同位模式挖掘

钱 烽
qf6101 at gmail.com

提纲

- 同位模式挖掘简介
- 相关工作
- 主要贡献
  - 空间同位模式挖掘
  - 区域同位模式挖掘
  - 时空同位模式挖掘
  - 同位轨迹挖掘
- 总结与展望

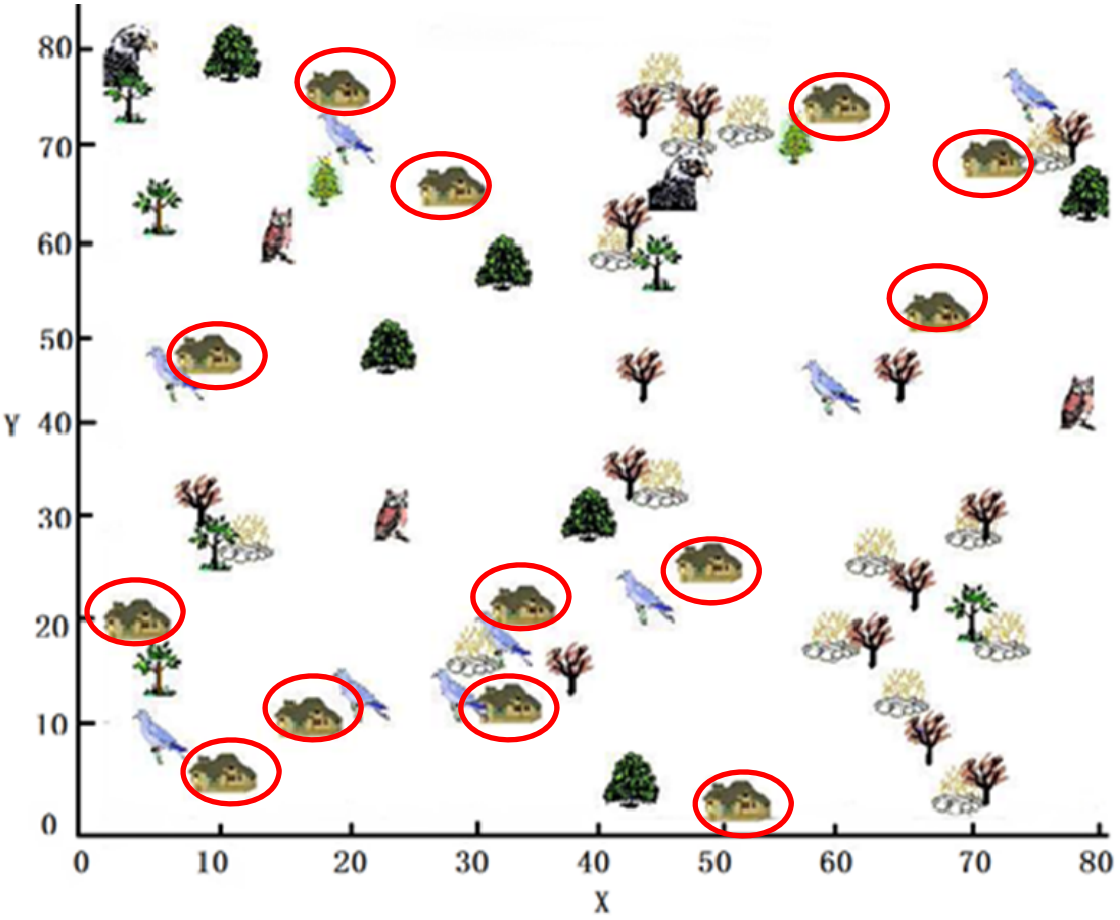
同位模式挖掘简介

- 空间数据挖掘的关键技术
  - 空间分类
  - 空间聚类
  - 空间孤立点检测
  - 同位模式挖掘（本文的研究内容）
- 同位模式挖掘：
  - 在空间数据集中发现一些空间特征的子集，这些子集的空间实例频繁地聚集在一起

空间数据集

- 空间特征：

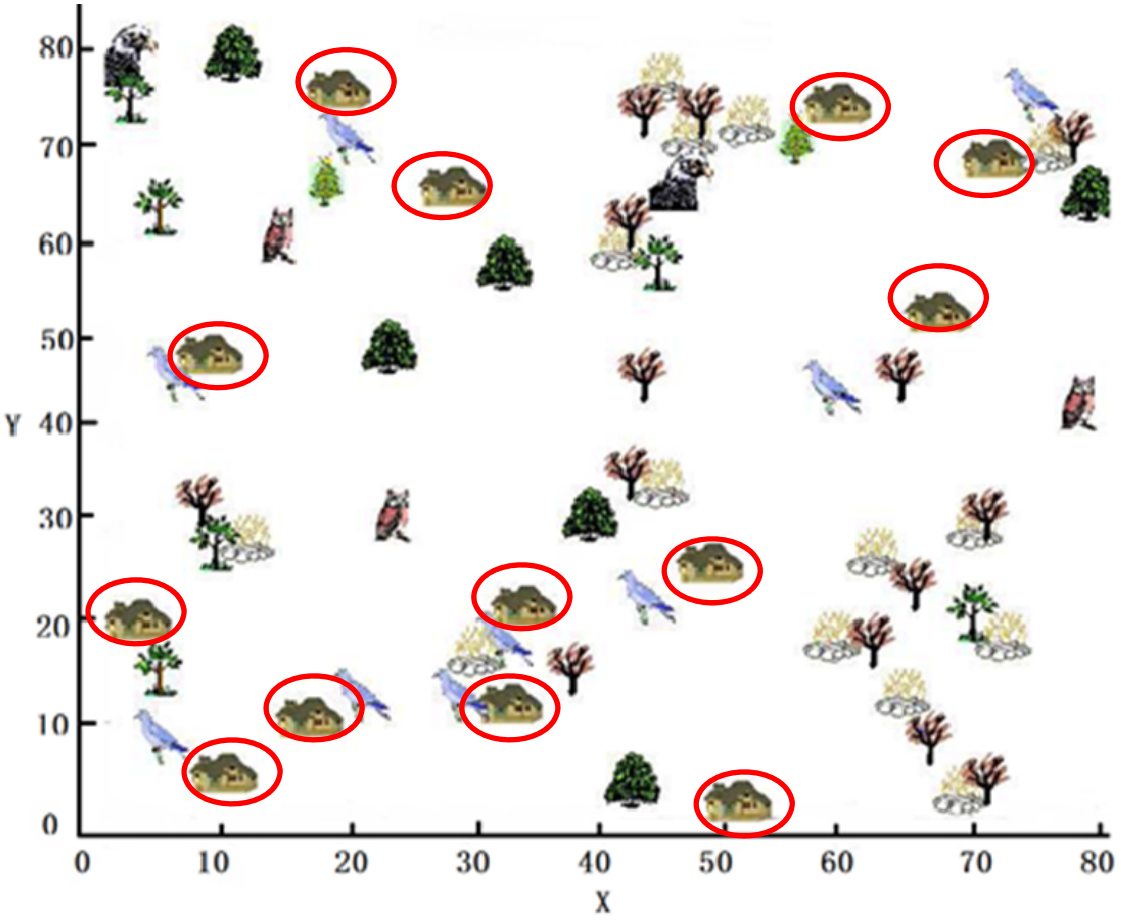
- 空间实例：
  - 空间特征在二维或三维空间中的具体位置
  - “房屋  ”为例



空间数据集的例子（摘自《空间数据库》）

空间数据集

- 其他应用中的具体空间特征：
  - 疾病类型、犯罪类型、商业类型、服务类型...



