第4章 面向流程挖掘算法评估的典型模型库构建

本章介绍面向流程挖掘算法评估的典型模型库的构建。首先我们对应用了典型模型库的评估框架进行总体概述，并重点阐述应用了典型模型库的评估框架与之前的评估框架相比的改进，然后分别介绍基于针对流程模型的特征选取和典型模型库的构建。

1. 1. 基于典型模型库的流程挖掘算法框架整体介绍

基于典型模型库的流程挖掘算法评估框架与基于专用模型子集与推荐技术的流程挖掘评估框架比较相似，同样的分为三部分:Rozinat评估框架、学习阶段和推荐阶段。即首先一部分模型作为训练模型集在Rozinat评估框架中进行经验分析，获得流程挖掘算法在这部分模型集合上的表现（表现是通过挖掘所得的流程模型与原始模型之间的相似度来衡量）。然后在学习阶段，对训练模型提取模型特征，并根据挖掘算法在流程模型上的表现与模型特征构建回归模型。最后在推荐阶段，对企业模型集中模型提取特征，并根据学习阶段获得的回归模型估算出流程挖掘算法在企业模型集上的表现。

但是基于典型模型库的流程挖掘算法评估框架相比基于模型子集与推荐技术的流程模型框架，主要有两点区别与改进:

1.经验评估的输入:如图4.1所示，虚线中的Rozinat框架的输入是典型模型库，而不是从企业中选取的一部分模型集合。这样对于多个企业模型集合，不必要每个企业模型都独立进行经验评估，而是一次统一训练与评估即可，减少了经验评估的时间，同时避免因为企业模型中提取训练集可能出现的训练集质量较差的情况。

2.模型特征提取: 在学习阶段和推荐阶段，模型特征提取从48维降至6维，并省去了降维的步骤。这个改变是因为48维模型特征中大部分特征的变化并不影响流程挖掘算法的挖掘结果，所以不再提取这些部分的特征。这样可以节省提取特征的时间，并增加推荐阶段的准确度。

