Journal of System Simulation

Volume 36 | Issue 1 Article 20

1-20-2024

Spatio-temporal Association Rule Mining of Traffic Congestion in a Large-scale Road Network Based on Trajectory Data

Qifan Zhou
School of Transportation and Logistics, Southwest Jiaotong University, Chengdu 611756, China, 2015114658@my.swjtu.edu.cn

Haixu Liu
School of Transportation and Logistics, Southwest Jiaotong University, Chengdu 611756, China; National Engineering Laboratory of Application Technology of Integrated Transportation Big Data, Southwest Jiaotong University, Chengdu 611756, China; National United Engineering Laboratory of Integrated and Intelligent Transportation, Southwest Jiaotong University, Chengdu 610031, China; Comprehensive Transportation Key Laboratory of Sichuan Province, Southwest Jiaotong University, Chengdu 611756, China, hxliu@swjtu.edu.cn

Zhipeng Dong
School of Transportation and Logistics, Southwest Jiaotong University, Chengdu 611756, China

Yin Xu
School of Transportation and Logistics, Southwest Jiaotong University, Chengdu 611756, China

Spatio-temporal Association Rule Mining of Traffic Congestion in a Large-scale Road Network Based on Trajectory Data

Abstract

Abstract: A K neighbor-RElim (KNR) algorithm and a sequential KNbr-RElim (SKNR) algorithm are proposed to mine traffic congestion association rules and congestion propagation spatio-temporal association rules by vehicle trajectory data in a large-scale road network. The KNR algorithm extends the spatial topology constraint based on the RElim algorithm. The KNR can be used to mine the road links prone to congestion from the large-scale trajectory dataset in a large-scale road network and quantify the strength of association for congested road links. The SKNR algorithm expands the time dimension in the form of sliding window and can be applied for mining the congestion propagation phenomenon which is difficult to observe directly in a large-scale road network and tracing the path of congestion propagation. The algorithms are illustrated and verified by the empirical results of the Chengdu road network with vehicle trajectory data. The results show the effectiveness and robustness of the proposed algorithms.

Keywords

data mining, association rules, congestion propagation, trajectory data, RElim algorithm

第 36 卷第 1 期 系统仿真学报© Vol. 36 No. 1
2024 年 1 月 Journal of System Simulation Jan. 2024

基于轨迹数据的大规模路网交通拥挤时空关联规则挖掘

周启帆[1]，刘海旭[1,2,3,4*]，董志鹏[1]，徐银[1]

(1. 西南交通大学 交通运输与物流学院， 四川 成都 611756；2. 西南交通大学 综合交通大数据应用技术国家工程实验室，四川 成都 611756；3. 西南交通大学 综合交通运输智能化国家地方联合工程实验室，四川 成都 610031；4. 西南交通大学 综合运输四川省重点实验室，四川 成都 611756)

摘要： *提出了K近邻RElim（K neighbor-RElim，KNR）算法和时序K近邻RElim（sequential KNbr-RElim，SKNR）算法，利用大规模路网的车辆轨迹数据来挖掘路段拥挤关联规则和拥挤传播时空关联规则。其中KNR算法在RElim算法基础上拓展了空间拓扑约束，可高效从大规模车辆轨迹数据集中挖掘路网中关联性拥挤易发路段，并量化这些路段间拥挤的关联性强度。而SKNR算法进一步以滑动窗口的形式拓展时间维度，可以挖掘出大规模路网中难以直接观测的拥挤传播现象，并追溯拥挤传播路径。*以成都路网和车辆轨迹数据的挖掘结果对所提出的算法进行了说明和验证，结果表明了算法的有效性和鲁棒性。

关键词： 数据挖掘；关联规则；拥挤传播；轨迹数据；RElim算法

中图分类号：TP301.6 文献标志码：A 文章编号：1004-731X(2024)01-0260-12

DOI: 10.16182/j.issn1004731x.joss.22-0898

引用格式: 周启帆, 刘海旭, 董志鹏, 等. 基于轨迹数据的大规模路网交通拥挤时空关联规则挖掘[J]. 系统仿真学报, 2024, 36(1): 260-271.

Reference format: Zhou Qifan, Liu Haixu, Dong Zhipeng, et al. Spatio-temporal Association Rule Mining of Traffic Congestion in a Large-scale Road Network Based on Trajectory Data[J]. Journal of System Simulation, 2024, 36(1): 260-271.

Spatio-temporal Association Rule Mining of Traffic Congestion in a Large-scale Road Network Based on Trajectory Data

Zhou Qifan[1], *Liu Haixu*[1,2,3,4*], *Dong Zhipeng*[1], *Xu Yin*[1]

(1. School of Transportation and Logistics, Southwest Jiaotong University, Chengdu 611756, China; 2. National Engineering Laboratory of Application Technology of Integrated Transportation Big Data, Southwest Jiaotong University, Chengdu 611756, China; 3. National United Engineering Laboratory of Integrated and Intelligent Transportation, Southwest Jiaotong University, Chengdu 610031, China; 4. Comprehensive Transportation Key Laboratory of Sichuan Province, Southwest Jiaotong University, Chengdu 611756, China)

Abstract: *A K neighbor-RElim (KNR) algorithm and a sequential KNbr-RElim (SKNR) algorithm are proposed to mine traffic congestion association rules and congestion propagation spatio-temporal association rules by vehicle trajectory data in a large-scale road network. The KNR algorithm extends the spatial topology constraint based on the RElim algorithm. The KNR can be used to mine the road links prone to congestion from the large-scale trajectory dataset in a large-scale road network and quantify the strength of association for congested road links. The SKNR algorithm expands the time dimension in the form of sliding window and can be applied for mining the congestion propagation phenomenon which is difficult to observe directly in a large-scale road network and tracing the path of*

收稿日期：2022-08-03 修回日期：2022-11-06

基金项目：国家自然科学基金(61873126)；湖北省交通运输厅科技项目(2022-11-1-5)

第一作者：周启帆(1996-)，男，硕士生，研究方向为交通建模与仿真。E-mail：2015114658@my.swjtu.edu.cn

通讯作者：刘海旭(1974-)，男，副教授，博士，研究方向为交通建模与仿真。E-mail：hxliu@swjtu.edu.cn

congestion propagation. The algorithms are illustrated and verified by the empirical results of the Chengdu road network with vehicle trajectory data. The results show the effectiveness and robustness of the proposed algorithms.

