



不确定室内移动数据的 分析挖掘方法研究

毕业论文答辩

姓名： 李环

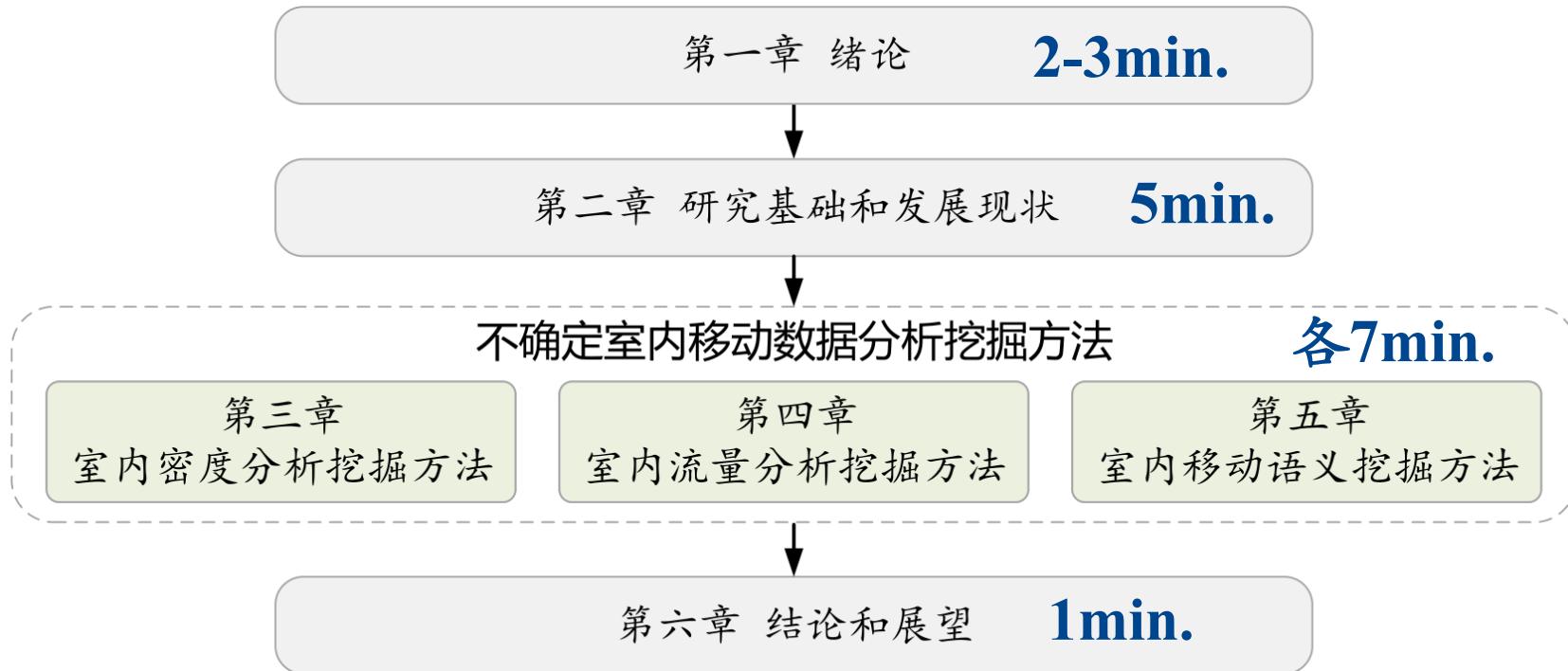
专业： 计算机科学与技术

导师： 陈刚教授，寿黎但教授

日期： 20180607



论文和答辩材料的结构组织





目录

1 研究背景和动机

2 研究基础和现状

3 室内密度分析挖掘

4 室内流量分析挖掘

5 室内移动语义挖掘

6 结论和展望

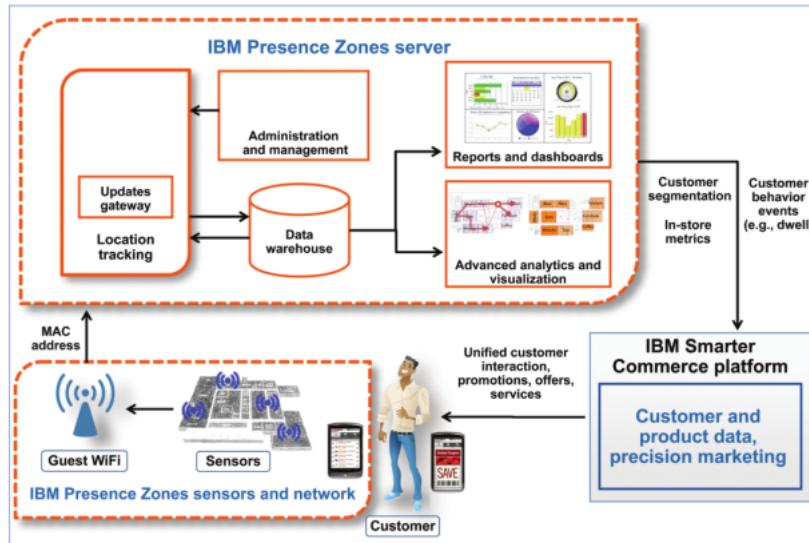


- 多项研究表明^[1-3]，人类日常生活近90%的时间在室内空间度过：
 - 办公楼、住房、购物中心、机场、车站……

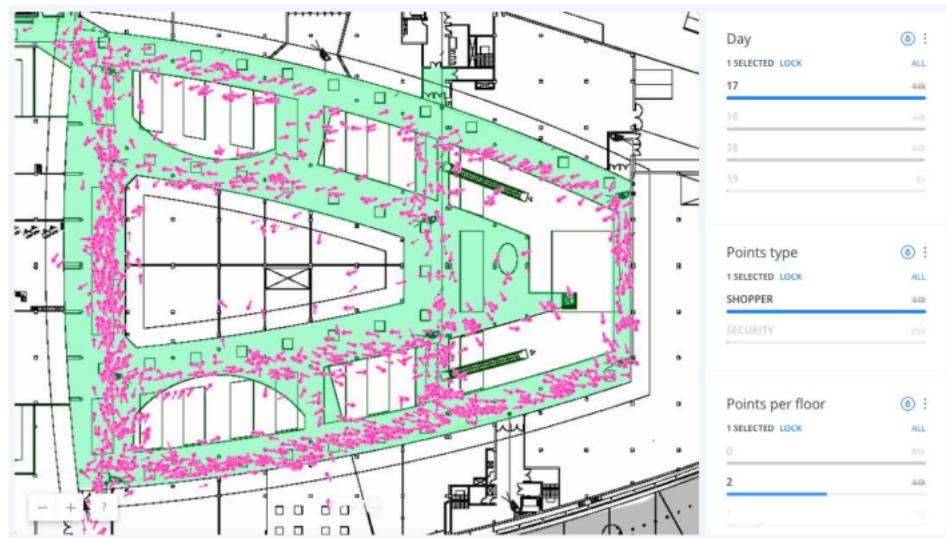


- 室内定位技术日趋成熟：
 - RFID、蓝牙、Wi-Fi、监控摄像头等传感元件渗透到室内环境的各个角落；
 - 2018年，将有8亿台设备频繁使用室内定位服务，与GPS服务达到同一量级。
- 室内移动数据开始爆发式增长，成为人类社会的又一笔伟大财富：
 - 在学界，移动数据挖掘方兴未艾，而室内场景更具挑战：
 - 移动模式挖掘^[8-12]；热点资源发现^[13-18]；行为推断预测^[19-22]等主题受到关注。
 - 在业界，indoor-LBS掀起了商业智能服务的一轮新浪潮：
 - 2013年起，苹果开始大力推广iBeacon；
 - IBM和思科投身移动用户追踪、行为分析和智能营销；
 - 支付宝和微信团队利用无线热点数据，进行场景推断和服务预测；
 - indoor.rs、carto、nextome等公司专注于室内移动数据的分析和智能应用。

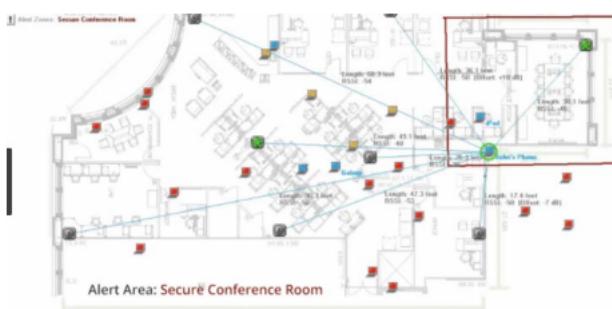
基于室内移动数据的智能服务



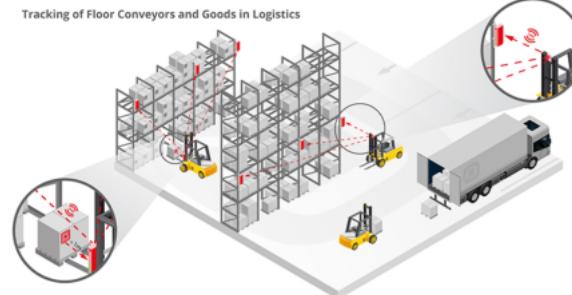
(a) 顾客行为及营销分析^[23]



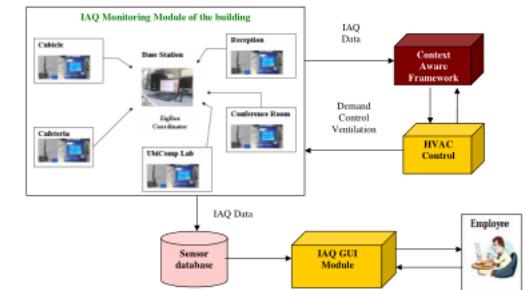
(b) 室内路径规划分析^[31]



(c) 异常入侵分析^[27]



(d) 物流监控数据分析^[29]