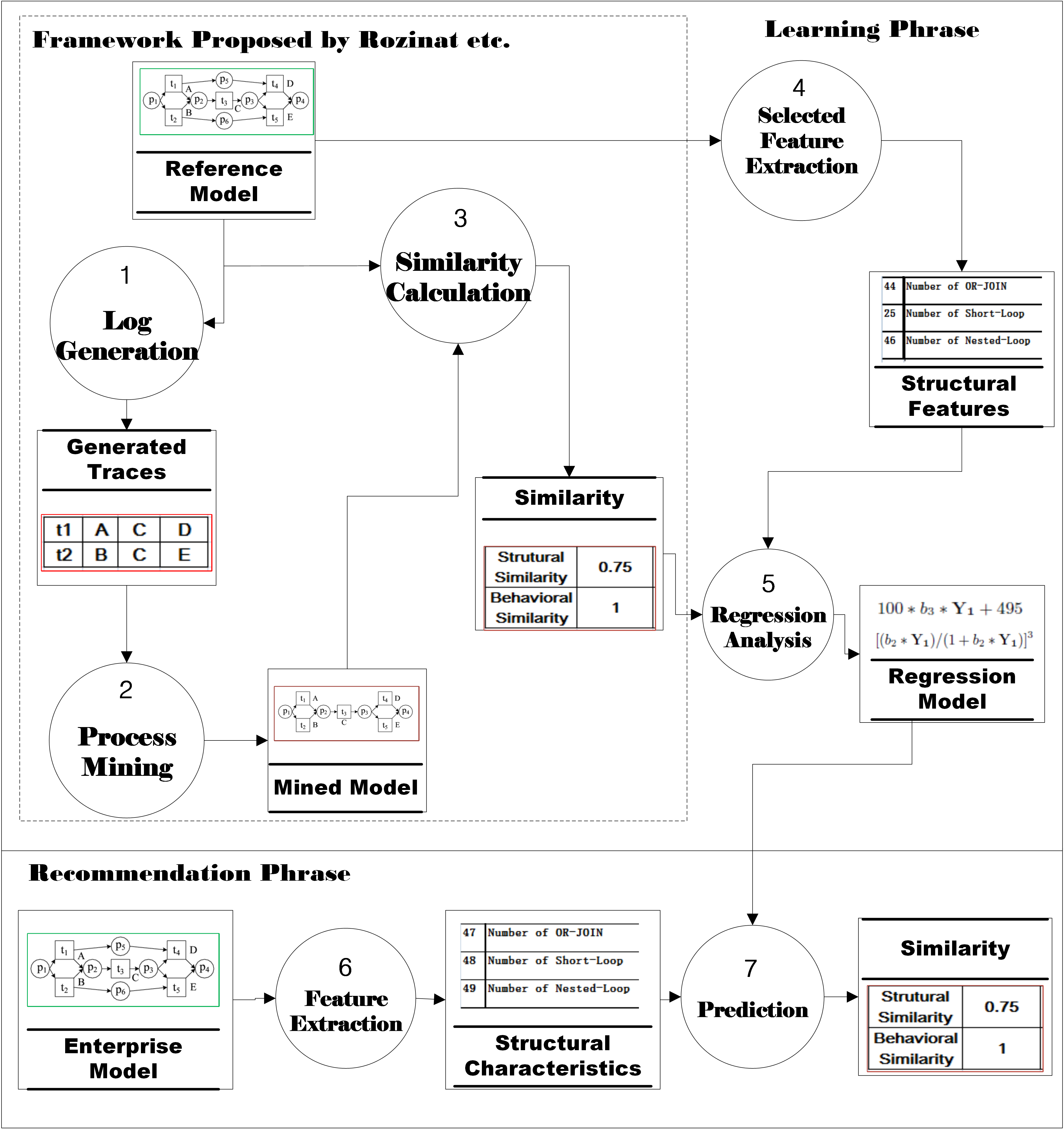


图4.1 基于典型模型库的流程挖掘算法评估框架

1. 1. 流程模型的特征选取

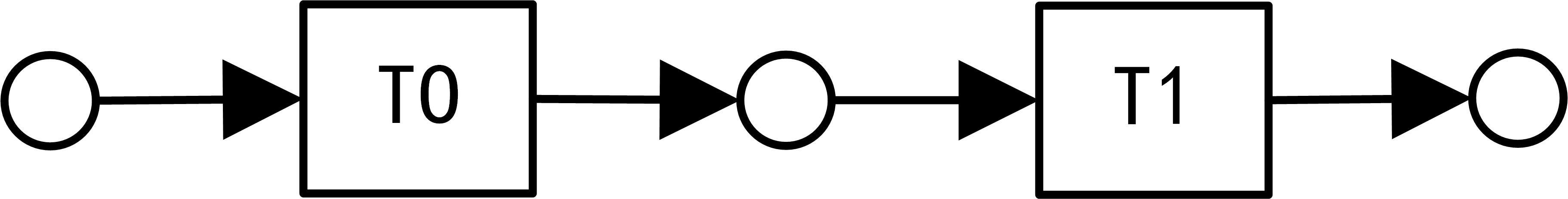
正如前文所述，描述流程模型的特征共有48维，但是其中很多特征对流程挖掘并没有影响，比如流程模型中*变迁个数*、*库所个数*等。这些不相关特征可能会在推荐过程中产生负面的影响。另外，在模型集中计算这些属性也会花费额外的时间，如果流程模型中包含的模型较多，而且模型较为复杂，那么抽取模型特征的时间开销就会比较大。在本节中，提出两个对流程模型特征选择的标准，并且针对这些标准，对要提取的模型特征个数进行约减。

* + 1. 特征选择标准

根据对现有流程挖掘算法的分析，本文提出的两个特征选择标准是:

1. 描述流程模型规模的特征应该被排除；
2. 描述流程模型Connector的特征应该被排除。

第一条标准将诸如*变迁个数*、*直径*（流程模型中起始库所到汇结库所的最长路径的长度）等特征排除。图4.2给出了关于这个标准的一个简单的例子。图中两个模型均为简单的只有顺序结构的流程模型，对应的事件日志分别是{TOT1}和{T0T1…T100}。这两个模型均可以被所有的主流流程挖掘算法正确挖掘，然而这两个模型的规模则十分不同：即一个流程模型的变迁个数为2，直径为5；而另一个流程模型的变迁个数为100，直径为201。