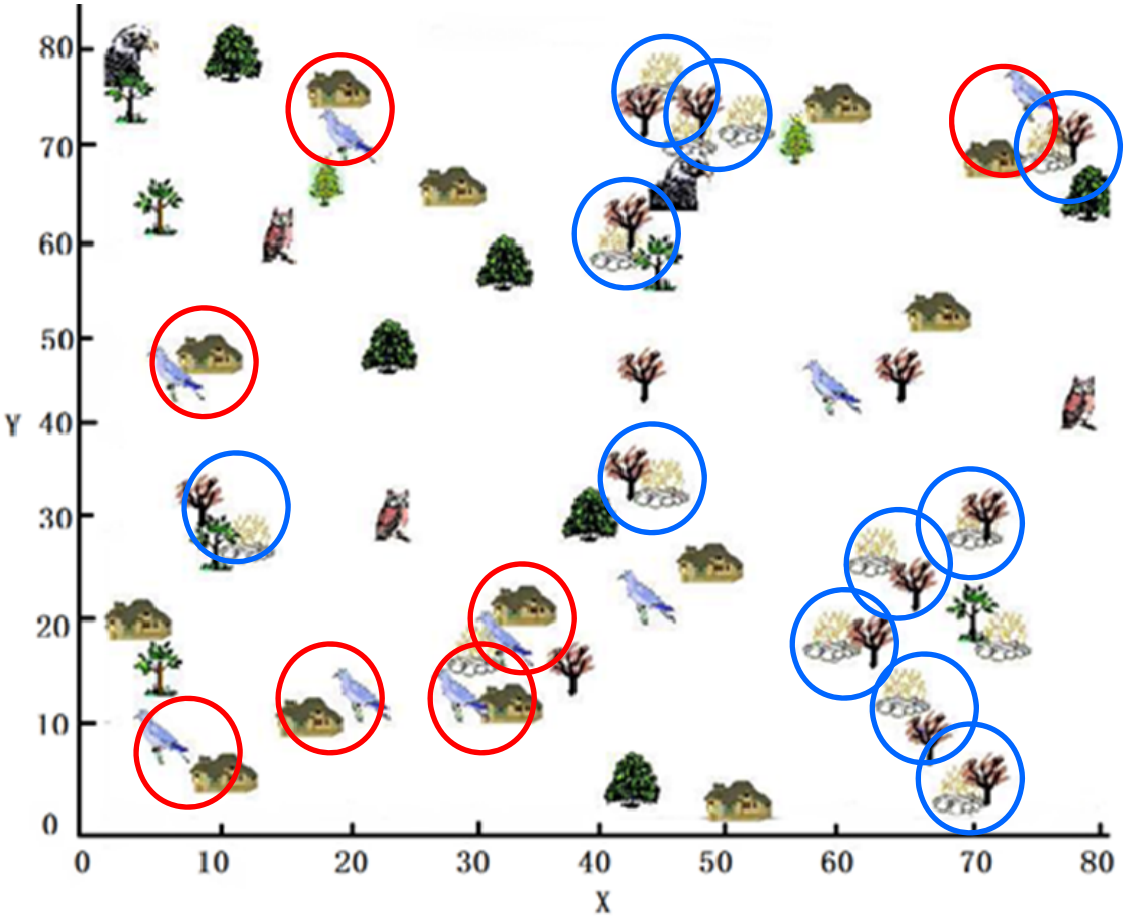
空间数据集的例子（摘自《空间数据库》）

同位模式

- 同位模式：
  - 空间实例频繁地聚集在一起的空间特征子集
  - 例如：

和

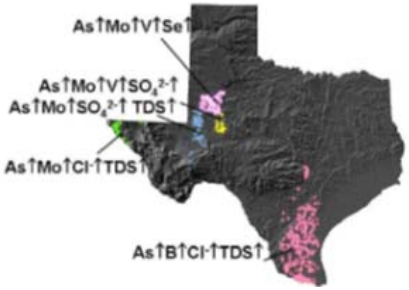
- 同位模式挖掘：
  - 在空间或时空数据集中挖掘同位模式的过程



空间数据集的例子（摘自《空间数据库》）

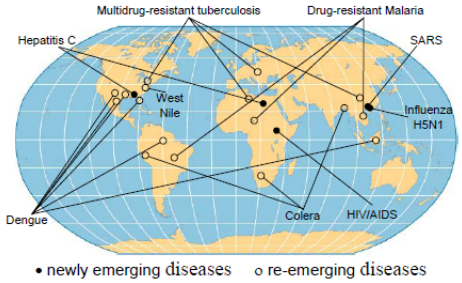
应用领域

移动商务

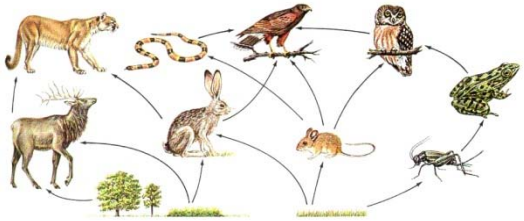


地球科学

公共安全



公共卫生



生物科学

交通物流

同位模式挖掘的四类问题

- 挖掘局部空间区域包含的相异同位模式集
- 关注同位模式的同时也关注局部空间区域

- 同位模式挖掘的的一般性任务
- 在纯粹的空间数据集中挖掘同位模式的过程

- 挖掘同位模式在时空数据集中的传播轨迹

- 在时空数据集中挖掘同位模式
- 为每个空间实例增加一个时间标签

本文从上述四个方面分别展开具体工作

提纲

- 同位模式挖掘简介
- 相关工作
- 主要贡献
  - 空间同位模式挖掘
  - 区域同位模式挖掘
  - 时空同位模式挖掘
  - 同位轨迹挖掘
- 总结与展望

相关工作

- 空间同位模式挖掘
  - [Y. Huang, TKDE 2004]
  - [J.S. Yoo, TKDE 2006]
  - [X. Xiao, SIGSPATIAL 2008]
  - …
- 区域同位模式挖掘
  - [M. Celik, ICDM 2007]
  - [C.F. Eick, SIGSPATIAL 2008]
  - …

相关工作

- 时空同位模式挖掘
  - [J.S. Yoo, SDM 2006]
  - [M. Celik, ICDM 2006]
  - [Y. Huang, TKDE 2008]
  - …
- 移动模式和轨迹模式挖掘
  - [N. Mamoulis, SIGKDD 2004]
  - [F. Giannotti, SIGKDD 2007]
  - …

参考文献

- [Y. Huang, TKDE 2004] Y. Huang, S. Shekhar, H. Xiong (2004) *Discovering colocation patterns from spatial datasets: a general approach*. IEEE Trans Knowl Data Eng 16(12):1472–1485.
- [J.S. Yoo, TKDE 2006] J.S. Yoo, S. Shekhar (2006) *A joinless approach for mining spatial colocation patterns*. IEEE Trans Knowl Data Eng 18(10):1323–1337.
- [X. Xiao, SIGSPATIAL 2008] X. Xiao, X. Xie, Q. Luo, W.Y. Ma (2008) *Density based co-location pattern discovery*. In: Proceedings of the 16th ACM SIGSPATIAL international conference on advances in geographic information systems, Irvine, USA, November 5–7, pp 1–10.
- [M. Celik, ICDM 2007] M. Celik, S. Shekhar, J.P. Rogers, J.A. Shine, J.S. Yoo (2006) *Mixed-drove spatio-temporal co-occurrence pattern mining: a summary of results*. In: Proceedings of the 6th international conference on data mining, Hong Kong, China, December 18–22, pp 119–128.

参考文献

- [C.F. Eick, SIGSPAITAL 2008] C.F. Eick, R. Parmar, W. Ding, T.F. Stepinski, and J.P. Nicot. *Finding regional co-location patterns for sets of continuous variables in spatial datasets*. In: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pages 1–110, Irvine, USA, November 5–7, 2008.
- [J.S. Yoo, SDM 2006] J.S. Yoo, S. Shekhar, S. Kim, M. Celik (2006) *Discovery of co-evolving spatial event sets*. In: Proceedings of the 6th SIAM international conference on data mining, Bethesda, USA, November 20–22, pp 306–315.
- [M. Celik, ICDM 2006] M. Celik, S. Shekhar, J.P. Rogers, J.A. Shine, J.S. Yoo (2006) *Mixed-drove spatio-temporal co-occurrence pattern mining: a summary of results*. In: Proceedings of the 6th international conference on data mining, Hong Kong, China, December 18–22, pp 119–128.

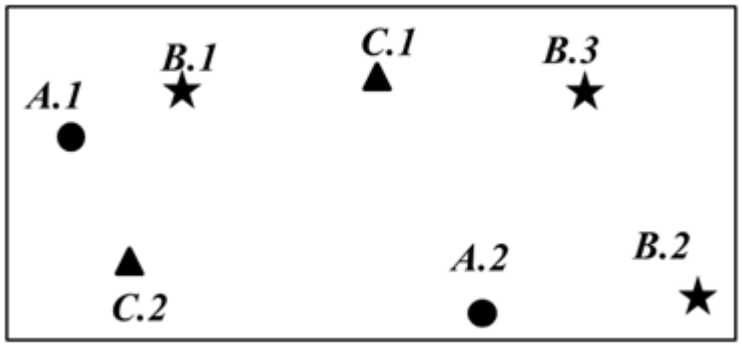
参考文献

- [Y. Huang, TKDE 2008] Y. Huang, L. Zhang, P. Zhang (2008) *A framework for mining sequential patterns from spatio-temporal event data sets*. IEEE Trans Knowl Data Eng 20(4):433–448.
- [N. Mamoulis, SIGKDD 2004] N. Mamoulis, H. Cao, G. Kollios, M. Hadjieleftheriou, Y. Tao, D.W. Cheung (2004) *Mining, indexing, and querying historical spatiotemporal data*. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 236-245, Seattle, USA, August 22-25, 2004.
- [F. Giannotti, SIGKDD 2007] F. Giannotti, M. Nanni, F. Pinelli, D. Pedreschi (2007) *Trajectory Pattern Mining*. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 330-339, San Jose, USA, August 12-15, 2007.
- ...

提纲

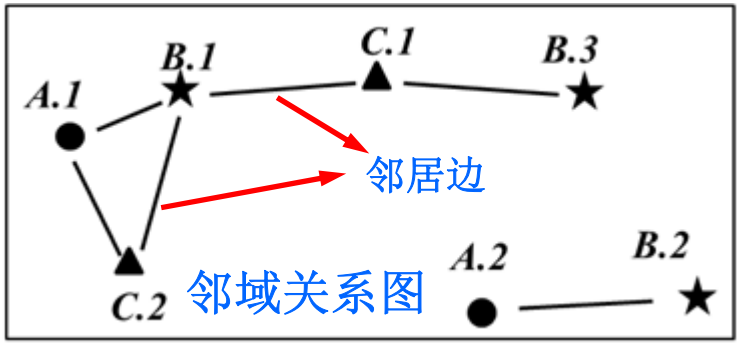
- 同位模式挖掘简介
- 相关工作
- 主要贡献
  - 空间同位模式挖掘
  - 区域同位模式挖掘
  - 时空同位模式挖掘
  - 同位轨迹挖掘
- 总结与展望

传统的三阶段策略



步骤 1

距离阈值



空间特征	空间实例
A ●	A.1, A.2
B ★	B.1,B.2,B.3
C ▲	C.1,C.2

步骤 2　收集团实例

{A,B}	{A,C}	{B,C}	{A,B,C}
A.1,B.1	A.1,C.2	B.1,C.1	A.1,B.1,C.2
A.2,B.2		B.3,C.1	
		B.1,C.2	

团实例（clique instances）

计算流行度

流行度阈值

步骤 3

空间同位模式	{A,B}	{B,C}

[J.S. Yoo, SDM 2006]
[M. Celik, ICDM 2006]
[X. Xiao, SIGSPATIAL 2008]

三阶段策略的局限性

- 距离阈值
  - 过小：忽略流行同位模式的团实例
  - 过大：团实例的收集带来巨大的工作量
- 流行度阈值
  - 过小：产生众多平凡同位模式项集
  - 过大：可能低估一些重要的同位模式
- 其他挖掘方法：
  - 距离阈值→其他邻域约束：相似的阈值影响
  - 空间统计方法：挖掘结果不完整

