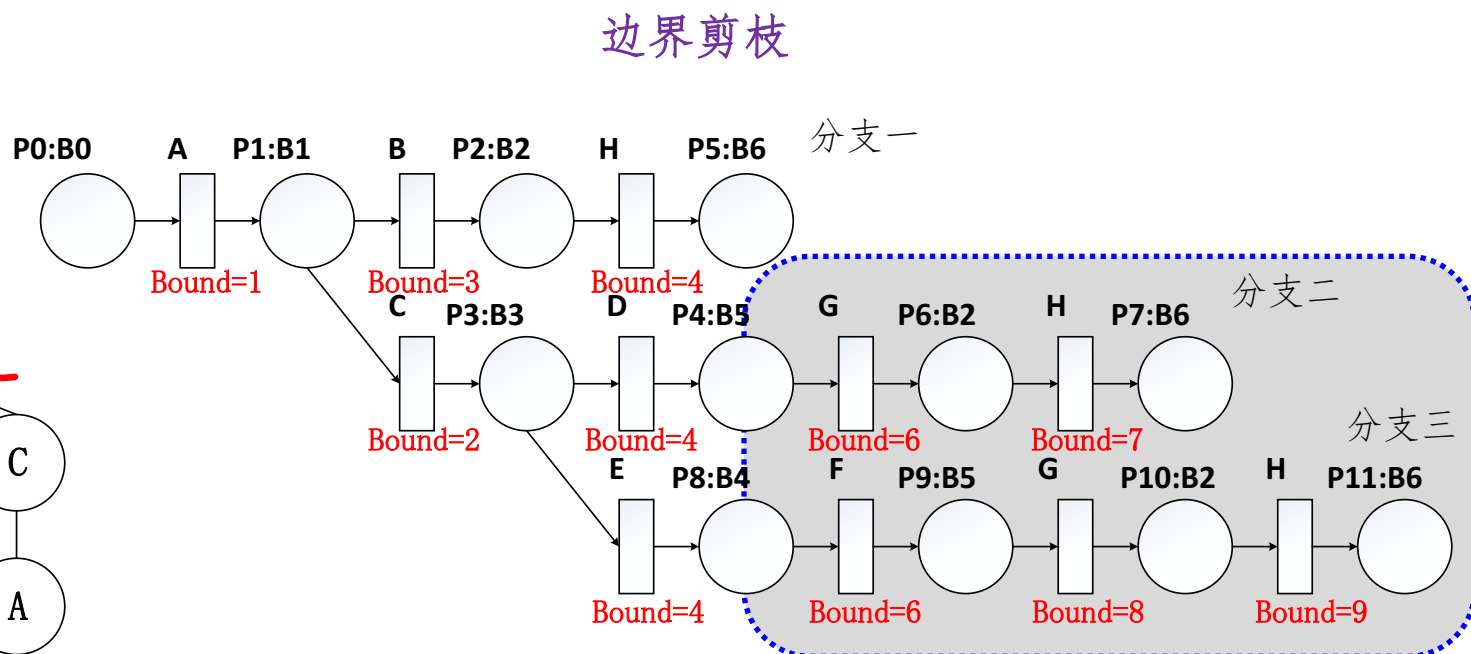
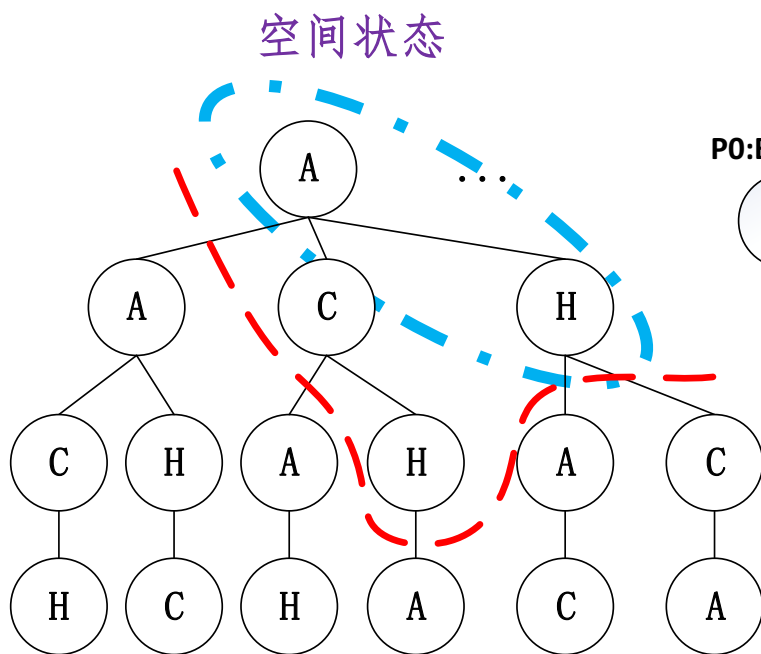