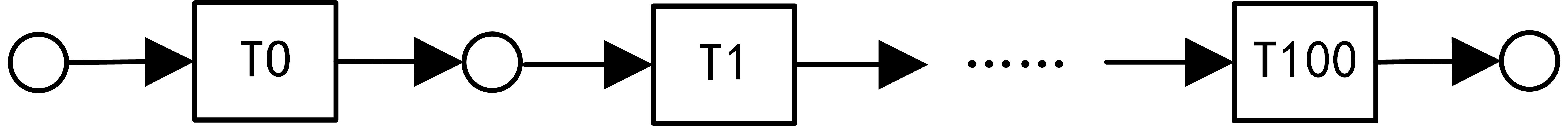
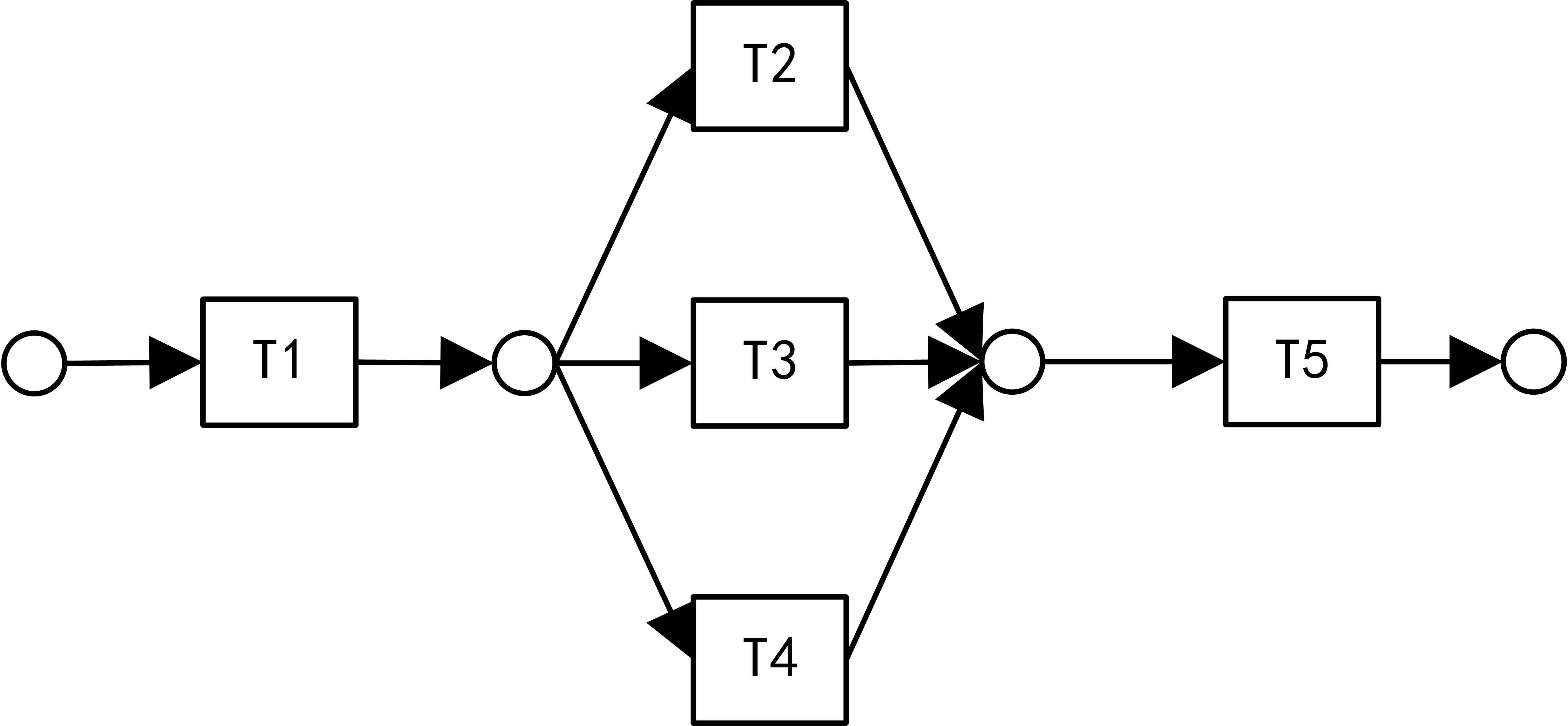


图4.2长度截然不同的两个顺序流程模型

Connector是指在流程模型中表达路由信息的元素。对于流程模型是工作流网时，Connector为*并行分裂*、*并行合并*、*选择分裂*和*选择合并*。第二条标准将诸如*并行分支个数、互斥分支个数*等特征排除。设计这条标准的原因是【A Universal里面的8】指出结构化的工作流网可以被所有的流程挖掘算法正确的挖掘出来，而Connector则是结构化的工作流网中的一部分。因此Connector相关的模型特征并不能够区分不同的流程挖掘算法。例如图4.3中两个流程模型的Connector结构完全不同，如上面的模型中只有一个只有三个分支的选择Connector，而下面的模型中则包含了一个有五分支的选择Connector和一个三分支的并发Connector。尽管这两个流程模型的Connector的信息十分不同，但是这两个流程模型都是结构化的工作流网，可以被所有的流程挖掘算法正确的挖掘出来。