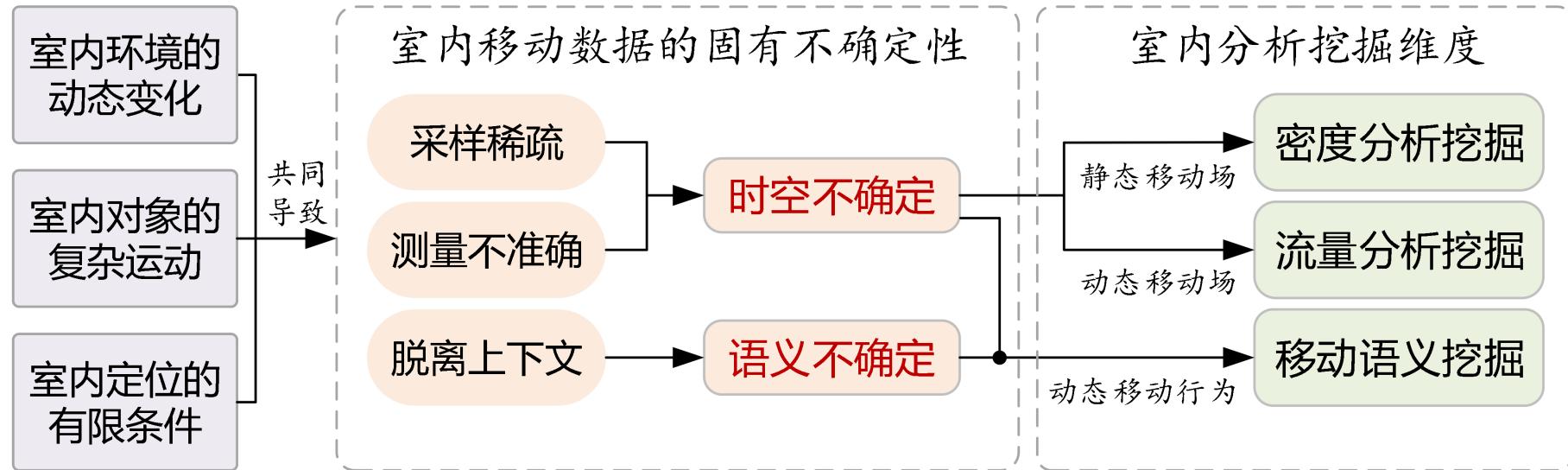


(e) 空气污染事件分析^[33]

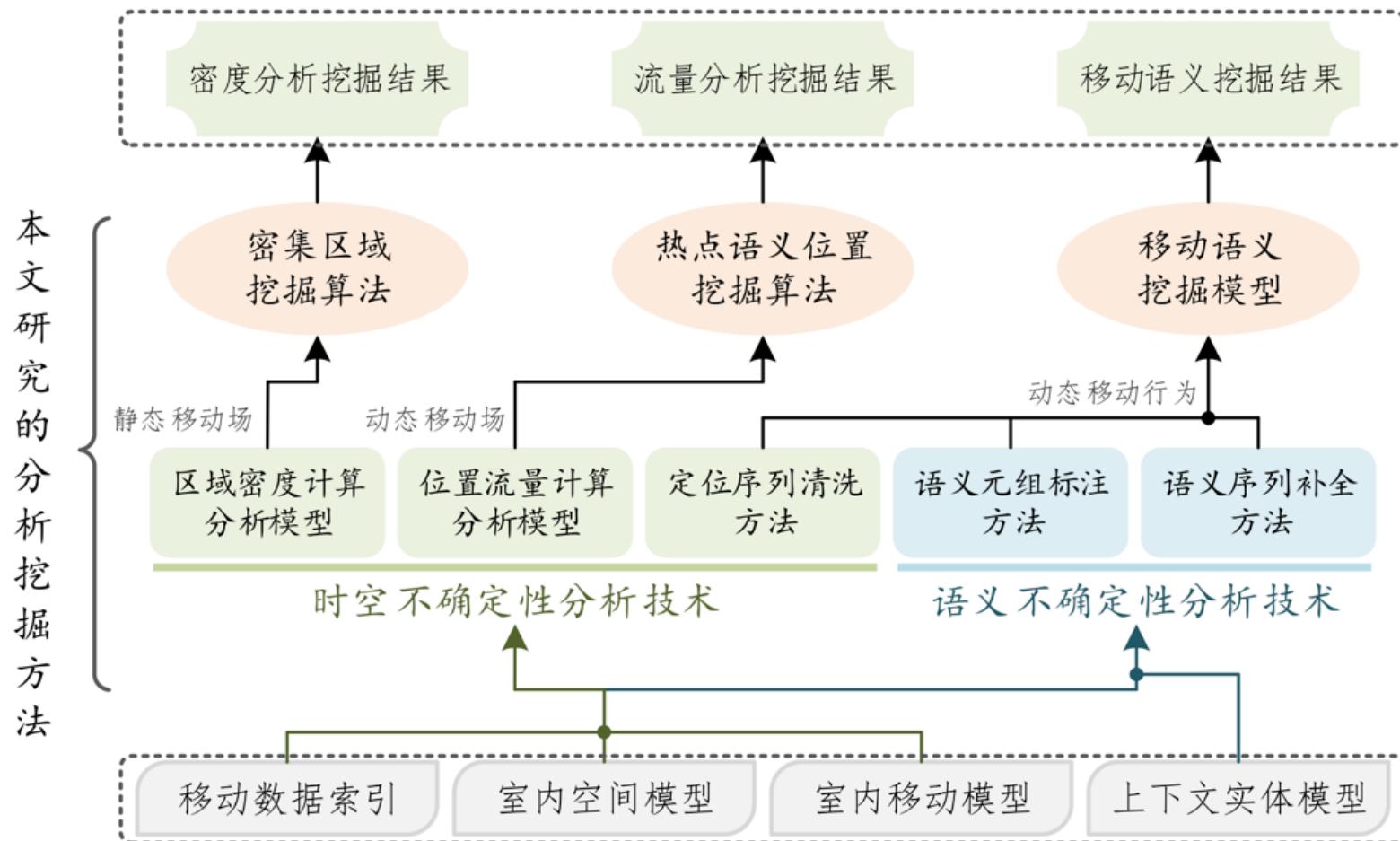
这些智能服务对政府机构、商业团体乃至个人而言，都具有重大的使用价值，彰显了其深远的社会经济影响力。



- 室内移动数据具有天然的**不确定性** —— 分析挖掘新的挑战。
 - 室内定位软硬件条件的局限性；
 - 室内环境的特殊空间结构和动态变化特点；
 - 室内对象运动的随机性和复杂性。
- **时空不确定性：**
 - 采样稀疏 – 室内移动数据通常是一组十分离散的位置报告；
 - 测量不准确 – 位置信息表达不准确、不充分。
- **语义不确定性：**
 - 脱离上下文 – 缺少对语义相关上下文的直接描述；
 - 位置不准确和精细复杂上下文实体的矛盾。
- 对数据不确定性进行**通用建模和分析**，有效支撑上层分析挖掘过程。
 - 降低室内智能服务开展的先决条件；
 - 拓宽实际应用的范畴；
 - 推动相关室内分析服务产业的发展。



- 从实际应用场景入手，考虑热门的室内移动知识分析挖掘问题。
- 移动场：对象在空间环境中的整体运动情况。
 - 静态移动场：快照时刻，对象密度。（时空不确定性）
 - 动态移动场：一定时间范围，对象流量。（时空不确定性）
- 移动行为：对象在语义层面上移动属性的体现。
 - 动态移动行为：一定时间范围，移动语义where-when-what。（时空+语义不确定性）



充分考虑室内空间拓扑、室内对象移动、室内定位机制的一般性特点。
底层设计/使用合理的数据结构和模型，分项输出分析挖掘的移动知识。



目录

1 研究背景和动机

2 研究基础和现状

3 室内密度分析挖掘

4 室内流量分析挖掘

5 室内移动语义挖掘

6 结论和展望



室内移动数据的相关课题

- ◆ 室内移动数据采集（定位追踪）
 - ✓ 基于无线信号的定位技术
 - ✓ 基于视觉的定位技术
 - ✓ 混合定位技术
- ◆ 室内移动数据管理技术
 - ✓ 室内空间建模及数据索引
 - ✓ 室内移动数据清洗
 - ✓ 室内移动数据查询
- ◆ 室内移动数据分析挖掘技术
 - ✓ 室内移动模式挖掘
 - ✓ 室内路线/区域发现
 - ✓ 室内移动预测

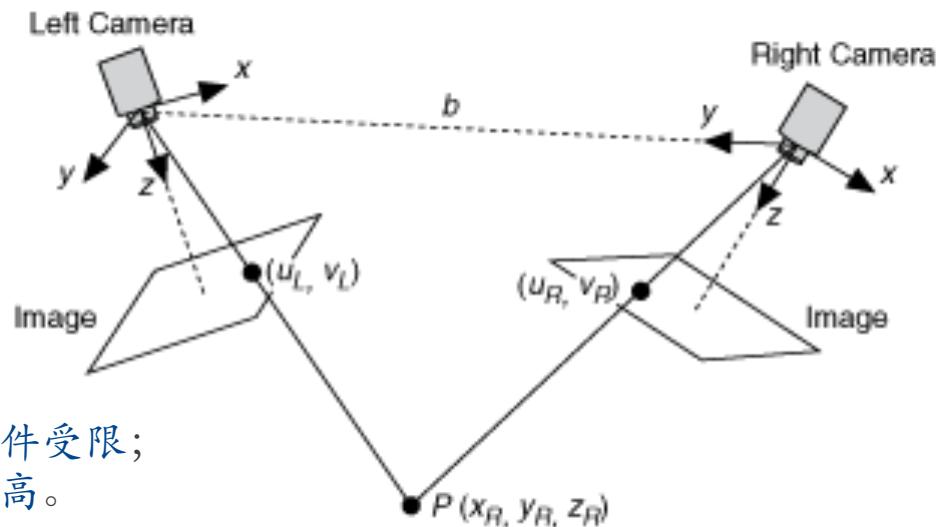
室外移动数据分析挖掘的相关课题

- ◆ 不确定轨迹建模和分析技术
- ◆ 密度分析挖掘技术
- ◆ 流量分析挖掘技术
- ◆ 语义提取和轨迹翻译技术

- 室内定位追踪^[34]的三大挑战：
 - 基础架构的差异 - 缺乏统一架构；
 - 动态环境的变化 - 传感元件精度影响严重、语义区域信息失准；
 - 精度要求的苛刻 - 充分的、细粒度位置是有效计算分析的保障。
- 基于无线信号的定位追踪^[35,36]：
 - 依赖价格昂贵的专用硬件设备（红外^[39]、超声波^[40]、UWB^[41]、VHF^[42]等）；
 - 移动端无线信号三边测距（Trilateration），AoA、ToA、RSSI^[43-45]等 - 信号波动大；
 - 基于统计学习的指纹（Fingerprinting）算法^[46-57] - 迭代离线训练；
 - 无线范围感应的临近定位（Proximity Analysis）算法^[12,30,58-60] - 粗粒度。
- 基于计算机视觉的定位追踪^[37]：

