

问题和关键定义：

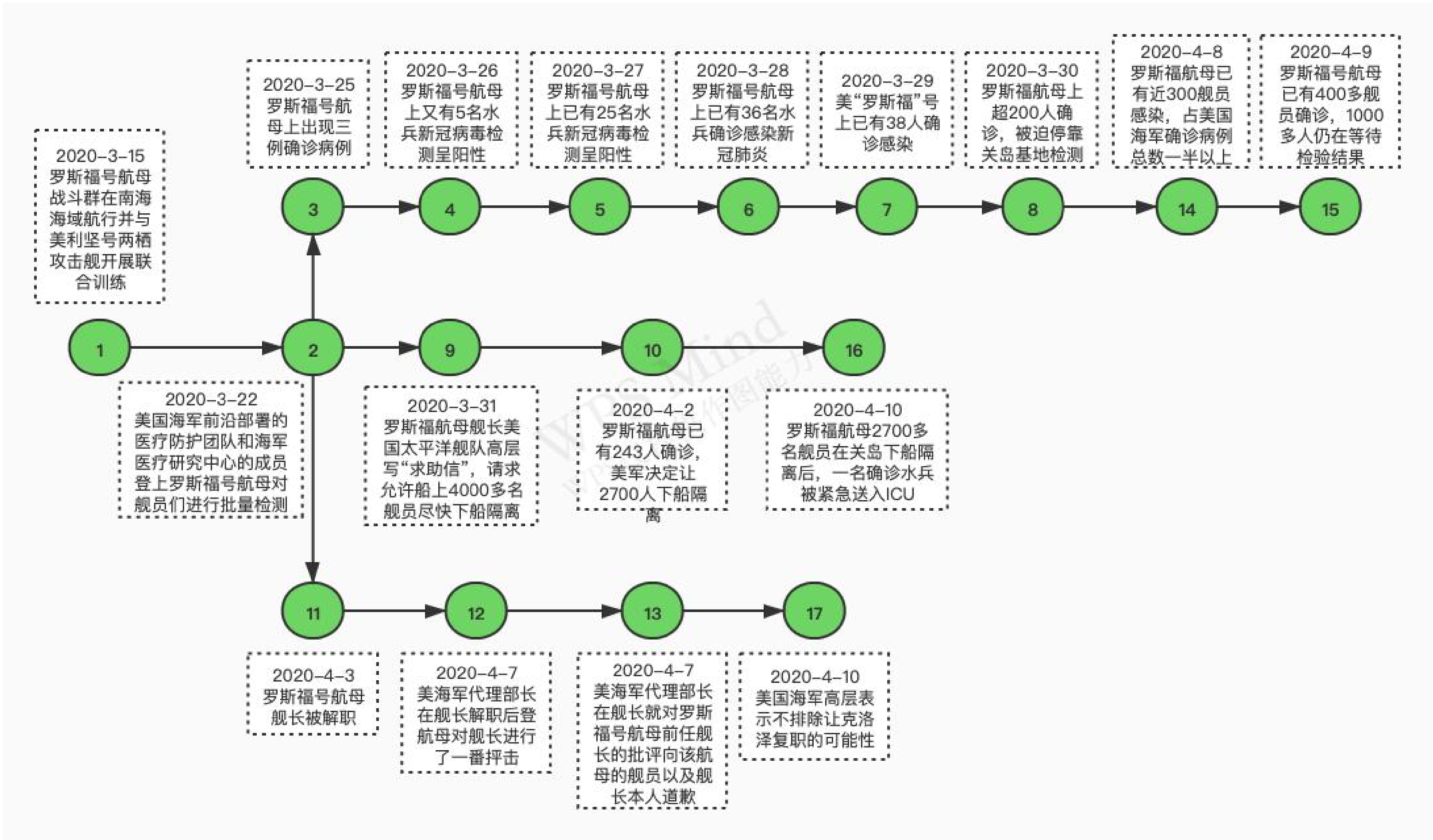
如何在涵盖不同主题和包含大量冗余信息的海量新闻简报中准确、快速提取热点事件。同时，在分析和理解简报上下文语义的基础上，以实时在线的方式描述事件的演变过程，并在不破坏已有事件演变轨迹的基础上，确保用户的浏览体验。