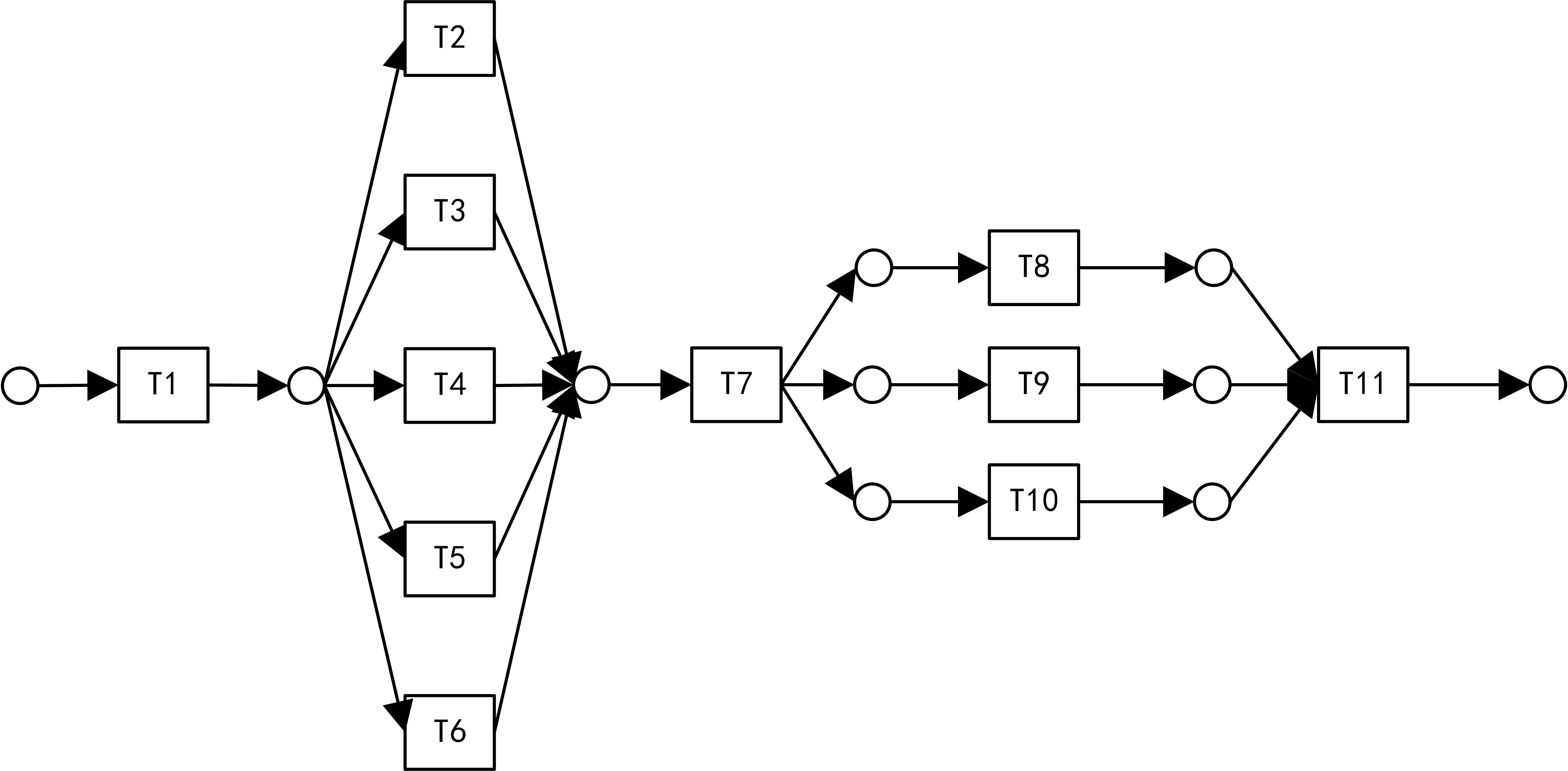


图4.3两个Connector截然不同的流程模型

* + 1. 选取的特征

基于上面提到的两个模型特征选择标准，分析了描述业务流程模型的48维特征，最后保留了六个特征:

1. 不可见任务个数：流程模型中不可见任务的个数；
2. 重名任务个数：流程模型中重名任务的个数；
3. 非自由选择结构个数：流程模型中非自由选择结构的个数；
4. 非结构化循环个数：流程模型中非结构化循环的个数；
5. 短循环个数：流程模型中短循环的个数；
6. 嵌套循环个数：流程模型中嵌套循环的个数。
   1. 一组通用的典型参考模型集合

我们手工构造了60个模型来作为这组通用的重要参考模型集合。受上一节中的6个从流程挖掘方面描述流程模型特征的启发，针对每个模型特征构建了10个流程模型。

图4.4(a)-(f)分别展示了这六个模型子集的基本属性：分别测量了每个模型子集的*变迁数目*、*库所数目*、*边数目*以及子集对应的模型特征。如对于不可见任务子集，测量了*不可见任务个数*这一属性。

* 1. 1. 不可见任务

不可见任务是指一个在流程模型中存在的任务但是不出现在事件日志中。因为这个任务并不出现在事件日志中，这使得挖掘流程模型的难度较大，流程挖掘算法需要根据日志中已有的任务来推断出没有在日志中出现的任务，并将它们放在流程模型的正确位置上。例如图4.5是一个包含不可见任务（图中黑色矩形）的业务流程模型。因为不可见任务并不保存在事件日志中，所以在图4.5所能产生的事件日志中，存在一些任务T5、T6直接相连（即T5T6这样的片段）的轨迹。

