

一种从时空数据中挖掘事件影响传播树的新方法

房滇武¹, 王丽珍², 邹目权^{1,3}, 邓飞¹

(1. 昆明学院信息工程学院, 云南 昆明 650214;

2. 滇池学院, 云南 昆明 650228;

3. 富滇银行股份有限公司博士后科研工作站, 云南 昆明 650200)

摘要: 时空事件连锁发生的现象, 在现实中很常见。为揭示这类现象背后的机制, 开启了影响传播模式挖掘研究。挖掘事件影响传播树是其中一项基础工作, 通常根据事件时空邻近关系, 生成事件邻居集, 用前缀树方法构建事件影响传播树。在时空事件较稠密时, 组合式增长使挖掘时间与空间成本剧增, 无法对大规模数据进行挖掘。为此, 提出一种新方法, 通过构建地理空间实体KD树, 检索事件之间的时空邻近关系。设计一种三层哈希图数据结构存储事件的时空邻近关系, 使事件影响传播树的信息虚拟化, 不创建树实体, 避免组合爆炸和大量树操作, 提高挖掘效率, 降低空间成本。在LSTW时空数据集上的实验结果, 验证了新方法的效果。

关键词: 时空数据挖掘; 事件影响传播树; KD树; 哈希图

中图分类号: TP391.4

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202446

A new method for mining event influence propagation trees from spatio-temporal data

FANG Dianwu¹, WANG Lizhen², ZOU Muquan^{1,3}, DENG Fei¹

1. School of Information Engineering, Kunming University, Kunming 650214, China

2. Dianchi College, Kunming 650228, China

3. Postdoctoral Research Workstation, FuDian Bank, Kunming 650200, China

Abstract: In reality, there was often a phenomenon where spatio-temporal events happened one by one. To uncover the mechanism behind phenomenon of this kind, the research of influence propagation pattern mining was initiated. One of fundamental tasks was to mine event influence propagation trees. A traditional way was used to generate a set of spatio-temporal event neighbors based on spatio-temporal proximity of events, and apply a prefix tree method to construct an event influence propagation tree. Once the spatio-temporal events were dense, the cost of mining time and space would be significantly increased by an explosive growth in the number of combinations, therefore, it would be difficult to mine large-scale data. To this end, a new method was proposed to construct a KD tree of geographic entities and retrieve the spatio-temporal proximity relationship between events. A three-layer Hashmap data structure was designed to store the spatio-temporal proximity relationship between events, virtualizing the information of event influence propagation trees

收稿日期: 2024-08-15; 修回日期: 2024-11-01

通信作者: 王丽珍, lzhwang@ynu.edu.cn

基金项目: 昆明学院引进人才科研项目 (No.XJ20230085); 云南省彩云博士后计划一创新项目 (No.YNBH24024); 云南省地方本科高校基础研究联合专项 (No.202101BA070001-152); 昆明学院重点科研团队项目 (No.XJ20230047); 云南省院士专家工作站项目 (No.202205AF150031)

Foundation Items: Talent Introduction and Research Project of Kunming University (No.XJ20230085), Yunnan Caiyun Postdoctoral Program-Innovation Project (No.YNBH24024), Yunnan Local Undergraduate University Basic Research Joint Special Fund (No.202101BA070001-152), Key Research Team Project of Kunming University (No.XJ20230047), Academician Expert Workstation of Yunnan Province (No.202205AF150031)

without creating entities of trees. Thus combinatorial explosion and a large number of tree operations were avoided, the mining efficiency was improved and spatial costs were cut down. The experimental results on the LSTW spatio-temporal dataset verify the effectiveness and efficiency of new method.

Key words: spatio-temporal data mining, event influence propagation tree, KD tree, Hashmap

0 引言

近年来,疫情扩散等现象频繁出现,给人类社会带来巨大冲击,也对公共治理等领域研究提出了新的要求。为深入研究时空事件影响传播现象背后的机制,人们开启了高影响传播模式挖掘研究,它是空间并置模式挖掘在高影响约束上的拓展。

空间并置模式挖掘研究在近20多年来取得了丰硕成果。基于地理学第一定律^[1],Shekhar等^[2]早于2001年开始了空间并置模式挖掘研究。空间并置模式挖掘关注不同特征的实例在空间中频繁聚集的一类现象,例如,在居民小区(设为特征A)附近,有餐馆(设为特征B)和药店(设为特征C)频繁聚集出现,集合{A,B,C}称为一个空间并置模式。其中,特定位置上的单个特征称为实例,第一个居民小区表示为A.1,第二家餐馆实例表示为B.2,第三家药店表示为C.3,若A.1、B.2和C.3两两邻近,{A.1,B.2,C.3}称为一个团(也称行实例),模式{A,B,C}的所有行实例组成一个表实例。对模式{A,B,C}进行挖掘,先计算模式{A,B,C}的表实例中单个特征(如A)的实例数量在其实例总数中所占比例(即特征参与率),得到最小参与率(即模式参与度),再根据用户给定的频繁参与度阈值,筛选空间频繁并置模式。由于频繁并置模式从两个特征(即二阶模式)向更高阶模式做组合式扩展,计算成本呈指数增长,因此挖掘效率是研究重点。由于模式阶数越高、参与度不变或更小(即满足向下闭合性),频繁高阶模式包含的低阶模式必然频繁,可对模式搜索空间剪枝以提升挖掘效率,Yoo和Wang等^[3-5]提出了无连接的Joinless算法来加速空间并置模式挖掘。Bao等^[6]将一个实例驱动模式与一个邻域驱动模式生成团再转换为两层C-Hash结构,通过创新数据结构,省略逐阶查找行实例的步骤,更有效地计算并置模式参与度,进一步提高挖掘效率。Yang等^[7]提出了一种基于启发式回溯搜索的候选搜索算法,以列计算取代生成模式表实例,挖掘效率相比于Joinless算法,有数倍乃至数量级上的提升。

随着大规模空间数据的涌现,基于分布式计算架构挖掘空间并置模式的方法,包括基于MapReduce或GPU的并行算法,在单机上把数据调配到内嵌的数千个GPU上并行处理^[8-11],在处理可计算的大规模数据上很有优势。

空间并置模式挖掘是一个蓬勃发展的研究热点。Huang等^[12]发现了以聚类方法挖掘空间并置模式的途径;Lu等^[13]研究了动态空间数据库中特征间的因果关系;Wang等^[14]在特征的模糊邻近关系的基础上,应用了模糊c-medoids聚类、模糊层次聚类与模糊密度峰值聚类,挖掘空间频繁并置模式;Zhao等^[15]定义了模式实例分布均匀系数,用来描述模式在空间中的分布特性,挖掘均匀并置模式;Wang等^[16]引入星形参与实例,度量表实例不成团的并置模式的频繁性,提出了极大亚频繁并置模式挖掘,以及基于前缀树的方法和基于划分的方法,在文献[17]中做了细化和拓展研究;Yu等^[18]利用核密度估计以及多边形聚类技术,提出了一种并置模式频繁区域识别的方法;Fang等^[19]使用密度峰值聚类技术对空间实例聚类,在空间实例聚类结果上设计了频繁并置模式提取方法;Bao等^[20]用本体表达领域知识,通过与用户交互筛选有趣模式,设计过滤器精简挖掘结果;Wang等^[21]提出用户交互加概率模型的二次挖掘方法,通过与用户迭代交互加速模式挖掘,基于交互的概率统计结果,筛选示例模式并对模式排序,精简挖掘结果。为区别特征的不同价值,Yang等^[22]为特征引入效用概念,定义了模式效用及模式效用率,提出了高效用并置模式挖掘算法。针对该方法会遗漏由高参与率低效用的特征组成的高效用并置模式的问题,王晓璇等^[23]提出特征效用参与率和确定特征效用权重的方法,以效用度度量高效用并置模式,设计了两个剪枝算法提升挖掘效率。Wang等^[24]将效用概念实例化,定义了内效用率和间效用率,识别特征在并置模式中的全局影响,基于实例效用,挖掘高效用并置模式。