Keywords: data mining; association rules; congestion propagation; trajectory data; RElim algorithm

0 引言

造成大规模路网拥挤的原因极为复杂，因此大规模路网的拥挤演变规律也极难把握。路网中空间上不相邻的路段有可能同时出现拥挤的现象，而路网中一条路段发生的拥挤往往会对周边路段产生不同程度的影响，且具有不易直接观测的传播路径。尽管拥挤的关联性和传播性是交通网络中客观存在的基本规律，但迄今为止还难以准确描述并分析这种规律。

拥挤相关性的研究方法可分为基于仿真模型的方法和基于数据驱动的方法。基于交通仿真模型的方法是研究拥挤相关性和传播规律的重要途径。该类方法的相关研究多集中在微观交通或中观交通领域，主要关注路段或交叉口内部，以及相邻交叉口之间的拥挤传播现象。仿真模型包括跟驰模型[1]、元胞传输模型[2]、流体力学模型[3]以及其他的动力学模型。比如，Long 等[4]采用改进的 CTM(cell transmission model)模型探究了路段内部以及路段之间的拥挤传播关系，并成功定位路网的拥挤瓶颈路段，研究发现路段在网络中的位置以及路段流量是拥挤形成的主要因素。杨泳[5]提出可变元胞传输模型(VCTM, variable cell transmission model)来刻画城市快速路的交通流特性、分析交通拥挤在突发事件下的消散模式和扩散规律。以跟驰模型、元胞传输模型为代表的仿真模型方法的有效性很大程度上依赖于模型的强假设，这些假设在实际交通环境下是否成立仍然需要验证；此外，仿真模型方法的参数标定困难。这些因素都在一定程度上影响了交通仿真模型方法的普适性。

而基于数据驱动的方法不需要设定强约束来对交通流特征抽象处理，其数据本身就是对客观交通状态的真实表达。目前基于数据驱动的拥挤相关性研究仍然以路段或交叉口为主要研究对象，其数据一般由感应线圈或交通摄像头采集得到，而从大规模区域范围探究网络拥挤相关性的研究较少，支持大规模区域拥挤相关特性研究的数据也均为车辆轨迹数据。

Saberi 等[6]和 Befekadu[7]均利用传染病模型(susceptible infected recovered, SIR)描述了交通整体性拥挤产生和消散的过程，但不足点在于 SIR 模型未能成功刻画畅通路段重复向拥挤状态转换的过程。总体而言，传染病模型主要用于观察路网整体的拥挤随时间演变的过程，对于具体路段间的拥挤相关关系并不关注。

另一部分数据驱动的研究则侧重拥挤在路网中的传播方向与过程。文献[8-9]以北京市道路为研究对象，进行改进的 Moran’s I 指数分析和识别道路交通拥挤事件，并构建改进的 SDM(spatial durbin model)模型分析拥堵空间传播特性。欧吉顺[10]通过时空聚类的方法追溯大规模城市路网中不同拥挤瓶颈的独立传播分支，同时利用时空注意力权重矩阵度量路网拥挤瓶颈路段的时空关联关系。王启燕[11]和谭星[12]通过 GNP(genetic network programming)算法和改进的 Apriori 算法挖掘路网中拥挤传播事件的时序关联规则并利用其传播规律预测拥挤的发生。

由于大规模路网中拥挤成因复杂，很难用单一理论模型去准确刻画整个路网的拥挤相关性规律，因此仿真模型仍具有很大局限性。随着交通相关大数据获取成本的降低，应用诸如轨迹数据、交通视频数据等进行整个路网的拥挤分析成为可能。在当前基于数据驱动的拥挤相关性研究方面，针对大规模路网以常发、偶发性拥挤识别为主[13-14]，部分研究更加侧重拥挤现象随时间周期的

DOI: 10.16182/j.issn1004731x.joss.22-0898

宏观演变规律[15]，而对于路网中不同时间或空间相邻的拥挤事件的彼此作用关系关注不多。此外，交通相关性分析的关联规则算法在交通领域的应用尚处于初级阶段，多是在原始Apriori算法基础上增加空间约束[16]，研究搭配使用栅格路网数据，仅探究相邻栅格之间的区域是否存在关联性的交通拥挤，得到的结果也并不指向具体的路段。

基于上述考虑，本文在RElim算法的基础上，增加了空间拓扑约束和时序约束提出了基于拓扑约束的KNR(K neighbor-RElim)和SKNR(sequential KNbr-RElim)算法，采用成都市车辆轨迹数据，识别路网中的拥挤关联路段，量化拥挤关联性强度并实现对拥挤传播路径的追溯。本文算法用以挖掘拥挤在路网中潜在的关联与传播特性，有助于从宏观层面解释拥挤的形成机理，从而为交通拥挤治理制定精准的管控策略。

1 关联规则挖掘概述

关联规则挖掘技术是数据挖掘领域众多研究方向中的一个重要分支，其合理利用能够帮助从大规模数据集中高效地寻找具有潜在价值的信息。

1.1 关联规则挖掘基本概念

关联规则挖掘分为两个步骤：发现频繁模式，以及从频繁模式中发现关联规则。频繁模式挖掘是寻找高频出现在数据集中的模式(如项集、子序列或子结构)[17]，也可以称作频繁项集挖掘(frequent item set mining)；频繁模式挖掘是关联规则挖掘的基础，这些频繁项集描述的强关联规则中可能存在有价值的相关关系，频繁项集的挖掘依赖于以下基础概念：

(1) 项、项集：单个研究对象的一次观测结果称为一个项，多个项的集合称为项集，k项集记为$X=\{x_1,x_2,\cdots,x_k\}$。项集是承载关联规则的载体。

(2) 事务、事务集：研究范围内全部对象的一次观测结果称为一个事务，记为$T=(TID,X)$，多个事务的集合称为事务集，记为$D=\{T_1,T_2,\cdots,T_n\}$。全部观测结果组成的事务集称为事务数据库。

(3) 支持度、支持度计数：支持度表示指定项集在整个事务集中出现的频率，该项集出现的频数称为支持度计数。项集A的支持度记为$support(A)$。

(4) 频繁项集：其目标是基于给定的支持度阈值，从事务集中寻找不小于该支持度要求的项集。

1.2 关联规则的生成与辨识

频繁项集挖掘算法根据频繁项集产生模式的不同可分为生成式挖掘算法和模式增长类算法，两类挖掘方法的代表算法分别是Apriori算法和FP-Growth算法。前者扫描数据库涉及大量IO操作，严重限制了该算法的挖掘效率，而后者FP-tree结构相对复杂储存在内存中对计算机要求较高。RElim(recursive elimination)[18-19]是一种受FP-Growth启发而设计的增长式频繁项集挖掘算法，并在前两类算法的缺陷处有不同程度的改善。

