DOI：10.13718/j.cnki.xdzk.2020.11.008

时空亚频繁 co-location 模式挖掘

李新源，　陈红梅，　肖　清，　王丽珍

云南大学 信息学院，昆明 650504

摘要：空间 co-location 模式挖掘是空间数据挖掘的重要分支，在环境保护、公共交通、位置服务和城市计算等领域得到广泛应用. 与基于团实例模型的传统模式相比，基于星型实例模型的空间亚频繁 co-location 模式可以揭示空间特征更丰富的空间关系. 然而，现有空间亚频繁模式没有考虑空间数据的时间特性，而时间却是空间数据的重要维度. 因此，该研究考虑空间实例的位置时变性，基于星型实例模型的时空亚频繁 co-location 模式进行挖掘. 首先，提出了时空亚频繁 co-location 模式及其度量指标：时间亚频繁度；其次，证明了时间亚频繁度的反单调性(向下闭合性)，提出了有效的时空亚频繁模式挖掘算法；最后，通过大量实验，验证了所提算法的有效性及时空亚频繁模式的实用性.

关　键　词：空间数据挖掘；时空数据；时空亚频繁 co-location 模式

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1673－9868(2020)11－0068－09

随着空间数据的激剧增多，空间数据挖掘在测绘、导航、环境保护、公共交通、位置服务和城市计算等领域得到广泛应用. 空间 co-location 模式挖掘是空间数据挖掘的重要分支，旨在挖掘其实例在空间中频繁关联的空间特征的子集，例如小丑鱼通常居住在海葵的触手之间，商场附近总是有 KTV[1]. 传统 co-location 模式采用团实例模型，该模型要求模式实例中的特征实例两两邻近形成团，从而忽略了空间特征间重要的非团的空间关系. 针对这个问题，文献[2-3]提出了更具普适性的星型实例模型及亚频繁 co-location 模式，该模型放松了团约束，可以发现空间特征间更丰富的空间关系.

然而，上述星型实例模型及亚频繁 co-location 模式没有考虑空间数据的时间特性，而时间却是空间数据的关键要素. 在现实世界中，空间实例的生存或位置会随着时间的变化而变化，例如，医院、学校等特征的实例有生命期，人、车等特征的实例会移动. 在这样的时空数据上挖掘 co-location 模式，不同的时间将会得到不同的模式，即 co-location 模式也会随着时间的变化而变化，例如：{人，车，餐厅}是早上 6 点到 8 点的一个 co-location 模式；而{人，办公楼}是上午 9 点到 12 点的一个 co-location 模式. 此外，分析不同时间的 co-location 模式及其随时间的变化规律，将有助于我们进一步了解空间特征间的时空关系，为 co-location 模式在环境保护、公共交通、位置服务和城市计算等领域的深入应用提供决策支持.

基于上述分析，本文考虑空间实例的位置随着时间变化而变化，研究基于星型实例模型的时空亚频繁模式挖掘. 本研究主要面临 2 个方面的挑战：① 基于星型实例模型及亚频繁模式，如何有效地融入时间信息并合理地定义时空亚频繁模式？② 面对计算更为复杂的时空数据，如何高效地挖掘时空亚频繁模式？

收稿日期：2020－10－14

基金项目：国家自然科学基金项目(61662086，61966036)；云南省创新团队项目(2018HC019).

作者简介：李新源(1995－)，女，硕士研究生，主要从事空间数据挖掘研究.

通信作者：陈红梅，博士，副教授.

具体地，本研究主要工作包括：

(1) 分析融入时间信息的方式，提出时空亚频繁 co-location 模式及其度量指标：时间亚频繁度.

(2) 证明时空亚频繁模式的反单调性(向下闭合性)，提出有效的时空亚频繁模式挖掘算法.

(3) 在合成数据集基础上进行大量实验，验证了所提算法的有效性及时空亚频繁模式的实用性.

1　相关工作

根据所采用的实例模型，空间 co-location 模式主要可以分为 2 类：基于团实例模型的 co-location 模式和基于星型实例模型的 co-location 模式.

1.1　基于团实例模型的 co-location 模式

基于团实例模型的 co-location 模式由文献[4]提出，该模型要求模式实例中的特征实例两两邻近形成团，并采用参与度(最小参与率)度量模式的频繁程度. 为了提高模式挖掘效率，文献[4]证明了参与度的反单调性(向下闭合性)，并提出了基于完全连接的 join-based 算法. 针对生成候选模式实例时大量连接操作导致算法效率低的问题，文献[5]通过划分特征实例，提出了基于部分连接的 partial join 算法；文献[6]采用星型邻居实例及模式实例查找方法，提出了无连接的 join-less 算法. 而文献[7-9]采用前缀树降低模式实例生成代价，分别提出了基于 CPI-tree 结构的无连接算法[7]，基于 iCPI-tree 结构的模式挖掘算法[8]，以及基于有序团的极大模式挖掘算法[9].

