# 1 引言

## 1.1 研究背景

随着互联网、物联网技术的飞速发展,近年来在许多管理信息系统及决策支持系统应用中出现了大量的由一系列值对(事件类型，发生时间)组成的时序数据，如网络监控日志、数据库日志、入侵检测序列、电信故障诊断、通信网络告警序列、Web交互日志、股票交易日志、道路交通监控、基因序列、用户行为预测等[1]。这些事件背后隐藏了大量的规律，对这些事件序列进行挖掘和分析，可以发现他们之间的关联关系，从而揭示用户或系统潜在的行为模式，这种反复出现的、规律的模式被定义为“情节”[1]。频繁情节就是事件序列上频繁出现且有一定关系的事件类型的集合。频繁情节挖掘已经成为数据挖掘领域，尤其是时序数据挖掘领域的热点之一，它具有广阔的应用前景，目前已被广泛应用在多个领域，为科学预测及商业决策提供了依据。

为了解决频繁情节挖掘问题，ManiLLa等人[1]首先给出了情节的定义，并在此基础上提出了两个经典算法WINEPI[2]、 MINEPI[3]。其中，WINEPI基于滑动窗口(Sliding Window)来定义支持度，支持度定义为出现情节的滑动窗口的个数；而MINEPI是基于情节的最小发生来定义支持度，最小发生指情节在一个时间区间中出现，但不在其任何真子区间上出现。这些算法在挖掘频繁情节时都存在一些不足，比如滑动窗口大小固定，不够灵活；在对情节进行计数时可能包含多次重叠的发生，存在“重复计数”和冗余问题，从而导致挖掘的质量和效率不高。

为此，Laxman等人在上述算法的基础上提出了“非重叠发生”的概念[4]，并且提出了一个高效、快速的情节挖掘算法NONEPI[5]，该算法基于非重叠发生的支持度定义，通过采用有限状态机来统计情节发生的次数，该算法的缺点是与Apriori算法类似，通过迭代产生大量的候选和频繁情节，且每次对候选情节计数时都需要扫描事件序列一次。有些学者提出了一些改进的算法[6,7]，在计数之前对候选情节预先进行“剪枝”，提高了算法的执行效率。然而，当事件类型较多、事件序列长度较大时，反复扫描序列仍然需要较高的时间和空间代价。

为了克服以上算法的不足，本文提出一个改进的频繁情节挖掘算法NONEPI++，该算法通过计算情节发生的时间戳来产生候选情节，然后通过“剪枝”挖掘满足支持度阈值的频繁情节，算法只需扫描事件序列一次，大大提高了频繁情节挖掘的效率。

## 1.2 研究平台

（1）Eclipse

Eclipse 是一个开放[源代码](http://baike.baidu.com/subview/60376/5122159.htm" \t "_blank)的、基于[Java](http://baike.baidu.com/subview/29/12654100.htm" \t "_blank)的可扩展开发平台。就其本身而言，它只是一个框架和一组服务，用于通过插件组件构建开发环境。Eclipse 附带了一个标准的插件集，即Java[开发工具](http://baike.baidu.com/view/1355803.htm" \t "_blank)（[JDK](http://baike.baidu.com/subview/25214/5047948.htm" \t "_blank)）。Eclipse 还包括插件开发环境（Plug-in Development Environment，PDE），这个组件主要针对希望扩展 Eclipse 的软件开发人员，因为它允许构建与 Eclipse 环境无缝集成的工具。

（2）JAVA语言

JAVA语言是一种可以撰写跨平台应用软件的面向对象的程序设计语言，是由Sun Microsystems公司于1995年5月推出的Java程序设计语言和Java平台（JavaSE, JavaEE, JavaME）的总称。Java 技术具有卓越的通用性、高效性、平台移植性和安全性，广泛应用于个人PC、数据中心、游戏控制台、科学超级计算机、移动电话和互联网，同时拥有全球最大的开发者专业社群。在全球云计算和移动互联网的产业环境下，Java更具备了显著优势和广阔前景。Java语言的跨平台特性也为算法设计提供了强有力的支持，目前很多数据挖掘算法都是基于Java语言实现。如Java数据挖掘包（JDMP）是一个开源的Java程序库，用于数据分析和机器学习。