空间并置模式挖掘领域对影响类主题的早期研究,是在考虑道路对其周边环境的噪声影响时,采

用地理信息系统领域缓冲区概念^[25]。Li等^[26]提出了一种基于网格的扩展空间对象的统计学方法,检测污染物排放点与儿童癌症病例之间的可能关联,根据统计检验(而非用户给定全局阈值)来验证候选并置模式的显著性。Bembenik等^[27]比较了扩展空间对象并置模式挖掘的各种方法。陈兰^[28]由小区周边道路设立防噪墙,引申出实例影响的概念,提出高影响并置模式挖掘,以超市和便利店等对消费者的重叠影响,印证了影响模式挖掘的可行性。Lei等^[29]将特征划分为影响特征和受影响特征两类,根据影响随距离衰减的性质,针对工厂污染源对癌症患者有影响的场景,挖掘高影响并置模式。Qian等^[30-31]研究了时空共现在不同区域上的传播模式,把特征间的时间间隔的影响引入兴趣度量,提出了一个带权重的滑动窗口模型来挖掘时空并置模式。Huang等^[32]针对流行病在不同物种间传播的场景,提出了序列模式及相应的挖掘方法。Chen等^[33]从连续一段时间内物体的轨迹中,挖掘到反复出现的共同运动模式。Hu等^[34]针对传统并置模式参与率的定义,不能以相同比例反映两个或更多特征的变化,会遗漏一些有意义的模式,提出了可反映特征间动态关系的动态空间并置模式,基于特征邻近图,挖掘极大动态并置模式。Shekhar等^[35]回顾了时空数据挖掘中最新的计算技术和工具,认为目前绝大多数研究仍然停留在欧几里得空间中,时空网络结构所要求的独特的非对称邻域、邻域关系的方向性,需要有新的时空统计基础和计算方法来支持时空网络数据挖掘。Ma等^[36]发现在星形物化模型中,中心实例通常比非中心实例具有更多的邻居,提出了一个挖掘带主导特征亚频繁并置模式的方法。

鉴于实例间的影响及其传播机制,也是社交网络和流行病传播研究领域的热点,一些研究借鉴了这些领域的研究成果。例如,Shang等^[37]使用层次分析法,构建多维属性的判断矩阵,获得属性的权重,再计算网络中的结点与边的权重,形成加权拓扑网络。然后计算有微博转发和不转发概率的结点的影响。Peng等^[38]提出了一种量化用户在移动社会网络中的社会影响力的机制,通过引入朋友熵和交互频率熵,计算用户在网络中的直接和间接影响,提出了一个算法来表征用户社会影响的动态传播过程。Li等^[39]根据与位置相关的属性构建层次树,用泰森图分区方法,将位置分散到不同段中,

然后应用四色映射定理对多边形进行着色,方便移动用户快速选择虚拟位置来保护隐私。Citraro等^[40]引入了一种自底向上的低复杂度方法,通过优化结构和属性同源聚类标准,识别网络中隐藏的中尺度拓扑。Subbian等^[41]认识到社交网络中的信息流趋势和影响者变得日益重要,提出了一个挖掘信息流模式的算法,确定网络中的关键影响者。Feng等^[42]应用属性感知分析,查找与指定区域在多维属性上最相似的其他区域。

受疫情扩散^[43-44]等应用场景的驱动,在上述研究成果的基础上,旨在从时空数据中挖掘既在空间上并置又在时间上共现的高影响传播模式被提出。时空事件的影响传播表现为树状传播,因此,挖掘事件影响传播树是高影响传播模式挖掘过程中的一项基础工作。

然而,常用的方法是根据事件两两时空邻近的关系,组合生成时空事件邻居集,再用前缀树方法构建事件影响传播树,这样能处理的数据规模有限。为解决该问题,本文提出一种从大规模时空数据中挖掘事件影响传播树的新方法。本文创新点如下。

(1) 通过构建地理空间实体KD树(K-dimensional tree),快速搜索天气源事件,建立以天气源事件为根结点的交通事件时空邻居及其交通事件队列,检索事件之间的时空邻近关系。

(2) 设计一种三层哈希图数据结构,存储检索事件的时空邻近关系,把事件影响传播树虚拟化,不再创建事件影响传播实体树,避免组合爆炸与大量树操作,从而提高挖掘效率,降低空间成本。

(3) 在开源数据集上的实验,验证了新方法的效果。

1 场景描述与相关定义

1.1 场景描述

为描述时空事件影响传播过程,给出一个三维时空场景,如图1所示。其中,不同类型的21个时空点事件,按时间先后次序,沿着给定的时空阈值内的时空点传播影响,时空事件的位置及出现时刻以三元组 $\langle x, y, t \rangle$ 表示。

1.2 相关定义

(1) 地理空间实体

地理空间实体(GE, geospatial entity) e ,是指特定地理空间区域内,由一个或多个离散空间点表征的对象,如位于某个固定位置上的气象站,或

拥有一组打卡点或监控点的道路。

基于美国 LSTW 时空数据集 2016 年 8 月至 2020 年 12 月的天气与交通数据, 关注两类地理空间实体, 一类是播报天气事件的气象站, 一类是发生交通事件的道路。前者用编码、GPS 经纬度坐标等表示, 后者用名称、道路方向 (左侧或右侧)、城市名、郡县名、州名等表示。

例如, 在数据集中, 气象站编码 KAEI, 纬度 43.6822, 经度 -93.3722, 在 MN 州 Freeborn 郡 Albert Lea 市, 邮编 56007; 道路名 Sonoma Hwy, 编号 4453, 有系列交通时空事件位置坐标 <38.254936, -122.340012; ...>, 右侧, 在 CA 州 Napa 郡 Napa 市, 邮编 94559。

图 1 场景不涉及地理空间实体。

假定事件传播的时间阈值 $T=d=1$, 其中 d 为距离阈值, 根据图 1 的时空事件数据, 可得到事件影响传播树 $T_1 \sim T_4$ 的森林, 如图 2 所示。

本文只讨论事件影响传播树挖掘, 后续的影响传播模式挖掘另文阐述。

(2) 时空事件

时空事件 (ST_e , 简称事件) o , 指在特定的时段在某个地理空间位置上出现的事物。有两类事件与前述地理空间实体相对应: 一类是由某个气象站

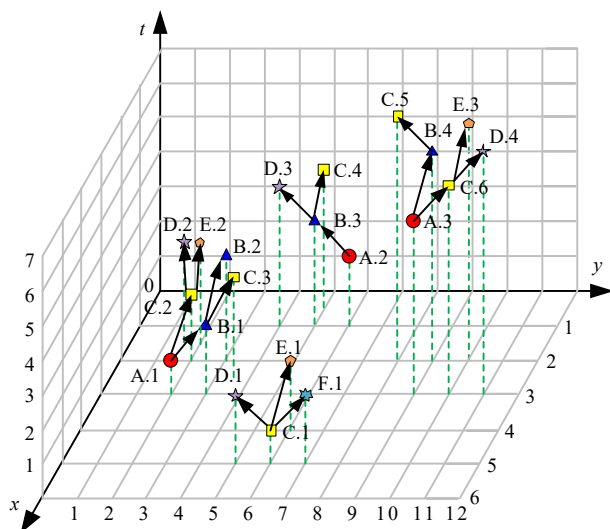
播报的, 当地发生于特定时段内的天气事件, 如下雪; 另一类是某条道路上发生于特定时段内的交通事件, 如拥堵。每个事件由一个四元组 $\langle f\text{-number}, o\text{-start}, o\text{-end}, location \rangle$ 表示, 其中, $f\text{-number}$ 指事件 o 所属特征 f 及其编号, $f \in F$, F 为时空特征集, $o\text{-start}$ 与 $o\text{-end}$ 指事件开始与结束时刻, $location$ 指事件位置。针对天气事件, 特征代表严寒、雾、冰雹等天气类型, 位置指播报天气事件的当地气象站位置; 针对交通事件, 特征代表事故、撞车、拥堵等交通类型, 位置指发生交通事件的 GPS 经纬度坐标、所处道路及其方向、距其最近的气象站编号。

例如, 在数据集中, 天气事件 $\langle F.1, 2016/8/1\ 00:16:00, 2016/8/1\ 00:56:00, K04V \rangle$ 指气象站 K04V 播报的本地该时段出现的第一场小雨; 交通事件 $\langle U.4, 2016/8/1\ 00:08:00, 2016/8/1\ 00:19:44, 37.105865, -121.613411, S\ Valley\ Fwy\ S, R, KE16 \rangle$ 指在该时该地发生于 S Valley Fwy S 道路右侧的第 4 个严重拥堵事件, 最近的气象站编号为 KE16。