迭代式同位模式挖掘

- 流行度回报：

$$R(l_i) = |\tilde{l}_i| \cdot (\lambda \tilde{B}(l_i) + (1-\lambda) I(l_i))$$

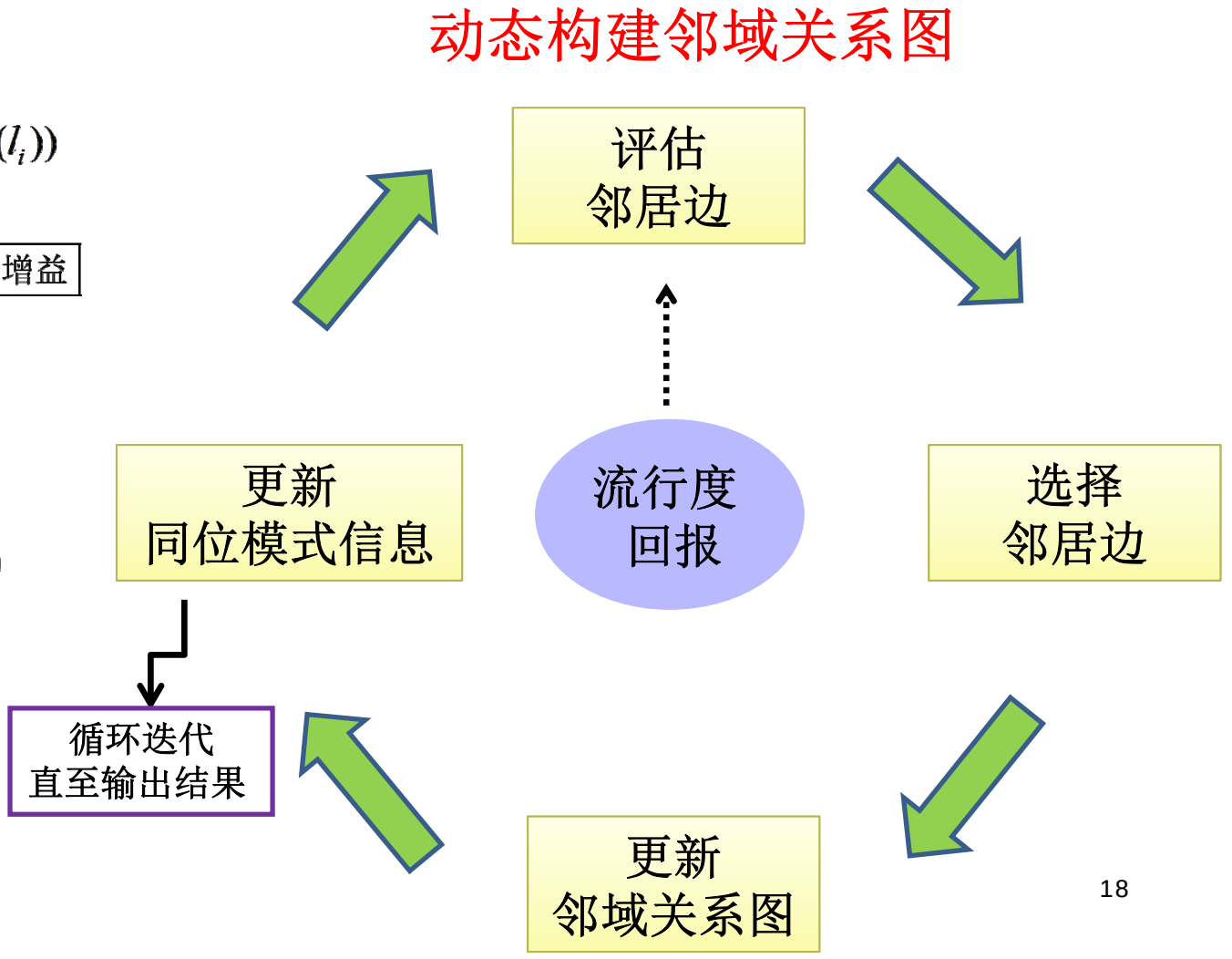
平衡系数；邻居边；不平衡增益；流行度增益

- 贪心选择：

$$l_t = \arg\max_{l_t \in L \setminus G_{t-1}} R(l_t)$$

$$S_t = \arg\max_{S_t \subset L \setminus G_{t-1}} \tilde{R}(S_t, \bigcup_{l \in S_t} R(l))$$

- 自然终止：
  同位模式与其子集的团实例数相同时



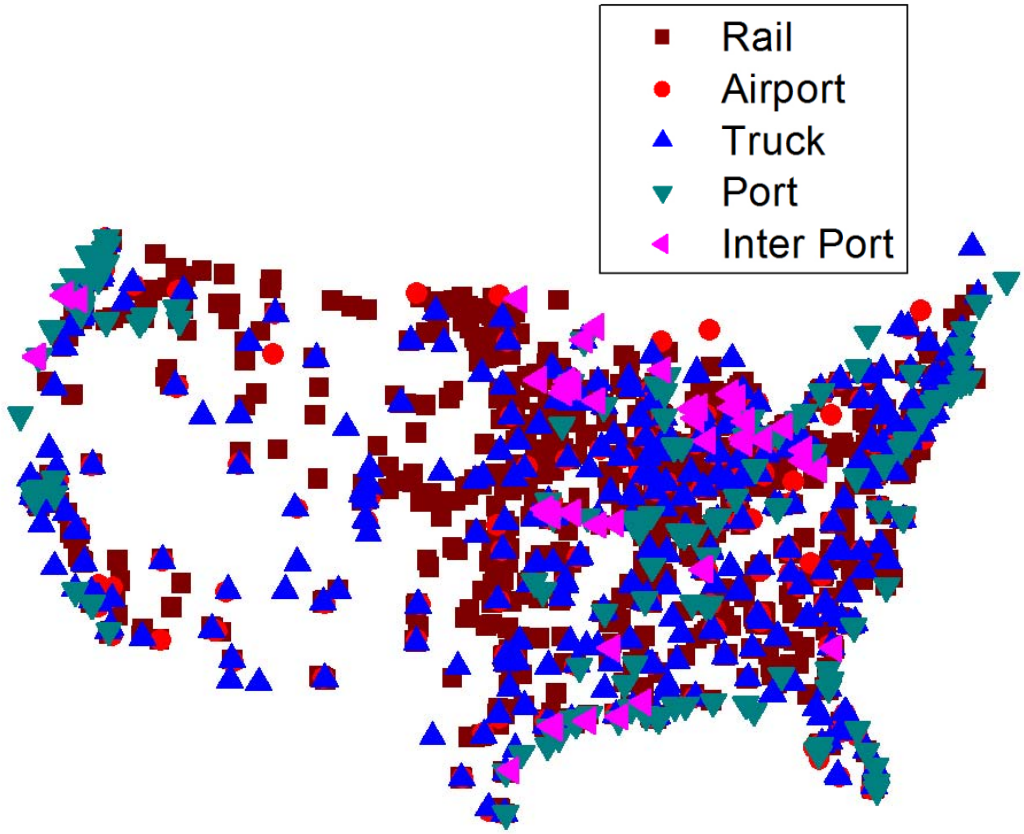
迭代式挖掘的优点

- 阈值无关
  - 避免由阈值引起的局限性
- 高质量的邻域关系图
  - 规模小、结构简单
  - 更多的同位模式信息
- 挖掘过程更具目的性
  - 选择性地向邻域关系图添加邻居边

真实数据集（一）

- NTAD-ITF数据集
  - 美国国家交通地图数据库

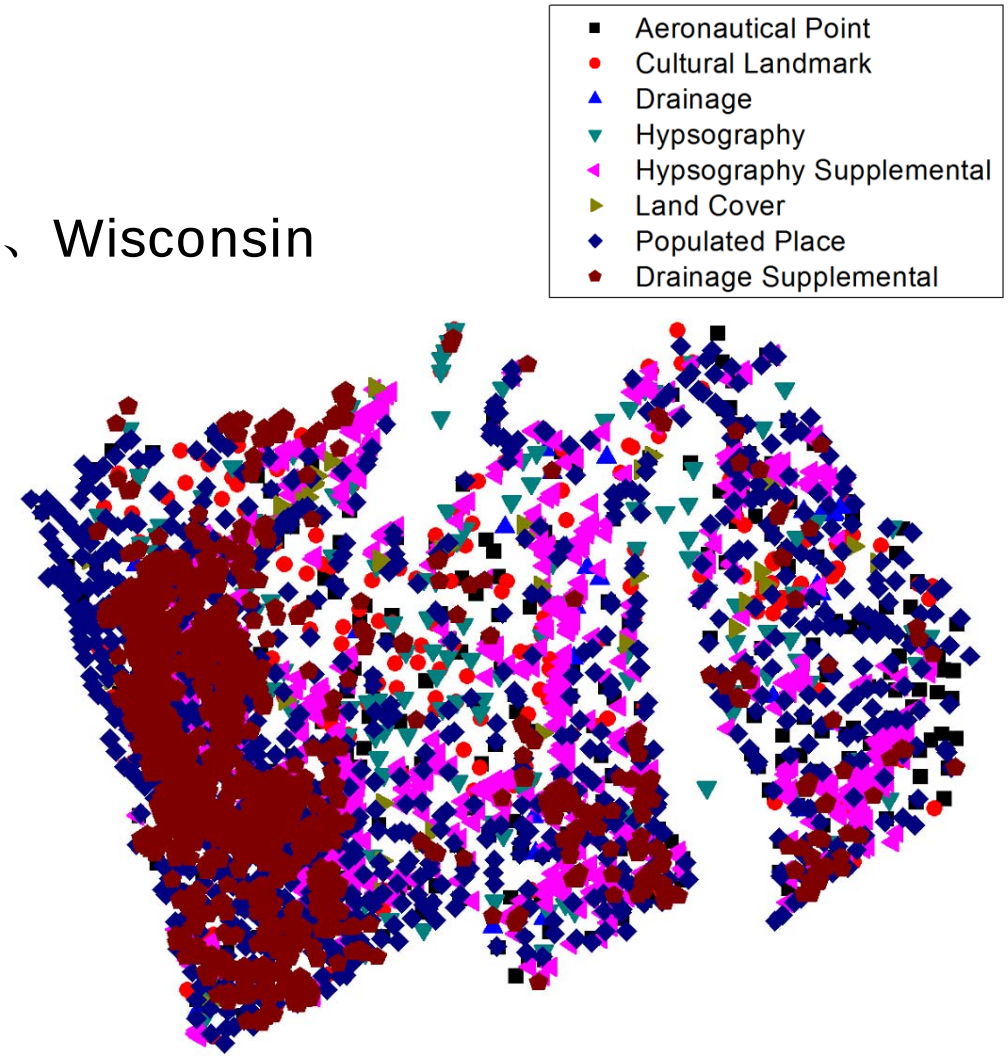
序号	空间特征（设备类型）	简称	空间实例数量
1	Rail	R	2,157
2	Airport	A	398
3	Truck	T	443
4	Port	P	172
5	Inter Port	I	87