频繁项集挖掘为关联规则挖掘的第一步，下一步利用已获得的频繁模式来发掘有价值的关联规则。频繁项集中各子项集之间的关联程度需要通过置信度来衡量。事务中出现项集A时，项集B同时出现的条件概率称作项集A对项集B的置信度，记作$confidence(A=>B)$：

$$confidence(A=>B=P(B|A)=\frac{support(A\cup B)}{support(A)}) \tag{1}$$

项集A与B之间存在强关联规则只能说明二者具有较高相关度，但相关关系与相关性强弱还需要提升度等指标刻画，项集A和项集B之间的$lift$值、$KULC$值、IR值分别定义如下：

$$lift(A=>B)=\frac{confidence(A=>B)}{support(B)}=\frac{support(A\cup B)}{support(A)\times support(B)} \tag{2}$$

$$KULC=\frac{P(B|A)+P(A|B)}{2}=\frac{confidence(A=>B)+confidence(B=>A)}{2} \tag{3}$$

$$IR=\frac{P(B|A)}{P(A|B)}=\frac{confidence(A=>B)}{confidence(B=>A)}=\frac{support(B)}{support(A)} \tag{4}$$

其中，*lift*提升度的值如果小于1，说明项集*A*的出现和项集*B*的出现是负相关的；反之亦有相应结论，*KULC*可以反映项集*A*和项集*B*之间的相关性强度，取值范围[0，1]。二者共同出现的频率越高则*KULC*值越大，说明相关性越强，当*KULC*值为0时*A*和*B*之间相互独立。

IR(imbalance ratio)指数反映的是项集*A*和项集*B*发生概率的平衡性，当*IR*值很大时，说明项集*B*描述事件发生的概率远高于项集*A*描述的事件。

2 路段拥挤关联规则挖掘

2.1 路段拥挤状态判别标准

基于融合矢量路网信息的轨迹数据，可方便地获取特定时段内的路段平均速度。路段(u,v,k)在t_1至t_2时段内的平均速度$v_{t_1t_2}^{-uvk}$计算方法为

$$v_{t_1t_2}^{-uvk}=\sum_{i=1}^{n}v_{it}^{uvk}/n,\ \ t_1\leqslant t\leqslant t_2 \tag{5}$$

式中：(u,v,k)为路段的唯一标识ID；v_{it}^{uvk}为匹配至路段(u,v,k)的第i个轨迹点的瞬时速度；t表示该轨迹点对应的时间戳。研究中将各路段历史轨迹速度的90%分位数视为该路段的自由流速度。路段的交通拥挤水平通过出行时间系数(travel time index, TTI)来定义：

$$TTI=v_{\text{free}}^{uvk}/v_{t_1t_2}^{-uvk} \tag{6}$$

式中：v_{free}^{uvk}为路段(u,v,k)的自由流速度。根据国内各城市对于交通拥堵程度的定义[20]，认为$TTI<1.3$、$1.3\leqslant TTI<1.6$、$1.6\leqslant TTI<1.9$、$1.9\leqslant TTI<2.2$、$TTI\geqslant 2.2$的路段分别处于畅通、基本畅通、缓行、轻度拥挤、重度拥挤状态。

2.2 *K*邻边拓扑约束和*K*邻边搜索

为合理约束拥挤路段之间的空间相邻关系，引入*K*邻边拓扑约束概念。*K*邻边的定义：以图中的某一条边为初始边，与该边两顶点距离*K*个拓扑距离以内的边(不含初始边)的集合。图1所示为初始边和近邻边所构成的*K*邻边(*K*=1，2)关系图。

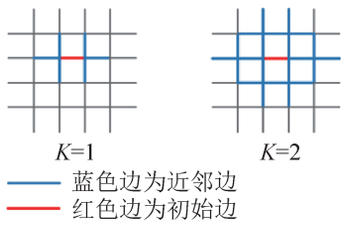

图1 *K*邻边关系图

Fig. 1 Graph of *K* neighbor edges

为最大程度地提升大规模路网中算法的挖掘效率，*K*近邻空间拓扑约束未选用*K*近邻搜索算法的常用距离类型，而是设置初始边拓扑距离为0，与其直接相连路段距离为1，每增加一条间隔路段，则拓扑距离加1。

使用*K*邻边拓扑关系作为城市路网拥挤事件频繁项集挖掘的空间约束，直接以路段为基本单位探究拥挤的关联关系，挖掘结果指向具体路段并且能够区分不同行驶方向上的拥挤关联事件。

搜索*K*邻边的算法为深度优先搜索(deep first search, DFS)算法或者广度优先搜索(breadth first search, BFS)算法，下面陈述*K*邻边搜索算法的DFS算法流程：

step 1：初始化。准备一个集合，用于存储已经确定会被搜索的边。将初始边添加到集合中，设定初始边为被搜索边，搜索深度为*K*。

step 2：递归搜索。检查搜索深度，当搜索深度为0时结束递归。获取指定边的邻边(1邻边)，遍历检查邻边是否已经位于集合中。如果是，则跳过此邻边；否则，将该邻边添加至集合中，并设置该邻边为被搜索边，搜索深度为*K*−1，执行递归搜索。

step 3：结束搜索。将集合中的初始边移除，*K*邻边搜索完毕。

2.3 基于拓扑约束的KNR算法

在RElim算法的基础上，使用*K*邻边概念定义路段之间的空间相邻关系，并进一步提出融合*K*邻边拓扑约束剪枝的KNR算法，能够在数据挖掘过程中施加空间和时间的剪枝操作，满足交通时空数据挖掘的需求，结果指向具体关联路段，并实现近邻路段之间拥挤关联强度的挖掘。

基于 K 邻边拓扑关系，设计具有空间约束的KNR频繁项集挖掘算法，算法流程如下：

step 1：首先对原始的事务数据库做第一次完整的扫描，扫描过程中统计各1项集的支持度计数。将扫描获得的1项集按支持度计数增序排列。

step 2：第二次扫描原始事务数据库，在扫描过程中将各个事务中支持度小于最小支持度阈值的事务元素剔除，并重新将每个事务元素按照第一次扫描后的支持度计数增序排列。

step 3：利用 step 2 过滤并排序后的事务，构建事务链表组(transaction link list)，事务链表组的链表头用频繁1项集初始化。

step 4：初始化一个频繁项集计数器，然后按照支持度由低到高的顺序，依次取出事务链表组中支持度最低的链表头及对应的事务链表，对每一个链表头和事务链表执行下面步骤：

step 4.1：利用取出的事务链表构建子事务链表组。子事务链表组的构建方式与RElim算法不同。

step 4.2：递归挖掘。如果该事务链表中的事务数量不小于最小支持度计数，则将当前链表头作为频繁1项集更新到频繁项集计数器中；然后根据链表头生成子邻边事务链表组，并对子邻边事务链表组从step 4开始递归地挖掘频繁项集；否则直接执行剩余步骤。