在团实例模型下，研究者从效用、不确定性、模糊性等方面对 co-location 模式进行了广泛研究[10-12]. 研究者也引入时间维度，进行了时空数据上的 co-location 模式研究. 文献[13]计算各个时间片上的模式参与度，然后通过标准化欧式距离计算最终的模式参与度；文献[14]定义了空间上邻近且时间上持久但不一定邻近的混合对象组，提出了一种混合驱动时空共现模式挖掘算法 MDCOPs；文献[15]提出了 top-k MDCOPs 算法；文献[16]进一步考虑对象的生存周期，提出了 PACOPs 算法；文献[17]采用加权滑动窗口模型，进一步考虑各时间片间的关联关系，在多个时间片上挖掘模式；文献[18]基于动态增量滑动窗口从时空事件流中挖掘时空共现模式. 文献[19]考虑交通数据的时空特性，提出采用参与度和传播影响力度量模式有趣性，挖掘频繁同现且具有强传播影响力的拥堵模式.

1.2　基于星型实例模型的 co-location 模式

基于星型实例模型的 co-location 模式由文献[2-3]提出，该模型放宽了团约束条件，仅需模式实例中的特征实例与中心特征实例邻近，从而可以发现更丰富的空间关系. 文献[2-3]针对传统团实例模型及 co-location 模式的不足，提出星型实例模型及亚频繁 co-location 模式，并证明了亚频繁模式的反单调性，提出了 2 种高效的极大亚频繁模式挖掘算法，即基于前缀树的算法 PTBA 和基于分区的算法 PBA；文献[20]考虑空间特征间的主导关系，定义了亚频繁模式中的主导特征及主导特征模式，提出了有效的挖掘算法.

与上述研究不同，本文考虑空间实例的时间特性，研究时空亚频繁 co-location 模式挖掘.

2　基本概念及问题定义

空间特征表示空间中不同种类的事物(如学校)，空间特征的集合记为 $F=\{f_1, f_2, \cdots, f_n\}$. 空间实例是空间特征出现在具体位置上的实例(如云南大学)，空间实例的集合记为 $S=S_1 \cup S_2 \cup \cdots \cup S_n$，其中 $S_i(1 \leqslant i \leqslant n)$表示特征 f_i 的实例集. 如果两个空间实例 $O_i, O_j \in S$ 的距离 $\mathrm{dis}(O_i, O_j)$小于给定距离阈值 d，即 $\mathrm{distance}(O_i, O_j) \leqslant d$，称这 2 个实例满足空间邻近关系 R，记为 $R(O_i, O_j)$. 空间 co-location 模式 P 是空间特征集合 F 的一个子集，即 $P \subseteq F$. 空间 co-location 模式 P 中空间特征的个数称为模式 P 的阶. 在图 1(a)中，A 是空间特征，A_1 是特征 A 的一个空间实例，实例 A_1 和 C_1 之间的连线表示它们满足空间邻近关系，即 $R(A_1, C_1)$. $\{A, B\}$是一个 2 阶模式. 空间 co-location 模式挖掘的目标是挖掘满足一定条件(如参与度大于最小参与度阈值)的频繁模式. 本研究考虑时空数据中空间实例的位置时变

性，即给定时间片集合 $T=\{t_0, t_1, \cdots t_{m-1}\}$，每个时间片上的空间特征集和空间实例集相同，但是空间实例在不同时间片上的位置可能不同. 图1所示的时空数据集显示了4个特征的18个实例在4个时间片上的位置变化.

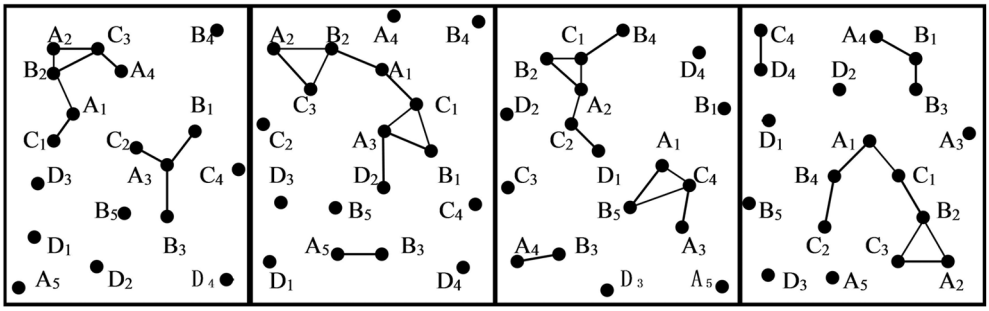


图1 一组4个时间片上的时空数据集

2.1 亚频繁 co-location 模式[2-3]

本节给出基于星型实例模型的亚频繁 co-location 模式相关定义.