真实数据集（二）

- DCW数据集
  - 世界数字地图数据库
  - Michigan、Minnesota、Wisconsin

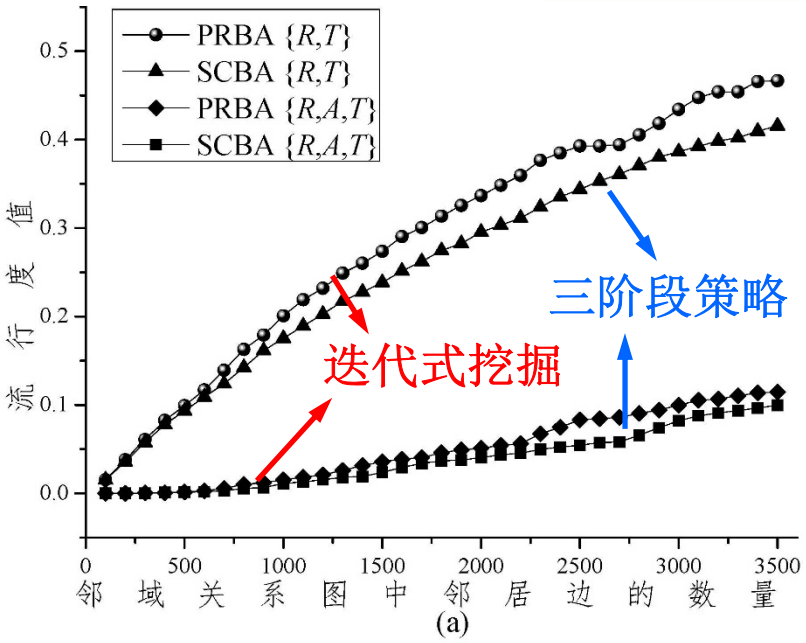
序号	空间特征（地标类型）	简称	空间实例数量
1	Aeronautical Point	AP	294
2	Cultural Landmark	CL	262
3	Drainage	DN	36
4	Drainage Supplemental	DS	1,541
5	Hypsography	HY	175
6	Hypsography Supplemental	HS	1,271
7	Land Cover	LC	49
8	Populated Place	PP	903



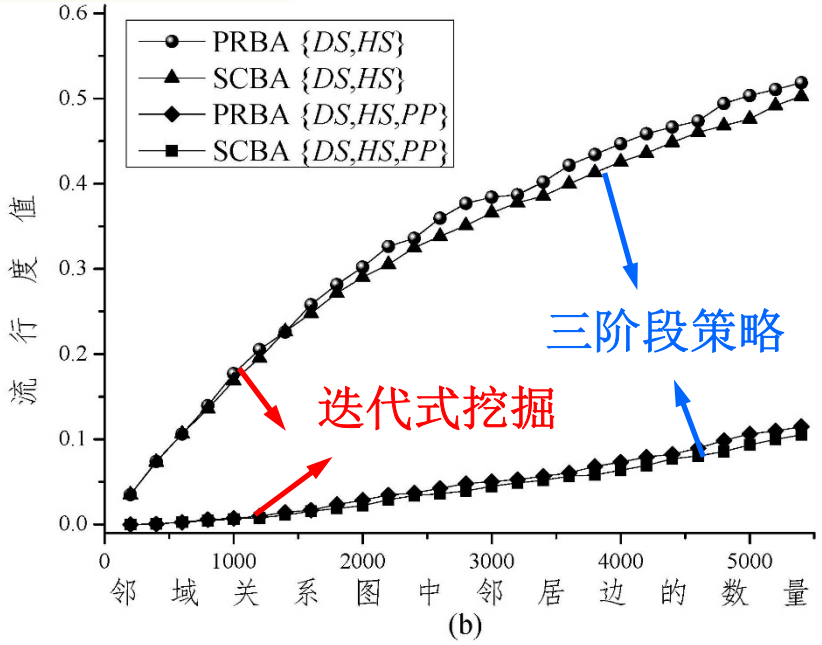
实验结果（挑选部分实验结果进行展示）

- 纵坐标：同位模式的流行度（值越大越好）

迭代式挖掘的值更大



NTAD-ITF数据集

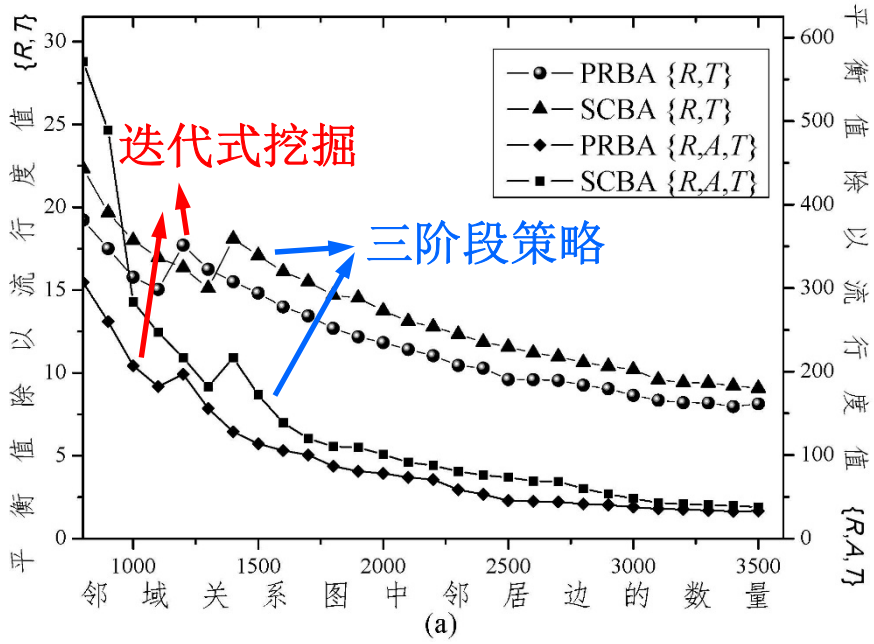


DCW数据集

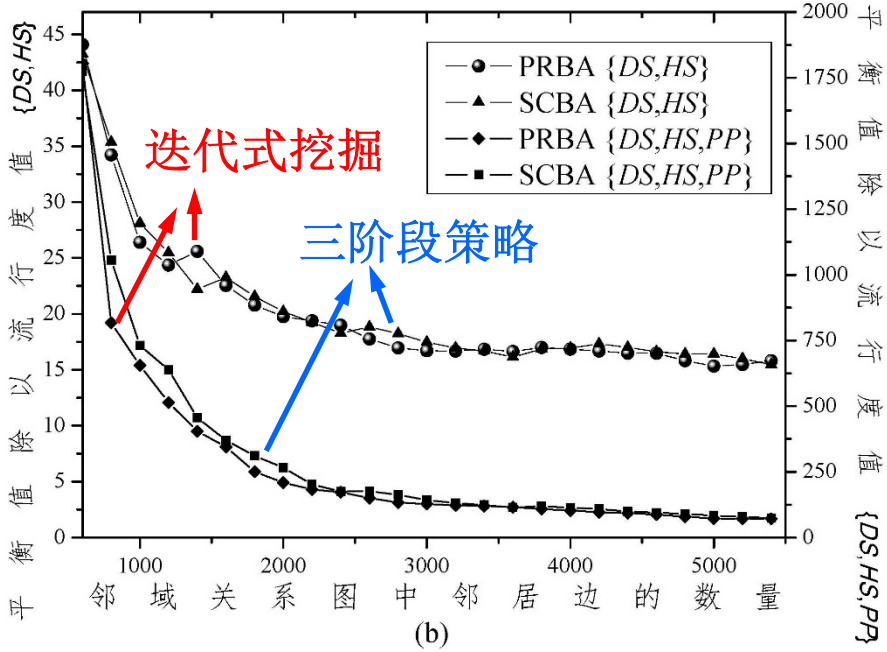
实验结果

- 纵坐标：同位模式的非显著性（值越小越好）

迭代式挖掘的值更小



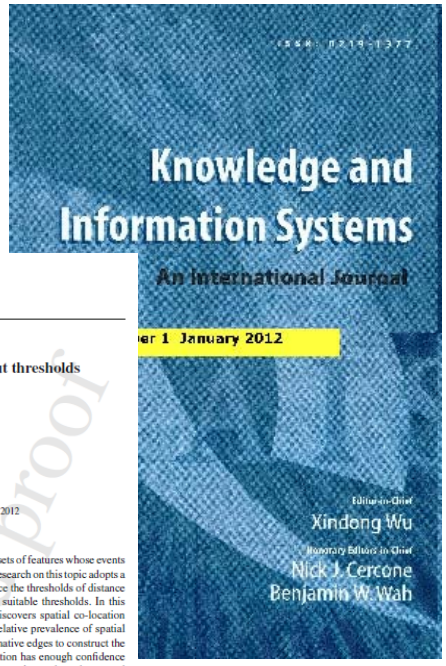
NTAD-ITF数据集



DCW数据集

研究成果

- 创新点：
  - 面向空间同位模式挖掘问题--提出了基于流行度回报的迭代式挖掘框架
  - Feng Qian, Qinming He, Kevin Chiew and Jiangfeng He. *Spatial Co-location Pattern Discovery without Thresholds* (*2012*). Knowledge and Information Systems (KAIS) Journal, 33(2):419-445. (SCI IF: 2.225)



Knowl Inf Syst
DOI 10.1007/s10115-012-0506-9

REGULAR PAPER

Spatial co-location pattern discovery without thresholds

Feng Qian · Qinming He · Kevin Chiew · Jiangfeng He

Received: 26 October 2010 / Revised: 14 October 2011 / Accepted: 1 April 2012

Abstract Spatial co-location pattern mining discovers the subsets of features whose events are frequently located together in geographic space. The current research on this topic adopts a threshold-based approach that requires users to specify in advance the thresholds of distance and prevalence. However, in practice, it is not easy to specify suitable thresholds. In this article, we propose a novel iterative mining framework that discovers spatial co-location patterns without predefined thresholds. With the absolute and relative prevalence of spatial co-locations, our method allows users to iteratively select informative edges to construct the neighborhood relationship graph until every significant co-location has enough confidence and eventually to discover all spatial co-location patterns. The experimental results on real world data sets indicate that our framework is effective for prevalent co-locations discovery.

Keywords Iterative framework · Threshold-free · Spatial co-location pattern · Prevalence reward

1 Introduction

The spatial co-location pattern mining [8] discovers the subsets of features (co-locations) of which the events are frequently located together in geographic space. It has been applied in many areas like M-commerce (mobile commerce), earth science, biology, public health, and transportation [26]. Taking M-commerce as an example of its application, to provide location-based services, service providers may tend to find services requested frequently located in spatial proximity.