(3) 时空邻近

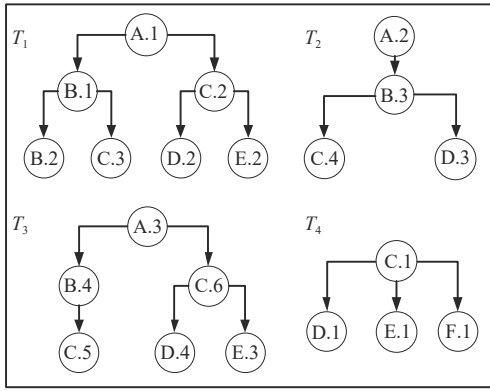
给定事件 $o_i, o_j (i \neq j)$, 若它们在空间上并置且在时间上共现, 则它们之间存在时空邻近关系, 记作 $ST_p(o_i, o_j)$ 。

事件 o_i, o_j 在空间上的并置分两种情况: 一种



序号	时空点	(x, y, t)
1.	A.1	(3, 2, 1)
2.	A.2	(1, 6, 2)
3.	A.3	(3, 9, 5)
4.	B.1	(3, 3, 2)
5.	B.2	(2, 3, 3)
6.	B.3	(1, 5, 3)
7.	B.4	(2, 9, 6)
8.	C.1	(5, 6, 1)
9.	C.2	(2, 2, 2)
10.	C.3	(2.5, 3.5, 3)
11.	C.4	(0.5, 5, 4)
12.	C.5	(2, 8, 7)
13.	C.6	(3, 10, 6)
14.	D.1	(5, 5, 2)
15.	D.2	(1.5, 1.5, 3)
16.	D.3	(1, 4, 4)
17.	D.4	(3, 11, 7)
18.	E.1	(4, 6, 2)
19.	E.2	(1.5, 2, 3)
20.	E.3	(2, 10, 7)
21.	F.1	(5, 7, 2)

图 1 时空事件影响传播的场景及其时空点

图2 由事件影响传播树 $T_1 \sim T_4$ 组成的森林

是天气事件与交通事件在空间上的并置，指离该交通事件最近的气象站播报该天气事件的情况，另一种是两个交通事件在空间上的并置，指的是这两个事件属于同一条同侧道路，除经纬度坐标之外的位置属性相同，且它们之间的哈弗林距离（注：球面上两点间的距离）不大于用户给定的距离阈值 d ，即 $\text{Haversine_distance}(o_i, o_j) \leq d$ 。事件 o_i, o_j 在时间上共现，指它们开始时间的间隔不大于给定的时间阈值 T ，即 $0 \leq |o_i.\text{start} - o_j.\text{start}| \leq T$ 。

由于天气事件成因复杂，它们之间的时空邻近情景不在此考虑之列。在图1中，假定时空阈值为 $\langle 1, 1 \rangle$ ，则 A.1 与 B.1 时空邻近。

(4) 事件的时空邻居集

给定 k 个事件集 $O = \{o_1, \dots, o_k\}$, $k \geq 2$ 、时空特征集 $F = \{f_1, \dots, f_m\}$, $m \geq 2$ 、事件 o_i 及其特征 $f_i \in F$ 。由先前出现的源事件 o_i 与其时空邻居组成的事件集，记作 $\text{ST}_p\text{N}(o_i)$ ，按式 (1) 计算。

$$\text{ST}_p\text{N}(o_i) = \{o_j \in O | o_j = o_i \vee (o_j.f_{\text{alphabet_order}} \wedge (o_i.\text{start} > o_j.\text{start}) \wedge \text{ST}_p(o_i, o_j))\},$$

$$1 \leq i, j \leq k, o_j.f_{\text{alphabet_order}} \in F \quad (1)$$

其中， $f_{\text{alphabet_order}}$ 表示提取时空邻居事件所属特征并按其名字的字典序。

例如，图2给出了不同事件的时空邻居集， $\text{ST}_p\text{N}(A.1) = \{A.1, B.1, C.2\}$ ， $\text{ST}_p\text{N}(B.1) = \{B.1, B.2, C.3\}$ 。

(5) 两个事件之间的影响

给定事件 o_i, o_j ($i \neq j$)， o_i 为天气事件（或交通事件）， o_j 为交通事件，假定有 $\text{ST}_p(o_i, o_j)$ ，且 $o_i.\text{start} > o_j.\text{start}$ ，可认为事件 o_i 的出现在一定程度上促成了事件 o_j 的发生，或事件 o_i 影响了事件 o_j ，记作 $\text{ST}_e\text{I}(o_i, o_j)$ ，以有向线段表示。

因此，一个天气事件可能对一个或多个交通事

件有影响，一个交通事件也可能对同一条道路上在其后发生的交通事件有影响，然而，不考虑天气事件之间的影响，以及交通事件对天气事件的影响。

例如，图1中，有向线段 $e(A.1, B.1)$ 表示事件 A.1 对 B.1 有影响。

(6) 事件的影响传播树

给定 k 个事件集 $O = \{o_1, \dots, o_k\}$, $k \geq 2$ 、时空特征集 $F = \{f_1, \dots, f_m\}$, $m \geq 2$ 、源事件 $o_r \in O$ 及其特征 $f_{o_r} \in F$ 。假定以 o_r 为根生成一棵树 $T(o_r) = (O, E)$ ， O 是顶点的集合， $E = \{(o_i, o_j) | o_i = o_r \vee (o_i, o_j \in O \wedge \text{ST}_e\text{I}(o_i, o_j)), i \neq j\}$ 是边的集合，标记函数 $l: O \rightarrow F$ ，将 O 中的每个事件映射到特征集 F 的某个特征上，则 $T(o_r)$ 称为事件 o_r 的一棵影响传播树，记作 $\text{ST}_e\text{T}(o_r)$ 。由于事件的时空邻居集，采用有序排列的形式， $\text{ST}_e\text{T}(o_r)$ 是一棵有序树。

例如，图2给出了以 A.1、A.2、A.3、C.1 为根的4棵影响传播树 $T_1 \sim T_4$ 。

2 事件影响传播树的挖掘算法

2.1 事件影响传播树的挖掘流程与基准算法

事件影响传播树的挖掘流程如图3所示，挖掘内容对应算法1的步骤如下。

遍历天气事件数据集，读取天气事件及其气象站信息，把天气事件添加到其气象站的时序队列中，再把气象站添加到气象站集 W 中（步骤1~3）。

根据气象站集 W ，生成带天气事件序列的气象站 KD 树（步骤4）。

遍历交通事件数据集，若离交通事件 o' 最近的气象站 w_j 在气象站集 W 中存在，则从该气象站的天气事件序列中，查找在交通事件 o' 之前最近发生的天气源事件 o_c （步骤5、6）。

否则，用气象站 KD 树找到离交通事件 o' 最近的、在距离阈值 d 内的气象站，再从该气象站的天气事件序列中，查找在交通事件 o' 之前最近发生的天气源事件 o_c （步骤7）。

在天气源事件 o_c 不为空时，把交通事件 o' 添加到天气事件的时空邻居集 N_w 中 o_c 的时空邻居集里（步骤8）。若交通事件 o' 所在道路 r 上的交通事件序列 Q_r 存在，则把交通事件 o' 添加到 Q_r 中（步骤9、10）；否则，创建交通事件 o' 所在道路 r 上的交通事件序列 Q_r 并把 o' 添加到 Q_r 中（步骤11）。

根据交通事件序列 Q_r 和时空阈值 $\langle T, d \rangle$ ，生成

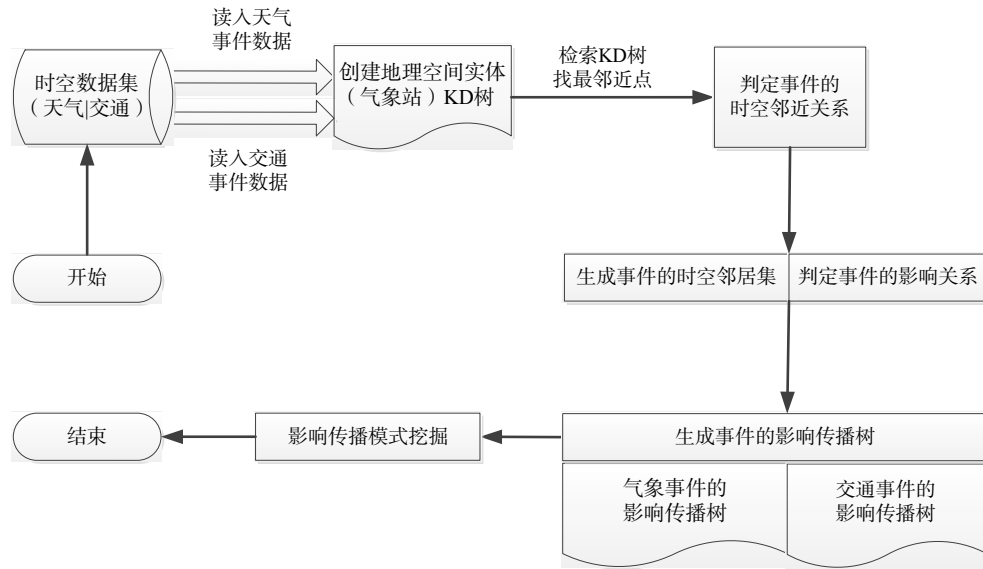


图3 事件影响传播树的挖掘流程

算法1：事件影响传播树的挖掘算法

输入: 天气事件集 $O = \{o_1, o_2, \dots, o_n\}$ 、气象站集 $W = \{w_1, w_2, \dots, w_u\}$ 、交通事件集 $O' = \{o'_1, o'_2, \dots, o'_n\}$ 、时空邻近关系阈值 $\langle T, d \rangle$ 、特征集 $F = \{f_1, f_2, \dots, f_m\}$;