## 1.3 本文组织结构

第1章：研究背景，介绍情节挖掘的国内外研究现状。

第2章：预备知识，介绍情节挖掘的相关概念。

第3章：频繁情节挖掘算法，介绍非重叠发生的频繁情节挖掘算法NONEPI。

第4章：改进的频繁情节挖掘算法，提出改进的频繁情节挖掘算法NONEPI++。

第5章：算法实现，采用Java语言实现NONEPI++算法。

第6章：实验评估，基于模拟及真实数据集对两种算法进行评估。

第7章：总结与展望。

# 2 预备知识

## 2.1 事件序列

序列由各种类型的事件组成，事件定义为一个二元组(*A*,*t*)，其中A为事件类型,事件类型的集合记为**ε**(*A*∈***ε***)，*t*表示事件发生时间，一般定义为正整数[1,8]。

事件序列**S**定义为一个三元组*(S,Ts,Te)*，其中*S*=<(*A*1, *t*1), (*A*2,*t*2), *…*, (*A*n, *t*n)>，且Ai∈**ε**(i=1,2,…n)，ti≤ti+1(i=1,2,…n-1)，Ts和Te是整数，其中Ts表示事件序列**S**的起始时间，Te表示事件序列**S**的终止时间，且Ts≤ti<Te。

例如，图2.1所示的事件序列**S**=(*s*,29,68)，其中**s**=<(E,31),(D,32),(F,33),(A,35), (B,37),(C,38),…(D,67)> , Ts=29, Te=68。

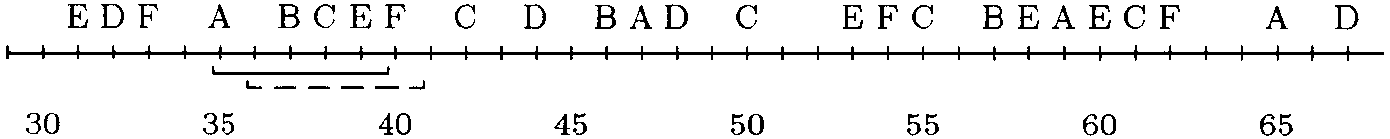


图2.1 事件序列

## 2.2 窗口

事件序列上的窗口定义为**w**=(*w*,*ts*,*te*)，*w*由所有满足*ts*≤ *t* <*te*的二元组(A,*t*)所组成。且满足（*ts*<*Te*，*te*>*Ts*）。其中，*te*-*ts*称为窗口w的宽度，记作*width(*w*)*。事件序列**S**上所有宽度为*win*的窗口的集合记作***W****(s,win)*。

根据定义，给定序列上的第一个窗口和最后一个窗口超出了序列，因此第一个窗口仅包含事件序列中的第一个时间点，最后一个窗口仅包含事件序列中的最后一个时间点。根据定义，在事件序列两端点附近的事件与在事件序列中间的事件都在同样多的窗口中得到考虑。给定事件序列**s**=(*s*,*Ts*,*Te*)和窗口宽度*win*，那么***W****(s,win)*中的窗口数目是*Te*-*Ts*+*win*-1。

例如，图2.1中显示了事件序列**S**上宽度为5的两个窗口，一个窗口开始于时间35，用实线表示，另一个紧跟其后的窗口开始于时间36，用虚线表示。开始于时间35的窗口是 ( <(*A*,35),(*B*,37),(*C*,38),(*E*,39)>,35,40)。注意事件(*F*,40) 发生在该窗口的结束时间上，但它不属于该窗口。同理，开始于36的窗口与此类似，区别是去除第一个事件(A,35)，新的事件(F,40)在最后。