定义1 星型邻居实例(SNsI). 给定空间实例 O_i，距离阈值 d，O_i 的星型邻居实例集合 SNsI(O_i)定义为：SNsI(O_i)=$\{O_j \mid \text{distance}(O_i, O_j) \leqslant d\}$.

在图1(a)所示，SNsI(A_1)=$\{A_1, B_2, C_1\}$，SNsI(A_3)=$\{A_3, B_1, B_3, C_2\}$.

定义2 星型参与实例(SPIns). 特征 f 在 co-location 模式 P 中的星型参与实例表示为 SPIns(f, P)，是由 f 的实例组成的集合，其星型邻居实例包含 P 中的所有特征.

在图1(a)中，SPIns(A, (A B C))=$\{A_1, A_2, A_3\}$，SPIns(B, (A B C))=$\{B_2\}$.

定义3 星型参与率(SPR). 特征 f 在 co-location 模式 P 中的星型参与率 SPR(f, P)是特征 f 的星型参与实例数与 f 的实例数的比率，即 SPR(f, P)=|SPIns(f, P)|/|S_f|，其中，S_f 是 f 的实例集合.

在图1(a)中，SPR(A, (A B C))=3/5.

定义4 星型参与度(SPI). 星型参与度 SPI(P)是 co-location 模式 P 中所有特征的最小星型参与率，即 $\text{SPI}(P)=\min\limits_{f\in P}\{\text{SPR}(f, P)\}$.

在图1(a)中，SPI(A B C)=min(3/5, 1/5, 1/4)=1/5.

定义5 亚频繁 co-location 模式(SCP). 如果 co-location 模式 P 的星型参与度不低于给定的空间亚频繁阈值 θ，即 SPI(P)$\geqslant\theta$，则模式 P 称为亚频繁 co-location 模式.

在图1(a)中，假设 θ=1/5，那么模式(A B C)为亚频繁 co-location 模式.

2.2 时空亚频繁 co-location 模式

本文关注空间实例的位置随时间变化而变化，研究基于星型实例模型的时空亚频繁 co-location 模式挖掘，下面给出相关定义.

定义6 时间亚频繁度(TSF). 给定时间片集合 $T=\{t_0, \cdots, t_{m-1}\}$ 上的时空数据集，亚频繁 co-location 模式 P 出现的时间片数占总时间片数的比例称为模式 P 的时间亚频繁度 TSF(P)，

$$\text{TSF}(P)=\frac{|\{t \mid \text{SPI}^t(P)\geqslant\theta, t\in T\}|}{|T|}$$

其中，SPI^t(P)表示时间片 t 上，模式 P 的星型参与度.

定义7 时空亚频繁 co-location 模式(STSCP). 如果亚频繁 co-location 模式 P 的时间亚频繁度不低于给定的时间亚频繁阈值 δ，即 TSF(P) $\geqslant\delta$，则模式 P 称为时空亚频繁 co-location 模式.

问题定义：给定时间片集合 $T=\{t_0, \cdots, t_{m-1}\}$ 上的时空数据集以及距离阈值 d、空间亚频繁阈值 θ、时间亚频繁阈值 δ，时空亚频繁 co-location 模式挖掘就是找出时空数据集中所有满足阈值的模式.

3　挖掘算法

首先，提出一种朴素算法，然后，证明时间亚频繁度的反单调性(向下闭合性)，并基于反单调性，提出新的高效的时空亚频繁模式挖掘算法(STSCP，SpatioTemporal Sub-prevalent Co-location Patterns mining algorithm).

3.1　朴素算法

朴素算法的基本思想是：首先，对每个时间片，使用空间亚频繁模式挖掘算法，挖掘所有空间亚频繁模式；然后，在所有时间片上，计算这些模式的时间亚频繁度，以挖掘时空亚频繁模式. 朴素算法会尽早删除空间上不满足空间亚频繁阈值的模式，但在生成所有时间片的空间亚频繁模式之前，它不会尽早删除时间上不满足时间亚频繁阈值的模式，这导致了不必要的计算成本.

3.2　STSCP 算法

根据文献[2-3]，星型参与度具有反单调性(向下闭合性)，即给定时间片 t 及其上的两个模式 P_i 和 P_j，如果 $P_i \subseteq P_j$，则 $\mathrm{SPI}^t(P_i) \geqslant \mathrm{SPI}^t(P_j)$. 因此，如果 P_i 不是空间亚频繁模式(即 $\mathrm{SPI}^t(P_i) < \theta$)，则 P_j 也不是空间亚频繁模式(即 $\mathrm{SPI}^t(P_j) < \theta$).

本研究所提的时间亚频繁度也具有反单调性(向下闭合性)，下面给予证明.

引理 1(时间亚频繁度 TSF 的反单调性)　给定时间片集 T 及其上的 2 个模式 P_i 和 P_j，如果 $P_i \subseteq P_j$，则 $\mathrm{TSF}(P_i) \geqslant \mathrm{TSF}(P_j)$.