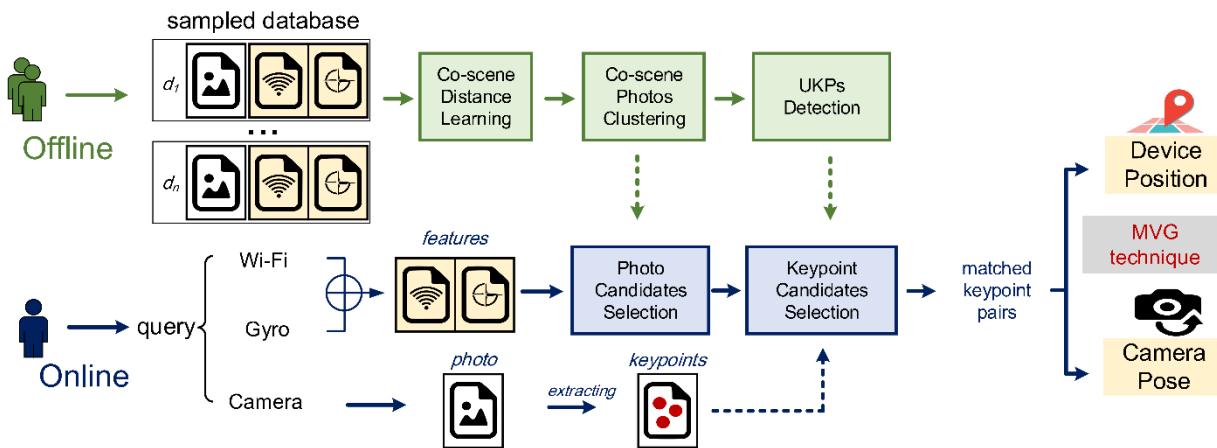
$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} X_0 \\ Y_0 \\ Z_0 \end{pmatrix} + \lambda \mathbf{R} \begin{pmatrix} x' \\ y' \\ -c \end{pmatrix}$$

- 参照图像特征点^[63-65] - 非鲁棒；
- 参照标识物（markers）^[66,67] - 部署条件受限；
- 无参照的单目SLAM^[68,69] - 计算代价高。



- 无线信号和/或视觉混合的定位追踪^[38]:

- filter-and-refine，粗粒度模型剪枝、高精度模型精细化位置（如蓝牙临近定位+指纹定位^[70]、RSSI信号+惯性测量单元^[71]等）；
- 多传感数据融合算法^[72,73] – 建立融合模型，对不同数据的相关性进行学习和构建。



[UbiComp15] 离线阶段，通过Wi-Fi指纹和陀螺仪度数对视觉（图片帧）共性场景进行融合特征表达。在线阶段，快速定位到视觉共性图片簇，进行相机配准，得到设备位置和相机朝向。

- 小结 对数据不确定性的讨论：

- 部分/完全依赖视觉的方法，对定位设备和环境有很高要求，不能普遍适用；
- 无线定位方法，门槛低、普及度高，能得到海量的室内移动数据；
- 无线方法，受限于信号传播和环境的波动，精度难以保障；
- 目前可广泛获取的室内移动数据，往往具有天然的不确定性。

- 室内空间建模及数据索引：
 - 室内空间充斥大量实体（门窗、墙体等），移动约束不同于自由空间或路网空间；
 - 室内空间模型：基于对象特征（属性、操作及关联）^[74]、几何坐标（位置、方向、距离）^[75]或符号（语义、几何和拓扑关系）^[76-80]；
 - 数据索引：面向临近定位（RFID-like）^[81-83]或几何定位（x,y-like）^[84-85]的静态数据（轨迹、序列）或移动对象；
 - 大部分工作未考虑移动数据的不确定性，少数考虑的工作面向特定问题和数据格式，带有较强前提假设。
- 室内移动数据清洗：
 - 符号（临近）定位中的（假阳性）消歧或（假阴性）复原，或二者兼之。如基于地图位置相关性的概率推断模型^[86]，时间、空间和拓扑约束的概率（图）模型^[87-90]，基于历史数据学习的RFID部署（为隐状态）的HMM模型^[91]等；
 - 几何定位序列中的随机误差或错误（楼层错误或离群点）识别和消除，如贝叶斯滤波^[92]、卡尔曼滤波^[93]、路网结构约束下轨迹的bootstrapping采样精化^[94]等；
 - 主要专注于符号定位序列清洗，更为常见的室内几何轨迹的不确定分析和清洗技术有很大空缺。

- 室内移动数据查询：

- 位置相关查询 - 实体对象^[95], 室内POI^[96], 活动事件^[97]等;
- 室内路径查询 - 基于路径长度^[98]、用户偏好^[99]、或上下文感知^[100]等;
- 距离敏感查询 - 空间范围^[80,82,85,102]、kNN^[80,83,85]等;
- 时空范围查询 - 空间范围可定义为单元空间ID^[101]或RFID序列^[81]等;
- 连接查询 - 距离敏感的对象对^[103]、历史轨迹的概率自连接^[104]等;
- 轨迹相似度查询 - 考虑空间和语义属性^[105];

	静态定位记录/轨迹	在线-移动对象	历史-移动对象
静态查询	空间范围 ^[80] k 近邻 ^[80]	空间范围 ^[85] k 近邻 ^[83,85] 连续范围监控 ^[82]	时空范围 ^[81] 拓扑关联 ^[81]
动态查询	连续范围监控 ^[102]	距离敏感连接 ^[103]	时空连接 ^[104]

- 小结 对数据不确定性的讨论：

- 提供了上层分析挖掘的基础，但面对移动数据固有的不确定性，仍有很大不足；
- 现有的空间建模、对象索引和查询处理均面向特定问题和数据模型，特殊假设前提；
- 流行的数据清洗技术主要面向符号定位数据的消歧和复原，缺少通用的框架对更为常见的几何定位序列进行不确定性分析和错误识别与恢复。

- 室内移动模式挖掘：
 - 从定位序列中找出频繁的、相似的移动模式可以帮助分析和理解室内的移动行为；
 - 频繁轨迹模式 (frequent pattern)^[8,11] – RFID元件ID序列；
 - 移动序列模式 (sequential pattern)^[9] - 购买行为（事务+路径遍历）；
 - 位置访问模式 (visiting pattern)^[12] – 蓝牙结点序列中找出位置访问事件；
 - 停留模式 (stop-by pattern)^[10] – 不确定的RFID序列中找出区域停留事件；
- 室内热点路线/区域发现：
 - 发现被移动对象频繁访问/使用的路线、位置或区域，帮助相关资源的规划或推荐；
 - 无线信号观测序列中抽取动态环境的语义位置^[13] (HMM，语义位置为隐状态)；
 - 最常用室内路径检测^[14] (路线拆解、聚类、簇内共性路径识别)；
 - 几何定位序列中提取热点位置^[15] (建立用户-位置关系矩阵并进行矩阵乘法迭代)；
 - 受限路径空间 (传送带) 的瓶颈点推理^[16] (拓扑图模型+数据的概率补全)；
 - 半受限路径空间 (住房) 的密集区域^[17] 和频繁访问地点^[18] – 确定性RFID序列；
- 室内移动预测：
 - 对室内移动行为进行预测可帮助室内资源的合理调配；
 - 下一位置预测^[19] (Wi-Fi AP序列表达为观测序列，二阶HMM模型)；
 - 时态行为即滞留时长和离开时间点预测^[20] (高精度、高精度Wi-Fi指纹定位数据)；
 - 位置和时态行为预测^[21] (决策树、SVM、kNN、GB进行集成学习)；
 - 语义行为级别的预测^[22] (时空阈值抽取语义行为，构建动态贝叶斯网络)。

- 小结 对数据不确定性的讨论：

- 现有的模式挖掘、热点资源发现和行为预测方法，均面向特定数据模型和问题定义。
- 和本课题研究的静态/动态移动场、动态移动行为问题存在极大差异：
 - 密集区域挖掘^[17] – 历史间隔的符号定位序列；自然分割的房间区域；确定性数据设定；
 - 频繁访问位置（最高流量）挖掘^[18] – 符号定位序列；确定性数据设定；考虑对象可能区域和查询位置的几何相交，忽略定位序列到空间流量的概率映射；
 - 语义行为提取 – 室内绝大多数未考虑语义层面的行为提取。
- 极少数工作考虑采样或者观测不确定性带来的干扰：
 - 使用特殊设计的预处理技术，扩展性不强。
- 本课题考虑对时空和语义不确定性的抽象建模分析，能扩展到普通环境下采集的移动数据。

- 不确定轨迹建模和分析技术：

- 同样面临采样点稀疏（sampling error）和定位精度低（measurement error）的问题；
- 不确定轨迹建模方法^[106-111] – 随时间演变的特定空间范围；
- 面向不确定位置/轨迹的查询^[109-114] – 概率的Range和kNN查询等；
- 不确定轨迹设定下的路径推断^[126]/补齐^[127]、轨迹对齐^[128]、目的地预测^[129]等。

- 密度分析挖掘技术：

- 欧氏空间下时空对象计数^[115]、静态对象聚类^[116]；路网空间下静态对象聚类^[117]、静态对象的（快照/时间段）概率密度查询^[118]；不适用：室内移动和拓扑的特殊性；
- 欧氏空间下线性运动物体的快照密度查询^[119-120]、连续密度查询^[121]；不适用：线性模型未考虑室内对象移动的复杂形态和拓扑限制；
- 欧式空间下的在线密集区域查询^[122]、路网空间下的密度最大路段查询^[123]、连续密集路段监控^[124]；不适用：空间建模方式不同、未考虑采样稀疏导致的时空不确定性；

- 流量分析挖掘技术：

- 基于时空因素的重要语义位置排序模型^[125]、欧氏空间下时空对象计数^[115]、基于转移概率的不确定轨迹构建计数^[129]；不适用：不支持室内拓扑、[115,125]仅处理确定数据、[129]利用历史数据计算转移概率；
- 利用轨迹进行运动不确定性的补全或推断^[126-128]；不适用：未考虑测量不确定性、不支持室内拓扑、未考虑室内定位机制对序列补全的影响；

- 语义提取和轨迹翻译技术：
 - 基于地理信息抽取停留和移动（stop/move）事件^[131]；
 - 对原始GPS序列进行清洗和关键运动特征抽取与组合^[132]；
 - 封装移动数据几何属性和语义信息的混合轨迹模型^[133]；
 - 按照移动行为分割轨迹并提取关键特征进行片段摘要^[134]；
- 和本课题中室内移动语义挖掘问题的不同：
 - 复杂室内拓扑结构下，室内定位数据极高的时空和语义不确定性（房间跳跃、楼层错误）；
 - 抽取的语义元组的表达形式不同，不同于混合轨迹^[133]或文本摘要^[134]；
 - 利用历史数据获取先验的移动知识，从而进行概率推断补全，其他工作未考虑。