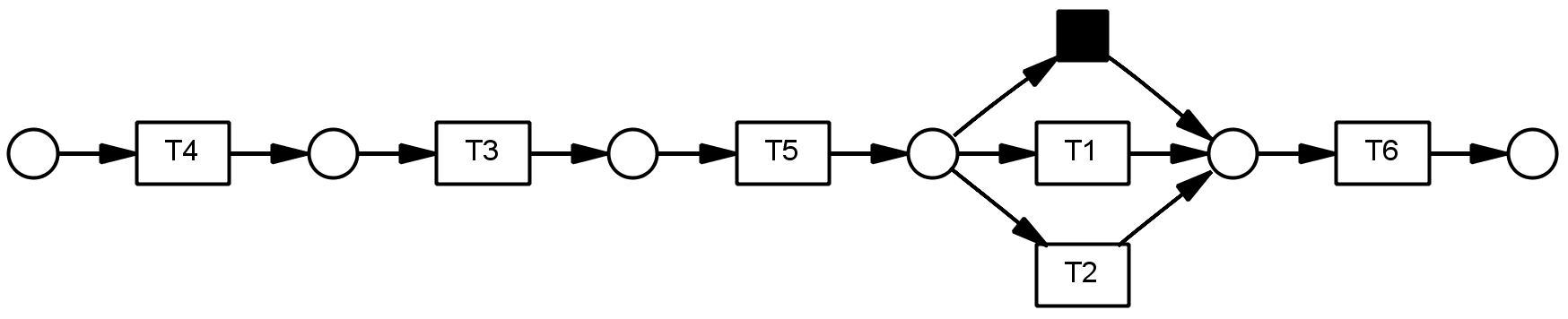


图4.5一个含有不可见任务的业务流程模型

* + 1. 重名任务

重名任务是指流程模型中包含两个或多个具有相同名称的任务。如图4.6是一个具有重名任务的流程模型。在这个模型中含有两个名称为T4的任务。在模型对应的事件日志中两个T4任务记录下来的事件是相同的,这增加了事件日志的挖掘难度：流程挖掘算法需要区分出具有相同名称的事件是否是来自相同的任务。

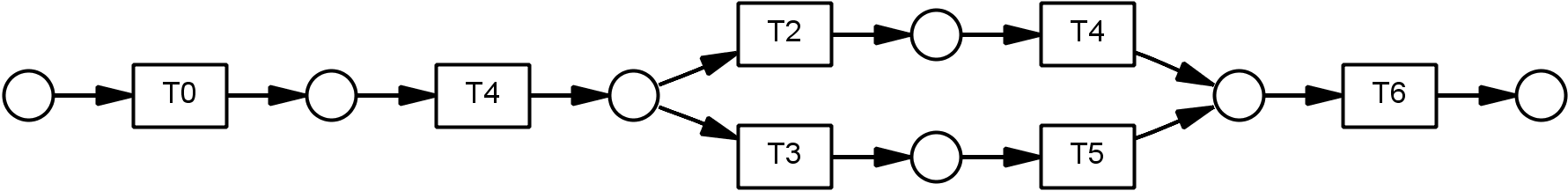


图4.6一个含有重名任务的业务流程模型

* + 1. 非自由选择

非自由选择结构是指流程模型中存在一对变迁，它们的输入库所集合存在交集但又不完全一样。例如图4.7是一个含有非自由选择结构的业务流程模型，图中T3和T4共享了一个输入库所P1,然而T3和T4的输入库所集合则不相同，T3的另一个输入库所为P2,T4的另一个输入库所为P3。库所P1、P2、P3以及变迁T3、T4构成了一组非自由选择结构，即当T2执行以后，接下来执行T3还是T4是取决于P2和P3中是否含有令牌，即取决于在T2执行之前是T0还是T1被执行了。