F. Qian (✉) · Q. He · J. He
College of Computer Science and Technology, Zhejiang University, Hangzhou, People's Republic of China
e-mail: qfeng@zju.edu.cn

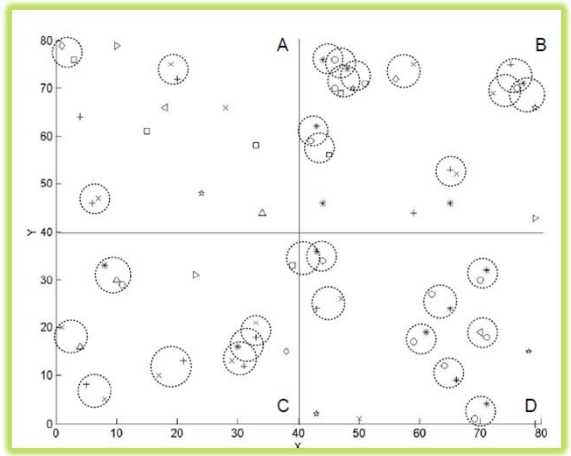
K. Chiew
School of Engineering, Tan Tao University, Long An Province, Vietnam

提纲

- 同位模式挖掘简介
- 相关工作
- 主要贡献
  - 空间同位模式挖掘
  - 区域同位模式挖掘
  - 时空同位模式挖掘
  - 同位轨迹挖掘
- 总结与展望

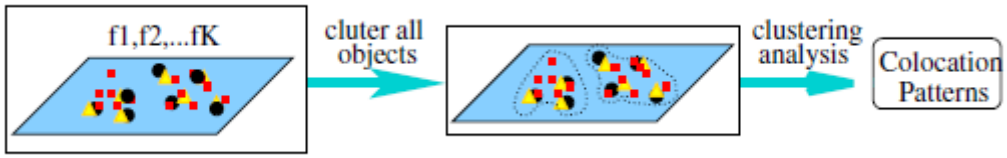
传统的区域同位模式挖掘

- 两种类型：
  - 在事先划定的区域内独立挖掘同位模式
    - 空间区域无法自动识别



[M. Celik, ICDM 2007]

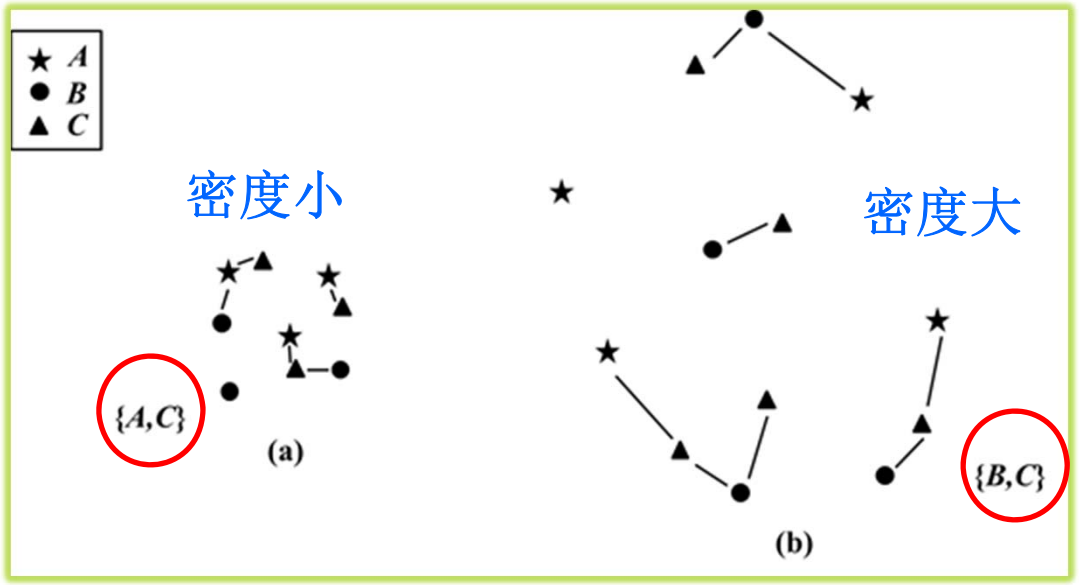
  - 采用划分聚类的方式搜索区域同位模式
    - 无法处理离散类型空间数据



[C.F. Eick, SIGSPATIAL 2008]

层次式挖掘的动机

- 邻域距离的多样性
  - 传统距离阈值无法处理
- 空间数据的异质性
  - 不同空间区域包含相异的同位模式集



↔
适合类簇(a)的距离阈值

⟷
适合类簇(b)的距离阈值

层次式区域同位模式挖掘

- 空间区域的相似度：

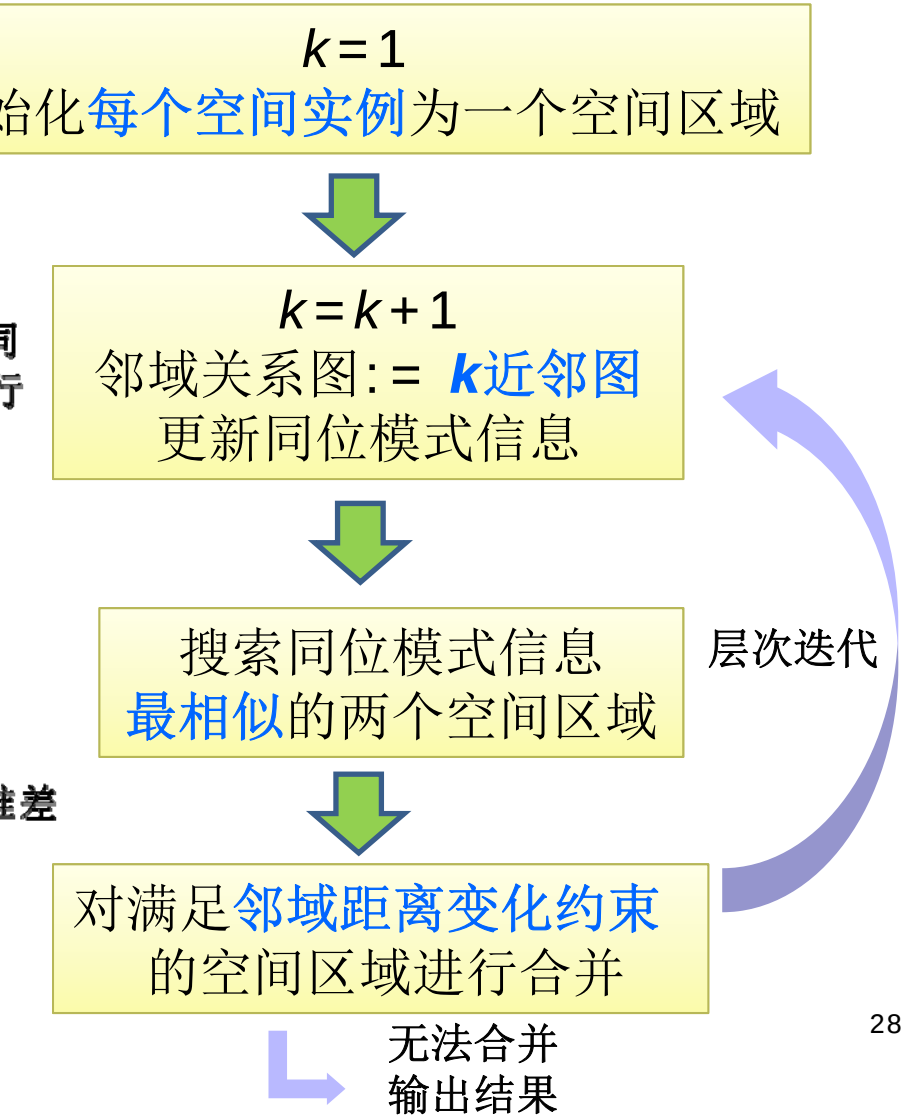
$$R(G_1, G_2) = \frac{1}{L-1}\sum_{i=2}^{L} J(C_{1i}, C_{2i})$$

- $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$
- C_{1i}是基于邻域关系图G_1的长度为i的空间同位模式集合，并且其中每个同位模式的流行度值都大于流行度阈值（$Pi(C_{1i}) \geq \theta$）
- L是$C_{1i} \cup C_{2i}$中同位模式的最大长度

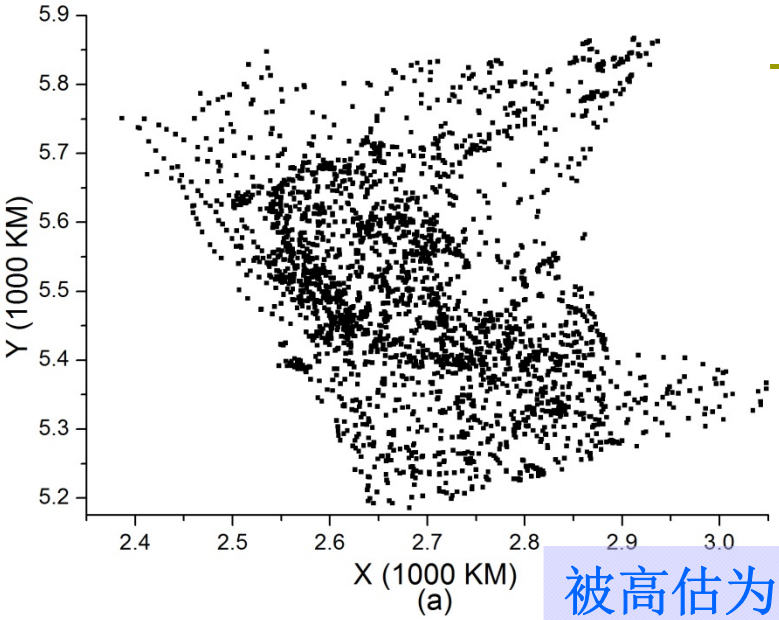
- 邻域距离变化约束：

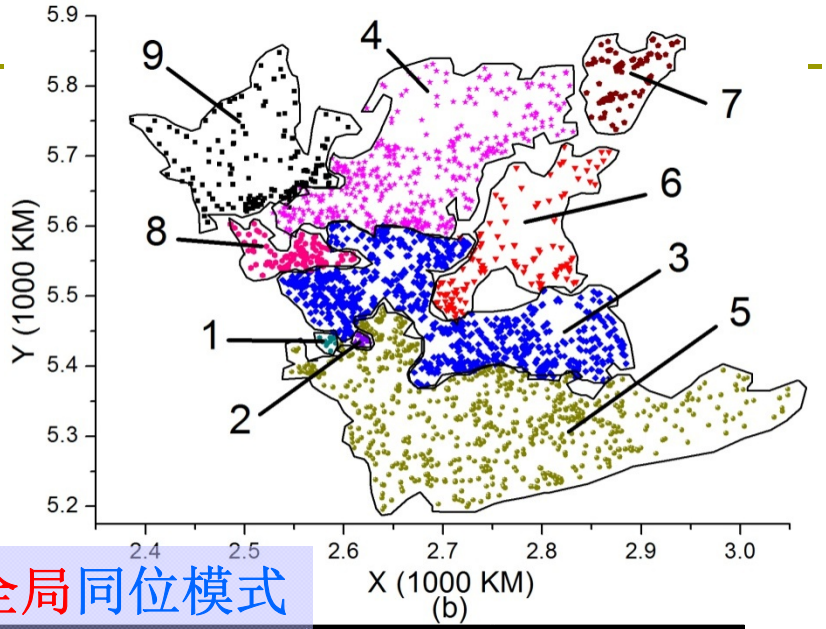
$$\Omega(G) = \frac{\sigma(E(G))}{\mu(E(G))} \qquad \Omega(G) \leq \varepsilon$$