step 4.3：将链表头、事务链表从事务链表组中删除。

step 4.4：将对应的子事务链表组与原事务链表组合并。

step 4.5：如果 step 4.2 执行了递归处理并获取到子频繁项集计数器，则利用子计数器更新 step 4 中初始化的频繁项集计数器。

step 5：返回更新后的频繁项集计数器。

KNR 算法流程如图 2 所示，其中虚线框标记的步骤为对原始 RElim 算法的修改节点。

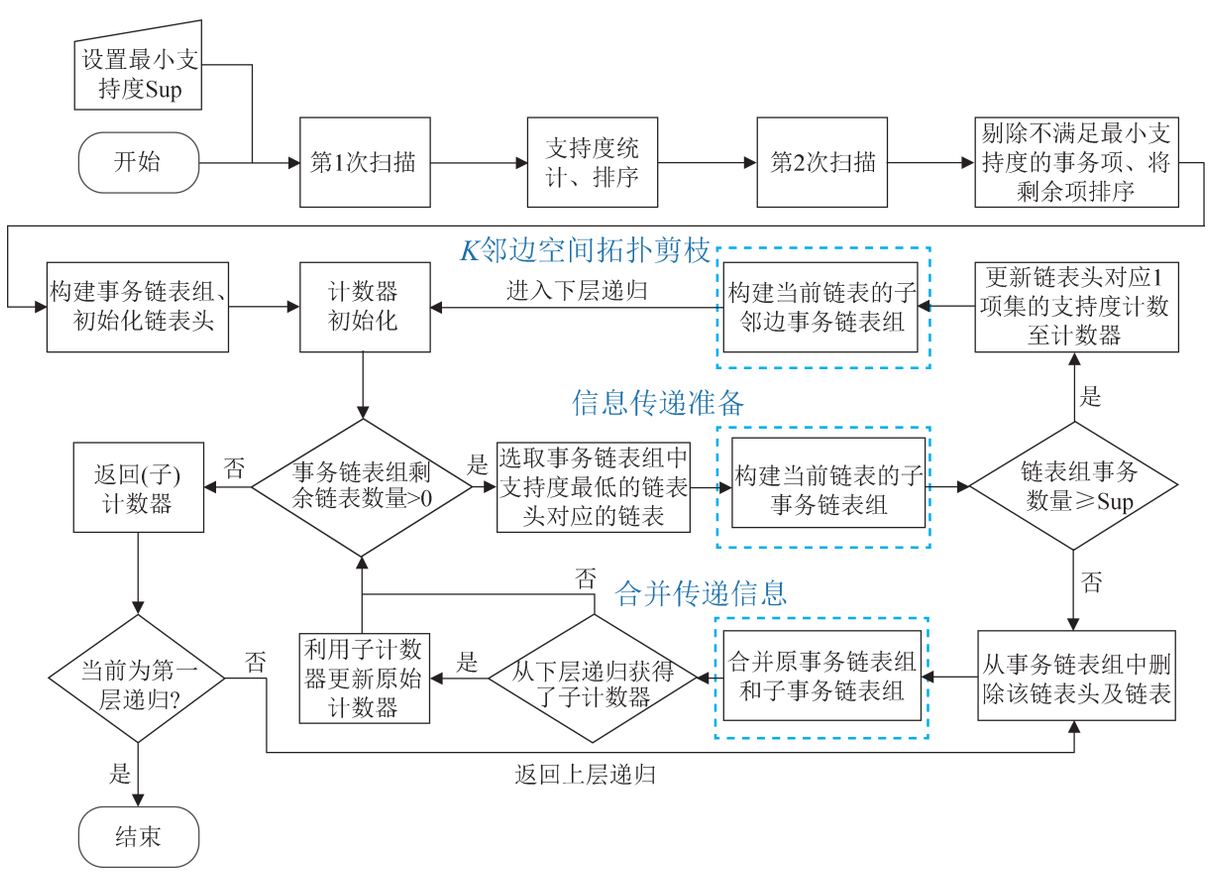

图2 基于拓扑约束的KNR算法流程图

Fig. 2 Flow of KNR algorithm based on topology constraints

K邻边空间拓扑剪枝需要在子邻边事务链表组的生成阶段实现。子邻边事务链表组与RElim算法的事务链表组的区别在于子邻边事务链表组的链表头必须是原链表头的K邻边。

子邻边事物链表组生成方法：以选定链表头为初始边获取其K邻边，将满足支持度要求的K邻边初始化为子邻边事务链表组的链表头；遍历该事务链表中的每一个事务，将当前链表头从事务中删除，计算事务中剩余元素与K邻边的交集，如果交集为空，则跳过该事务，否则选取交集中支持度最低的一个元素作为该事务的链表头；当整个事务链表中事务的遍历完毕之后，完成子邻边事务链表组的生成。

3 路段拥挤传播时空关联规则挖掘

在空间关联规则挖掘算法的基础上增加时序约束，将KNR算法拓展为SKNR算法，算法可用于挖掘路网中频繁发生的拥挤传播路径。通过滑动窗口将原始事务转换为时序事务，同时增加时序关联约束，挖掘具备时空约束性的关联规则。

基础的频繁项集挖掘仅关注同时发生的事务项，无法发掘前一时刻事务项与后续时刻事务项的关联关系。而交通拥挤是典型的具备时序特性的事务数据，一条路段的拥挤随着时间演变可能传播到相邻路段，为了探究前一时刻某路段的拥挤在后一时段是否向相邻路段产生传播，采用时序关联规则挖掘算法挖掘数据中的时序频繁项集。

时序RElim挖掘在原有算法的概念上进一步丰富，下面介绍时序RElim算法的相关概念。

(1) 时序项集：对于一个时序跨度为n的时序事务集$D=\{T_1, T_2, \cdots, T_n\}$，$T_i$代表在时刻$i$发生的事务，若某$k$项集$X=\{x_1, x_2, \cdots, x_k\}$中的项均归属于$D$中任意$m(m \leqslant k)$个时间连续的事务称其为时序跨度为$m$的时序$k$项集。

(2) 时序频繁项集：以窗口尺寸为m的滑动窗口观察时序事务集$D=\{T_1, T_2, \cdots, T_n\}$，若时序跨度为$m$的时序$k$项集，在$n-m+1$次观察中出现次数满足支持度计数的要求，那么称该项集为时序频繁k项集。