证　对于任一时间片 $t \in T$，因为 $\mathrm{SPI}^t(P_i) \geqslant \mathrm{SPI}^t(P_j)$，所以如果 $\mathrm{SPI}^t(P_j) \geqslant \theta$，则 $\mathrm{SPI}^t(P_i) \geqslant \theta$. 因此有 $\{t \mid \mathrm{SPI}^t(P_j) \geqslant \theta, t \in T\} \subseteq \{t \mid \mathrm{SPI}^t(P_i) \geqslant \theta, t \in T\}$. 所以，$\mathrm{TSF}(P_i) \geqslant \mathrm{TSF}(P_j)$.

根据引理 1，有如下引理 2.

引理 2　给定时间片集 T 及其上的 2 个模式 P_i 和 P_j，如果 $P_i \subseteq P_j$ 且 P_i 不是时空亚频繁模式(即 $\mathrm{TSF}(P_i) < \delta$)，则 P_j 也不是时空亚频繁模式($\mathrm{TSF}(P_j) < \delta$).

利用引理 2，本研究提出新的高效的时空亚频繁模式挖掘算法 STSCP. 算法 STSCP 的基本思想是采用“候选生成—亚频繁测试”方式，从 1 阶开始，逐阶挖掘时空亚频繁模式，并在候选生成中，利用引理 2 删除不可能满足条件的候选模式. 具体地，仅用 k 阶亚频繁模式，连接生成 $k+1$ 候选模式；如果 $k+1$ 阶候选模式的某个 k 阶子模式不是亚频繁模式，则删除该 $k+1$ 阶候选模式；测试剩余候选模式是否是亚频繁模式. 算法 STSCP 描述如表 1 所示.

在算法中，步骤 1－2 用于为每个时间片生成星型邻居实例集；步骤 5－12 给出一个迭代过程，直至生成的高阶时空亚频繁模式集为空. 该迭代过程的功能解释如下.

生成候选时空亚频繁模式(步骤 6)：仅用 k 阶时空亚频繁模式，连接生成 $k+1$ 阶候选时空模式，如果 $k+1$ 阶候选时空模式的某个 k 阶子模式不是时空亚频繁模式，则删除该 $k+1$ 阶候选时空模式.

生成候选空间亚频繁模式(步骤 8)：对时间片 t，初始时，$k+1$ 阶候选时空模式为其上的 $k+1$ 阶候选空间模式. 如果 $k+1$ 阶候选空间模式的某个 k 阶子模式不是其上的空间亚频繁模式，则从时间片 t 上删除该 $k+1$ 阶候选空间模式.

生成星型参与实例(步骤 9)：对时间片 t，生成其上候选空间模式的星型参与实例.

生成空间亚频繁模式(步骤 10)：对时间片 t，计算其上 $k+1$ 阶候选空间模式的星型参与度，生成 $k+1$ 阶空间亚频繁模式.

生成时空亚频繁模式(步骤 11)：采用时间亚频繁表记录时间片 0～t 上的 $k+1$ 阶空间亚频繁模式及其当前的时间亚频繁度. 如果 $k+1$ 阶空间亚频繁模式的时间亚频繁度小于 $\delta-(1-(t+1)/|T|)$，则删除该 $k+1$ 阶空间亚频繁模式(即 $k+1$ 阶候选时空模式)，后继时间片不需再处理，从而尽早删除不满足时间亚频繁阈值的模式，减少不必要的计算，提高挖掘效率.

表1 STSCP算法描述

STSCP算法

输入：ST：时空数据集，其包含空间特征集F，空间实例集S，时间片集T
d：距离阈值
θ：空间亚频繁阈值
δ：时间亚频繁阈值
输出：时空亚频繁模式集
变量：k：模式的阶
t：时间片
$D(t)$：时间片t上的星型邻居实例集合
$T_k(t)$：时间片t上的k阶空间亚频繁模式的星型参与实例集合
$CSSCP_k(t)$：时间片t上的k阶候选空间亚频繁模式
$CSTSCP_k$：k阶候选时空亚频繁模式
$SSCP_k(t)$：时间片t上的k阶空间亚频繁模式
$STSCP_k$：k阶时空亚频繁模式
1. for each t in T {
2. D(t)=gen_n_inst(ST, d) //生成星型邻居实例
3. $SSCP_1(t)=F$}
4. $STSCP_1=F$, $k=1$
5. while ($STSCP_k$ is not empty) {
6. $CSTSCP_{k+1}$=gen_cand_st($STSCP_k$) //生成$k+1$阶候选时空亚频繁模式
7. for each t in T {
8. $CSSCP_{k+1}$(t)=gen_cand_s($CSTSCP_{k+1}$, $SSCP_k(t)$) //生成$k+1$阶候选空间亚频繁模式
9. $T_{k+1}(t)$=gen_p_inst($CSSCP_{k+1}(t)$, $D(t)$) //生成$k+1$阶星型参与实例
10. $SSCP_{k+1}$(t)=find_prev_s($T_{k+1}(t)$, $CSSCP_{k+1}(t)$, θ) //生成$k+1$阶空间亚频繁模式
11. $STSCP_{k+1}$=find_prev_st($SSCP_{k+1}(t)$, δ) //生成$k+1$阶时空亚频繁模式 }
12. $k=k+1$ }
13. return union ($STSCP_2$, …, $STSCP_{k+1}$)

4 实验结果与分析

4.1 实验设置

数据集：根据文献[2-3]提供的方法，生成8个不同规模的合成数据集，这些数据集的生成参数如表2第1—4行所示.