- $\sigma(\cdot)$和$\mu(\cdot)$是邻居边长度集合的均值和标准差
- ε 是邻域距离变化阈值



实验结果（DCW数据集：Minnesota）





被高估为全局同位模式

执行算法	$Pi(\{DS, HS, PP\})$	$Pi(\{AP, DS, PP\})$	$Pi(\{HS, CL, PP\})$	$Pi(\{AP, CL, PP\})$
join 算法（30KM）	0.65	Null	Null	Null
join 算法（35KM）	0.72	0.70	Null	Null
join 算法（40KM）	0.77	0.78	0.64	Null
RCMA(层次式区域同位模式挖掘算法)	0.98（区域 3） 0.95（区域 7） 0.69（区域 5） 0.90（区域 6）	0.66（区域 5） 0.66（区域 8） 0.75（区域 9）	0.90（区域 5）	0.81（区域 6）

不同距离阈值

三阶段策略

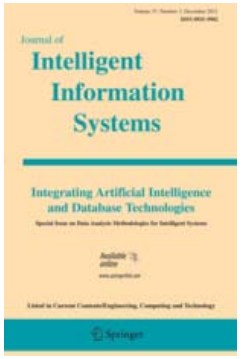
层次式挖掘

被忽视

在准确区域找到相应的区域同位模式

研究成果

- 创新点：
  - 面向区域同位模式挖掘问题--提出了基于*k*近邻图的层次式挖掘框架



- Feng Qian, Qinming He and Jiangfeng He. *Mining Spatial Co-location Patterns with Dynamic Neighborhood Constraint*. Proceedings of 2009 European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Lecture Notes in Computer Science, September 2009: 238-253. (EI: 20094312390753)

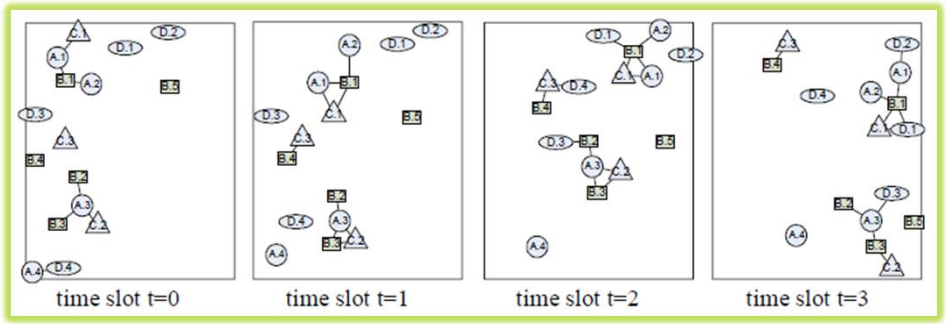
- Feng Qian, Kevin Chiew, Qinming He and Hao Huang. *Discovery of Regional Co-location Patterns with k-Nearest Neighbor Graph*. Journal of Intelligent Information Systems (JIIS), online first. (SCI 0.833)

提纲

- 同位模式挖掘简介
- 相关工作
- 主要贡献
  - 空间同位模式挖掘
  - 区域同位模式挖掘
  - 时空同位模式挖掘
  - 同位轨迹挖掘
- 总结与展望

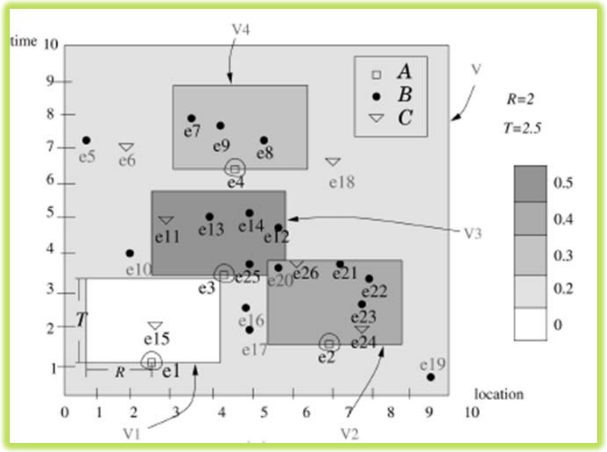
传统的时空同位模式挖掘

- 在划分好的时间片中独立挖掘同位模式
  - 忽略时间片之间的模式信息



[J.S. Yoo, SDM 2006]
[M. Celik, ICDM 2006]

- 时间＝额外的空间维度
  - 忽略时间的方向性
  - 较高的时间复杂度



[Y. Huang, TKDE 2008]

加权滑动窗口模型

从基本长度的同位模式（size=2）开始搜索

沿着时间轴移动窗口保持一定数量时间片

计算“时空团实例”
计算“时空流行度”

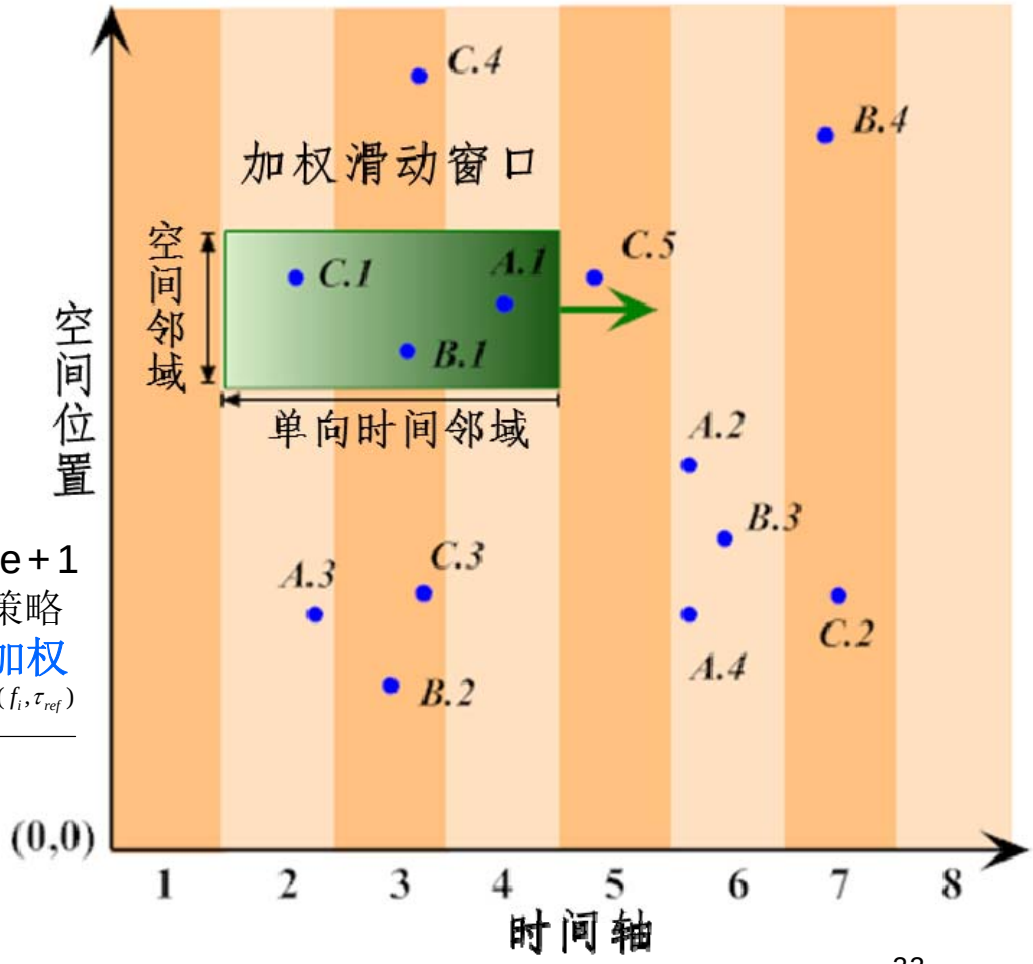
size=size+1
Apriori策略

根据时间间隔远近做加权

$$\frac{\sum_{\tau_{ref}\le\tau_{obj}} P(\tau_{obj},\tau_{ref})\times Rpr(C,f_i,\tau_{obj},\tau_{ref})\times F(f_i,\tau_{ref})}{\sum_{\tau_{ref}\le\tau_{obj}} P(\tau_{obj},\tau_{ref})\times F(f_i,\tau_{ref})}$$