定义1（新闻简报）是指以某个热点事件为核心而触发的新闻报道，时态是突发性或跃进性的。特点是言简意赅、紧扣主题，篇幅长短不一以及写作风格各异。

定义2（热点事件）是指在最近发生并具有重大影响力和极受关注的事件。热点事件E表示一组报道相同事件的一篇或多篇新闻简报。

定义3（事件演变轨迹）是指围绕一个或多个特定的人物或命名实体，在特定的时间发生在特定地点的事件发展结构图。事件E1到E2的有向边缘表示事件E1到E2的时间演变关系或逻辑连接关系。

定义4（主题）是指由一组高度相关或彼此相似的事件演变轨迹集合。



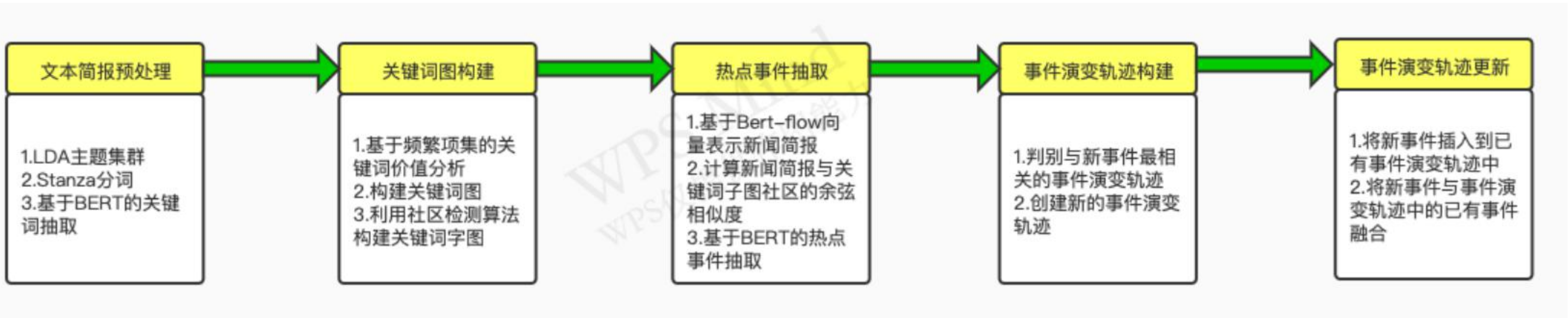
事件演变轨迹模型：

构建事件演变轨迹模型的方法概述如下图所示，由热点事件智能检测和事件演变轨迹实时构建两大部分组成，共分为5个步骤。

首先，通过主题建模方法和NLP工具处理新闻简报流，模型在预处理阶段主要包括的是新闻简报的主题集群创建、分词和关键词提取等。

步骤2-3，笔者提出一种三阶段的方法将新闻简报分组到热点事件中。对于每个时间段t的简报语料库D_t，模型计算出关键词的频繁项集，选取重要性高的关键词来构建关键词图。然后，再利用社区检测算法从关键词图中提取子图作为子主题。其中，关键词少的子主题将被剔除。获得子主题后，找到与子主题内容相关的新闻简报，并通过一个能分析上下文语义的语言模型将简报聚类成细粒度的热点事件。该过程由一个预先训练好的新闻简报对关系分类器辅助进行。

步骤4-5，实时构建和更新事件演变轨迹，模型是将每个新事件插入到现有轨迹中的正确位置，或者创建一条新的演变轨迹（前提是新事件不属于任何现有轨迹）。与现有方法不同，我们判别与事件关联的轨迹时，融入了上下文语义分析，改进了事件间连接强度的计算方法。



性能评估：

- 事件聚类方法的比较结果
- 与其他事件发展结构描述方法的对比

表 3 事件聚类方法的比较结果			
Table 3 Comparison results of event clustering methods			
Algorithm	Homogeneity	Completeness	V-measure
Our approach	0.968	0.970	0.969
LDA+AP	0.642	0.932	0.760
KeyGraph	0.582	0.977	0.730
Story Forest	0.952	0.946	0.949