对比算法：为了评估本研究所提的朴素算法和STSCP算法，选用空间极大亚频繁模式挖掘算法PTBA[2-3]，以及将PTBA与时间亚频繁度结合形成的算法PTBA* 作为基准算法.

表2 数据集生成参数及实验参数

参数	实验一	实验二	实验三	实验四	实验五	实验六	实验七	实验八
空间范围	5 000×5 000	5 000×5 000	5 000×5 000	5 000×5 000	5 000×5 000	5 000×5 000	5 000×5 000	3 000×3 000
实例数	20 000～120 000	20 000	20 000	20 000	20 000	20 000	80 000	20 000
特征数	25	20～50	25	25	25	25	30	20
时间片数	10	10	5～50	10	10	10	10	10
空间亚频繁阈值	0.4	0.4	0.4	0.3～0.6	0.4	0.4	0.5	0.5
时间亚频繁阈值	0.6	0.6	0.6	0.6	0.2～1.0	0.6	0.1～1.0	0.1～1.0
距离阈值	60	60	60	60	60	40～120	60	60

4.2　算法效率分析

首先，从实例数、特征数、时间片数、空间亚频繁阈值、时间亚频繁阈值、距离阈值 6 个方面，分析它们对算法效率的影响.

4.2.1　实例数量的影响

采用表 2 实验一所示数据集及参数，评估实例数对算法效率的影响，结果如图 2 所示. 朴素算法的执行时间始终高于 STSCP 和 PTBA*. 当实例数量较小时，数据集比较稀疏，STSCP 比 PTBA* 快. 随着实例数越来越多，数据集越来越密集，相比 PTBA*，STSCP 执行时间增加加快. 这是因为 STSCP 是从 1 阶(低阶)开始，自下而上生成所有时空亚频繁模式，而 PTBA* 算法是从 n 阶(高阶)开始，自上而下生成极大空间亚频繁模式，在稀疏数据集上，极大模式倾向于低阶模式，所以 STSCP 比 PTBA* 更快. 而当数据密集时，极大模式倾向于高阶模式，所以 PTBA* 比 STSCP 更快.

4.2.2　特征数量的影响

采用表 2 实验二所示数据集及参数，评估特征数对算法效率的影响，结果如图 3 所示. 随着特征数的增加，3 个算法的执行时间增加，但是 STSCP 均比 PTBA* 和朴素算法快. 在实例数不变，特征数增加的情况下，每个特征的平均实例数减少，使得每个时间片上的数据稀疏，从而 STSCP 比 PTBA* 快.

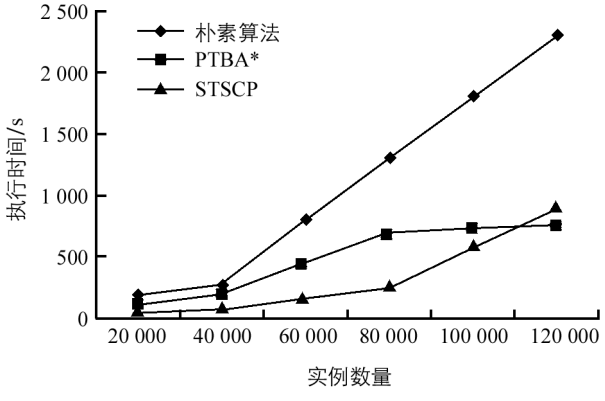


图 2　实例数量对 3 种算法的影响

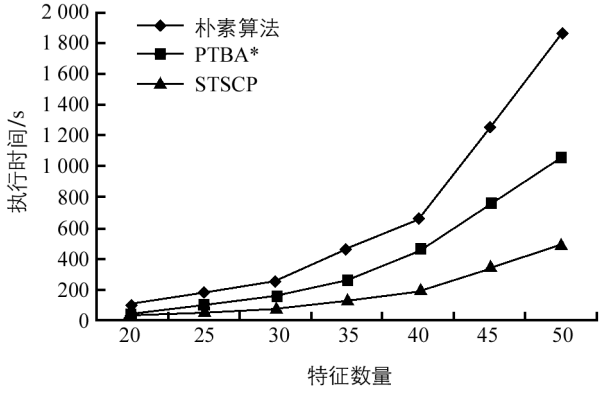


图 3　特征数量对 3 种算法的影响

4.2.3　时间片数量的影响

采用表 2 实验三所示数据集及参数，评估时间片数对算法效率的影响，结果如图 4 所示. 3 个算法的执行时间随着时间片数的增加而增加，STSCP 比朴素算法快是因为 STSCP 会尽早删除非时空亚频繁模式. 而 STSCP 比 PTBA* 快是因为随着时间片数的增加，每个时间片上的数据变得稀疏.