目录

1 研究背景和动机

2 研究基础和现状

3 室内密度分析挖掘

4 室内流量分析挖掘

5 室内移动语义挖掘

6 结论和展望



- 研究动机：
 - 公共室内环境，容易在短时间内出现密集人群；
 - 及时有效地计算室内区域的人数密度并找出当前空间中的密集区域，将在**拥塞控制、安防管理**方面起到十分关键有利的作用；
 - 大型机场 - 当前航站楼最拥堵区域，打开更多通道帮助旅客快速通行；
 - 商场运营 - 向当前拥堵区域加派人员，进行巡逻疏散确保顾客安全。
- 给每个基本分区（房间）安装计数器？
 - 缺点：需要对专用硬件进行大量投资；拓扑结构变更时需重新部署；缺乏门或物理边界的区域，很难找到合适安装位置；
 - 目标：无需安装特定硬件、不受复杂动态的室内拓扑结构影响，并灵活适用于用户自定义区域的室内密度计算方法。
- 研究挑战：
 - 室内拓扑结构限定了移动对象运动，必须考虑到**复杂室内拓扑结构**的影响；
 - 室内定位系统的采样频率较低，仅报告离散的室内位置。因此，用于计算室内密度的移动数据可能是过时的，即留有相当大的时空不确定性；在室内拓扑下处理则变得更具挑战性。



- 输入 在线室内定位表 (OIPT) :

- 有限存储和较低吞吐量;
- 仅记录每个移动对象的最新位置，无其他信息;
- 分析时刻，位置数据老旧 (out-of-date)。

- 不确定区域和对象出现度 (Object Presence) :

- 不确定区域：最大速度限制下能到达的室内空间部分;
- 对象出现度：和查询区域的相交面积占据的比例。

- 负载量 (Load) 和区域密度 (Region Density) :

- 对某个查询区域，所有对象在其中出现度的总和;
- 区域密度：负载量除以区域的面积。

- 问题定义

问题 3.1 (Top- k 室内密集区域挖掘) 给定室内对象集合 O 、保存 O 中对象最新定位记录的 OIPT 和查询区域的集合 Q , top- k 室内密集区域挖掘找出一个包含 k 个最密集区域的 k -子集 $Q_k \subseteq Q$, 其中 $\forall r \in Q_k, \forall r' \in Q \setminus Q_k, \tau_O(r) \geq \tau_O(r')$ 。

objectID	location	t
o_1	l_1	t_1
o_2	l_3	t_1
o_3	l_6	t_6

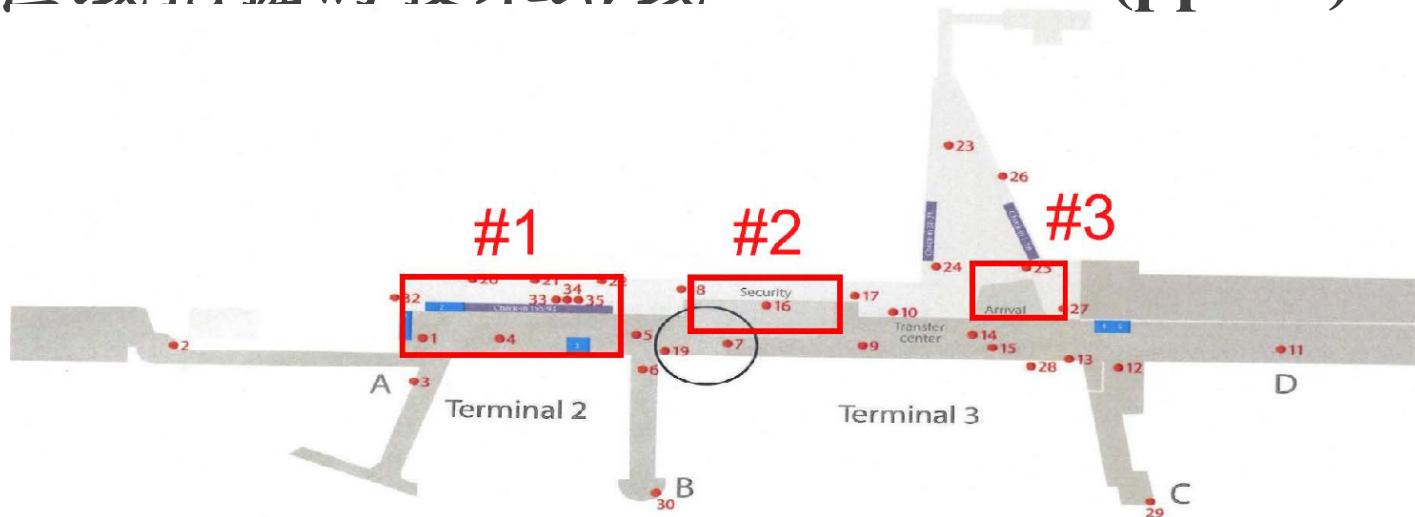
$$\phi_r(o) = \frac{\text{Area}(UR_I(loc) \cap r)}{\text{Area}(UR_I(loc))}$$

$$\lambda_O(r) = \sum_{o \in O} \phi_r(o)$$

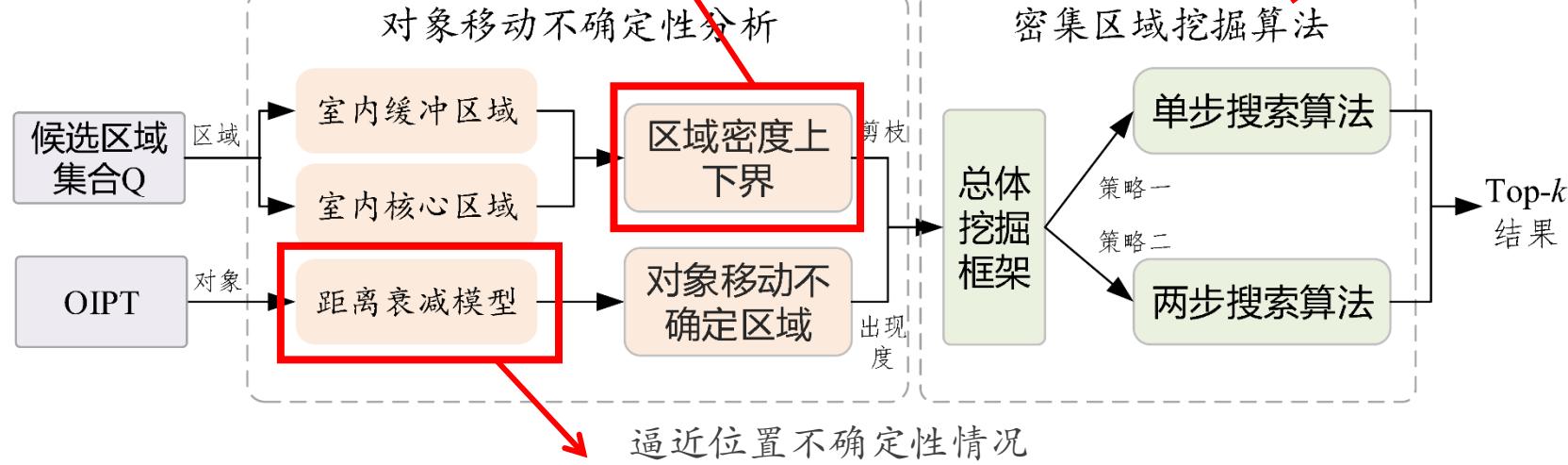
$$\tau_O(r) = \frac{\lambda_O(r)}{\text{Area}(r)}$$

密集区域挖掘的技术路线

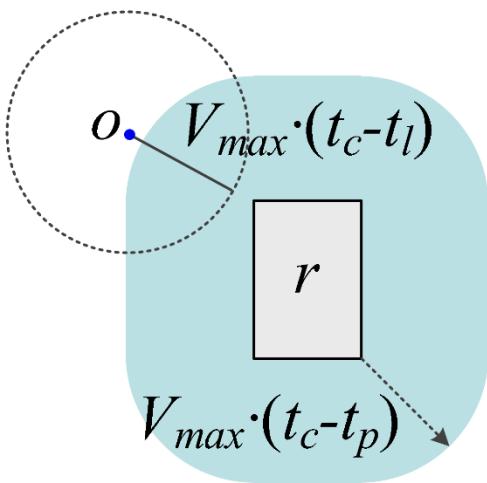
(pp. 33)



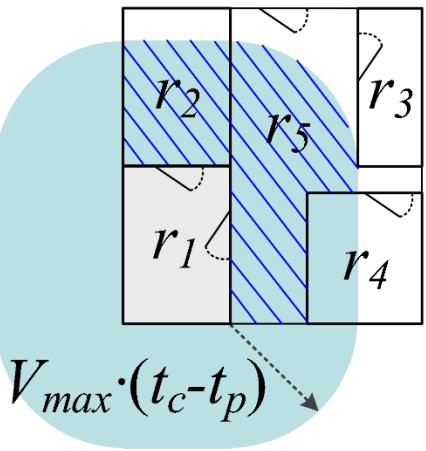
确定密度的有效界，仅关注有关对象



室内缓冲区域的意义



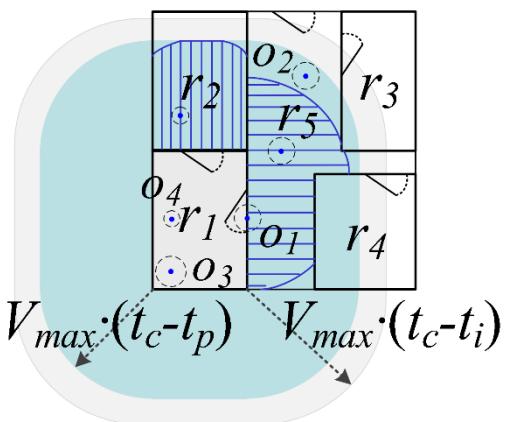
一般缓冲区域



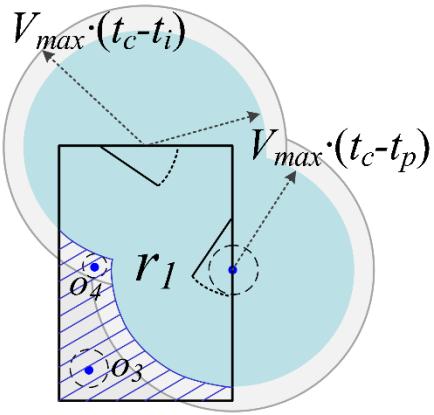
室内缓冲区域

t_p 时刻为某固定过去时刻，若对象 O 在 t_l 时刻($t_l > t_p$)的观测在缓冲区域外部，则在 t_c 时刻无法到达区域 r 内部。

室内缓冲/核心区域的计算判定



室内缓冲区域（准确）

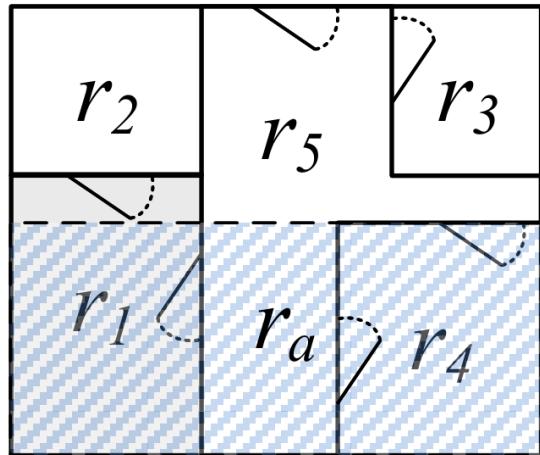


室内核心区域

- 引理3.1：根据一般缓冲区域排除 r_3 ；
- 引理3.2：根据门拓扑排除 r_4 ；
- 引理3.3：根据Indoor Range Query纳入 r_2/r_5 的部分，加上 r_1 ，构成室内缓冲区域；
- 引理3.4：根据缓冲区域的反向差集去除 r_1 内部能够在特定距离内离开的部分（计算室内核心区域）。