(3) 时序关联规则的置信度：当项集A发生时，在随后的时间内项集B跟随发生的条件概率称时序关联规则中项集A对项集B的置信度。

SKNR算法在KNR算法的基础上将原始事务数据集转换为时序事务数据集，并增加了时间约束。算法步骤如下：

step 1：首先确定滑动窗口尺寸，利用滑动窗口对原始的事务数据库做第一次完整的扫描。扫描过程中为窗口内分属于不同时序事务的项增加时序编号，即窗口中第一个窗格内的事务项时序编号为1，第二个窗格内的事务项时序编号为2，以此类推。统计增加时序编号的各1项集的支持度计数并将获得的1项集按照支持度计数增序排列。

step 2：第二次利用滑动窗口以相同的方式扫描、转换原始事务数据库。每次滑动都将滑动窗口内添加时序编号后的项组合成一个新的时序事务。并将各时序事务中支持度小于最小支持度的事务元素剔除，最后重新将每个时序事务中的事务元素按照时序编号增序排列，相同时序编号的元素按第一次扫描后的支持度计数增序排列。

step 3：利用step 2过滤并排序后的事务，构建事务链表组，事务链表组的链表头用时序编号为1的频繁1项集初始化。

step 4：初始化一个频繁项集计数器，然后按照支持度由低到高的顺序，依次取出事务链表组中支持度最低的链表头及对应的事务链表，对每一个链表头和事务链表执行下面步骤。

step 4.1：利用取出的事务链表构建子事务链表组。子事务链表组的构建方式与KNR算法一致。

step 4.2：递归挖掘。如果该事务链表中的事

DOI: 10.16182/j.issn1004731x.joss.22-0898

务数量不小于最小支持度计数，则将当前链表头作为频繁 1 项集更新到频繁项集计数器中，然后根据链表头生成子时序邻边事务链表组，并对子邻边事务链表组从 step 4 开始递归地挖掘频繁项集；否则，执行剩余步骤中除 step 5 以外的步骤。

step 4.3：将链表头、事务链表从事务链表组中删除。

step 4.4：将对应的子事务链表组与原事务链表组合并。子事务链表组与原事务链表组合并的方法与 KNR 算法一致。

step 4.5：如果 step 4.2 执行了递归处理并获取到子频繁项集计数器，则利用子计数器更新步骤 4 中初始化的频繁项集计数器。频繁项集计数器的更新方法与 RElim 算法一致。

step 5：返回更新后的频繁项集计数器。

SKNR 算法流程图如图 3 所示。

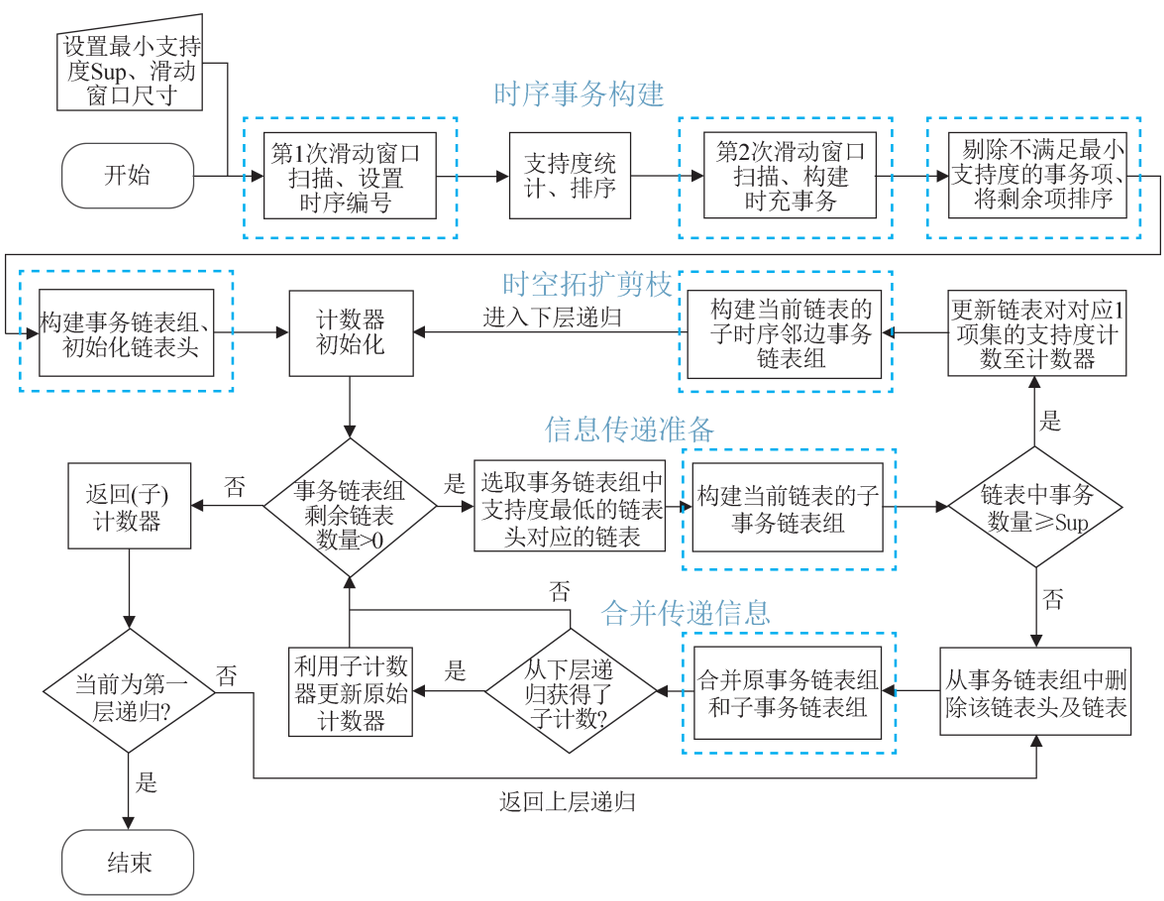

图 3　SKNR 算法流程图

Fig. 3　Flow of SKNR algorithm

4　实例分析

本节以成都市车辆轨迹数据为实例数据集挖掘交通路网中潜在的拥挤时空关联关则，对提出的 KNR 算法有效性进行验证。

研究中使用的车辆轨迹数据出自滴滴出行“盖亚”数据开放计划，数据集包括 2018 年 10 月 1 日—11 月 30 日成都区域路网的轨迹数据，轨迹覆盖范围为 104.03°E~104.14°E、30.64°N~30.74°N。每 10 天的轨迹数据以 csv 文件格式记录，单个 csv 数据文件约 20 G，含出行订单约 250 万条。

使用的成都市路网矢量数据出自开放街道地图(open street map, OSM)。实际使用过程中对道路矢量数据进行了一定的简化，将冗余的路段与相

邻路段合并，最终构建与轨迹数据范围相对应的城市道路交通拓扑路网。

4.1 路网拥挤空间关联规则实例挖掘与分析

研究采用轨迹数据进行关联规则挖掘，首先筛选并剔除异常轨迹点，以连续轨迹段为基本单位，将每个连续轨迹段和地图匹配的问题转化为隐马尔可夫模型(hidden Markov model，HMM)的解码问题，实现轨迹点和路网的匹配。

基于以上工作，选取 2018 年 10 月 11—19 日间每天 06:00—24:00 的轨迹数据，以 15 min 为事务周期，共获得事务 648 组，将事务周期内的非畅行路段作为事务项进行频繁项集挖掘。首先设置一个相对适中的最小支持度计数 80，综合考虑支持度和频繁项集数量，选择使用 2 邻边及 0.15 最小支持度分析城市路网拥挤路段的空间关联关系。