输出: 所有满足 $\langle T, d \rangle$ 的事件影响传播树集合 T_{events} ;

变量: Q_w/Q_t 指天气事件/交通事件的序列集, N_w/N_t 指天气事件/交通事件的时空邻居集。

- 1: 对于 O 中的每一个天气事件 o , 执行
- 2: 把天气事件 o 添加到气象站 w_i 的队列 Q_{w_i} 中;
- 3: 把气象站 w_i 添加到气象站集合 W 中;
- 4: $\text{KDtree}_{\text{wq}} = \text{gen_KDtree_for_weather_stations}(W)$; //生成气象站集合 W 的KD树
- 5: 对于交通事件集 O' 中的每一个交通事件 o' , 执行
- 6: 若距离 o' 最近的气象站 w_j 在 W 中存在, 则在 w_j 的序列集 Q_{w_j} 中搜索最近天气源事件 o_c ;
- 7: 否则 $o_c = \text{KDtree}_{\text{wq_search}}(o', d)$;
- 8: 若 o_c 不为空, 则把交通事件 o' 添加到 N_w 中的天气事件 o_c 的时空邻居;
- 9: 若交通事件 o' 所在的道路 r 的交通事件队列 Q_{t_r} 存在, 则执行
- 10: 把交通事件 o' 添加到队列 Q_{t_r} 中;
- 11: 否则为道路 r 生成一个交通事件队列 Q_{t_r} , 并把交通事件 o' 添加到 Q_{t_r} 中;
- 12: $N_t = \text{gen_STneighbors_for_traffic_events}(Q_t, \langle T, d \rangle)$;
- 13: $T_{\text{events}} = \text{gen_prefixTrees_for_events}(N_w, N_t)$;
- 14: return T_{events} //返回事件影响传播树的集合

交通事件的时空邻居集 N_t (步骤12)。

基于天气事件的时空邻居集 N_w 和交通事件的时空邻居集 N_t , 采用文献[5]创建前缀树的方式, 生成并返回所有事件影响传播树的集合 T_{events} (步骤13、14)。

2.2 三层哈希图结构

基于事件时空邻居集生成前缀树, 再做水平或垂直编码^[45-47]创建事件影响传播树的方法, 占用空间较多, 时间开销大。为此, 提出一个带优化策略的改进算法, 以减少空间占用并提高效率, 即采用三层哈希图结构 $\text{hashmap}\langle \text{Character}, \text{hashmap}\langle \text{String}, \text{hashmap}\langle \text{Character}, \text{ArrayList}\langle \text{String} \rangle \rangle \rangle \rangle$ 来存储和检索天气和交通事件的时空邻居。

事件影响传播树的挖掘 (event influence propagation tree mining, EIPTM) 算法如算法1所示。

通过调用事件之间的时空邻近关系, 可获取以特定事件为根结点的、虚拟的影响传播树的信息, 而无须构建影响传播树的实体树。

针对图1的示例数据, 给出该数据结构的对应表示, 如图4所示。

三层哈希图结构可实现对任意事件时空邻居的快速检索, 多重调用可获得特定根结点事件的影响传播子树下指定特征的事件数, 无须像传统方法那样只有对事件影响传播树做水平或垂直编码, 才可获得这些指定特征的事件数。这有利于后续的影响传播模式挖掘, 能够减少树检索次数, 降低空间与时间耗费。

2.3 算法的时间复杂度分析

基于EIPTM对事件影响传播树挖掘的时间耗费 T 进行分析。为叙述方便, 令 S 为输入数据源, 事件总数 $n = n_1 + n_2$, n_1 为天气事件总数量, n_2 为交通事件总数量。

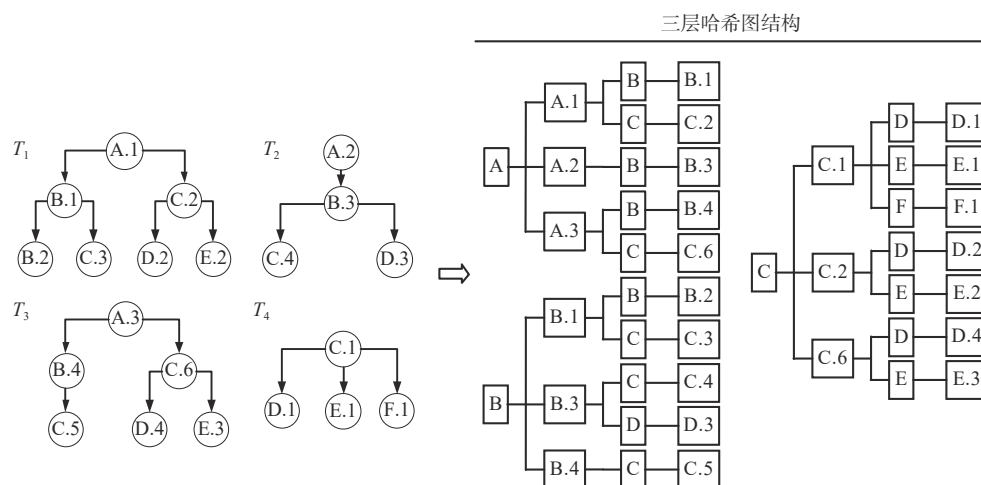


图4 以三层哈希图结构存储事件时空邻居集

首先，遍历天气事件集，把每个天气事件添加到其气象站的时序队列，把该气象站添加到气象站集中，时间开销为 $O(n_1)$ 。

其次，根据气象站位置创建气象站KD树，时间开销为 $O(h \log_2 h)$ ，其中 h 为气象站的总数量。

再次，遍历交通事件集。其中，若离交通事件最近的气象站在气象站集中存在，则从气象站的天气事件序列中找到最近发生的天气事件最多需要 $O(\log_2 n_1)$ 时间；否则用气象站KD树搜索离交通事件最近的气象站需要 $O(\sqrt{h})$ 时间，再从气象站的天气事件序列中找到最近发生的天气事件需 $O(\log_2 n_1)$ 时间；若天气事件非空，把交通事件添加到其时空邻居集中需 $O(1)$ 时间。

此外，把交通事件添加到其所在道路的交通事件序列，或创建其所在道路的交通事件序列并把交通事件插入该序列最多需要 $O(\log_2 n_2)$ 时间。

然后，根据所有道路上的交通事件队列，创建交通事件的时空邻居集，扣除交通事件时空邻居集的重叠情况，时间开销为 $O(n_2)$ 。

最后，根据天气及交通事件的时空邻居集，构建所有事件的影响传播树，这些树最多包含全部事件的线性倍数的事件，时间开销为 $O(n)$ 。

综上所述，生成事件影响传播树的时间耗费，可由式(2)得出。

$$T_{\text{gen_event_trees}}(S) = O(n_1) + O(h \log_2 h) + O(n_2)O(\sqrt{h}) + O(\log_2 n_1) + O(\log_2 n_2) + O(n_2) + O(n) \quad (2)$$

在事件影响传播场景中，因 $n_1, n_2 < n$, $h \ll n$ ，事件影响传播树的挖掘算法的时间耗费 T 为 $O(n)$ 。不同时空数据集的事件分布和稠密性有所不同，算法

提升效率的效果可能会有差异。

3 实验评估

3.1 数据集描述

LSTW 开源数据集是近年来时空影响传播研究中常用的一个大型时空数据集，涵盖美国 50 个州 2016 年 8 月至 2020 年 12 月的 560 万个天气事件和 3 140 万个交通事件。