4.2.4　时间亚频繁阈值的影响

采用表 2 实验四所示数据集及参数，评估时间亚频繁阈值对算法效率的影响，结果如图 5 所示. 从图中可以看出，3 个算法的执行时间随着时间亚频繁阈值的增加而减少，且 STSCP 比 PTBA* 和朴素算法具有更高的效率. 此外，由于朴素算法不能尽早删除不满足时间亚频繁阈值的模式，它的执行时间减少较缓慢.

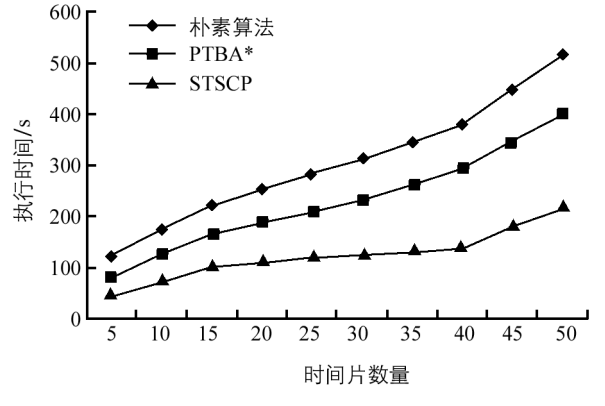


图 4　时间片数量对 3 种算法的影响

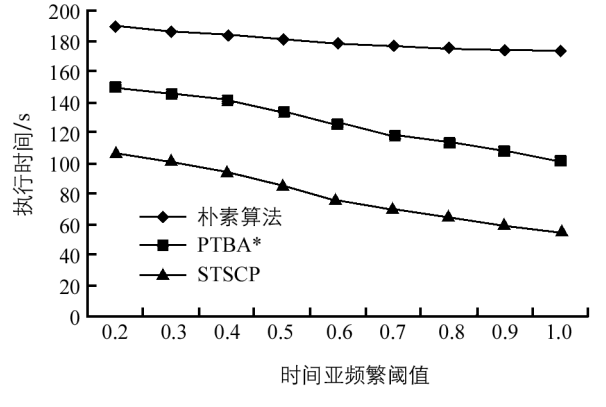


图 5　时间亚频繁阈值对 3 种算法的影响

4.2.5 空间亚频繁阈值的影响

采用表2实验五所示数据集及参数，评估空间亚频繁阈值对算法效率的影响，结果如图6所示. 随着空间亚频繁阈值的增加，STSCP、PTBA*、朴素算法的执行时间减少. 空间亚频繁阈值较小时，朴素算法的执行时间远高于STSCP和PTBA*，这是因为朴素算法往往会生成大量非时空亚频繁模式. 对于STSCP和PTBA*，当空间亚频繁阈值较小时，STSCP的执行时间比PTBA*减少速度更快，而当空间亚频繁阈值达到0.5以后，PTBA*的执行时间比STSCP减少速度更快.

4.2.6 邻近距离阈值的影响

采用表2实验六所示数据集及参数，评估邻近距离阈值对算法效率的影响，结果如图7所示. 随着邻近距离阈值增大，STSCP、PTBA*、朴素算法的执行时间呈上升趋势. 当距离阈值小于100时，虽然STSCP比朴素算法和PTBA*更快，但3种算法的执行时间都在快速增加，这是因为邻近距离阈值增加可能导致2阶模式空间亚频繁模式增加. 当距离阈值大于100后，PTBA*比STSCP快，这是因为邻近区域足够大，PTBA*可以比STSCP更早停止.