并输出相应的同位模式

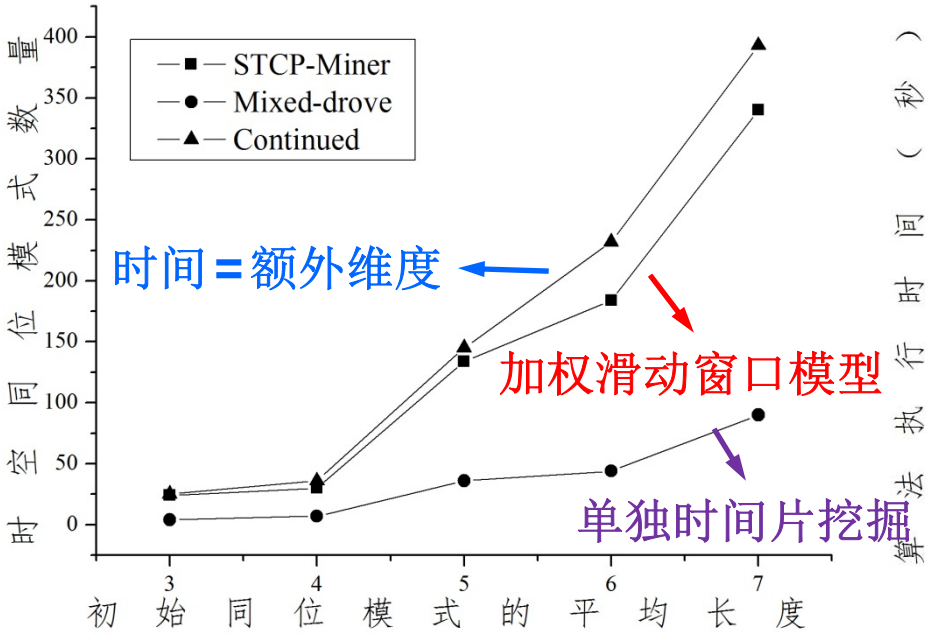
没有输出
算法终止

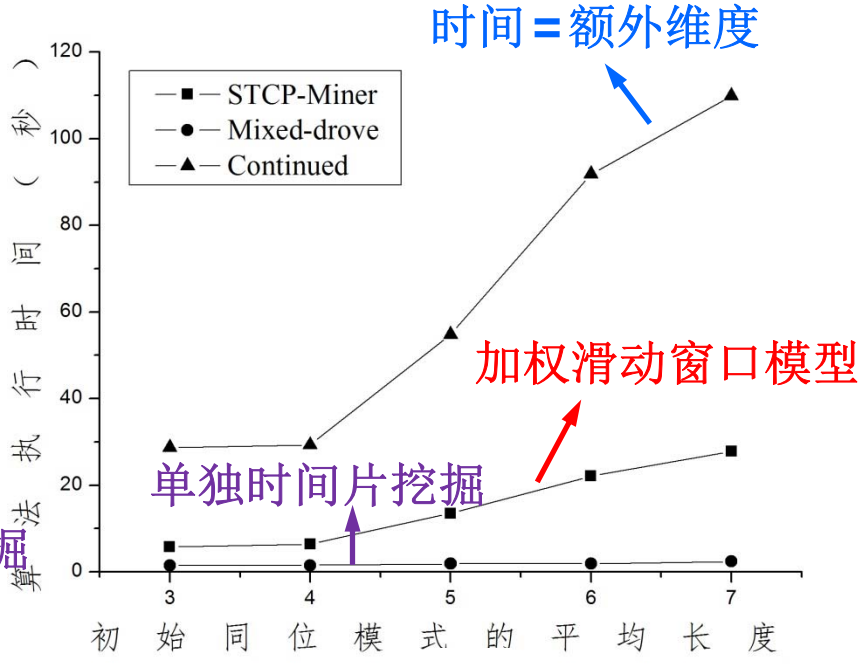


实验结果（基于合成数据集）

- 纵坐标：
  - 挖掘结果（同位模式数量）
  - 挖掘效率（算法运行时间）

加权滑动窗口模型
--较好的挖掘结果
--较高的挖掘效率

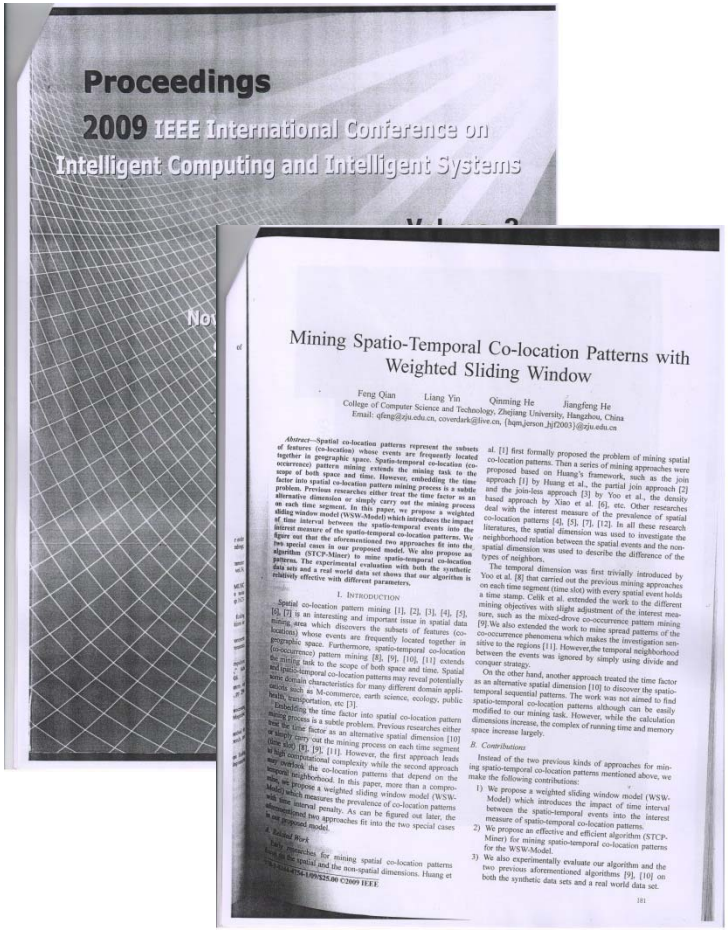




传统方法是“加权滑动窗口模型”退化后的两个特例

研究成果

- 创新点：
  - 面向时空同位模式挖掘问题--提出了加权滑动窗口模型
    - Feng Qian, Liang Yin, Qinming He and Jiangfeng He. *Mining Spatio-temporal Co-location Patterns with Weighted Sliding Window*. Proceedings of 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS), IEEE Computer Society Press, November 2009: 181-185. (EI: 20101212793419)

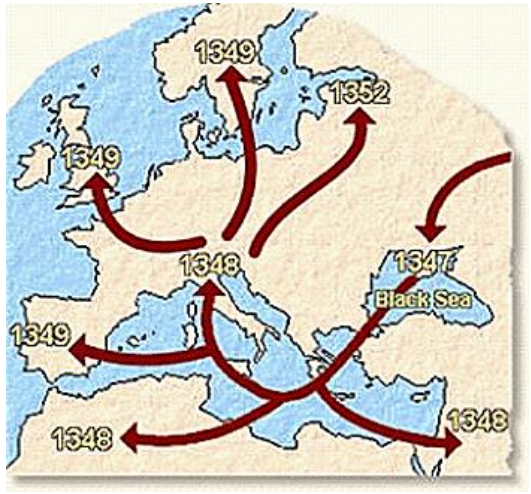


提纲

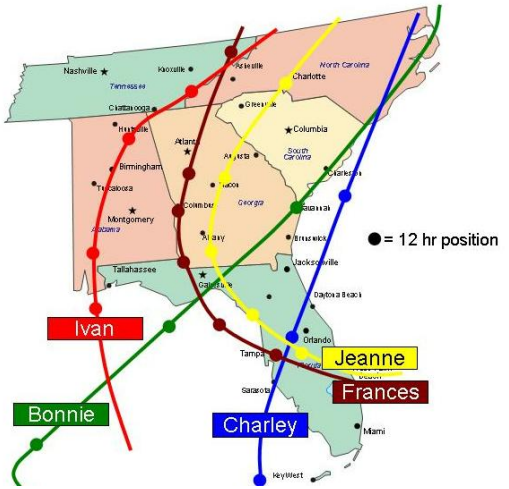
- 同位模式挖掘简介
- 相关工作
- 主要贡献
  - 空间同位模式挖掘
  - 区域同位模式挖掘
  - 时空同位模式挖掘
  - 同位轨迹挖掘
- 总结与展望