室内缓冲/核心区域判定算法

复杂查询区域 (pp.37)



- 任一边可以为：
 - 没有门的墙壁（左右下边）；
 - 带有门的墙壁（上边r₄部分）；
 - 位于某一分区内的开放边（上边r₁和r₄部分）；
 - 前三者组合（上边）。

判定算法 (pp.37)

同时计算缓冲/核心区域（并操作/差操作）

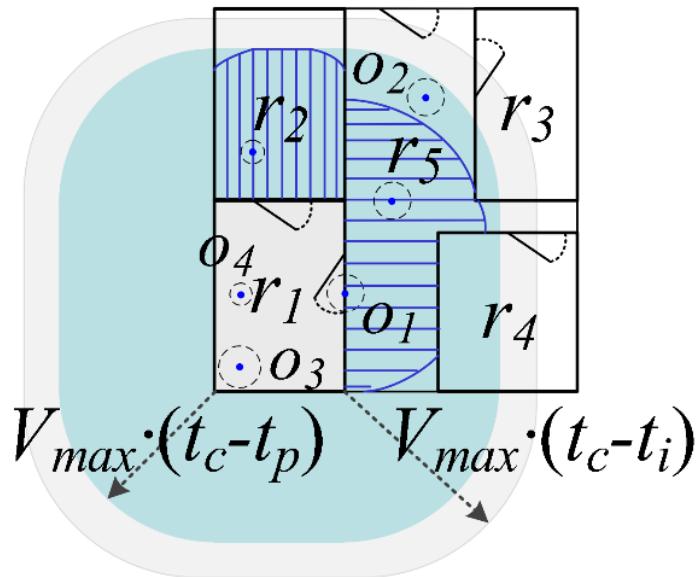
算法 3.1 DetermineIbcRs(Region r , Distance δ)

```
1  $\Theta_I^{\triangleright} \leftarrow r; \Theta_I^{\triangleleft} \leftarrow r$ 
2 for  $r$ 's each side  $\lambda$  do
3   for each door  $d$  on  $\lambda$  do
4      $r_m \leftarrow Range_I(d, \delta)$ 
5      $\Theta_I^{\triangleright} \leftarrow \Theta_I^{\triangleright} \cup r_m; \Theta_I^{\triangleleft} \leftarrow \Theta_I^{\triangleleft} \setminus r_m$ 
6   for each open segment  $g$  on  $\lambda$  do
7     find the indoor partition  $p$  that contains  $g$ 
8     get  $g$ 's general buffer region  $\Theta^{\triangleright}(g) \leftarrow \mathcal{M}(g, \delta)$ 
9      $r_m \leftarrow p \cap \Theta^{\triangleright}(g)$ 
10     $\Theta_I^{\triangleright} \leftarrow \Theta_I^{\triangleright} \cup r_m; \Theta_I^{\triangleleft} \leftarrow \Theta_I^{\triangleleft} \setminus r_m$ 
11    for each door  $d \in P2D(p)$  and in  $\Theta^{\triangleright}(g)$  do
12      get the shortest indoor distance  $\delta'$  from  $d$  to  $g$ 
13       $\Theta_I^{\triangleright} \leftarrow \Theta_I^{\triangleright} \cup Range_I(d, \delta - \delta')$ 
14 return  $\Theta_I^{\triangleright}, \Theta_I^{\triangleleft}$ 
```

按照门的拓扑结构迭代向外扩散计算

密度上下界和不确定区域的距离衰减(pp. 38-39)

基于缓冲/核心区域的密度上下界



- 引理3.5 室内区域密度上（下）界：
上界包含 r_2 和 r_5 中的对象；
- 引理3.6 时态松弛密度上（下）界：
更老的 t_i 时刻计算的上界包含对象 O_2 ；

对象移动的距离衰减效应

缩写	基本形式	名称
LDL	$\Gamma(\delta) = 1 - \delta/D$	Linear Decay Law
I1PL	$\Gamma(\delta) = 1/(\delta + 1)$	Inverse 1 st Power Law
I2PL	$\Gamma(\delta) = 1/(\delta + 1)^2$	Inverse 2 nd Power Law
EDL	$\Gamma(\delta) = e^{-\delta}$	Exponential Decay Law
CL	$\Gamma(\delta) = C$	Constant Law

目的地据当前位置越远，对象到达该位置可能性越低。（距离衰减函数-单调非增）

距离衰减对象出现度

$$\phi_r^{\Gamma}(o) = \frac{\int_{l \in (UR_I(loc) \cap r)} \Gamma(dist_I(loc, l)) dl}{\int_{l \in UR_I(loc)} \Gamma(dist_I(loc, l)) dl}$$

距离函数 $dist_I()$ 使用最短室内移动距离^[80]计算。



Top- k 室内密集区域挖掘算法

基于上下界的剪枝 (pp.40)

剪枝规则 3.1 (上下界剪枝) 给定两个室内区域 r_1 和 r_2 , 以下关于密度的性质始终成立:

- 1) 当 $LowerBound(\tau_O(r_1)) > UpperBound(\tau_O(r_2))$ 时, 有 $\tau_O(r_1) > \tau_O(r_2)$ 。此外, 当区域 r_1 的密度下界 $LowerBound(\tau_O(r_1))$ 在所有区域中为第 k 高时, 区域 r_2 可被安全剪除, 无需计算其密度 $\tau_O(r_2)$;
- 2) 当 $\tau_O(r_1) > UpperBound(\tau_O(r_2))$ 时, 有 $\tau_O(r_1) > \tau_O(r_2)$ 。当区域 r_1 的密度 $\tau_O(r_1)$ 为当前 $top-k$ 结果中最低时, 区域 r_2 同样可被安全剪除。

算法 3.2 TopkIDRs(Indoor query region set Q , Partition R-tree R_P , Online indoor positioning table OIPT, Current time t_c)

```

1 initialize a max-heap  $H$ 
2 initialize a hash table  $h_Q : Q \rightarrow \{(2^{ObjectID}, 2^{ObjectID})\}$ 
3 initialize a lower bound density set  $S_{\perp}$ 
4  $t_{min} \leftarrow \min\{rec.t \mid rec \in OIPT\}$ 
5  $\delta \leftarrow V_{max} \cdot (t_c - t_{min})$ 
6 for each region  $r \in Q$  do
7    $ibr, icr \leftarrow \text{DetermineIbcRs}(r, \delta)$ 
8    $(set_{\top}, set_{\perp}) \leftarrow \text{COUNT4ibcRs}(ibr, icr, R_P)$ 
9    $h_Q[r] \leftarrow (set_{\top}, set_{\perp})$ 
10  add  $\frac{|set_{\perp}|}{Area(r)}$  to  $S_{\perp}$ 
11  $kbound \leftarrow$  the  $k$ -th highest in  $S_{\perp}$ 
12 for each region  $r \in Q$  do
13    $(set_{\top}, set_{\perp}) \leftarrow h_Q[r]$ 
14   if  $\frac{|set_{\top}|}{Area(r)} \geq kbound$  then
15     enheap( $H, \langle r, OE\_IBR, \frac{|set_{\top}|}{Area(r)} \rangle$ )
16 return Search( $H, h_Q$ )

```

总体框架 (pp.40-41)

- 使用最大堆控制候选区域（密度估计值）处理顺序；
- R-tree索引室内分区；
- 使用哈希表维护与区域密度相关的对象引用集合；
- 使用OIPT最旧时间戳进行时态松弛上下界的估计；
- $kbound$ 过滤（规则1）；
- 调用搜索函数：当前 k 个密集区域找到后返回（规则2）。

总体框架后不同的搜索策略

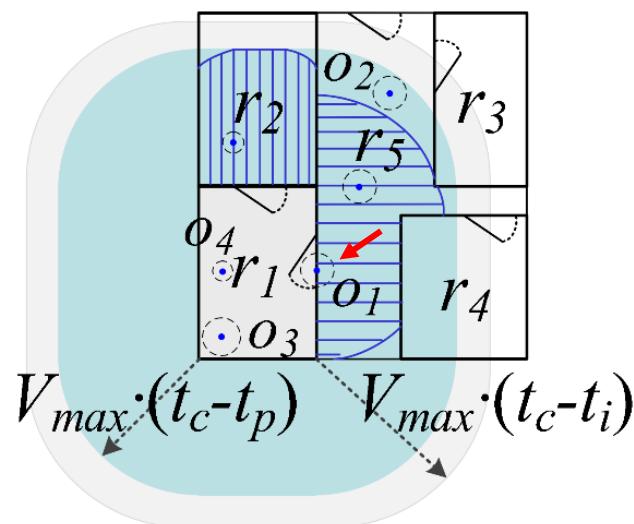
(pp. 41-45)

单步 (one-pass) 搜索 - 算法3.4

两个标记 (Flag) :

- OE_IBR: 当前密度是基于室内缓冲区域上估 (overestimate) ;
- IR: 当前密度是根据室内区域进行精确计算的;

```
for each object  $o \in set_{\top} \setminus set_{\perp}$  do
    if  $UR_I(o.loc)$  is fully contained in  $r$  then
        count  $\leftarrow$  count + 1
    else if  $UR_I(o.loc) \cap r \neq \emptyset$  then
        count  $\leftarrow$  count +  $\phi_r^{\Gamma}(o)$ 
```



两步 (two-passes) 搜索 - 算法3.5

三个标记 (Flag) :