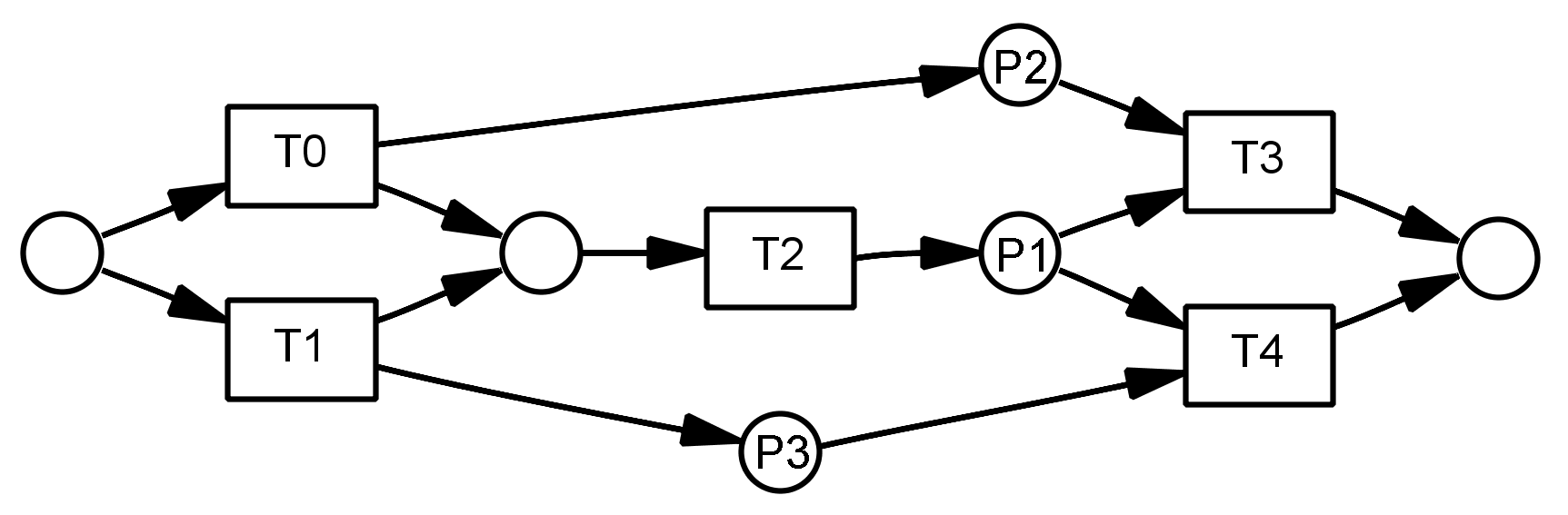


图4.7一个含有非自由选择结构的业务流程模型

* + 1. 非结构化循环

非结构化循环是指一个循环包含不止一个循环入口或者不止一个循环出口。通常一个结构化的循环中只会有一个循环入口和一个循环出口。非结构化循环会增加循环挖掘的难度（类似变成语言中goto语句往往会影响程序的可理解性），因为每增加一个循环的出口或者入口都会使得这个循环被执行的可能性增加。例如图4.8是一个含有非结构化循环结构的业务流程模型。任务T1、T3、T5构成了一个循环结构，在这个循环结构中有两个循环入口(任务T0、任务T7)和两个循环出口（任务T4、任务T6）。这两个循环入口和两个循环出口导致这个循环可以有多种组合行为，增加了流程挖掘的难度。

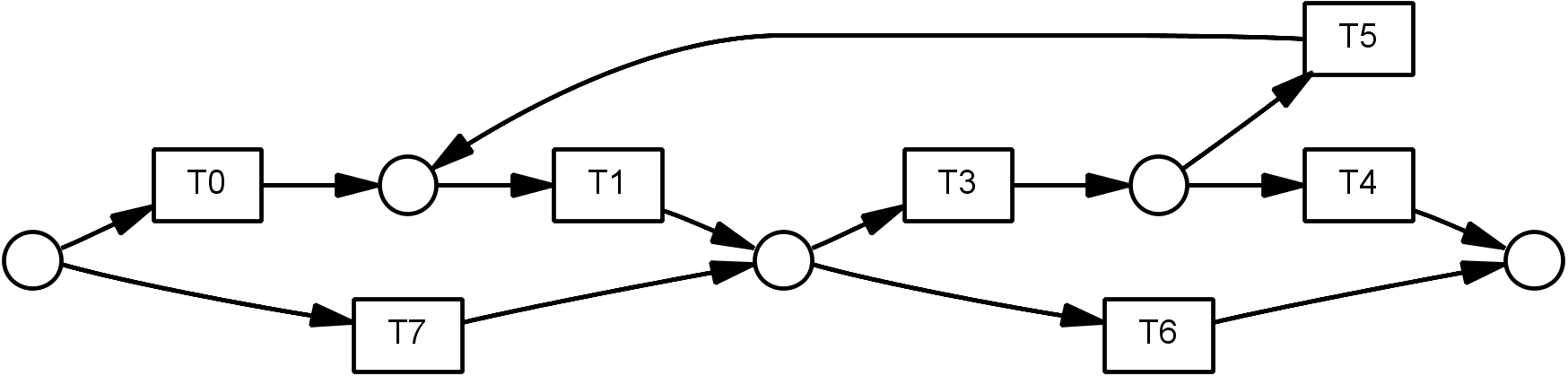


图4.8一个含有非结构化循环结构的业务流程模型

* + 1. 短循环

在一个流程中，一个任务可能会被连续的执行多次，这种情形在其对应的业务流程模型中的反映就是一个短循环结构。图4.9给出了一个含有短循环结构的业务流程模型。在这个模型中，任务T4处于一个短循环中，即T4可以被连续多次的执行。一个处于短循环中的任务在流程模型对应的事件日志中的表现为任务在一条事件轨迹中多次连续出现。

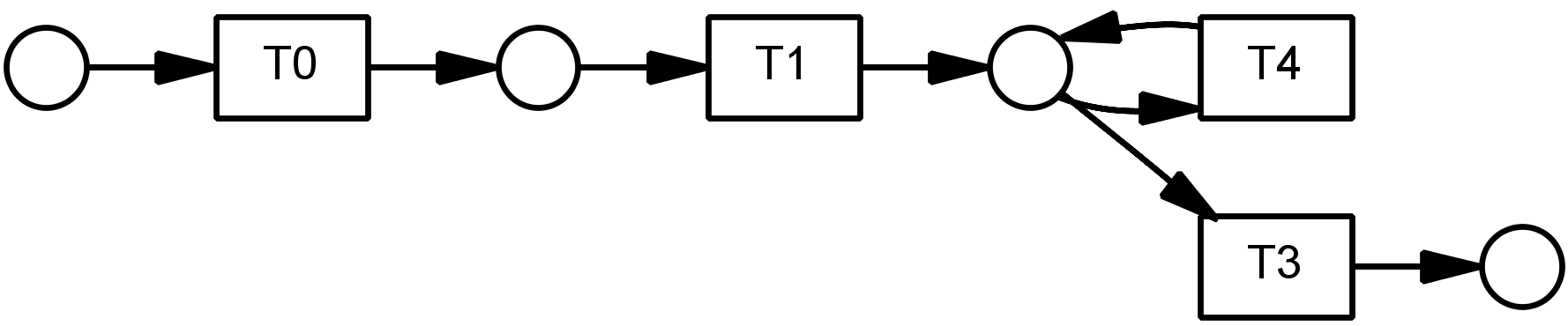


图4.9一个含有短循环结构的业务流程模型

* + 1. 嵌套循环

一个含有嵌套循环的业务流程模型是指这个模型中含有不少于两个循环，且这些循环共享一部分节点。例如图4.10是一个含有嵌套循环的业务流程模型。在这个模型中一共有两个循环结构：任务T1、T2和T6构成一个循环结构，任务T2和T7构成另一个循环结构。其中任务T2同时出现在这两个循环结构中。因为这两个循环中包含公共部分，这增加了挖掘的难度，因为流程模型的执行中可能会从一个循环中跳转到另一个循环中。