# 研究内容与方案

## 算法设计

### 加速算法

- 针对对齐修复枚举使用分支边界进行剪枝





# 研究内容与方案

## ➤ 算法设计

### ➤ 贪心算法

- 基于A\*算法的状态空间
- 对问题进行归纳分析，寻找问题的最优子结构
- 通过优化估计代价函数 $h(n)$ ，使算法的求解过程更快地向目标靠近的同时具备无后效性，加快求解速度。

### ➤ 局限性

- 产生的解可能为局部最优解





# 研究内容与方案

## ➤ 算法评估

### ➤ 实验方案设计

#### ➤ 对于实验输入数据规模

#### 模型结构

- 多循环
- 多并发
- 多互斥

#### 模型大小

- 包含一般数量事件
- 包含较多数量事件

#### 活动多集大小

- 包含一般数量任务
- 包含较多数量任务

#### ➤ 对于算法的参数调整

- 分析算法中的参数，通过控制变量方法对参数进行调整以观察其对算法结果的影响

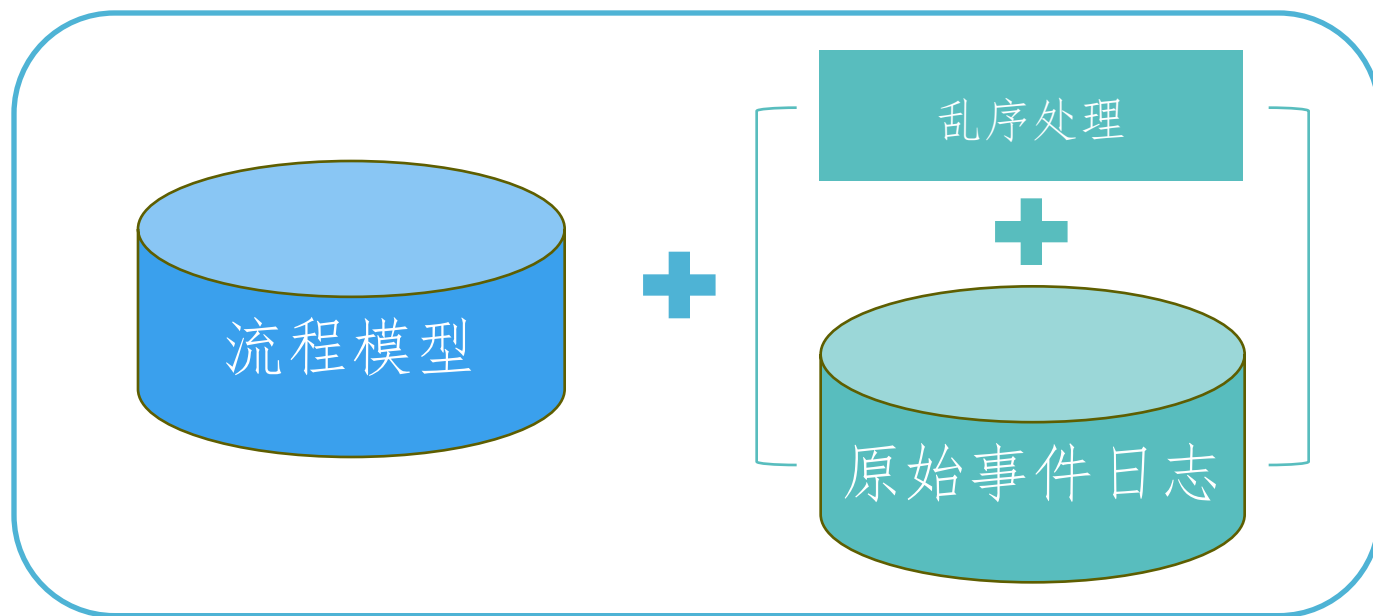


# 研究内容与方案

- ▶ 算法评估