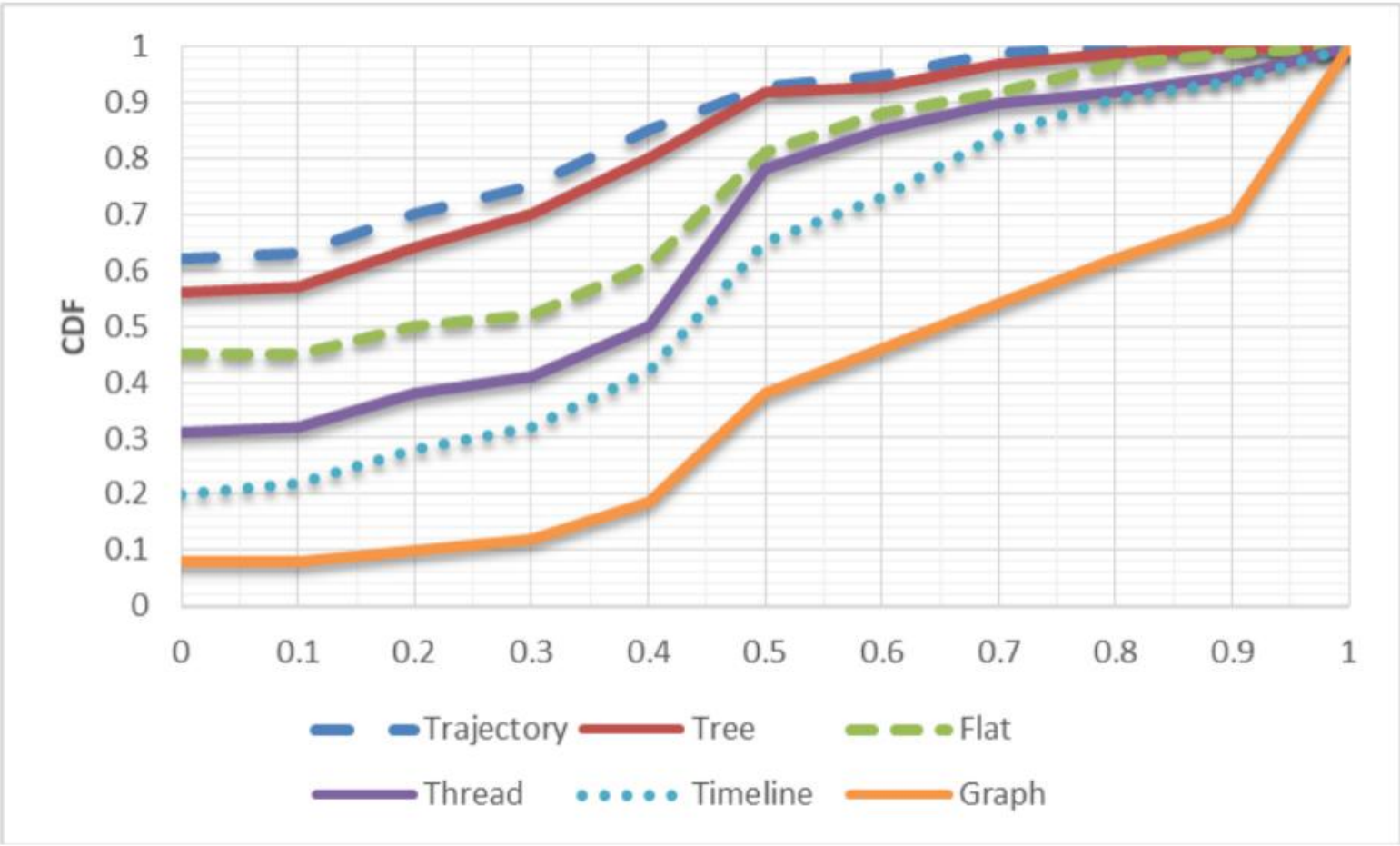


图 4 不正确边缘的百分比
Fig.4 Percentage of incorrect edges

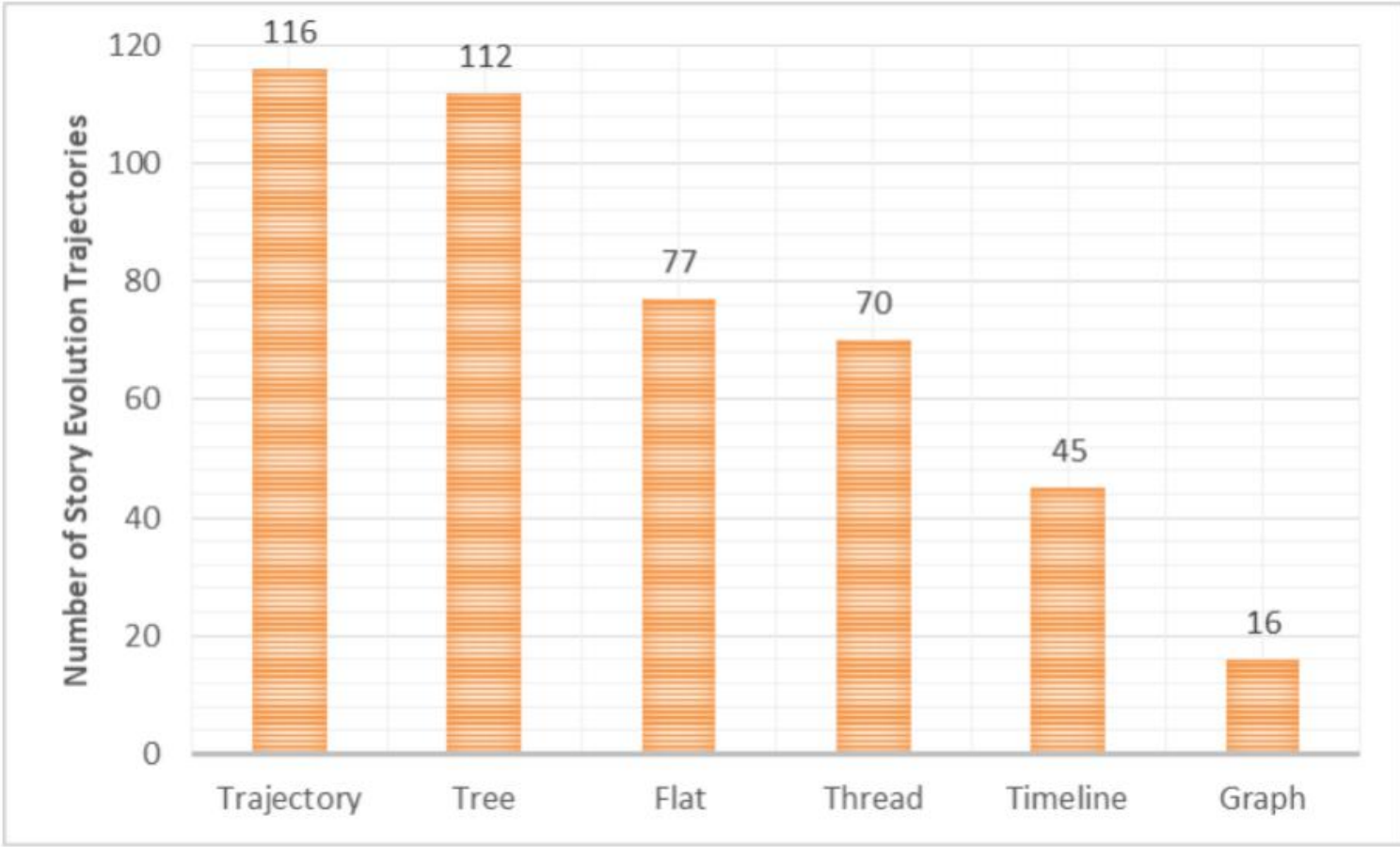


图 6 被评为最易读结构的次数
Fig.6 Number of times rated as the most readable structure

表 4 比较不同的事件演变结构生成算法						
Table 4 compares the different event evolution structure generation algorithms						
Algorithm	Traject	Tree	Flat	Thread	Timeline	Graph
Correct edges	86.8%	83.1%	70.4%	66.6%	60.2%	42.4%
Consistent paths	82.3%	78.5%	-	52.1%	31.6%	-
Best structure	116	112	77	70	45	16