天气事件的类型包括下雨、下雪、冰雹、暴风雨、严寒 (-23.7°C 及以下)、雾霾等，交通事件类型包括事故、车辆损毁、拥堵、施工、社会事件 (如马拉松赛、钢琴会、展览)、封路、流量事件 (如信号灯故障、动物在公路上) 等。

天气事件数据集的信息包括事件 ID、事件类型 (如下雨)、程度指标 (从轻微到严重)、事件开始时间、事件结束时间、事件所在美国时区、GPS 纬度、GPS 经度、播报事件的气象站编号、城市名、郡县名、州名、邮政编码。事件开始时间和事件结束时间采用协调世界时 (universal time coordinated, UTC) 记录。

交通事件数据集的信息包括：事件 ID、事件类型 (如拥堵)、程度指标 (0~4)、交通信息编码 (traffic message channel, TMC)、事件描述、事件开始时间、事件结束时间、事件所在美国时区、GPS 纬度、GPS 经度、事件在道路上影响波及的长度 (采用英里记录, 1 mi=1.61 km)、离事件最近的气象站编号、街道号、街道名、道路方向 (左侧/右侧)、城市名、郡县名、州名、邮政编码。

原始数据集中的数据倾斜较为严重，例如拥堵

事件在交通事件总数中占80.18%，下雨、起雾在天气事件总数中分别占63.54%、20.87%。为缓解数据倾斜现象以挖掘更精细的影响传播模式，在研究中对原数据集中的时空事件所属特征做了更多细化，例如把交通事故特征按程度不同划分为3个特征，把车流异常特征按不同等级划分为5个特征。由于区分特征主要用于影响传播模式挖掘，本文只讨论事件影响传播树挖掘，在此不再赘述。

鉴于原始真实数据集规模较大，为了在现有计算资源上快速获得实验评估结果，针对不同实验目的，对原始真实数据集做了随机截取，整理了10个数据集，在其上做了实验，来验证新算法的效果。这些数据集摘要信息，如表1所示。

表1 采用的时空数据集摘要信息

数据集名称	天气事件实例数	特征数	数据集名称	交通事件实例数	特征数
Real-1	100 000	36	Real-6	100 000	36
Real-2	200 000	36	Real-7	200 000	36
Real-3	300 000	36	Real-8	300 000	36
Real-4	400 000	36	Real-9	400 000	36
Real-5	500 000	36	Real-10	500 000	36

3.2 实验环境

使用Java语言和IntelliJ IDEA 2024.1(Community Edition)编程实现，在Intel Core(TM) i7-10510U四核CPU @ 1.80 GHz 2.30 GHz、16.0 GB RAM、Windows 10 Pro (64位)的计算机上运行实验。

3.3 与高效算法的挖掘效果对比

为衡量EIPTM算法的树结构挖掘的效果，参阅了从空间或时空对象中挖掘树结构的高效算法，例如先挖掘序列再从中挖掘频繁树的SLEUTH算法^[45-48]，先挖掘有序邻居集，再从中挖掘并置模式实例树的CPI-tree算法^[5]及其先导的FP-CM算法^[49]与扩展的iCPI-tree算法^[8-9,50]。鉴于iCPI-tree算法对CPI-tree算法扩展是为了更快挖掘空间并置模式，在物化空间邻居集生成树结构方面，两者的树挖掘效率相同，因此选择CPI-tree、SLEUTH算法做对比。

为适应CPI-tree算法对特征差异化的要求，需要把一个时空事件划分到某一类型的特征下，例如A.8，表示数据集中第8个严重寒冷事件。根据事件类型的不同情形，创建对应特征，对事件类型所做的特征细化如表2所示。

表2 对事件类型所做的特征细化

事件类型	程度	对应特征	事件类型	程度	对应特征	事件类型	程度	对应特征
寒冷	严重	A	事故	1	M	建造	4	?
雾霭	中等	B	事故	2	N	群体事件	1	~
雾霭	严重	C	事故	3	O	群体事件	2	{
冰雹	其他	D	事故	4	P	群体事件	3	!
降雨	未知	E	撞车	2	Q	车流异常	0	@
小雨	轻微	F	撞车	3	R	车流异常	1	#
中雨	中等	G	拥堵	0	S	车流异常	2	\$
大雨	严重	H	拥堵	1	T	车流异常	3	^
小雪	轻微	I	拥堵	2	V	车流异常	4	*
中雪	中等	J	拥堵	3	Y	匝道拥堵	2	aa
大雪	严重	K	建造	2	Z	匝道拥堵	3	bb
风暴	严重	L	建造	3	&	匝道拥堵	4	cc

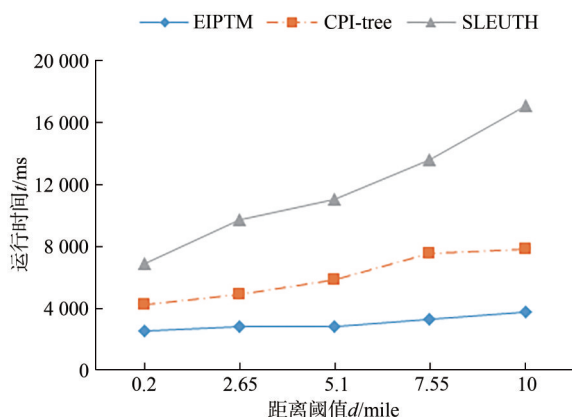
对比实验采用数据集Real-1和Real-6组合而成的组合数据集1，有20万个实例。

(1) 距离阈值 d 对算法效率的影响

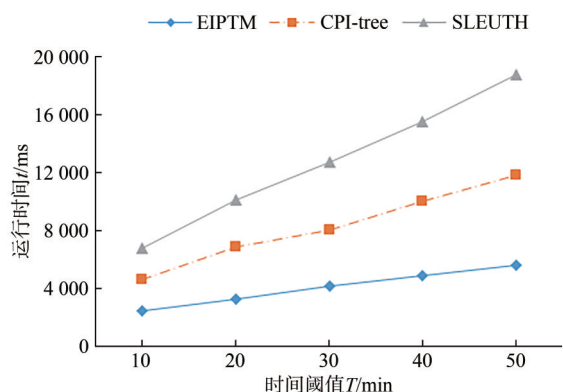
基于组合数据集1，设时间阈值 $T=10$ min保持不变，距离阈值 d 对算法效率的影响如图5所示，当 d 的值在0.2~10 mi以等差方式递增时，EIPTM算法在2 521~3 769 ms范围缓慢递增，CPI-tree算法在4 231~7 808 ms较缓提升，SLEUTH算法在6 853~17 148 ms范围大幅度提升，随着距离阈值的增加，事件之间的时空邻近关系呈非线性单调不递减。通常，EIPTM算法和CPI-tree算法只需扫描一遍数据集，SLEUTH算法需要多次扫描数据集。在计算时空邻居对方面，EIPTM算法充分利用数据集特点，天气事件找在其发生时段的交通事件，交通事件按同侧道路排成队列，找时空邻居对无冗余，计算量最少；CPI-tree、SLEUTH算法根据时空阈值查找时空邻居，算法不是针对数据集特点设计的，产生较多冗余的时空邻居对。在创建树结构方面，EIPTM算法以三层哈希图结构存储时空邻居，不创建树结构，索引路径即树结构，同样计算量最小；CPI-tree算法以事件所属特征的字母序及实例编号大小形成的次序，创建有序树；SLEUTH算法在时空邻居对基础上，先要创建事件序列，再根据水平/垂直编码，以前缀树方式创建无序树。不创建树快于创建有序树，有序树计算量小于无序树，故EIPTM算法最快，CPI-tree算法次之，SLEUTH算法最慢。

(2) 时间阈值 T 对模式数量的影响

基于组合数据集1，设距离阈值 $d=0.2$ mi保持

图5 距离阈值 d 对算法效率的影响

不变, 时间阈值 T 对算法效率的影响如图6所示, 当 T 的值在10~50 min以等差方式递增时, EIPTM算法在2 415~5 639 ms范围缓慢递增, CPI-tree算法在4 636~11 814 ms范围较缓提升, SLEUTH算法在6 723~18 749 ms范围大幅度提升, 随着时间阈值的增加, 事件之间的时空邻近关系也呈非线性单调不递减, 即在空间阈值范围内的更长时间跨度内的事件将被包括进来, 事件时空邻近关系增多。

