判断循环的时候需要先存储循环，不能在最小case集合之后，否则主路径上的就不包含在循环部分中了

还没有写并发部分

原论文并没有写并发关系，我们可以加上去

论文写作思路：

1. 首先定义过程模型，过程模型中存在的特征关系
2. 然后定义Token，Token与过程模型的关系（Token Log能够完整的表现整个模型的行为结构……）
3. Token之间不同的关系代表原模型结构中的关系
4. 对比Token Log的相似度可以代替对比过程模型的相似度
5. 从tokenLog中提取出||，自循环，顺序，选择等结构，符合那几种情况的分别对应是XX结构。||通过XX判断，自循环的直接就是顺序结构，而没有体现出来的就是选择结构。从tokenLog的case中挖掘出这些信息，写入矩阵。【具体过程见代码写作部分】
6. 对矩阵命名一个好的名字，对此进行解释
7. 解释矩阵可以完全取代过程模型，可以通过矩阵对过程模型进行还原
8. 对比矩阵求取相似度

代码编写思路：

在读取tokenlog文件时先去重，【去重的必要性】去重的具体过程见【todolist4】

1. 首先定义Token的结构，通过readFile()函数将日志中的所有Token读入Token类型的list
2. 通过getTask()函数使用TreeSet获得日志中所涉及的所有的任务并有序排序，并通过HashMap将任务的名称与位置一一对应，从而根据任务集合构建模型的矩阵
3. 遍历所有的tokenLog，先将顺序关系写入矩阵，然后将tokenLog按照生产任务的生产时间排序，对比所有拥有相同生产时间的token，如果符合并发结构的条件就写入矩阵。然后将矩阵写入writeFile()文件
4. 对比相似度时，countSimilarity()函数通过计算矩阵中所有的非空的数目相加作为分母，两个矩阵相对应的作为分子，但是在对比相似度前需要先将矩阵同维化。【同维化过程见todolist（1）】 ，使用jaccord算法进行计数比较矩阵的相似度，可以将jaccord的那个公式列出来。
5. 矩阵记录了过程模型的弧线关系以及并发关系，所以通过实际操作，可以通过矩阵还原过程模型

To Do List

1. 将不同维度的矩阵同维化，然后对比相似度

不同维度的矩阵同维化：首先将两个模型矩阵的任务集合分别提取出来，求取并集（这可以画图表示，新的模型矩阵是旧的模型矩阵……根据论文（基于任务最短跟随距离矩阵的过程模型行为相似性算法））

（1）通过union()函数使用TreeSet数据类型获得两个模型矩阵的所有任务集合的并集并且按顺序排好，返回空的同维矩阵

（2）通过transform()函数将待比较的两个矩阵中的关系分别存储进新的同维矩阵

（3）通过countSimilarity()函数，遍历两个同维矩阵，计算不为空的总数以及相对应关系相同的个数并×2

（4）最后不为空的总数作为分母，相应关系相同的个数作为分子，求相似度

1. \*对比相似度的时候，实验判断是否为并发结构赋予不同的权重
2. 对比两个tokenLog的相似度（求最小case集合，求case的触发序列，case之间分别对应，求相似度）【可以通过实验证明这样的处理方式比处理矩阵对比慢】 或者是求出来所有case的关系序列，然后在case之间映射的时候用单向距离的平均值。

使用双向距离来表示相似度的大小

首先根据去重后的tokenLog<List<Token>, String>，进一步进行去除循环结构，只保留一个循环节，同时从这里边提取出Token的消耗关系<P, C, R>（利用set进行存储，可以去除循环，指保留一个）hashMap<List<R>, int>用来存储每一个case的内的发生关系以及该case的数目。对两个tokenlog求相似度，就是求最小case之间的相似度，通过一一遍历求取距离最小的，最大时间复杂度为n^2，然后双向距离求平均。

这个方法的计算结果受case的影响，相同的两个模型，不同的运行case，计算结果可能就不同，所以不太适合，此外，除去了循环结构，使得k=1或者0，但是因此也会使得忽略循环结构，使得计算结果有误差。并且计算的复杂度也较大

1. 对比tokenLog是否去重对计算时间的影响，比较两个的计算时间

去重：构建数据结构<生产者，生产时间，消费者，消费时间>，将每一个case中的所有token存为list的形式，然后创建hashMap，key为caseList，value为对应该caseList的caseID，最后获得的hashMap即为<caseList<Token>, String<caseID>>的形式，使用hash进行检查是否重复，最后获得去重后的结果。去重之后，遍历hasnMap，对所有的case进行遍历，将其关系存为hashMap<String, String> 的形式，

遇到的问题：

1. 在比较两个String[][] 的元素是否相等时，用了==，导致矩阵同维化错误，找bug找了一下午。

可以使用hashtable存储矩阵，这样操作起来比较方便………………hashtable内部是无序的

使用矩阵的好处：

1. 包含了过程模型结构中所有的特殊结构：顺序，选择，并发，单步循环，多步循环（这个是直接对比token log相似性所不具备的有点）
2. 不随case数目的增加而增大计算量，最终矩阵的大小只与任务数量有关
3. 计算比较简单，时间复杂度比较小