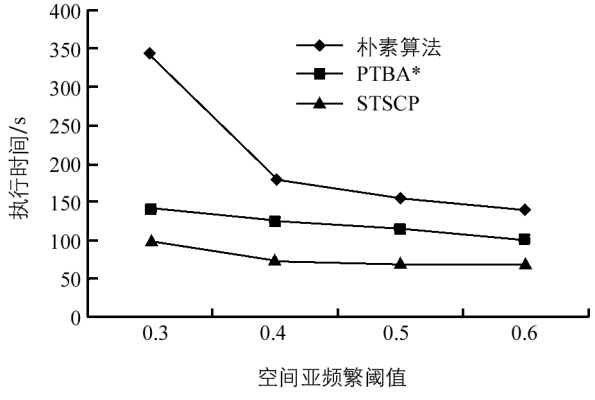


图6 空间亚频繁阈值对3种算法的影响

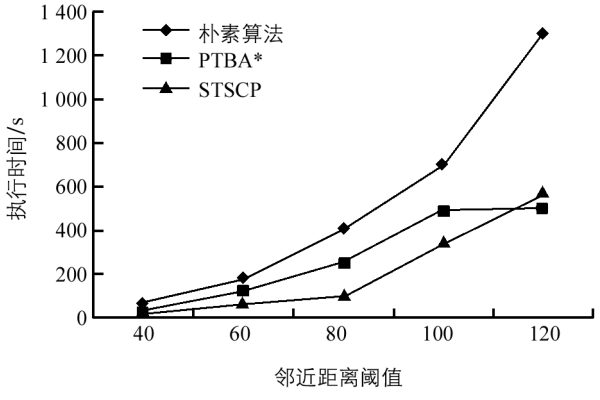
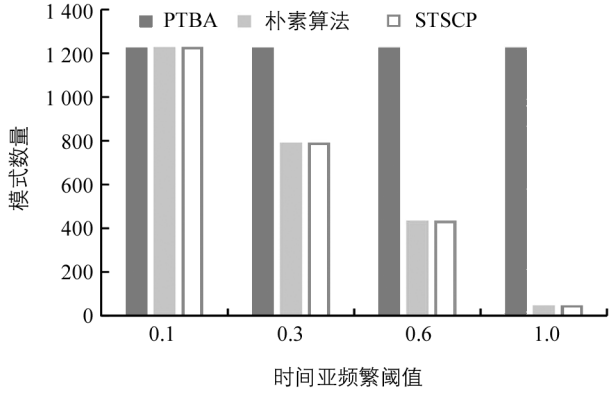


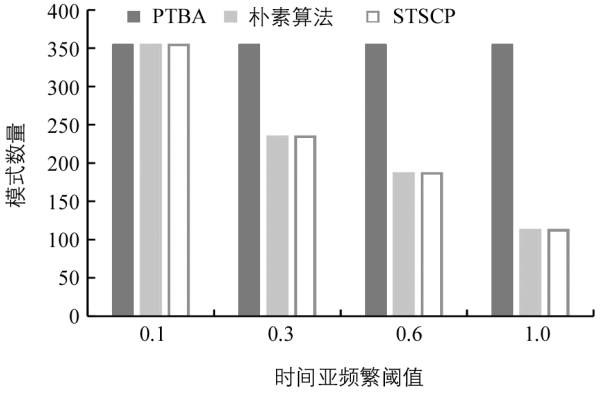
图7 邻近距离阈值对3种算法的影响

4.3 挖掘结果分析

为了分析本研究所提时空亚频繁模式与文献[2-3]所提空间亚频繁模式的差异，采用表2实验七和实验八所示数据集及参数，评估时间亚频繁阈值对本文所提算法STSCP与文献[2-3]所提算法PTBA挖掘结果的影响，结果如图8所示. 随着时间亚频繁阈值的增加，PTBA挖掘出的模式数量一直保持不变，而STSCP挖掘出的模式数量一直在减少，这是因为PTBA是在空间上挖掘亚频繁模式，挖掘结果不受时间亚频繁阈值的影响，而对于STSCP随着时间亚频繁阈值的提高，要求模式出现在更多的时间片上，从而符合要求的模式越来越少. 从图8可以看到，当时间亚频繁阈值为0.1时，模式只需要在一个时间片上出现，这相当于没有时间约束，此时STSCP挖掘的模式数量与PTBA相同，而当时间亚频繁阈值为1时，要求模式在所有时间片上出现，符合要求的模式数量最少，所以STSCP挖掘的模式数量远低于PTBA.



a. 合成数据集1



b. 合成数据集2

图8 不同时间亚频繁阈值下的模式数量

在现实中，用户需求的模式可能是多次出现的模式，也可能是瞬时出现的模式. 本研究所提算法 STSCP 可以根据用户设定的时间亚频繁阈值，挖掘不同出现频度的时空亚频繁模式，而 PTBA 只能挖掘一种出现频度的时空亚频繁模式，因此 STSCP 可以挖掘更符合用户需求、更具有丰富语义也更具有实用价值的模式.

5　结　论

时间是空间数据的重要维度，在现实世界中，空间实例的生存或位置会随着时间的变化而变化，而现有星型实例模型及亚频繁 co-location 模式没有考虑空间数据的时间特性. 因此，本研究基于星型实例模型的时空亚频繁 co-location 模式挖掘. 首先分析了融入时间的方式，定义了时空亚频繁模式的指标，提出了时空亚频繁 co-location 模式. 然后，基于星型参与度的反单调性证明了时间亚频繁度满足向下闭合性，提出了有效的时空亚频繁模式挖掘算法. 最后，在合成数据集基础上进行大量实验，验证了所提算法能够挖掘到更丰富、更有价值的时空亚频繁 co-location 模式.