图6 时间阈值 T 对算法效率的影响

基于前述原因, EIPTM算法计算量最少, 用时最少, 效率最高; CPI-tree算法次之, SLEUTH算法效率排在最后。

从图5、图6可以看出, 时间阈值 T 对3个算法的效率影响比距离阈值 d 更大一些, 由此可见, LSTW数据集在天气事件和交通事件的时间分布上, 比在空间分布上更为密集。

3.4 挖掘效率评估

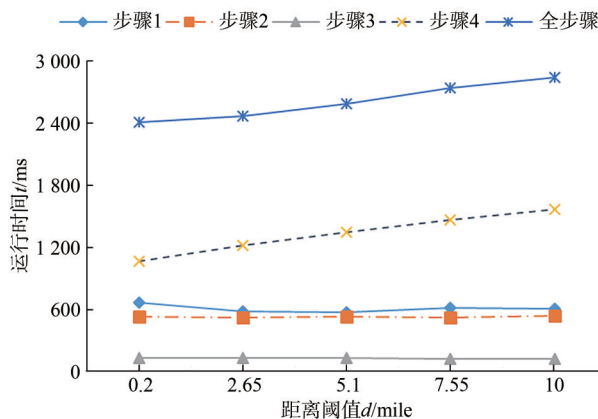
以下基于Real-1~Real-10数据集, 根据距离阈值 d 、时间阈值 T 变化, 评估EIPTM算法的挖掘效率。

3.4.1 在Real-1和Real-6数据集上的挖掘效率评估

在此对组成EIPTM算法各挖掘步骤的效率情况做出评估。其中, 步骤1是指创建一棵气象站KD树并把天气事件插入气象站天气事件序列的时间, 步骤2是指得到街道集、交通事件列表、交通事件特征集的时间, 步骤3是指完成天气事件时空邻居集的时间, 步骤4是指完成交通事件时空邻居集的时间, 全步骤是指EIPTM算法挖掘过程花费的全部时间。

(1) 距离阈值 d 对各挖掘步骤效率的影响

设时间阈值 $T=10$ min, 保持不变, 距离阈值 d 对各挖掘步骤效率的影响如图7所示。当距离阈值 d 在0.2~10 mi范围递增时, 步骤1在566~669 ms范围波动, 步骤2在528~542 ms范围波动, 步骤3在126~136 ms范围波动。步骤1构建气象站KD树, 步骤2挖掘街道集及其交通事件集, 步骤3挖掘天气事件时空邻居集, 该时空邻居集主要与搜索到的气象站相关, 不受距离阈值变化的影响。EIPTM算法挖掘时间主要受步骤4花费时间的影响, 该步骤挖掘交通事件时空邻居集的时间在1 073~1 566 ms范围持续增加, 这是由于随着距离阈值 d 的增加, 越来越多的交通事件之间具有时空邻近关系, 交通事件时空邻居集增大导致时间成本上升。

图7 距离阈值 d 对各挖掘步骤效率的影响

(2) 时间阈值 T 对各挖掘步骤效率的影响

设距离阈值 $d=0.2$ mi保持不变, 时间阈值 T 对各挖掘步骤效率的影响如图8所示, 当时间阈值 T 在10~50 min范围递增时, 步骤1在566~669 ms范围波动, 步骤2在518~543 ms范围波动, 步骤3在125~136 ms范围波动。EIPTM算法挖掘时间主要受步骤4所需时间的影响, 该步骤挖掘交通事件时

空邻居集的时间在1 073~4 378 ms范围持续增加,这是由于随着时间阈值 T 的增加,越来越多的交通事件之间具有时空邻近关系,交通事件时空邻居集增大导致时间成本上升。

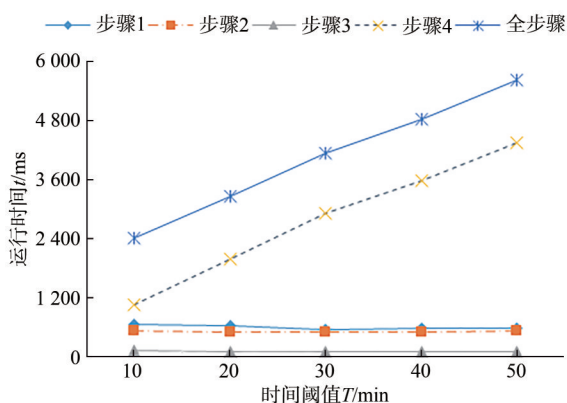


图8 时间阈值 T 对各挖掘步骤效率的影响

因此,从步骤4挖掘交通事件时空邻居集的时间,在时间阈值 T 不变、距离阈值 d 在1 073~1 566 ms范围持续增加,以及在距离阈值 d 不变、时间阈值 T 在1 073~4 378 ms范围持续增加,可以看到,EIPTM算法挖掘时间主要受步骤4的影响,在特定数据集上,交通事件时空邻居集的挖掘受时间阈值 T 的影响,比受距离阈值 d 的影响更大。

3.4.2 在Real-1~Real-10数据集上的挖掘效率评估

本文对数据集的实例规模予以扩展,扩大天气事件与交通事件的总体规模,从而对EIPTM算法的挖掘效率进行评估。

组合数据集1有20万个实例,是数据集Real-1和Real-6的组合;组合数据集2有40万个实例,是数据集Real-2和Real-7的组合;组合数据集3有60万个实例,是数据集Real-3和Real-8的组合;组合数据集4有80万个实例,是数据集Real-4和Real-9的组合;组合数据集5有100万个实例,是数据集Real-5和Real-10的组合。

图9、图10的纵坐标运行时间指EIPTM算法挖掘全过程的花费时间。

(1) 距离阈值 d 对EIPTM算法挖掘效率的影响

设时间阈值 $T=10$ min保持不变,距离阈值 d 对EIPTM算法挖掘效率的影响如图9所示,当距离阈值 d 在0.2~10 mi范围递增时,EIPTM算法在组合数据集1上的挖掘时间在2.42~2.84 s范围持续递增,在组合数据集2上的挖掘时间在4.53~5.63 s范围先升后降,在组合数据集4上有时间成本最大值

5.77 s。类似地,在组合数据集3上的挖掘时间在8.06~10.2 s范围先升后降,在组合数据集4上有时间成本最大值10.3 s。在组合数据集4上的挖掘时间在11.85~14.33 s范围持续递增,在组合数据集5上的挖掘时间在15.03~19.18 s范围持续递增。

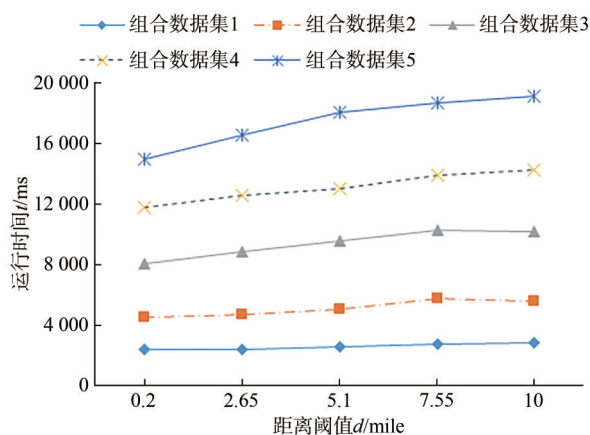
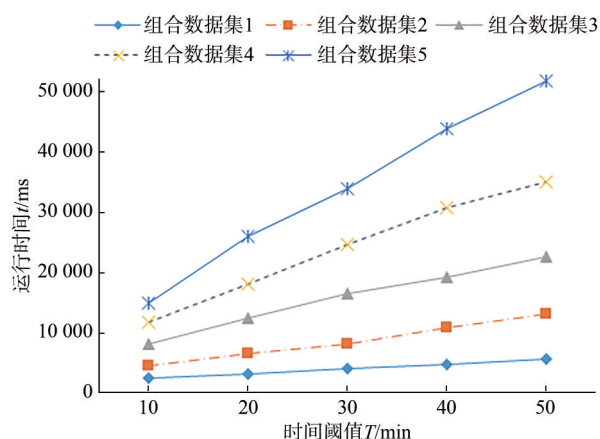


图9 距离阈值 d 对EIPTM算法挖掘效率的影响

这说明,EIPTM算法的挖掘效率受距离阈值的影响,随着距离阈值的增加,越来越多的交通事件之间形成时空邻近关系,交通事件时空邻居集增大导致时间成本上升。然而,交通事件之间的时空邻近关系,是沿着道路延伸的时空邻近关系,而不是空间离散点之间因时空距离而形成的邻近关系。显然,在组合数据集4上依道路形成的交通事件时空邻近关系是最多的。