## 2.3 情节

情节是指相邻发生的有着全序关系的事件类型的集合。情节可以用有向无环图来描述，如图2.2所示的情节α和β。

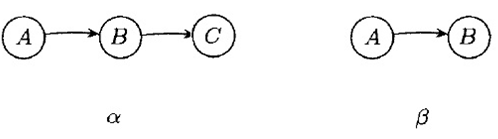
 

图2.2 情节示例

情节α出现在图2.2所示的序列中，事件类型A和B按照先后顺序发生。当然，在这两个事件之间也可以有其它的事件类型发生。情节的形式化定义为一个三元组(V, <,g)，其中V为节点集，<为V上的全序关系，g:V→E为从节点到事件类型的映射。情节α的含义就是：g(V)中的事件以全序关系<所指定的次序发生。情节α所含的事件数目记作|α|，值为|V|。如果没有某个事件类型在一个情节中出现两次，即映射g是单射的，这样的情节是单射情节[2,9]。

例如，考虑图2.2中的情节α=(V, <,g)，节点集V包含两个节点，记为x,y。映射g表示图中的事件类型和节点的关系：g(x)=A，g(y)=B，A在B之前发生，也就是x在y前发生，有x <y。情节α是单射的，由于它没有包含重复的事件类型。在α发生的窗口中，有两个事件类型A和B，因此这里只计算α出现的窗口数，不是每个窗口中的发生个数。

如果存在单射f:V’→V，对任意v∈V’，都有g’(v)=g(f(v))，对任意v,w∈V’，有v<’w，则也有f(v)≤f(w)，那么称情节β=(V’,<’,g’)是情节α =(V,<,g)的子情节，或情节α =(V,<,g)是情节β=(V’,<’,g’)的超情节，记为β⊆α或α⊇β。若β⊆α且β≠α，则β是α的真子情节，或α是β的真超情节，记为β⊂α并且α⊃β。

例如，图2.2中可以看出α=<ABC>，β=<AB>，β是α的子情节，同时也是α的真子情节，但是β’=<BA>不是α的子情节。

## 2.4 情节的发生

情节α=(V,<,g)在事件序列**s**=(<(A1,t1),(A2,t2),……,(An,tn)>,Ts,Te)中发生，当且仅当存在一个从V的节点到**s**中事件的单射h:V→{1,……,n}，满足对于所有的x∈V，有g(x)=Ah(x)，并且对任何x,y∈V，如果有x≤y(x≠y)，则有th(x)<th(y)。例如，图2.1中的窗口(w,35,40)包含了事件A、B、C、E。

给定序列S和情节α=<E1E2…Ek>，若至少存在序列S'=<(E1,T1),(E2,T2,…(Ek,Tk)>，满足Ti<Ti+1(1≤i<k)，且S'是S的一个子序列，称情节α在S上发生。区间[Ti,Tk]称为α在S上的一次发生，其中Ti和Tk分别为起始时间和终止时间。

设[Ts,Te]是情节α在S上的一次发生，若不存在另一次发生[Ts',Te']，使得Ts<Ts'且Te'≤Te，则[Ts,Te]是α在S上的一次最小发生。[Ts,Te] 为α的一次发生的时间戳，α的所有最小发生的时间戳集合记为α.mo[2,10]。

设[Ts,Te]和[Ts',Te']是情节α在事件序列S上的两次发生，若Te< Ts'，则称[Ts,Te]和[Ts',Te']是α在序列S上的非重叠发生。α所有非重叠发生的时间戳集合记为α.no[11]。

若[Ts,Te]和[Ts',Te']是α在S上的两次最小发生且Te< Ts'，称[Ts,Te]和[Ts',Te']是α在S上的两次最小且非重叠发生[12,13]。α所有最小且非重叠发生的时间戳集合记为α.no。