参考文献：

[1] 范高峰. 带时间约束的 co-location 模式挖掘 [D]. 昆明：云南大学，2012.

[2] WANG L Z，BAO X G，ZHOU L H，et al. Maximal Sub-prevalent Co-location Patterns and Efficient Mining Algorithms [M] //Lecture Notes in Computer Science. Cham：Springer International Publishing，2017：199-214.

[3] 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.

[4] HUANG Y，SHEKHAR S，XIONG H. Discovering Colocation Patterns from Spatial Data Sets：a General Approach [J]. IEEE Transactions on Knowledge and Data Engineering，2004，16(12)：1472-1485.

[5] YOO J S，SHEKHAR S，SMITH J，et al. A Partial Join Approach for Mining Co-location Patterns [C] //Proceedings of the 12th annual ACM international workshop on Geographic information systems - GIS '04. November 12-13，2004. Washington DC，USA. New York：ACM Press，2004：241-249.

[6] 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.

[7] WANG L Z，BAO Y Z，LU J，et al. A New Join-less Approach for Co-location Pattern Mining [C] //2008 8th IEEE International Conference on Computer and Information Technology. July 8-11，2008，Sydney，NSW，Australia. IEEE，2008：197-202.

[8] 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.

[9] WANG L Z，ZHOU L H，LU J，et al. An Order-clique-based Approach for Mining Maximal Co-locations [J]. Information Sciences，2009，179(19)：3370-3382.

[10] 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.

[11] LIU Z，HUANG Y. Mining Co-locations under Uncertainty [M] //Advances in Spatial and Temporal Databases. Berlin，Heidelberg：Springer Berlin Heidelberg，2013：429-446.

[12] 雷　乐，王丽珍，肖　清. 空间 co-location 模式挖掘中的模糊技术初探 [J]. 计算机工程与应用，2019，55(21)：158-166.

[13] YOO J S，SHEKHAR S，KIM S，et al. Discovery of Co-evolving Spatial Event Sets [C] //Proceedings of the 2006 SIAM International Conference on Data Mining. Philadelphia，PA：Society for Industrial and Applied Mathematics，2006：306-315.

[14] CELIK M，SHEKHAR S，ROGERS J P，et al. Mixed-Drove Spatiotemporal Co-Occurrence Pattern Mining [J]. IEEE Transactions on Knowledge and Data Engineering，2008，20(10)：1322-1335.

[15] CELIK M，SHEKHAR S，ROGERS J P，et al. Mining at most Top-K% Mixed-drove Spatio-temporal Co-occurrence Patterns：a Summary of Results [C] //2007 IEEE 23rd International Conference on Data Engineering Workshop. April

17-20, 2007, Istanbul, Turkey. IEEE, 2007: 565-574.

[16] CELIK M. Discovering Partial Spatio-temporal Co-occurrence Patterns [C] //Proceedings 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services. June 29 - July 1, 2011, Fuzhou, China. IEEE, 2011: 116-120.

[17] QIAN F, YIN L, HE Q M, et al. Mining Spatio-temporal Co-location Patterns with Weighted Sliding Window [C] // 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems. November 20-22, 2009, Shanghai, China. IEEE, 2009: 181-185.

[18] HUO J T, ZHANG J Z, MENG X F. On Co-occurrence Pattern Discovery from Spatio-temporal Event Stream [M] // Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013: 385-395.

[19] YANG L, WANG L Z. Mining Traffic Congestion Propagation Patterns Based on Spatio-temporal Co-location Patterns [J]. Evolutionary Intelligence, 2020, 13(2): 221-233.

[20] 马 董, 陈红梅, 王丽珍, 等. 空间亚频繁 co-location 模式的主导特征挖掘 [J]. 计算机应用, 2020, 40(2): 465-472.

Spatiotemporal Sub-prevalent Co-location Pattern Mining

LI Xin-yuan, CHEN Hong-mei, XIAO Qing, WANG Li-zhen

School of Information Science and Engineering, Yunnan University, Kunming 650504, China

Abstract: Spatial co-location pattern mining is an important research area, and has been widely applied in various fields such as environment protection, public transport, location-based services and urban computing. Compared with the traditional pattern based on clique instances, the spatial sub-prevalent pattern based on star instances can reveal richer spatial correlations among features. However, the temporal characteristic of spatial data, which is an important dimension of spatial data, is not considered by the current spatial sub-prevalent pattern. Hence, in this paper, a spatiotemporal sub-prevalent pattern based on star instances is presented by analyzing spatial instances whose locations change over time. Firstly, a metric to measure the pattern called "time sub-frequency" is proposed. Then, the anti-monotonicity of the metric is proven, and an efficient algorithm for mining the pattern is designed. Finally, the validity of the algorithm presented herein and the applicability of the pattern are verified through a number of experiments.

Key words: spatial data mining; spatiotemporal data; spatiotemporal sub-prevalent co-location pattern

责任编辑 包 颖