0.15 最小支持度下频繁 K 项集合计 841 项，其中频繁 1 项集 644 项、频繁 2 项集 174 项、频繁 3 项集 22 项、频繁 4 项集 1 项。可认为频繁 1 项集路段对应该路网区域范围内的拥挤常发路段，提取频繁 1 项集路段显示如图 4，图中路段颜色深度代表其支持度。

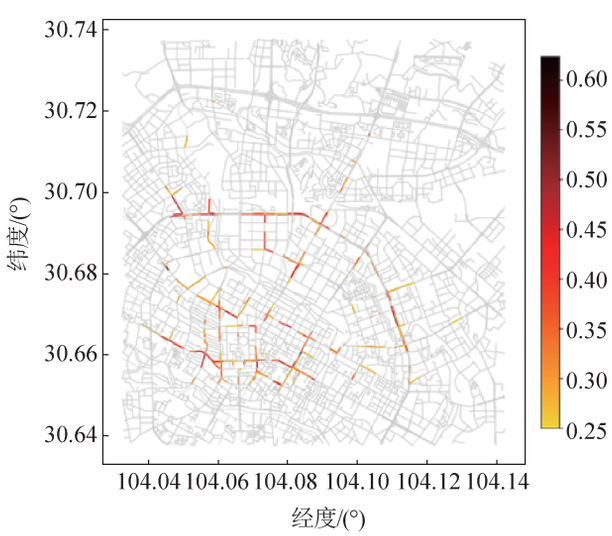

图 4　拥挤常发路段支持度等值图
Fig. 4　Support contour map of frequent congestion roads

由图 4 可知，成都市的拥挤常发路段主要分布在二环内中心城区，其中显著拥挤的路段集中分布在天府广场及春熙路片区。此外，一、二环快速路及邻近路段拥挤的发生频率也相对较高，总体符合环形放射路网的基本规律。

设置最小置信度为 0.7，观察频繁项集中满足最小支持度和置信度限制的空间关联规则分布，其中包括由频繁 2 项集生成的关联规则 45 项、3 项集生成的关联规则 23 项、4 项集生成的关联规则 2 项，生成相应的关联规则分布图，其中 2 项集关联规则展示图如图 5 所示。可以观察到，具备关联性拥挤的路段主要分布在春熙路、天府广场、人民公园、青羊宫片区。除此之外，西南交大九里校区西南侧路段、解放路一段至二环路北四段立交等二环路段也是路段关联性拥挤的常发区域。

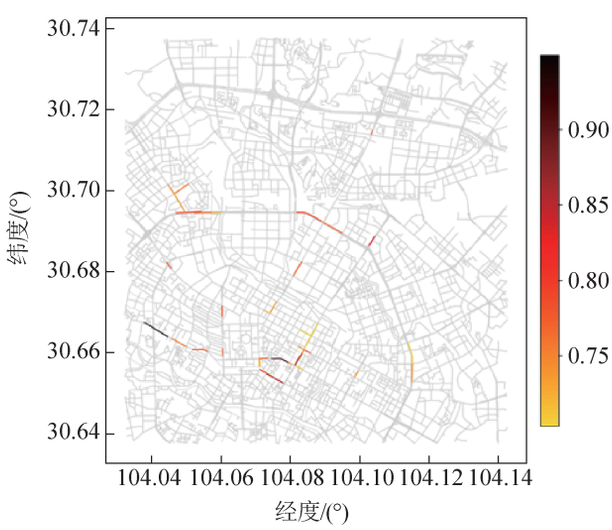

图 5　空间关联规则置信度等值图(2 项集)
Fig. 5　Confidence contour map of spatial association rules (2 itemset)

由 KNR 算法输出的挖掘结果如表 1 所示。表中 AB 为挖掘得到的关联规则项集，其中 $conf\_A$ 是项集 A 对项集 AB 的置信度，$kulc$ 和 ir 是用于判定 $conf\_A$ 有效性的指标。$A(u,v,k)$ 和 $B(u,v,k)$ 为存在拥挤关联性的两条路段，u、v、k 分别是路段的起点 ID、终端 ID、路段序号(相同两节点之间可能存在不止一条路段，具有相同节点的路段需要通过序号区分)。

https://dc-china-simulation.researchcommons.org/journal/vol36/iss1/20
DOI: 10.16182/j.issn1004731x.joss.22-0898

第 36 卷第 1 期 系统仿真学报 Vol. 36 No. 1
2024 年 1 月 Journal of System Simulation Jan. 2024

表 1　关联规则表(节选)

Table 1　Association rules (excerpt)

ID	*AB*	*conf_A*	*KULC*	*IR*
1	(4548211352, 5530491877, 0), (7366295632, 4548211352, 0)	0.94	0.76	1.66
2	(671600440, 6224263690, 0), (6224263690, 6224263706, 0)	0.76	0.70	1.19
3	(5530491877, 4548211421, 0), (4548211352, 5530491877, 0), (7366295632, 4548211352, 0)	0.99	0.70	2.47
4	(963638438, 5535297029, 0), (5535297029, 1159223970, 0)	0.74	0.62	1.48
5	(963638438, 5535318948, 0), (5535318948, 5218240010, 0)	0.76	0.68	1.27
6	(1339701401,963638438,0), (4549993385,1339701401, 0)	0.73	0.71	1.07

表 1 中第 1、2 条关联规则属于频繁 1 项集支持频繁 2 项集的关联规则。结合图 6 分析可知：当上游路段(7366295632，4548211352，0)出现拥挤时，下游路段(4548211352，5530491877，0)出现拥挤的置信度为 0.94。

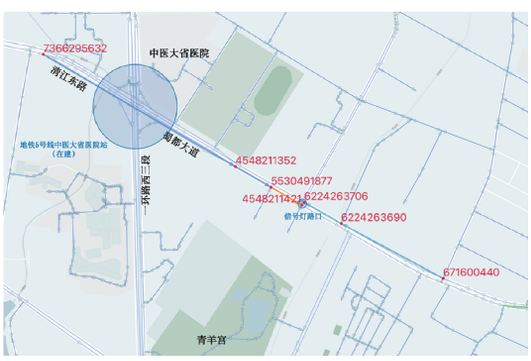

图 6　拥挤路段空间关联图(青羊宫片区)

Fig. 6　Spatial association map of congestion roads (Qingyanggong area)

当上游路段(671600440，6224263690，0)出现拥挤时，下游路段(6224263690，6224263706，0)出现拥挤的置信度为 0.76，KULC 和 IR 指标均无异常。而从下游路段指向上游路段的关联规则均不存在。因此，可能存在从下游路段指向上游路段的拥挤传播行为。

关联规则 3 表明，当上、下游路段(5530209192，1305816694，0)和(1305816694，6125970016，0)均处于拥挤状态时，中游路段(4548211352，5530491877，0)也处于拥挤状态的条件概率高达 0.99。因此，中游路段(4548211352，5530491877，0)可能是上下游拥挤的交汇路段。此时，上、中、下游路段处于同一条拥挤传播路径，但因拥挤传播和消散并不同步，拥堵路段处于不同环节，故路段标号不连续。