应用背景

- 同位现象揭露了应用领域背后的隐含特征
- 同位轨迹挖掘在跟踪传染性疾病、生态灾难等传播现象的应用领域中具有重要意义



包含发烧、头痛、关节肿痛、恶心呕吐等空间特征的同位现象：黑死病的传播轨迹

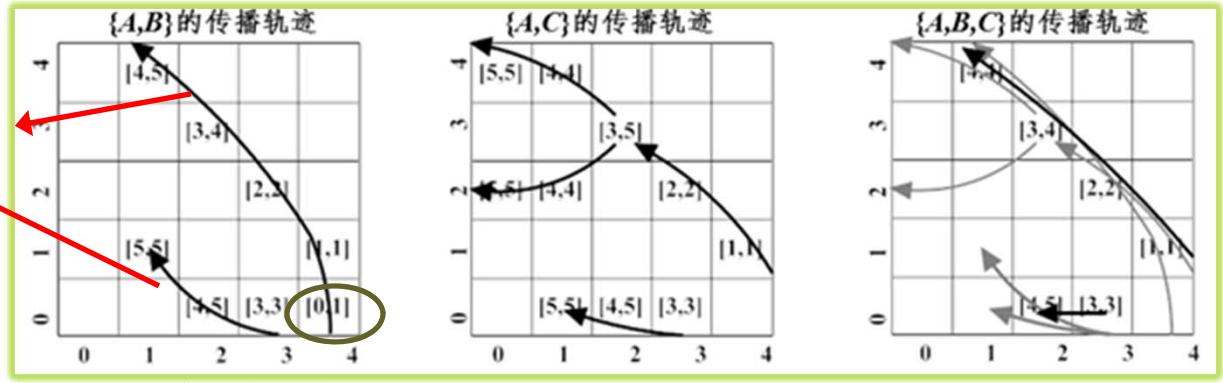


包含暴雨、洪水、停电、人员伤亡等空间特征的同位现象：飓风的传播轨迹

同位轨迹的例子

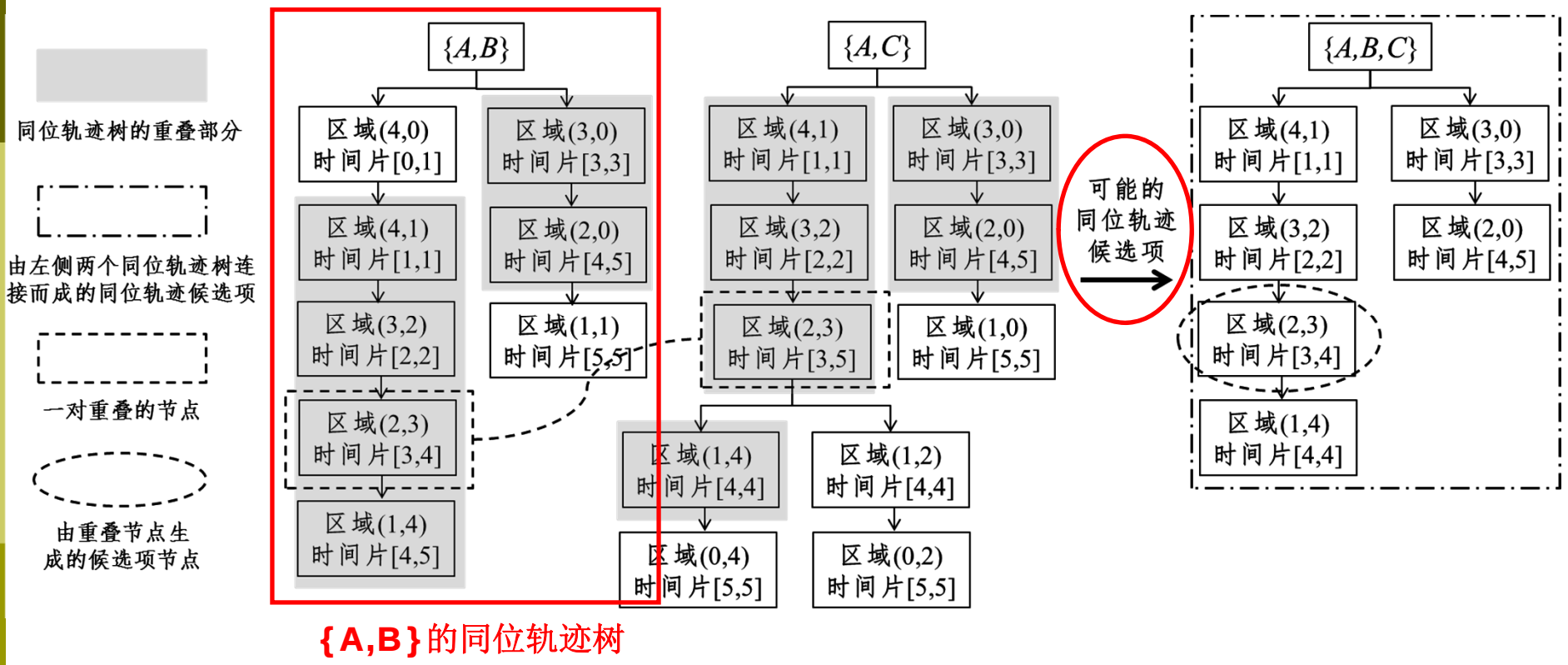
空间区域　局部区域（小格子）
包含的同位模式{A,B}

同位模式{A,B}
的传播轨迹



{A,B}在 时间片0至1
间在该格子内的持续

同位轨迹树



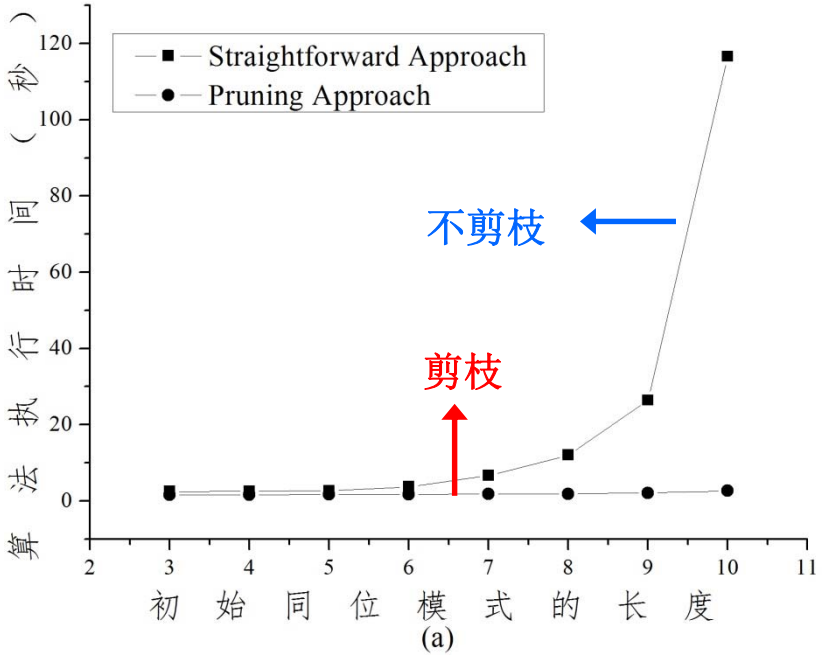
{A,B}的同位轨迹树

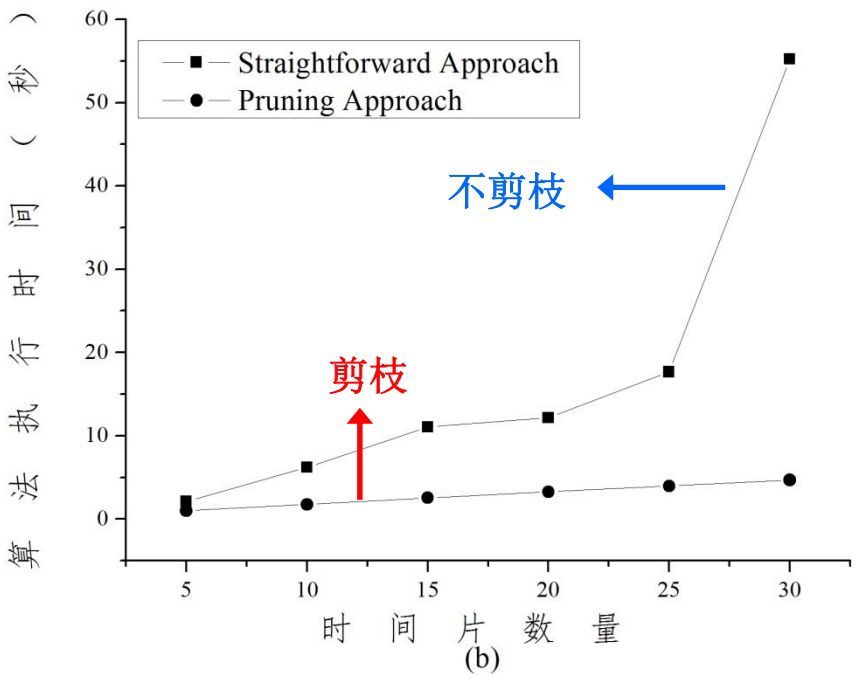
- 挖掘过程
  - 同位轨迹挖掘：找到所有的同位轨迹树
  - 同位轨迹树：具有单调递减特性
  - 挖掘算法采用Apriori策略进行剪枝
    - 为每个区域生成基本长度的同位模式集（size=2）
    - 生成基本长度（size=2）的同位轨迹树
    - 采用Apriori策略生成其他长度的同位轨迹树

实验结果（基于合成数据集）

- 纵坐标：算法运行时间

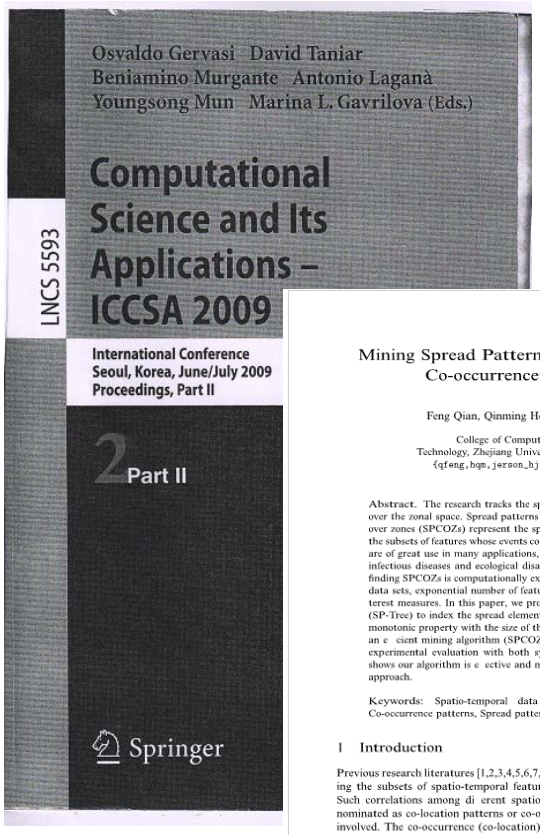
较高的挖掘效率





研究成果

- 创新点：
  - 首次提出了同位轨迹模式挖掘问题
  - 并进一步提出基于同位轨迹树的挖掘算法
  - Feng Qian, Qinming He and Jiangfeng He. *Mining Spread Patterns of Spatio-temporal Co-occurrences over Zones*. Proceedings of 2009 International Conference on Computational Science and Its Application (ICCSA), Lecture Notes in Computer Science, June 2009: 677-692. (EI: 20094612441099)



Mining Spread Patterns of Spatio-temporal Co-occurrences over Zones

Feng Qian, Qinming He, and Jiangfeng He

College of Computer Science and Technology, Zhejiang University, Hangzhou, China
{qfeng,hqm,jerson_hjf2003}@zju.edu.cn

Abstract. The research tracks the spread of co-occurrence phenomena over the zonal space. Spread patterns of spatio-temporal co-occurrences over zones (SPCOZs) represent the spread structures over the zones for the subsets of features whose events co-locate in space and time. SPCOZs are of great use in many applications, such as tracking the evolutions of infectious diseases and ecological disasters in space and time. However, finding SPCOZs is computationally expensive due to large size of history data sets, exponential number of feature combinations, and complex interest measures. In this paper, we propose a novel Spread Pattern Tree (SP-Tree) to index the spread elements of the SPCOZs which holds the monotonic property with the size of the co-occurrences. We also propose an e cient mining algorithm (SPCOZ-Miner) for mining SPCOZs. The experimental evaluation with both synthetic and real-world data sets shows our algorithm is e ective and much more e cient than a straight approach.

Keywords: Spatio-temporal data mining, Co-location patterns, Co-occurrence patterns, Spread patterns.

1 Introduction

Previous research literatures [1,2,3,4,5,6,7,8,9] have explored the problem of mining the subsets of spatio-temporal features whose events co-locate frequently. Such correlations among di erent spatio-temporal features were formally denominated as co-location patterns or co-occurrence patterns with time factor is involved. The co-occurrence (co-location) patterns are useful for revealing special characteristics behind the co-located phenomena. Its domain applications include earth science, biology, public health, transportation, etc [3]. However, these literatures did not take into account the locality of the co-occurrence patterns in space and time together. For example, a set of features may co-locate in a subset of space at some time slots while not meet the thresholds of the measures defined by the previous methods to mine co-occurrence patterns in a global vast scale of spatio-temporal framework.

In this paper, we track the spread of co-occurrence phenomena over the zones of space which deals with the locality problem. Spread patterns of spatio-temporal co-occurrences over zones (SPCOZs) represent the spread structures

O. Gervasi et al. (Eds.): ICCSA 2009, Part II, LNCS 5593, pp. 677–692, 2009.
© Springer-Verlag Berlin Heidelberg 2009

提纲

- 同位模式挖掘简介
- 相关工作
- 主要贡献
  - 空间同位模式挖掘
  - 区域同位模式挖掘
  - 时空同位模式挖掘
  - 同位轨迹挖掘
- 总结与展望

对本文工作的总结

- 迭代式空间同位模式挖掘框架
- 层次式区域同位模式挖掘框架
- 基于加权滑动窗口模型的时空同位模式挖掘
- 基于同位轨迹树的同位轨迹模式挖掘

对未来工作的展望

- 同位模式挖掘算法的效率提升
- 同位模式挖掘结果的进一步过滤和分析
- 同位模式的增量式挖掘
- 空间和时空数据集的同位模式信息索引

谢谢！