  - ▶ 实验数据来源



中国移动通信集团公司



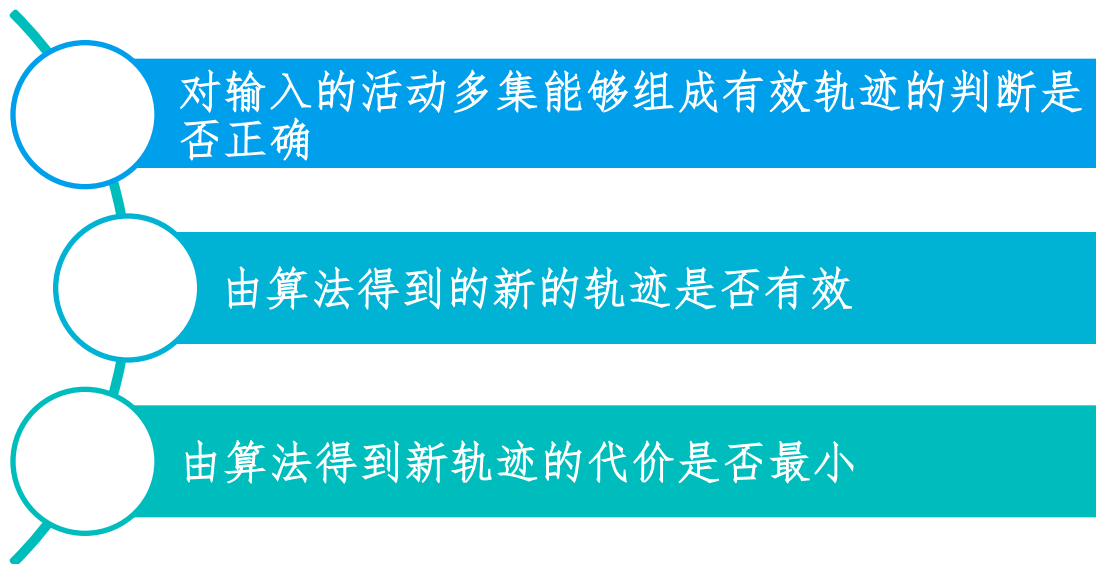


# 研究内容与方案

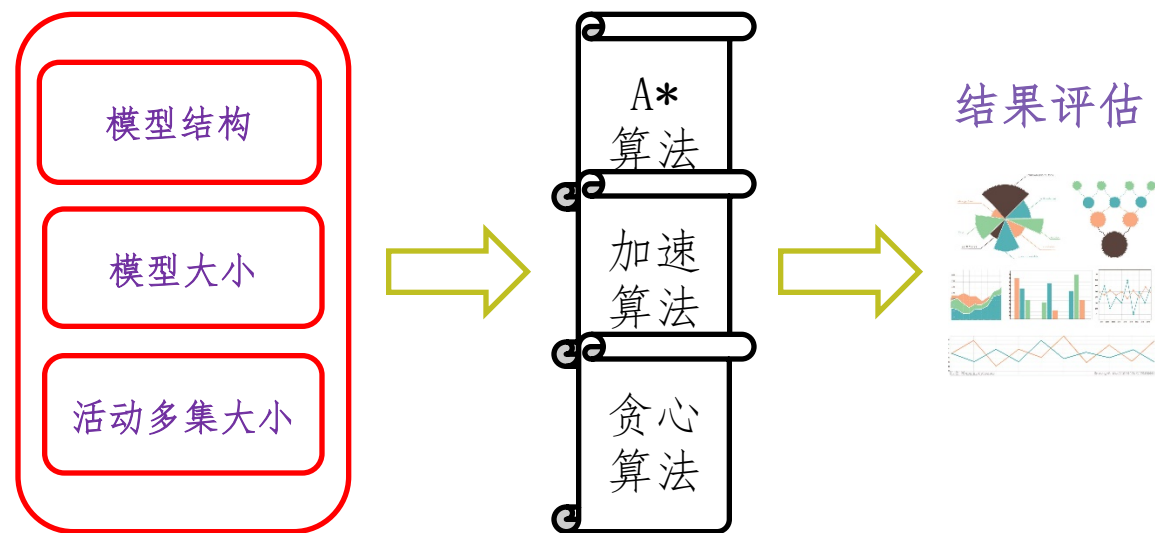
## ➤ 算法评估

### ➤ 实验结果评估

#### ➤ 准确性



#### ➤ 算法性能



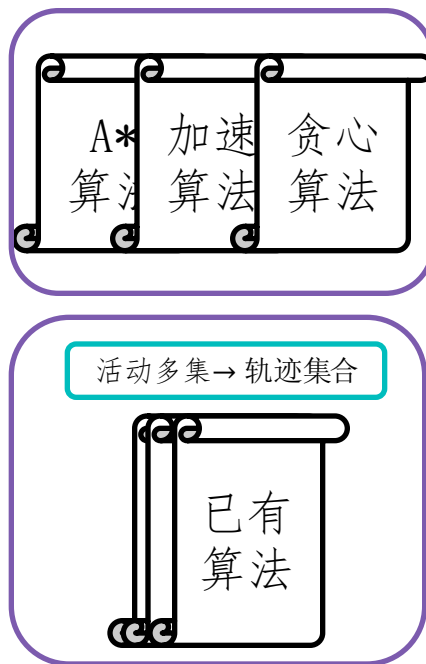
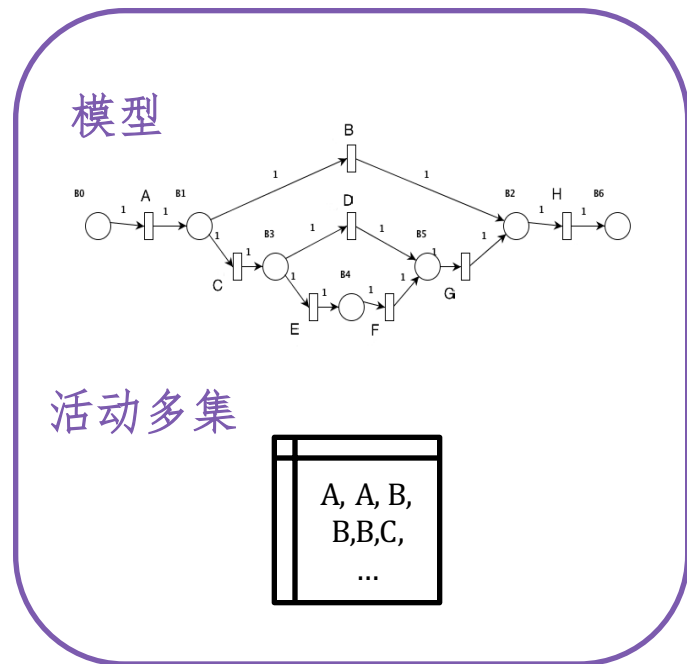


# 研究内容与方案

## ➤ 算法评估

### ➤ 对比实验设计

- 将活动多集转换为轨迹集合，以此作为输入来对已有日志修复算法进行实验





# 预期成果及可能的创新点

21/23

## ▶ 预期科研成果

- ▶ 发表学术论文1-2篇
- ▶ 基于完全前缀展开的无序事件日志修复A\*算法
- ▶ 基于完全前缀展开的无序事件日志修复加速算法
- ▶ 基于完全前缀展开的无序事件日志修复贪心算法

## ▶ 可能的创新点

- ▶ 证明对于无序事件日志的修复问题是NP-hard问题
- ▶ 无序事件日志修复问题的启发式代价估计



# 研究计划

时间	工作项目
2015.07 – 2015.09	文献调研阶段调研了符合性检测、流程分支技术、完全有限前缀提取技术、日志模型对齐方法的相关知识，深入理解现有的日志修复研究成果，并重点关注完全有限前缀提取方法和日志模型对齐算法。
2015.10 – 2016.02	基于大量阅读文献和已有成果，为解决完全松散日志修复问题的复杂性进行相关验证。
2016.03 – 2016.06	基于完全有限前缀和日志模型对齐方面已有的技术，对完全松散日志的修复操作设计代价计量算法，为每一步修复的代价估计设计启发式预估算法，最终完成A*算法、加速算法和贪心算法的设计。
2016.07 – 2016.09	使用大量预设模型和实际模型事件日志测试并评估算法准确度和性能。
2016.10 – 2016.12	总结日志修复算法在设计、理论验证和实际新能分析中的创新点与贡献点，完成一到两篇科技论文。
2017.01 – 2017.03	总结研究过程出现的问题和技术难点，撰写毕业论文。



?

欢迎各位老师提问