经查证，该路段位于青羊宫文化公园及成都中医大省医院片区，靠近清江东路、蜀都大道和一环路西三段立交。拥挤路段(4548211352，5530491877，0)衔接蜀都大道及其辅路，两路车流在此汇行，车流交汇点与前方信号灯路口距离较近，交通状况复杂多变；此外，2018 年内，附近的 5 号线地铁站口长期处于施工状态，也是该路段关联性拥挤频繁发生的原因之一。

4.2　路网拥挤传播实例挖掘与分析

选取 2018 年 10 月 11—19 日间每天 6:00—24:00 的轨迹数据，以 15 min 为事务周期，即每天包含事务 72 组，将事务周期内的非畅行路段作为事务项，采用 SKNR 算法进行拥挤传播规则的挖掘。由于拥挤无法从前日 24 时传播至次日 6 时，故每天的数据可构建时序事务 67 组，合计 603 组。

为控制合理的挖掘耗时，设定滑动窗口尺寸为 6，即挖掘最长在连续 1.5 h 内传播的拥挤规则。基于空间关联规则挖掘过程中的经验，设置拓扑约束为 2 邻边、最小支持度为 0.15，另外，要求时序频繁项集中不同路段的数量应不少于项集数量的 60%，避免因同一路段连续拥挤而产生过多低价值的关联规则。挖掘过程耗时约 15 min，获得频繁项集 1 462 个，各支持度下不同时序跨度的频

繁项集数量变化趋势如图7所示。

同样设置最小置信度为0.7，观察频繁项集中满足最小支持度和置信度限制的时序关联规则分布。其中包括时序跨度为2的时序关联规则142项，时序跨度为3的时序关联规则50项，时序跨度为4的时序关联规则16项。

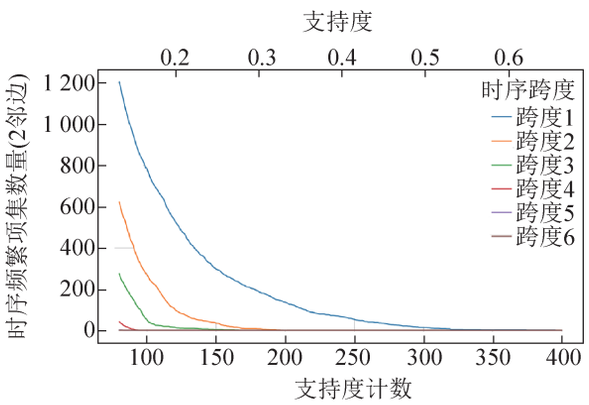

图7　时序频繁项集数量–支持度图
Fig. 7　Number of sequential frequent itemsets-support

完整的时序关联规则空间分布如图8所示，图中路段颜色的深度代表其置信度的大小。由SKNR算法输出的挖掘结果：部分关联规则如表2所示。

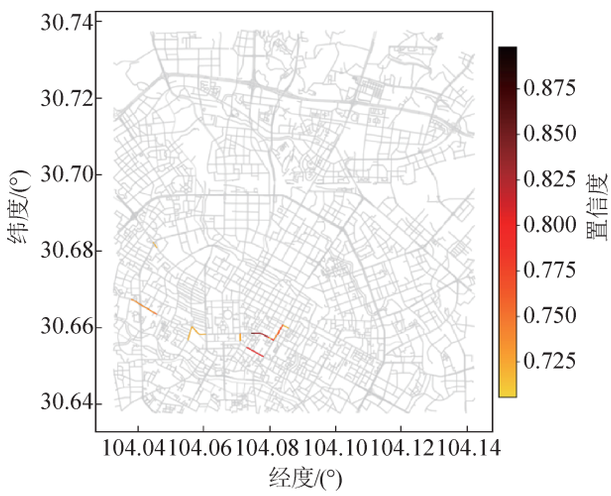

图8　拥挤传播时空关联规则置信度等值图
Fig. 8　Confidence contour map of spatio-temporal association rules for congestion propagation

表2　时序关联规则表(节选)
Table 2　Sequential association rules (excerpt)

ID	*AB*	*conf_A*	*KULC*	*IR*
1	(7366295632, 4548211352, 0, 1),(4548211352, 5530491877, 0, 2)	0.74	0.59	1.66
2	(7366295632, 4548211352, 0, 1), (5530491877, 4548211421, 0, 1), (4548211352, 5530491877, 0, 2)	0.76	0.53	2.47
3	(5530491877, 4548211421, 0, 1), (4548211352, 5530491877, 0, 2), (4548211352, 5530491877, 0, 3)	0.71	0.56	1.76
4	(7366295632, 4548211352, 0, 1), (4548211352, 5530491877, 0, 2), (4548211352, 5530491877, 0, 3)	0.73	0.53	2.26
5	(314622918, 672287905, 0, 1), (1159945489, 4548211258, 0，1), (314622918, 672287905, 0, 2)	0.71	0.57	1.61
6	(1159945489, 4548211258, 0, 1), (314622918, 672287905, 0, 2), (314622918, 672287905, 0, 3)	0.72	0.59	1.61

表2中时序关联规则1~4对应青羊宫附近的拥挤时空传播规律，其中规则1和规则2的时序跨度为2，规则3和规则4的时序跨度为3。规则1表明此路段拥挤从清江东路向蜀都大道方向传播，规则2则说明拥挤从上游的清江东路和下游的信号路口向中游路段(4548211352，5530491877，0)传播，规则3和规则4则表现更加复杂的拥挤传播关系，图9为拥挤关联规则的集中区域。总体而言，时序关联规则挖掘能够与4.1节空间关联规则挖掘结果相印证，同时表现更明确的拥挤传播过程。

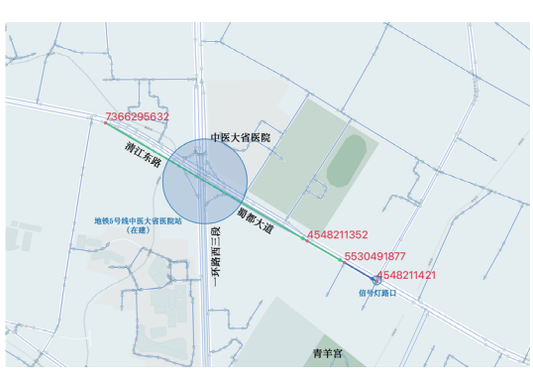

图9　路段拥挤传播图(青羊宫片区)
Fig. 9　Road congestion propagation map (Qingyanggong area)

https://dc-china-simulation.researchcommons.org/journal/vol36/iss1/20
DOI: 10.16182/j.issn1004731x.joss.22-0898