- OE_IBR: 当前密度是基于室内缓冲区域上估;
- OE_IR: 更严格的数据上界，基于超量计数上界进行上估;
- IR: 当前密度是根据室内区域进行精确计算的;

```
if  $UR_I(o.loc)$  is fully contained in  $r$  then
    countc  $\leftarrow$  countc + 1
else if  $UR_I(o.loc) \cap r \neq \emptyset$  then
    countu  $\leftarrow$  countu + 1; add  $o$  to  $set_u$ 
```

$OverCount(r)$: 不确定区域与区域 r 相交的对象的数量。 (引理3.7)

$$\tau_O(r) \leq \frac{OverCount(r)}{Area(r)} \leq \frac{COUNT(\Theta_I^\triangleright(r))}{Area(r)}$$



对比方法

- 基于单步搜索和两步搜索的密集区域挖掘方法分别记为**TopkIDRs1Pass**和**TopkIDRsImprd**；
- 查询集Q越大，后者的性能增益越明显，k值越大，性能增益将下降；(pp.45 改进方法的性能增益分析)
- DC：直接计数，统计OIPT中被查询区域r包含的报告位置数量来作为r的负载量，并进行全排序返回top-k结果。
- NLRegion：查询区域为导向的循环嵌套；
- NLObject：移动对象为导向的循环嵌套；
- NLwgbr：采用一般缓冲区域剪枝对象；
- NLwibr：采用室内缓冲区域剪枝对象；
- 除DC外，都为不确定模型(UM)；
- 后五种方法的复杂度分析(pp.46)。

Java实现；PC主机运行。

度量模型

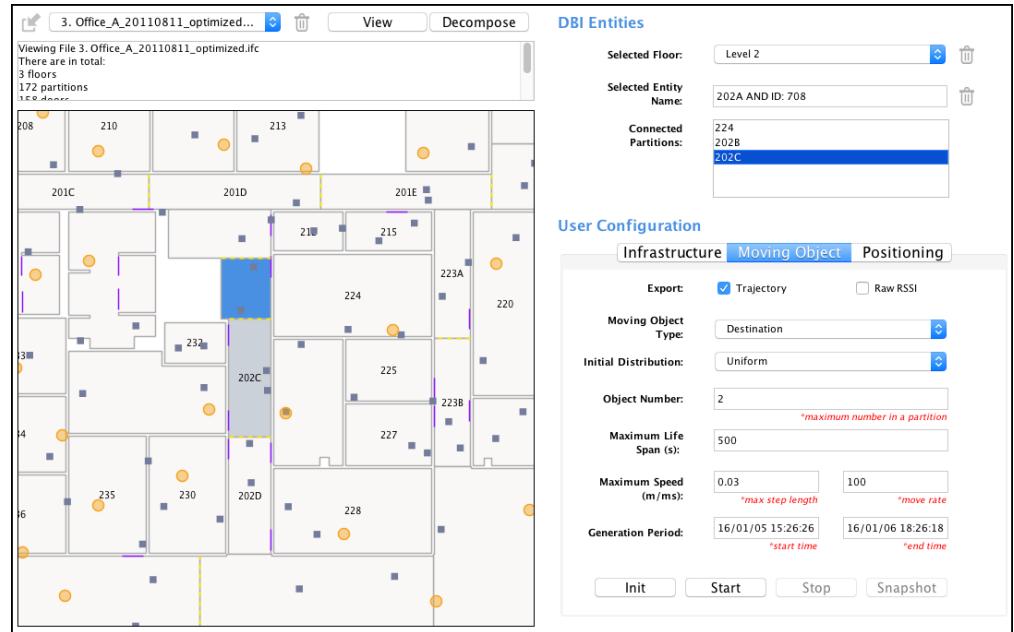
- 计算效率 (Efficiency)
 - 多次运行的平均时间开销；
 - 剪枝率：返回结果后无需计算出现度的对象的比例；
- 结果有效性 (Effectiveness)
 - 召回率：top-k挖掘结果中为真实top-k密集区域的比例；
 - Kendall系数：挖掘结果和真值top-k结果的排序一致性度量；1代表完全吻合，-1代表完全反序；

$$\tau = \frac{cp - dp}{0.5k(k - 1)}$$

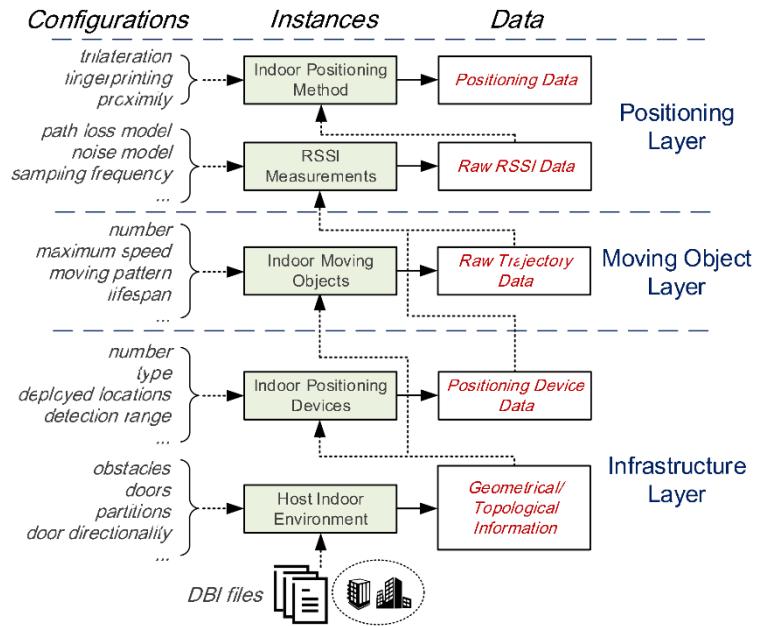
合成数据集实验



模拟数据工具集 – Vita (<http://longaspire.github.io/vita/>)



[VLDB16] 模拟真实的室内定位环境并生成各类室内移动数据，分为基设层（解析CAD等文件构建模拟环境；生成各类定位元件）、对象层（生成移动轨迹真值）、定位层（支持多种算法，生成无线信号观测值和定位结果观测值）。开源项目，可扩展外部模型。



在本工作中的使用：

- 根据真实商场地图，生成10层室内空间；
- 完成2小时移动对象运动模拟；
- 仿真Wi-Fi指纹定位，生成OIPT；
- 产生实验真值（每秒记录的轨迹值）。

实验设定

(pp. 48-49)

表 3.4 查询室内区域的类型

Table 3.4 Types of Indoor Query Regions in Q

类型	意义
ir_1	非完整的室内分区
ir_2	完整的室内分区
ir_3	两个及以上 ir_1 区域，或两个及以上 ir_2 区域，或多个 ir_1 与 ir_2 区域的组合。各组成分量为强连通。

参数	设定
k	1, 3, 5, ..., 15
$ O $	5K, 10K , 15K, 20K
$ Q $ (占所有室内分区比例)	2%, ..., 10% , ..., 14%
Q 中 ir_1, ir_2, ir_3 占比	40%, 50%, 10%
Δt (s)	1, 2, ..., 5 , ..., 10
距离衰减函数	CL, LDL, I1PL, I2PL, EDL

算法	运行时间 (millisec.)	剪枝率 (%)	内存开销 (MB.)
TopKIDRs1Pass	399.7	81.56	147.8
TopKIDRsImprvd	365.7	85.02	156.1
DC	68.5	-	2.2
NLRegion	148386.2	0	342.5
NLObject	2248.1	0	321.3
NLwibr	1082.2	60.74	68.6
NLwgbr	1597.7	34.85	51.2

▷ 默认参数下计算效率对比

DC: 很低的时间与内存开销, 但挖掘结果质量极低, 相关结果可参见pp.51 (UM vs. DC);

UM方法中: 提出方法比其它NL方法快几个数量级; 高效的密度上下界剪枝;

两步搜索方法优于单步搜索方法, 且其剪枝效率更高。

高效性和有效性验证

高效性验证 (pp.49-51)

- 通过对单步和两步搜索方法比较发现， k 值越大，后者的性能增益越小；
- 移动对象数量达到20K时，两个方法仍可在毫秒级完成（普通PC）；
- $|Q|$ 越大，两步搜索算法的性能增益越大；
- 越复杂的自定义查询区域需要越大计算代价，受到其缓冲/核心区域复杂性影响；
- OIPT中数据越老旧，越较大幅度提升挖掘算法的时间开销，但两个算法依然维持在毫秒级；
- 越复杂的距离衰减函数建模，需要时间越长。

有效性验证 (pp.51-53)

- 相比于确定性模型的DC算法，UM算法的有效性度量都较高；
- $|Q|$ 值固定的情况下， k 增大时两项有效性度量都提高；
- 有效性随着OIPT中数据的老旧程度提升而下降，但UM方法在实验调整中始终有高于0.91的召回率；高于0.72的Kendall系数；
- 采用复杂的指数衰减函数，能取得最高的有效性结果，Kendall系数高达0.82；逆一次幂律 (I1PL) 函数效果也较好；均匀分布不能很好反映移动对象的不确定移动情况。

部署定位系统：土木科技楼5F，Wi-Fi指纹算法，长期定位精度~4m

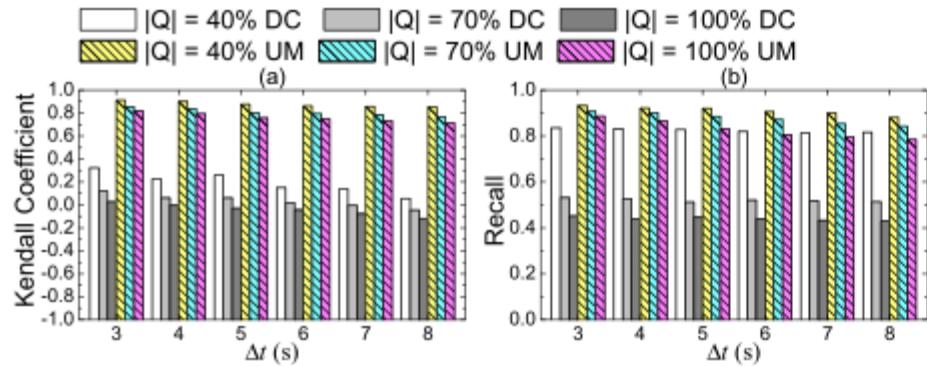


图 3.17 Δt 和 $|Q|$ 对结果效力的影响 [真实数据集]