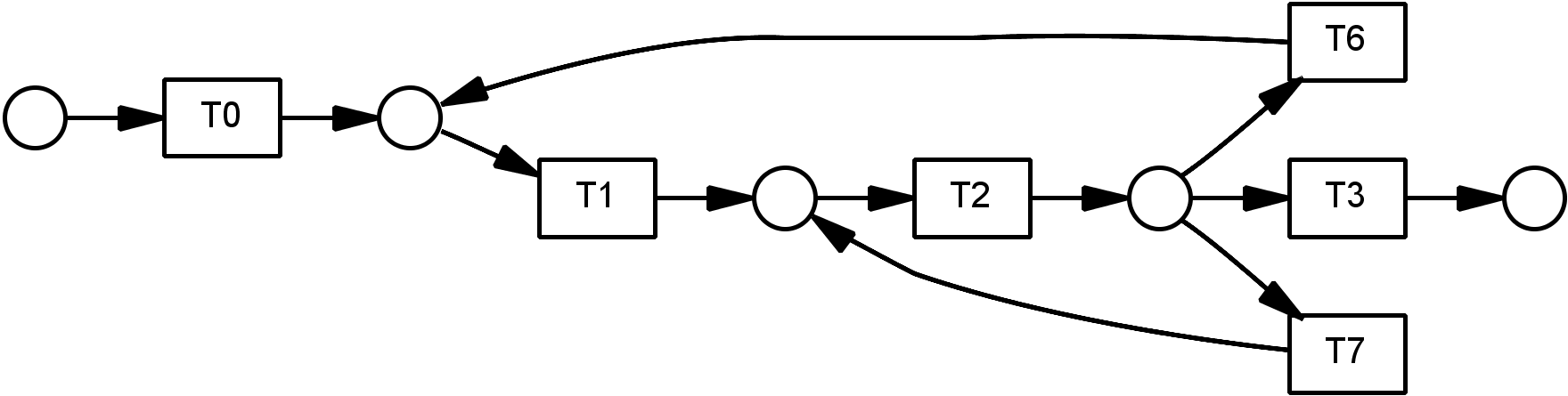


图4.10一个含有嵌套循环结构的业务流程模型

* 1. 实验评估

本节对面向流程挖掘算法的典型流程挖掘模型库进行实验。实验分为两个部分，分别评估流程模型特征选取和构建的典型模型库。本节进行的实验均实现在开源的业务流程模型与实例管理系统BeehiveZ【引用】。实验用计算机为Intel®Core™i7-2600@3.4GHz,8GB内存。

3. 4. 1. 数据集

为了保证实验的有效与完整，分别对人工模型集合和真实数据集进行了实验。表4.1中展示了各个数据集中模型的统计信息。

表4.1数据集基本统计信息

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 数据集 | 规模 | 平均值 | | | 最小值 | | | 最大值 | | |
| 变迁个数 | 库所个数 | 边个数 | 变迁个数 | 库所个数 | 边个数 | 变迁个数 | 库所个数 | 边个数 |
| 人工 | 270 | 6.1 | 6.2 | 13.2 | 2 | 3 | 4 | 13 | 14 | 30 |
| 东锅 | 108 | 7.2 | 7.6 | 14.7 | 3 | 4 | 6 | 12 | 11 | 24 |
| 高铁 | 243 | 16.0 | 14.7 | 32.6 | 6 | 6 | 12 | 36 | 32 | 72 |

人工模型集合（简称“人工”）总共有270个业务流程模型，模型来自学术论文、SAP参考模型，以及一部分作者手工创建的模型。

东方锅炉厂模型集合（简称“东锅”）是来自东方锅炉制造厂的共108个业务流程模型。东方锅炉制造厂是是中国一流的火力发电设备、核电站设备、电站辅机、环保设备、化工容器、煤气化设备等的制造商和服务提供商，产品占国内市场份额约30%。

高速铁路模型集合（简称“高铁”）是由来自中国北车唐山轨道客车有限责任公司（简称唐车公司）的243个模型构成。唐车公司始建于1881年，是中国第一家轨道装备制造企业。2008年，拥有完全自主知识产权的国产首列时速350公里“和谐号”CRH3动车组在唐车下线，中国由此成为世界上仅有的几个掌握时速350公里高速铁路装备技术的国家之一。

* + 1. 对流程模型特征选取效果的评估

对流程模型特征选取效果方面的评估，本文将选取获得的6维属性与没有经过筛选的48维属性进行对比。实验中对每个数据集选择其中的一部分数据来作为参考数据(分别选择了10%、20%、30%、40%、50%比例的数据作为参考数据集)。这部分参考数据用来训练获得回归模型，然后利用回归模型对没有当做训练集（即测试集）的模型推荐最适宜的流程挖掘算法。