【My question：】

1. 并发结构，有没有可能是三个甚至是四个任务并发？还是说只有两个任务可能会并发？

2. 对于一个几个G的文件，其中包含的最小运行实例集合可能就几个，但是其case却可能是上万个，所以重复率是很高的，去重是很有必要的。

在去重的过程中我可不可以认为，我通过前（比如）1000个case就可以获得所有的case集合，通过后边case来验证，如果重复率达到100%，那我就不在读入其他的case，而利用当前的case集合去处理后续的步骤。

1. 为什么log中的序号不是按序的，中间有间断

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型 | Case数目 | 消耗时间 | 文件大小 |  |
| 1 | 1000 | 3s | 221k |  |
| 1 | 10 000 | 5min～6min | 2.42m |  |
| 1 | 100 000 | 未跑出来 |  |  |
|  |  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | Case数目 | 最小case集合数目 | 运行时间 |
| 无去重 | 1000 |  | 1102ms |
| 去重 | 10000 |  | 751ms |

在判断是否符合五条性质的时候，TLP矩阵并不能符合五条性质，仅能符合性质1，3，所以需要重新定义TLP矩阵。

重新定义三条性质：漂移不变性，跨度负相关性，

顺序结构的调整，即顺序结构所在位置不同

添加顺序结构的位置不同，添加顺序结构的个数负相关性

并行结构，添加并行结构的位置

添加并行结构的数目

添加并行结构的跨度

循环结构的位置

循环结构的数目

循环结构的跨度

互斥结构……

循环结构可以通过判断同一case中重复出现的就是循环结构

选择结构，求得最小case集合，最小case集合并集-交集就是选择结构上的路径，最小case集合包含运行路径最大而且循环结构运行次数为1，交集代表了主路径上的弧线，可以画图来表示

并发结构

实验数据

性质1：

模型： "M1.tlog" VS "M2.tlog"

TLP同维化矩阵为：

A B C D P0 P3

A 1

B 1

C 1

D

P0 1

P3

TLP同维化矩阵为：

A B C D P0 P3

A 1

B 1

C 1

D 1

P0 1 1

P3

8 / 10

相似度结果为： 0.80000

155ms

模型： "M1.tlog" VS "M3.tlog"

TLP同维化矩阵为：

A B C D P0 P3

A 1

B 1

C 1

D

P0 1

P3

TLP同维化矩阵为：

A B C D P0 P3

A 1 1

B 1

C 1

D 1

P0 1

P3

8 / 10

相似度结果为： 0.80000

160ms

模型： "M1.tlog" VS "M4.tlog"

TLP同维化矩阵为：

A B C D P0 P3

A 1

B 1

C 1

D

P0 1

P3

TLP同维化矩阵为：

A B C D P0 P3

A 1

B 1 1

C 1

D 1

P0 1

P3

8 / 10

相似度结果为： 0.80000

145ms

**Reviewer's confidence (\*)**. 审稿者的信心

|  |
| --- |
| 5: (expert) |
| 4: (high) |
| 3: (medium) |
| 2: (low) |
| 1: (none) |

**Is the paper topic relevant for ICSSP (\*)**. Please indicate if you consider this paper to be relevant for the ICSSP audience主题是否与ICSSP相关

|  |
| --- |
| 1: strongly relevant |
| 0: moderately relevant |
| -1: little or weak relevance |

**Is the literature presented sufficient to support the context of the paper? (\*)**. Please indicate if you consider the literature background presented to be sufficient to understand the context of the problem/research/study described in the paper. 文献是否足以支持论文的背景？（\*）。 请说明您是否认为所提供的文献背景足以了解论文中描述的问题/研究/研究的背景。

|  |
| --- |
| 1: good / sufficient |
| 0: moderate |
| -1: poor / insufficient |

**Is the experimental / theoretical methods / methodology well described, credible and defensible? (\*)**. Please indicate to what degree you consider the experimental / theoretical method / methodology well described, credible and defensible实验/理论方法/方法是否被很好地描述，可信和可辩护？（\*）。 请说明您认为在何种程度上所描述的实验/理论方法/方法，可信和可靠

|  |
| --- |
| 1: good / sufficient |
| 0: moderate |
| -1: poor / insufficient |

**Are interpretations and/or conclusions justified by the results? (\*)**. Please indicate if you consider the interpretations and/or conclusions presented in the paper are justified by the results? 解释和/或结论是否证明了结果？（\*）。 请说明您是否认为论文中提出的解释和/或结论有结果证明合理？

|  |
| --- |
| 1: yes |
| 0: somewhat |
| -1: no |

**Overall evaluation (\*)**. Please provide a detailed review, including a justification for your scores. Both the score and the review text are required. 全面的评估 （\*）。 请提供详细的评论，包括评分的理由。 评分和评论文本都是必需的。

|  |
| --- |
| 2: accept |
| 1: weak accept |
| 0: borderline paper |
| -1: weak reject |
| -2: reject |