(2) 时间阈值 T 对EIPTM算法挖掘效率的影响

设距离阈值 $d=0.2$ mi保持不变,时间阈值 T 对EIPTM算法挖掘效率的影响如图10显示,当时间阈值 T 在10~50 min递增时,EIPTM算法在所有的组合数据集上都呈现持续上升趋势,上升增幅比图9距离阈值影响算法挖掘效率的程度更大。其中,在组合数据集1上,挖掘时间为2.42~5.64 s;在组合数据集2上,挖掘时间为4.53~13.18 s;在组合数据集3上,挖掘时间为8.06~22.58 s;在组合数据集4上,挖掘时间为11.85~35.13 s;在组合数据集5上,挖掘时间为15.03~51.87 s。没有出现之前随距离阈值增大而升中有降的情况,这是因为时间阈值不受交通道路线路的影响,随着时间阈值 T 的增加,越来越多的交通事件之间形成时空邻近关系,交通事件时空邻居集增大导致时间成本上升。显然,EIPTM算法挖掘效率受时间阈值 T 的影响,比受距离阈值 d 的影响更大。

图10 时间阈值 T 对 EIPTM 算法挖掘效率的影响

4 结束语

在研究时空事件影响传播模式挖掘的过程中，笔者发现现有的挖掘事件影响传播树的方法没有充分地利用时空事件之间的内在联系，树操作较多，空间与时间耗费大，难以满足在大规模时空数据集上做事件影响传播树挖掘的需要。为解决该问题，本文提出了一种新方法，针对天气交通数据集的特点，构建气象站KD树，快速搜索气象站查找天气源事件，建立天气事件影响交通事件、交通事件影响后继交通事件的影响队列，确定事件之间的时空邻近关系。本文设计了一种三层哈希图结构存储事件时空邻近关系，把事件影响传播树的信息隐含其中，无须构建事件影响传播树实体，减少了计算量，提高了挖掘效率。实验验证了新方法的效果。

本文是基于LSTW数据集进行的，研究方法和优化策略具有特异性，在与LSTW数据集类似的交通网络应用场景中，可拥有较高的数据挖掘效率。然而，这些特异性是与LSTW数据集紧密联系的，这使得新方法和优化策略在泛化应用上存在局限。主要体现在：一是LSTW数据集没有天气事件的风向、风速等数据，天气事件在时空邻域内影响交通事件发生是概然事件，有些交通事件可能是其他因素促成的，不一定是天气事件导致的；二是LSTW数据集没有交叉路口的道路转向信息，交通事件的时空队列只能是在一条单向道路上、先后发生的交通事件序列。因此，针对当下的具有更高分辨率的车道级交通数据以及带交叉路口信息的导航系统，新方法不适用，需要基于新的时空数据集进行拓展深入研究，例如需要考虑一条道路的不同车道拥堵

经交叉口转向带来的拥堵扩散影响的效果。

随着人工智能技术突飞猛进地发展，越来越多的学科领域开始引入人工智能深化研究与应用。在影响传播的研究中，出现了运用智能体对影响传播过程进行建模^[51]、针对城市交通道路网络进行动态短时推理与精准预测^[52]、融合增强的社交媒体数据感知交通态势^[53]、考虑时空关联的深度学习短时交通流预测感知^[54]等方面的研究成果，这些成果展现了新的视角或结论，对影响传播模式挖掘研究有极大的促进作用。后续研究考虑结合图神经网络技术，针对不同天气事件或交通事件发生背景下的、更精细的车道级城市交通网络进行学习，预测或推理时空事件高影响传播模式，探索相应的优化管控方法，为交通治理等决策提供参考。

参考文献：

- [1] TOBLER W. On the first law of geography: a reply[J]. Annals of the Association of American Geographers, 2004, 94(2): 304-310.
- [2] SHEKHAR S, HUANG Y. Discovering spatial co-location patterns: a summary of results[C]//Proceedings of the 7th International Symposium. Berlin: Springer, 2001: 236-256.
- [3] YOO J S, SHEKHAR S, CELIK M. A join-less approach for co-location pattern mining: a summary of results[C]//Proceedings of the 5th IEEE International Conference on Data Mining (ICDM'05). Piscataway: IEEE Press, 2006: 813-816.
- [4] YOO J S, SHEKHAR S. A joinless approach for mining spatial colocation patterns[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(10): 1323-1337.
- [5] WANG L Z, BAO Y Z, LU J, et al. A new join-less approach for co-location pattern mining[C]//Proceedings of the 2008 8th IEEE International Conference on Computer and Information Technology. Piscataway: IEEE Press, 2008: 197-202.
- [6] BAO X G, WANG L Z. A clique-based approach for co-location pattern mining[J]. Information Sciences, 2019, 490: 244-264.
- [7] YANG P Z, WANG L Z, WANG X X, et al. SCPM-CR: a novel method for spatial co-location pattern mining with coupling relation consideration[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(12): 5979-5992.
- [8] ANDRZEJEWSKI W, BOINSKI P. A parallel algorithm for building iCPI-trees[C]//Proceedings of the 18th East European Conference on Advances in Databases and Information Systems. Cham: Springer International Publishing, 2014: 276-289.
- [9] ANDRZEJEWSKI W, BOINSKI P. Parallel GPU-based plane-sweep algorithm for construction of iCPI-trees[J]. Journal of Database Management, 2015, 26(3): 1-20.
- [10] ANDRZEJEWSKI W, BOINSKI P. Efficient spatial co-location pattern mining on multiple GPUs[J]. Expert Systems with Applications, 2018, 93: 465-483.
- [11] ANDRZEJEWSKI W, BOINSKI P. Parallel approach to incremental co-location pattern mining[J]. Information Sciences, 2019, 496: 485-505.
- [12] HUANG Y, ZHANG P S. On the relationships between clustering and