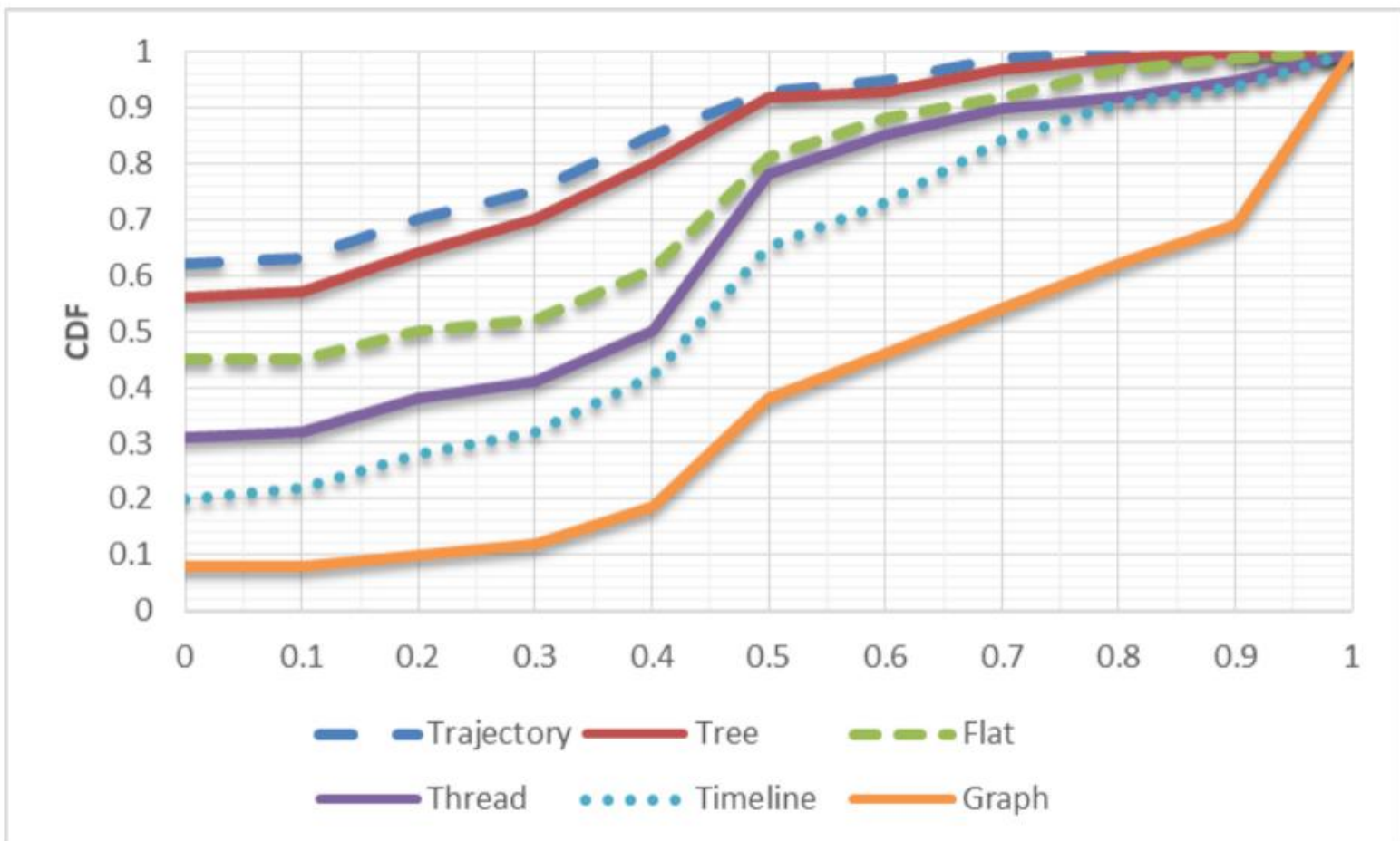


图 5 不一致路径的百分比
Fig.5 Percentage of inconsistent paths