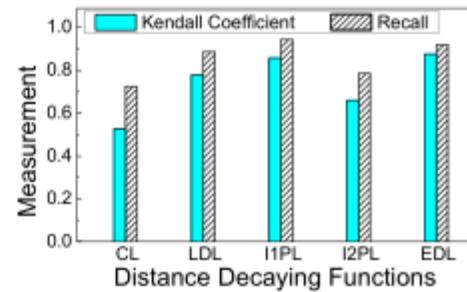


图 3.18 DDF 对结果效力的影响 [真实数据集]

有效性验证：

- UM方法显著优于DC方法， $|Q|$ 增大，两个方法有效性都降低，但UM受到影响更小；
- 随着OIPT中定位数据的老旧程度提升（时空不确定性增大），移动对象的不确定区域变大，有效性降低，当OIPT中最老的数据来自8s前时，UM的Kendall系数仍可达到0.7；
- 在真实数据集中，逆一次幂律(I1PL)函数取得和指数衰减函数(EDL)相当的有效性结果。这是因为测试空间障碍物过多，创建许多固定长度的室内通道，使得I1PL更擅长捕捉数据收集者的移动与距离衰减关系。



本项工作小结

- 使用了具有时空不确定性的快照定位数据，仅包含了每个对象的最新室内位置报告。
- 首先对室内密度的**定义和计算模型**进行了合理设计，可以适应由**离散的、老旧的**室内定位结果引起的对象位置的不确定性。
- 对密集区域挖掘计算中涉及数据的不确定性进行了全面分析：
 - 推导得出了室内区域密度的有效上界和下界；
 - 在密度计算中引入了**新型的距离衰减函数**。
- 利用分析的结果，我们设计了**高效的**top-k密集区域挖掘算法。
- 利用合成和真实数据集进行了**全面实验评估**：
 - 提出的室内密集区域挖掘算法是高效、可扩展及有效的。
 - 只使用存在不确定性的在线移动数据，不对对象移动的额外知识进行假设，得到的top-k密集区域仍能与真实情况保持高度一致。
- 相关长文发表在TKDE上。



目录

1 研究背景和动机

2 研究基础和现状

3 室内密度分析挖掘

4 室内流量分析挖掘

5 室内移动语义挖掘

6 结论和展望

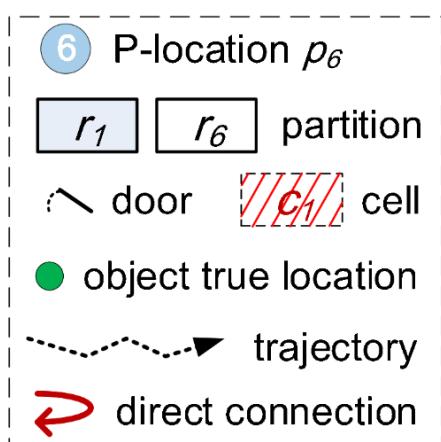
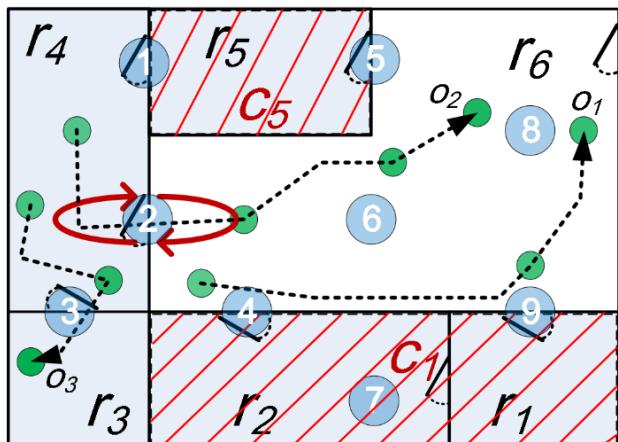


- 研究动机：
 - 室内流量 - 过去特定时间范围内通过特定室内区域的人数；
 - 流量信息具有重大价值，如位置相关的广告推荐^[30]和室内资源优化^[24]等场景；
 - 找出特定时间段内具有最大流量的热点室内语义位置 (semantic locations)；
 - 大型展览 - 每个展品区域为一个语义位置，流量高受欢迎，可进行展品推荐和优化；
 - 商场运营 - 每个店铺为一个语义位置，了解流量可辅助租金收取方案制定。
- 移动数据的时空不确定性 - 概率化采样：
 - 对象在过去时刻t的定位信息由格式为 (loc, prob) 的一组样本 (samples) 表示；
 - 为应对定位元件测量精度和环境波动更为鲁棒的一种格式选择；
 - 常见于加权kNN的Wi-Fi指纹算法^[46]，及蓝牙beacon、Wi-Fi 探针系统^[30,58,60]中。
- 研究挑战：
 - 如何对语义位置的流量值进行可靠有效的分析计算？
 - 室内移动数据的不确定特点使得区域内的对象个数无法直接算出；
 - 计算室内流量时必须适当考虑室内拓扑结构的特殊性。
 - 流量分析中繁重的计算量：
 - 需考虑所有被观测对象在特定时刻的多个位置样本，时/空复杂度极高；
 - 必须适时地找到仅与目标计算对象相关的数据部分，借助高效的计算策略来加快热点语义位置的分析挖掘过程。

- 输入 室内不确定定位表 (IUPT) :

oid, X, t	oid, X, t
$o_1, \{(p_4, 1.0)\}, t_1$	$o_1, \{(p_8, 1.0)\}, t_4$
$o_2, \{(p_1, 0.5), (p_2, 0.5)\}, t_1$	$o_2, \{(p_5, 0.3), (p_6, 0.6), (p_8, 0.1)\}, t_5$
$o_3, \{(p_2, 0.6), (p_3, 0.4)\}, t_2$	$o_3, \{(p_2, 0.4), (p_3, 0.6)\}, t_5$
$o_1, \{(p_9, 1.0)\}, t_3$	$o_2, \{(p_5, 0.2), (p_6, 0.3), (p_8, 0.5)\}, t_6$
$o_2, \{(p_2, 0.7), (p_4, 0.3)\}, t_3$	$o_3, \{(p_3, 1.0)\}, t_8$

$$\sum_{e \in X} e.prob = 1$$



语义位置 (**S-location**) : region, 与分析人员兴趣相关的区域位置, 如会议室、商场里的急救站等;

定位位置 (**P-location**) : point, 通常为指纹定位的离线参考点或者临近定位里 (RFID、蓝牙) 设备的位置;

室内单元 (**cell**) : 被一组P-location 分隔出来的单元, 从一个单元到另一个单元必须经过这组P-location之一 (称为 **分隔P-location**) ;

出现P-location: 在一个室内单元内部, 仅表示对象出现在这个单元内部。

流量计算模型

出现度计算方法 (pp.59-61)

- 对每个对象 o , 获取其样本集合序列, 通过笛卡尔积获得所有可能候选室内路径及其相应概率 pr_i 。
- 通过拓扑结构过滤无效候选路径 (p_3, p_4)。
- 每条有效路径 (P-location 序列) 通过语义位置 q 的概率:
 - 连续P-location对 (p_2, p_2) 通过位置 (q) 的概率

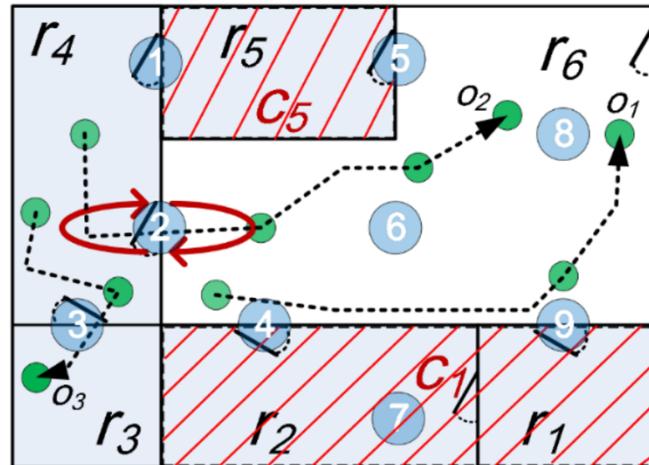
$$pr_{loc_j, loc_{j+1} \rightsquigarrow q} = \frac{|\{c \in C | c \text{ covers } q\}|}{|C|}$$

- 排除所有P-location对通过位置的概率
- $$pr_{\phi \rightsquigarrow q} = 1 - \prod_{1 \leq j \leq n-1} (1 - pr_{loc_j, loc_{j+1} \rightsquigarrow q})$$
- 计算对象 o 在位置 q 的**出现度 (presence)** :

$$\Phi_{t_s, t_e}(q, o) = \frac{\sum_{\phi_i \in P} (pr_{\phi_i \rightsquigarrow q} \cdot pr_i)}{\sum_{\phi_i \in P} pr_i}$$

流量定义 (pp.61)

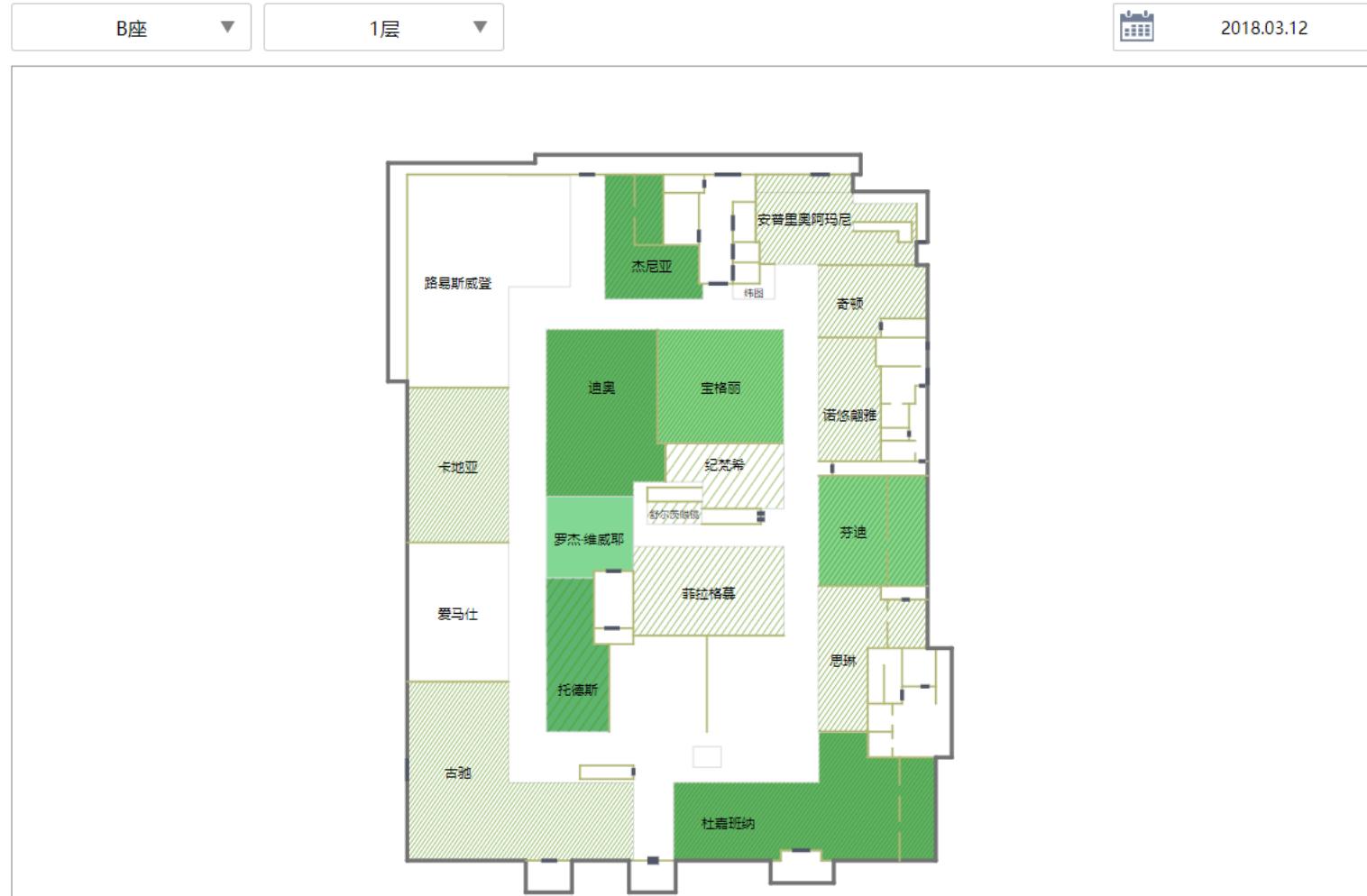
$$\Theta_{t_s, t_e, O}(q) = \sum_{o \in O} \Phi_{t_s, t_e}(q, o)$$