- spatial co-location pattern mining[C]//Proceedings of the 2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06). Piscataway: IEEE Press, 2006: 513-522.
- [13] LU J L, WANG L Z, FANG Y. Mining causal rules hidden in spatial co-locations based on dynamic spatial databases[C]//Proceedings of the 2016 International Conference on Computer, Information and Telecommunication Systems (CITS). Piscataway: IEEE Press, 2016: 1-6.
- [14] WANG X X, LEI L, WANG L Z, et al. Spatial colocation pattern discovery incorporating fuzzy theory[J]. IEEE Transactions on Fuzzy Systems, 2022, 30(6): 2055-2072.
- [15] ZHAO J S, WANG L Z, BAO X G, et al. Mining co-location patterns with spatial distribution characteristics[C]//Proceedings of the 2016 International Conference on Computer, Information and Telecommunication Systems (CITS). Piscataway: IEEE Press, 2016: 1-5.
- [16] WANG L Z, BAO X G, ZHOU L H, et al. Maximal sub-prevalent co-location patterns and efficient mining algorithms[C]//Proceedings of the 18th International Conference on Web Information Systems Engineering. Cham: Springer International Publishing, 2017: 199-214.
- [17] WANG L Z, BAO X G, ZHOU L H, et al. Mining maximal sub-prevalent co-location patterns[J]. World Wide Web, 2019, 22(5): 1971-1997.
- [18] YU W H. Identifying and analyzing the prevalent regions of a co-location pattern using polygons clustering approach[J]. ISPRS International Journal of Geo-Information, 2017, 6(9): 259.
- [19] FANG Y, WANG L Z, HU T. Spatial co-location pattern mining based on density peaks clustering and fuzzy theory[C]//Proceedings of the Second International Joint Conference on Asia-Pacific Web and Web-Age Information Management. Cham: Springer International Publishing, 2018: 298-305.
- [20] BAO X G, WANG L Z. Discovering interesting co-location patterns interactively using ontologies[C]//Proceedings of the International Conference on Database Systems for Advanced Applications Workshops. Cham: Springer International Publishing, 2017: 75-89.
- [21] WANG L Z, BAO X G, CAO L B. Interactive probabilistic post-mining of user-preferred spatial co-location patterns[C]//Proceedings of the 2018 IEEE 34th International Conference on Data Engineering (ICDE). Piscataway: IEEE Press, 2018: 1256-1259.
- [22] YANG S S, WANG L Z, BAO X G, et al. A framework for mining spatial high utility co-location patterns[C]//Proceedings of the 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). Piscataway: IEEE Press, 2015: 595-601.
- [23] 王晓璇, 王丽珍, 陈红梅, 等. 基于特征效用参与率的空间高效用 co-location 模式挖掘方法[J]. 计算机学报, 2019, 42(8): 1721-1738.
- WANG X X, WANG L Z, CHEN H M, et al. Mining spatial high utility co-location patterns based on feature utility ratio[J]. Chinese Journal of Computers, 2019, 42(8): 1721-1738.
- [24] WANG L Z, JIANG W G, CHEN H M, et al. Efficiently mining high utility co-location patterns from spatial data sets with instance-specific utilities[C]//Proceedings of the International Conference on Database Systems for Advanced Applications. Cham: Springer International Publishing, 2017: 458-474.
- [25] XIONG H, SHEKHAR S, HUANG Y, et al. A framework for discovering co-location patterns in data sets with extended spatial objects[C]//Proceedings of the 2004 SIAM International Conference on Data Mining. Philadelphia A: Society for Industrial and Applied Mathematics, 2004: 78-89.
- [26] LI J D, ADILMAGAMBETOV A, MOHOMED JABBAR M S, et al. On discovering co-location patterns in datasets: a case study of pollutants and child cancers[J]. GeoInformatica, 2016, 20(4): 651-692.
- [27] BEMBENIK R, JÓŹWICKI W, PROTAZIUK G. Methods for mining co-location patterns with extended spatial objects[J]. International Journal of Applied Mathematics and Computer Science, 2017, 27(4): 681-695.
- [28] 陈兰. 空间高影响 co-location 模式挖掘[D]. 昆明: 云南大学, 2017.
- CHEN L. Spatial high-impact co-location pattern mining[D]. Kunming: Yunnan University, 2017.
- [29] LEI L L, WANG L Z, ZENG Y M, et al. Discovering high influence co-location patterns from spatial data sets[C]//Proceedings of the 2019 IEEE International Conference on Big Knowledge (ICBK). Piscataway: IEEE Press, 2019: 137-144.
- [30] QIAN F, HE Q M, HE J F. Mining spread patterns of spatio-temporal co-occurrences over zones[C]//Proceedings of the International Conference on Computational Science and Its Applications. 2009: 677-692.
- [31] QIAN F, YIN L, HE Q M, et al. Mining spatio-temporal co-location patterns with weighted sliding window[C]//Proceedings of the 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems. Piscataway: IEEE Press, 2009: 181-185.
- [32] HUANG Y, ZHANG L Q, ZHANG P S. A framework for mining sequential patterns from spatio-temporal event data sets[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(4): 433-448.
- [33] CHEN L, GAO Y J, FANG Z Q, et al. Real-time distributed co-movement pattern detection on streaming trajectories[J]. Proceedings of the VLDB Endowment, 2019, 12(10): 1208-1220.
- [34] HU X, WANG G Y, DUAN J L. Mining maximal dynamic spatial colocation patterns[C]//Proceedings of the IEEE Transactions on Neural Networks and Learning Systems. Piscataway: IEEE Press, 2021: 1026-1036.
- [35] SHEKHAR S, JIANG Z, ALI R, et al. Spatiotemporal data mining: a computational perspective[J]. ISPRS International Journal of Geo-Information, 2015, 4(4): 2306-2338.
- [36] MA D, CHEN H, WANG L, et al. Dominant feature mining of spatial sub-prevalent co-location patterns[J]. Journal of Computer Applications, 2019, 40: 465-472.
- [37] 尚焱, 樊欣唯, 于洪. 考虑用户和传播属性的节点影响力评估算法[J]. 计算机工程与科学, 2015, 37(11): 2105-2111.
- SHANG Y, FAN X W, YU H. A novel node influence measurement algorithm based on characteristics of users and propagation[J]. Computer Engineering and Science, 2015, 37(11): 2105-2111.
- [38] PENG S C, YANG A M, CAO L H, et al. Social influence modeling using information theory in mobile social networks[J]. Information Sciences, 2017, 379: 146-159.
- [39] LI W H, LI C, GENG Y L. APS: attribute-aware privacy-preserving scheme in location-based services[J]. Information Sciences, 2020, 527: 460-476.
- [40] CITRARO S, ROSSETTI G. Eva: attribute-aware network segmentation[C]//Proceedings of the Eighth International Conference on Complex Networks and Their Applications. Berlin: Springer, 2020: 141-151.
- [41] SUBBIAN K, AGGARWAL C, SRIVASTAVA J. Content-centric flow mining for influence analysis in social streams[C]//Proceedings of the 22nd ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2013: 841-846.
- [42] FENG K Y, CONG G, JENSEN C S, et al. Finding attribute-aware similar regions for data analysis[J]. Proceedings of the VLDB Endowment, 2019, 12(11): 1414-1426.
- [43] BROCKMANN D, HUFNAGEL L, GEISEL T. The scaling laws of

- human travel[J]. Nature, 2006, 439(7075): 462-465.
- [44] BONACCORSI G, PIERRI F, CINELLI M, et al. Economic and social consequences of human mobility restrictions under COVID-19[J]. Proceedings of the National Academy of Sciences of the United States of America, 2020, 117(27): 15530-15535.
- [45] ZAKI M J. Efficiently mining frequent trees in a forest: algorithms and applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(8): 1021-1035.
- [46] ZAKI M J. Efficiently mining frequent embedded unordered trees[J]. Fundam Informaticae, 2004, 66: 33-52.
- [47] JIMÉNEZ A, BERZAL F, CUBERO J C. Mining induced and embedded subtrees in ordered, unordered, and partially-ordered trees[M]. Berlin: Springer, 2008.
- [48] MOOSAVI S, SAMAVATIAN M H, NANDI A, et al. Short and long-term pattern discovery over large-scale geo-spatiotemporal data[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2019: 2905-2913.
- [49] HUANG Y, ZHANG L Q, YU P. Can we apply projection based frequent pattern mining paradigm to spatial co-location mining? [C]// Proceedings of the 9th Pacific-Asia Conference on the Methodologies for Knowledge Discovery and Data Mining (PAKDD 2005). Berlin: Springer, 2005: 719-725.
- [50] WANG L Z, BAO Y Z, LU Z Y. Efficient discovery of spatial co-location patterns using the iCPI-tree[J]. The Open Information Systems Journal, 2009, 3(2): 69-80.
- [51] 项凤涛, 苏炯铭, 谷学强, 等. 基于智能体建模的新冠肺炎疫情传播问题研究[J]. 智能科学与技术学报, 2023, 5(1): 51-57.
XIANG F T, SU J M, GU X Q, et al. Research on the spread and countermeasures of COVID-19 using agent-based modeling[J]. Chinese Journal of Intelligent Science and Technology, 2023, 5(1): 51-57.
- [52] 覃缘琪, 季青原, 葛俊, 等. 城市交通路网动态短时推理与精准预测研究[J]. 智能科学与技术学报, 2022, 4(3): 380-395.
QIN Y Q, JI Q Y, GE J, et al. Short-term traffic state reasoning and precise prediction in urban networks[J]. Chinese Journal of Intelligent Science and Technology, 2022, 4(3): 380-395.
- [53] 陈苑文, 王晓, 李灵犀, 等. 基于社交媒体数据增强的交通态势感知研究及进展[J]. 智能科学与技术学报, 2022, 4(1): 1-13.
CHEN Y W, WANG X, LI L X, et al. Traffic situational awareness research and development enhanced by social media data: the state of the art and prospects[J]. Chinese Journal of Intelligent Science and Technology, 2022, 4(1): 1-13.
- [54] 张阳, 胡月, 辛东嵘. 一种考虑时空关联的深度学习短时交通流预测方法[J]. 智能科学与技术学报, 2021, 3(2): 172-178.
ZHANG Y, HU Y, XIN D R. A deep learning short-term traffic flow prediction method considering spatial-temporal association[J]. Chinese Journal of Intelligent Science and Technology, 2021, 3(2): 172-178.

[作者简介]



房滇武 (1971-), 男, 博士, 昆明学院讲师、硕士生导师, 主要研究方向为时空数据挖掘、模式识别和图神经网络等。



王丽珍 (1962-), 女, 博士, 滇池学院教授、博士生导师, IEEE 高级会员, 中国计算机学会 (CCF) 杰出会员, 主要研究方向为空间数据挖掘、交互式数据挖掘和大数据分析及应用等。



邹目权 (1986-), 男, 博士, 富滇银行股份有限公司博士后科研工作站博士后, 昆明学院硕士生导师, 主要研究方向为空间数据挖掘和数字金融。



邓飞 (1986-), 男, 西安交通大学博士生, 昆明学院硕士生导师, 主要研究方向为神经网络和机器学习。