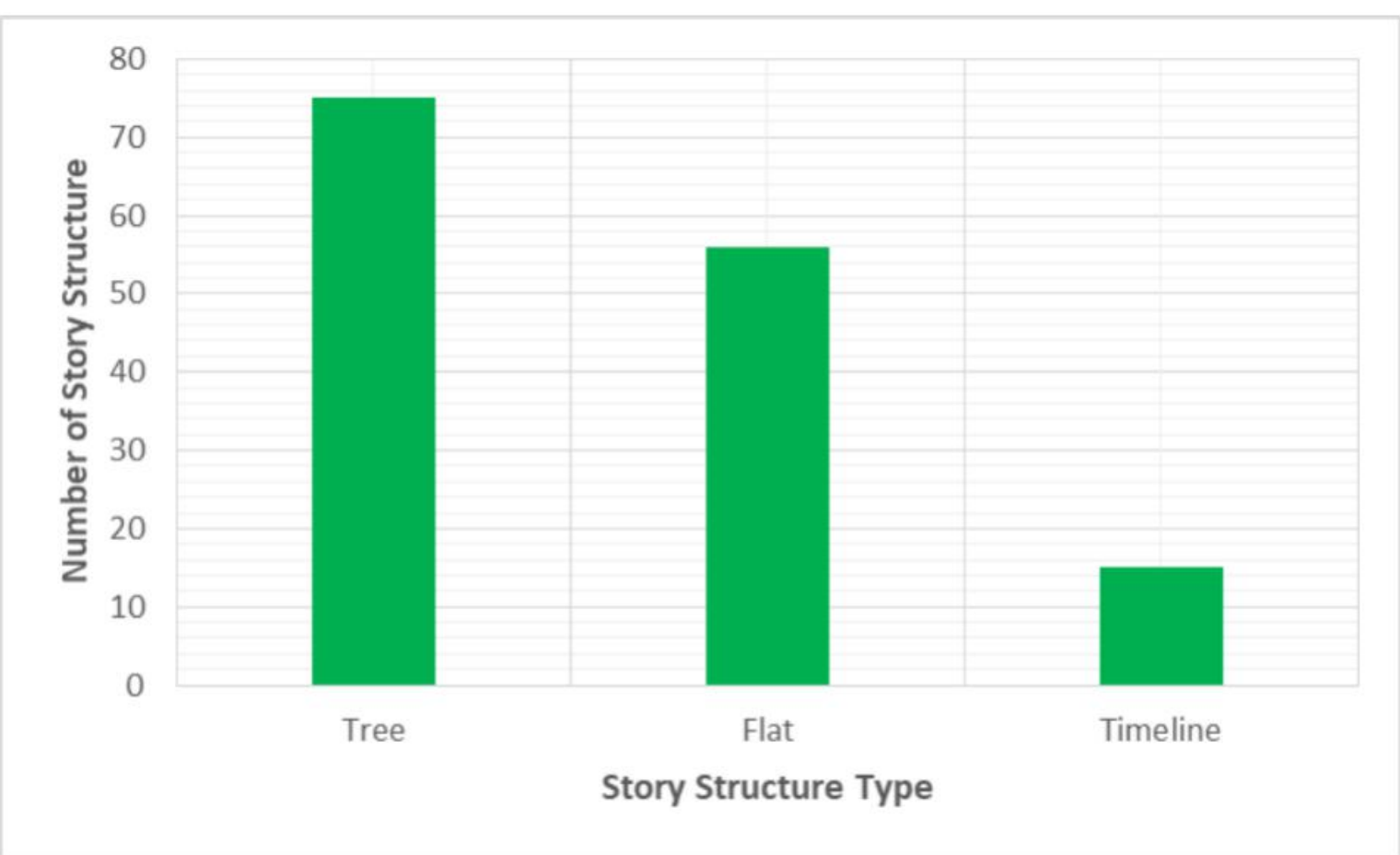


图 7 不同表示结构的数量
Fig.7 Numbers of different story structures