5 结论

本文提出了KNR空间关联规则挖掘算法，成功量化拥挤路段的空间关联性强度。为RElim算法拓展空间邻接的约束规则，使用K邻边概念定义路段之间的空间相邻关系。一方面符合拥挤仅可能在一定范围内路段间传播的基本特性，另一方面规避了大量无价值的频繁项集从而加快了频繁项集挖掘的速度。应用于轨迹数据挖掘，结果显示路网中的拥挤常发路段主要分布在二环内中心城区。其中天府广场、春熙路、青羊宫、西南交大九里校区等片区路段的空间相关性较强，个别路段发生拥挤的关联性强度超过95%。

进一步提出改进SKNR算法，挖掘路网中频繁出现的拥挤传播行为，量化路网拥挤传播的关联强度。拓展KNR算法以满足时序关联规则挖掘的需求，使用提出的算法挖掘车辆轨迹数据中的拥挤传播现象。结果显示拥挤传播现象主要发生在天府广场、春熙路、人民公园、青羊宫、一环路西三段–花牌坊街交接处，且绝大部分拥挤传播现象的时序跨度在30 min以内，长时间、大范围的拥挤传播现象在城市交通网络中并不会频繁发生。

参考文献：

[1] Chandler R E, Herman R, Montroll E W. Traffic Dynamics: Studies in Car Following[J]. Operations Research, 1958, 6(2): 165-184.

[2] Daganzo C F. The Cell Transmission Model: a Dynamic Representation of Highway Traffic Consistent with the Hydrodynamic Theory[J]. Transportation Research Part B: Methodological, 1994, 28(4): 269-287.

[3] Richards P I. Shock Waves on the Highway[J]. Operations Research, 1956, 4(1): 42-51.

[4] Long Jiancheng, Gao Ziyou, Ren Hualing, et al. Urban Traffic Congestion Propagation and Bottleneck Identification[J]. Science in China Series F: Information Sciences, 2008, 51(7): 948-964.

[5] 杨泳. 城市快速路交通拥堵形成、传播规律及消散控制策略研究[D]. 成都: 西南交通大学, 2014.
Yang Yong. Study on Congestion Forming、Propagation Properties and Dissipation Control Strategies of Urban Expressway[D]. Chengdu: Southwest Jiaotong University, 2014.

[6] Saberi M, Hamedmoghadam H, Ashfaq M, et al. A Simple Contagion Process Describes Spreading of Traffic jams in Urban Networks[J]. Nature Communications, 2020, 11(1): 1616.

[7] Befekadu G K. Rare Event Simulation in a Dynamical Model Describing the Spread of Traffic Congestions in Urban Network Systems[M]//Bracken P, Uzunov D I. A Collection of Papers on Chaos Theory and Its Applications. Rijeka: IntechOpen, 2021.

[8] 韦伟. 基于实测数据的道路交通状态特征及拥堵传播规律分析方法[D]. 北京: 北京交通大学, 2017.
Wei Wei. Analytical Method of Traffic Condition Characteristics and Congestion Propagation Rules Based on Practical Measured Data[D]. Beijing: Beijing Jiaotong University, 2017.

[9] 韦伟, 毛保华, 陈绍宽, 等. 基于当期事件识别的拥堵传播特征研究[J]. 交通运输系统工程与信息, 2016, 16(4): 165-170.
Wei Wei, Mao Baohua, Chen Shaokuan, et al. Spatial Propagating Study of Urban Traffic Congestion Based on Current Event Recognition[J]. Journal of Transportation Systems Engineering and Information Technology, 2016, 16(4): 165-170.

[10] 欧吉顺. 异质性城市路网交通拥堵传播模式发现与可视化研究[D]. 南京: 东南大学, 2019.
Ou Jishun. Traffic Congestion Propagation Pattern Recognition and Visualization for Heterogeneous Urban Networks[D]. Nanjing: Southeast University, 2019.

[11] 王启燕. 基于时间序列关联规则挖掘的交通拥堵预测研究[D]. 北京: 北京交通大学, 2020.
Wang Qiyan. Traffic Congestion Prediction Model Based On Time Series Association Mining Rules[D]. Beijing: Beijing Jiaotong university, 2020.

[12] 谭星. 城市主干路交通状态评价与关联规则挖掘研究[D]. 哈尔滨: 哈尔滨工业大学, 2018.
Tan Xing. Study on Traffic State Evaluation and Association Rules Mining on Arterial Roads[D]. Harbin: Harbin Institute of Technology, 2018.

[13] Xu Lin, Yue Yang, Li Qingquan. Identifying Urban Traffic Congestion Pattern from Historical Floating Car Data[J]. Procedia - Social and Behavioral Sciences, 2013, 96: 2084-2095.

[14] Li Qiang, Ge Qian, Miao Lixin, et al. Measuring Variability of Arterial Road Traffic Condition Using Archived Probe Data[J]. Journal of Transportation Systems Engineering and Information Technology, 2012,

12(2): 41-46.

[15] Oruc Altintasi, Hediye Tuydes-Yaman, Kagan Tuncay. Detection of Urban Traffic Patterns From Floating Car Data (FCD) [J]. Transportation Research Procedia, 2017, 22: 382-391.

[16] Xie Dongfan, Wang Meihong, Zhao Xiaomei. A Spatiotemporal Apriori Approach to Capture Dynamic Associations of Regional Traffic Congestion[J]. IEEE Access, 2020, 8: 3695-3709.

[17] Han Jiawei, Kamber M, Pei Jian, 等. 数据挖掘: 概念与技术[M]. 范明, 孟小峰, 译. 3版. 北京: 机械工业出版社, 2012.
Han Jiawei, Kamber M, Pei Jian, et al. Data Mining: Concepts and Techniques[M]. Translated by Fan Ming, Meng Xiaofeng. 3rd ed. Beijing: China Machine Press, 2012.

[18] Christian Borgelt. Keeping Things Simple: Finding Frequent Item Sets by Recursive Elimination[C]// Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations. New York, NY, USA: Association for Computing Machinery, 2005: 66-70.

[19] 刘喜苹, 刘彩苹, 谭义红. 一个新的不需要候选集的挖掘关联规则算法—Relim算法的研究[J]. 计算技术与自动化, 2006, 25(2): 81-84.
Liu Xiping, Liu Caiping, Tan Yihong. Study on a New Mining Association Rule Without Candidate Generation Relim Algorithm[J]. Computing Technology and Automation, 2006, 25(2): 81-84.

[20] Kong Xiangfu, Yang Jiawen, Yang Zhongyu. Measuring Traffic Congestion with Taxi GPS Data and Travel Time Index[C]//Measuring Traffic Congestion with Taxi GPS Data and Travel Time Index. Reston, VA, USA: ASCE, 2015: 3751-3762.

DOI: 10.16182/j.issn1004731x.joss.22-0898