**准确度**

为了评估准确度，首先将所有用来测试的模型利用Rozinat提出的评估框架经验的计算出模型最适宜的流程挖掘算法（称其为“实际最适宜流程挖掘算法”）。然后利用回归模型对测试模型估算出最适宜的的流程挖掘算法（称其为“估算最适宜流程挖掘算法”）。当 “实际最适宜流程挖掘算法”与“估算最适宜流程挖掘算法”相同时，认为估算的结果是正确的。



（c）高铁



（a）人工



（b）东锅

图4.11评估流程模型特征选取的准确度结果

图4.11(a)-(c)分别给出了人工、东锅和高铁三个数据集下的准确度的计算结果。对于三个数据集，在训练模型的比重增加的情况下，推荐的准确度也相应的增加。整体来说，利用6维模型特征的回归分析较利用48维模型特征来说推荐的准确程度较好。另外，因为高铁数据集的模型相比其他两个数据集的模型复杂，大部分流程挖掘算法无法在这个数据集中有好的表现。这一现象致使各个流程挖掘算法的在高铁数据集上的挖掘效果类似（都表现的不好），因此一个高铁模型可能会有多个最优流程挖掘算法，因而推荐算法在高铁模型集上准确度会较其他两个模型集好（因为一个模型可能有多个最优算法，评估框架有更大的可能推荐出其中一个最优的流程挖掘算法）。也因为高铁模型集的推荐准确度较其他两个模型集高，6维和48维的推荐准确度的差距较其他两个模型集小一些。

**时间效率**

在实验中,我们对比了6维特征和48维特征的特征提取时间。图12分别展示了人工、东锅和高铁模型集合的时间效率结果。普遍来看，在同样的数据集和训练集的比例中，6维模型特征时间开销小于48维特征。另外，因为高铁数据集的流程模型较为复杂，高铁模型集合的时间开销相比其他两个模型集大。东锅数据集合因为包含的流程模型较其他两个模型少，提取特征用的时间开销也是这三个数据集中最小的。



（c）高铁



（b）东锅



（a）人工

图4.12评估流程模型特征选取的时间效率结果

* + 1. 对典型模型库构建效果的评估

在对典型模型库构建效果的评估中，将本文提出的典型模型库作为参考模型集的评估（Universal Reference Model,简称URM）与传统的从测试数据集中抽取参考模型集(Original Reference Mode简称ORM)的评估进行对比。对于ORM，从测试模型集中选取三分之一的模型作为参考模型集，因为不同数据集中的模型不同，需要对应于每个测试数据集提取其对应的模型集来做训练。而对于URM来说，因为通用的典型模型库是与测试模型集合无关的，因此典型模型库可以只被训练一次就在多个测试集中做预测。

**准确度**

准确度的衡量标准与对流程模型特征选取效果进行评估的衡量标准一致，即对每一个模型，根据其真实最优算法与预估最优算法是否相同来判断推荐的准确度。

表4.2展示了准确度的评估结果，可以看出无论是人工数据还是真实数据，URM的方法均相比ORM的方法准确程度高一些。

表4.2对典型模型库构建的准确度评估

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据集 | 规模 | ORM | | URM | |
| 正确模型数 | 准确度 | 正确模型数 | 准确度 |
| 人工 | 180 | 158 | 87.78% | 166 | 92.22% |
| 东锅 | 72 | 60 | 83.33% | 66 | 91.67% |
| 高铁 | 162 | 146 | 90.12% | 154 | 95.06% |

**时间效率**

表4.3展示了对典型模型库构建的时间效率的评估。对于ORM，时间开销由三个部分组成：1.挖掘（对三分之一参考模型的训练数据集的经验评估）；2.训练（从训练集中构建回归模型）；3.推荐（对测试模型集提取特征，并利用回归模型计算推荐最优模型）。

对于URM，时间开销只有推荐一个部分组成。这是因为URM中的 训练集是通用的，只需要被挖掘和训练一次，对于不同的模型集只需要进行推荐即可。因此，相比ORM，URM的时间开销大幅度的减少。

表4.3对典型模型库构建的时间效率评估

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据集 | ORM(s) | | | | URM(s) |
| 挖掘 | 训练 | 推荐 | 总时间 | 推荐 |
| 人工模型 | 2789 | 83 | 14 | 2886 | 10 |
| 东锅 | 1393 | 31 | 11 | 1435 | 10 |
| 高铁 | 18722 | 42 | 22 | 18786 | 15 |

* 1. 本章小结

本章介绍了面向流程挖掘算法评估的典型模型库构建。介绍了基于典型模型库的流程挖掘算法评估框架，指出这个框架与基于专用模型子集和推荐技术的框架之间的两个区别。然后分别介绍这两个区别，即流程模型的特征约减，将描述流程模型的48维特征约减至6维，和根据这6维特征构建的典型模型库。实验证明在同样流程模型数据集合和流程挖掘算法的情况下，基于典型模型库的流程挖掘算法评估框架在时间开销和准确程度上均超越基于专用模型子集和推荐技术的流程挖掘算法评估框架。