slide3
如今，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。
通过符合性检测，系统可以检查实际业务操作同业务模型之间发生的偏离程度，而日志与模型修复技术则可以对坏损日志和过时模型进行修复。
slide4
系统记录的日志时常会遇到各种问题，其中分布式计算等应用场景下的日志很容易产生更不可靠的日志（如乱序轨迹），然而目前学术界缺少针对这类无序事件日志的有效修复算法。
因此，本课题旨在针对无序事件日志提出一种日志修复算法。
slide5
文献调研部分主要针对流程模型、符合性检测、模型展开技术和日志与模型修复技术四个方面进行了调研。

通过调研发现，流程模型类型目前主要有三种：一种非结构化的、只规定约束规则的 **Declarative** 模型、严格描述流程如何执行的 **Imperative** 模型和混合使用上述两种特点来进行模型表达的 **Hybrid** 模型

符合性检查
学术界主要从 **Fitness**、**Pesicion**、**Generalization**、**Structureness** 四个角度来做符合性检测
一般检测方法：日志重演和日志模型对齐

展开技术主要调研了分支流程和完全有限前缀这个内容
分支流程是一种基于偏序的 **Petri** 网结构的展开形式，对模型进行流程分支，可以大大简化对模型的分析难度
完全有限前缀是 **Petri** 网展开形成的出现网的前缀部分，它包含了出现网的所有状态并在状态重复的位置产生截断

日志与模型修复技术方面调研了控制流维度的修复技术和面向多维度的修复技术。
控制流维度的修复技术主要分为模型修复、日志修复和标签修复
多维度的修复技术主要综合考虑了控制流、数据和资源三个方面的约束

通过调研，也可以发现现有的日志修复算法难以解决对无序事件日志的修复问题
slide6
下面，我将从以下四个方面来对上述问题进行分析归纳，并介绍今后的研究内容和主要方法。
slide7
首先，再次对本文提出的问题进行一个描述：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？
slide8
通过对问题的输入进行分析可以发现：

第一阶段：活动多集可以组成轨迹的空间具有爆炸性
第二阶段：每条轨迹修复方案的空间具有爆炸性
而上述两个阶段的枚举相乘将会使搜索空间具有更大的爆炸性

因此，本文试图对问题的复杂性进行分析，通过使用归约的方法将问题转换为经典 NP-Hard 问题，验证解决该问题是否是 NP-Hard 问题、是否有多项式解以及是否有启发式解

slide9

在算法设计方面，算法的基本思路是分析模型、轨迹枚举和对其修复三个部分。
首先通过使用展开技术（流程分支、完全前缀展开等）来对输入模型进行分析
然后通过枚举来获得轨迹的状态空间
最后再对每种轨迹进行对齐修复并找出最优解

slide10

首先我希望通过设计 A*的方法来解决上述问题。
对于给定的输入活动多集和输入模型，考虑算法的搜索状态空间如右下图所示。

slide11

在获得搜索状态空间后，需要为修复定义实际代价函数和估计代价函数来分别记录当前状态下的实际代价和当前状态到目标状态的估计代价。
实际代价函数需要考虑三种不同的操作响应的代价：
在轨迹枚举阶段：从活动多集中选择一个活动加入到轨迹中，
在对齐修复阶段：将在模型中存在但在活动多集中不存在的活动加入到轨迹中
在对齐修复阶段：将在活动多集中存在但在模型中不存在的活动加入到轨迹中

slide12

加速算法是在 A*算法的搜索状态空间的基础上，添加各种辅助索引来达到加速的算法。
辅助索引等方法包括可达索引、任务索引和分支边界等
对于可达索引，考虑左图的状态空间，通过利用可达索引可以对其剪枝，剪枝结果如右图所示。

slide13

对于任务索引，考虑左图的状态空间和路径<A, C, H>，通过利用任务索引，在对该路径进行对齐时，可以跳过分支一

slide14

对于分支边界，考虑左图的状态空间和路径<A, H>，通过利用分支边界进行剪枝，可以发现在分支一结束时，代价值为 4，而对于分支二和分支三，其最终的代价将超过 4，因此可以终止对分支二、三的计算

slide15

除去 A*算法和加速算法，本文希望继续验证解决修复无序事件日志的问题是否可以通过贪心策略来实现求解过程。
通过设计合适的贪心策略来对 h(n)函数进行修改，使该启发式函数在使用贪心策略估算当前状态到目标状态距离的同时具备无后效性，最终得到高效的算法。
为此，本文将试图寻找问题的最优子结构（即问题的最优解包含其子问题的最优解），验证由贪心选择得到的子问题的最优解与贪心选择组合在一起能否生成原

问题的最优解。

slide16
在实验方案设计方面本文将主要考虑两个层面：**1**， 实验输入数据规模 **2**. 算法的可调参数

对于实验的输入数据规模，需要从模型结构、模型大小和活动多集大小三个方面来进行考虑。
对于可调参数，需要分析算法中的参数，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

slide17
本文将从中国移动来获取流程模型和原始事件日志。同时要对原始事件日志进行乱序处理来满足本文提出问题的输入需求。

slide18
对于实验结果的评估，需要从两个方面进行考虑：**1**. 算法结果的准确性；**2**. 算法的性能
算法结果的准确性可以分为三个层面：
对输入的活动多集能够组成有效轨迹的判断是否正确
由算法得到的新的轨迹是否有效
由算法得到新轨迹的代价是否最小

对于算法性能，需要针对不同的输入数据规模来分别对 **A***算法、加速算法和贪心算法作性能比较。
slide19
除了上述实验，本文还希望通过改造现有的日志修复算法来与本文提出的算法进行比较。
因此，本文徐璈对活动多集进行改造，是它能够满足现有算法的输入需求，然后进行相同环境下的对比实验。
slide20
本课题的预期科研成果为
发表学术论文 **1-2** 篇
设计基于完全前缀展开的无序事件日志修复 **A***算法、加速算法和贪心算法

本课题可能的创新点为证明对于无序事件日志的修复问题是 **NP-hard** 问题和为无序事件日志修复问题而设计的启发式代价估计方法
slide21