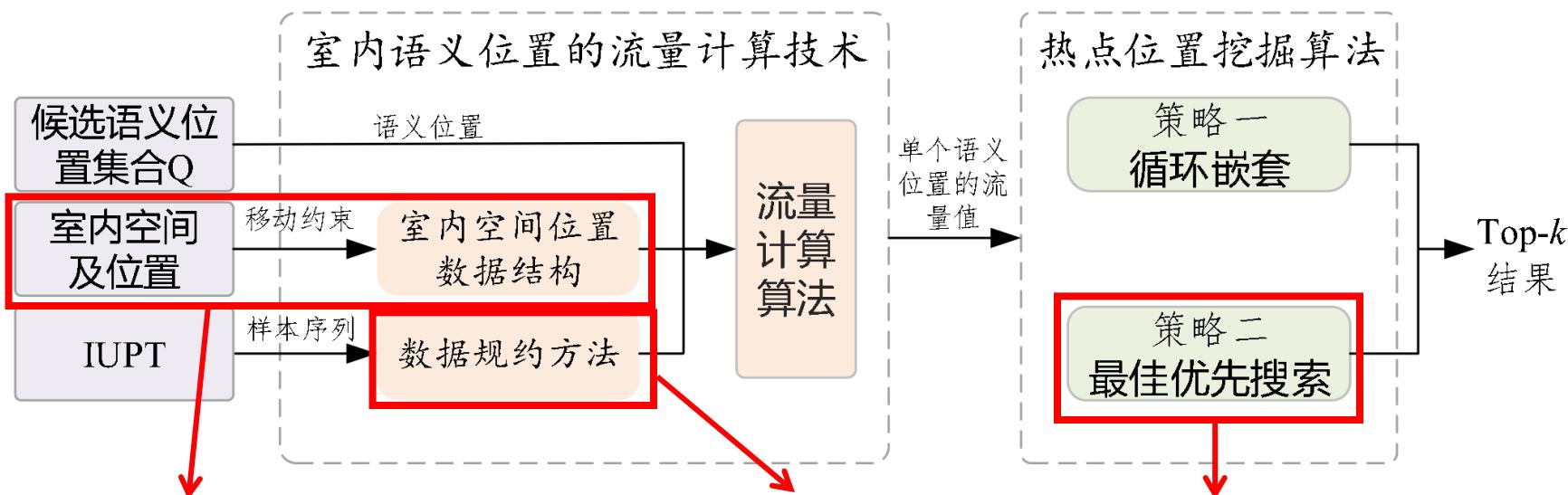


oid, X, t	oid, X, t
$o_1, \{(p_4, 1.0)\}, t_1$	$o_1, \{(p_8, 1.0)\}, t_4$
$o_2, \{(p_1, 0.5), (p_2, 0.5)\}, t_1$	$o_2, \{(p_5, 0.3), (p_6, 0.6), (p_8, 0.1)\}, t_5$
$o_3, \{(p_2, 0.6), (p_3, 0.4)\}, t_2$	$o_3, \{(p_2, 0.4), (p_3, 0.6)\}, t_5$
$o_1, \{(p_9, 1.0)\}, t_3$	$o_2, \{(p_5, 0.2), (p_6, 0.3), (p_8, 0.5)\}, t_6$
$o_2, \{(p_2, 0.7), (p_4, 0.3)\}, t_3$	$o_3, \{(p_3, 1.0)\}, t_8$

问题定义

问题 4.1 (Top- k 热点语义位置挖掘) 给定室内语义位置的查询集合 Q , 室内不确定定位结果表 $IUPT$, 室内对象集合 O 和特定时间段 $[t_s, t_e]$, top- k 室内热点语义位置挖掘找出一个包含 k 个室内语义位置的子集 $Q_k \subseteq Q$, 其中 $\forall q \in Q_k, \forall q' \in Q \setminus Q_k, \Theta_{t_s, t_e, O}(q) \geq \Theta_{t_s, t_e, O}(q')$ 。





观测到 P-location，计算的是 S-location 的流量，建立两者在室内拓扑下的关联。

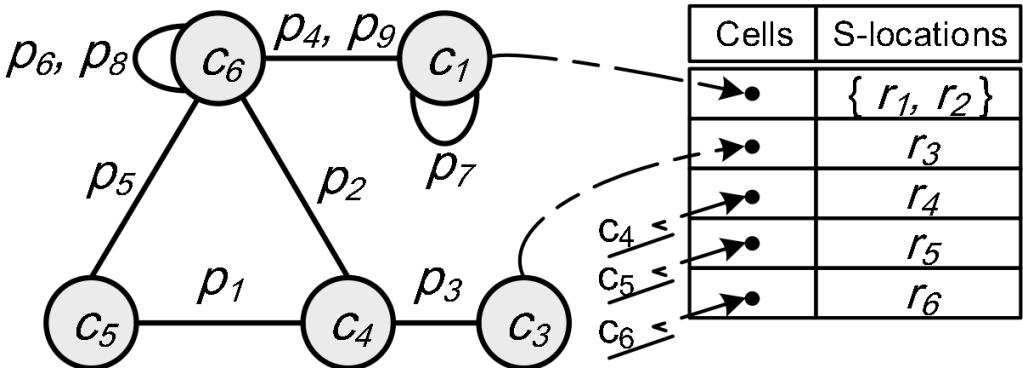
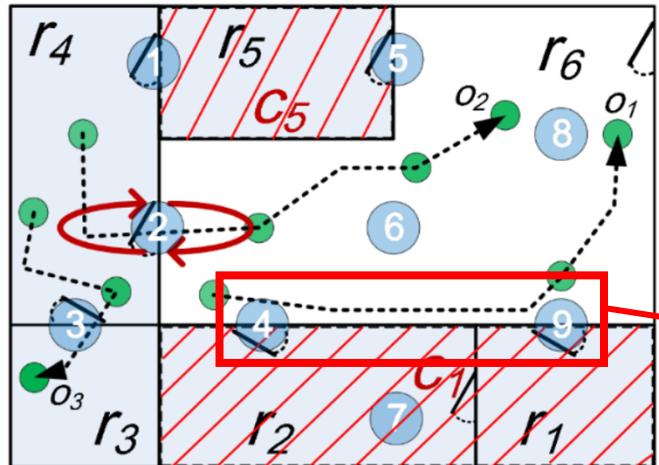
降低移动对象的样本序列规模。

快速剪枝和 top-k 无关的区域和对象。

单个室内语义位置的流量技术方法

加速数据读取和计算的数据结构 (pp.62-64)

室内空间位置图 (构建定位和语义位置拓扑关系)

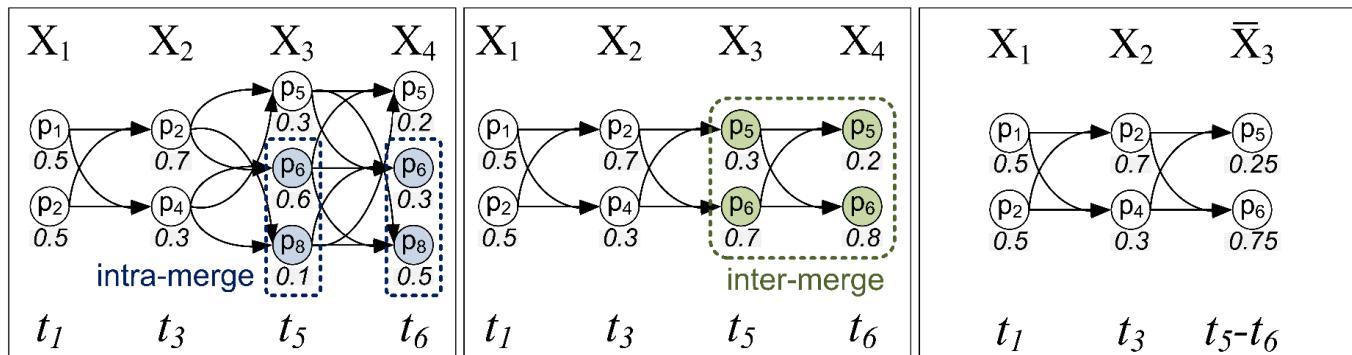


p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9
$\{c_4, c_5\}$		c_4	\emptyset	c_5	\emptyset	\emptyset	\emptyset	\emptyset
	$\{c_4, c_6\}$	c_4	c_6	\emptyset	c_6	\emptyset	c_6	c_6
		$\{c_3, c_4\}$	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset
			$\{c_1, c_6\}$	c_6	c_6	c_1	c_6	c_6
				$\{c_5, c_6\}$	c_6	\emptyset	c_6	c_6
					c_6	\emptyset	c_6	c_6
						c_1	\emptyset	c_1
							c_6	c_6
								$\{c_1, c_6\}$

室内位置矩阵 (通过定位观测映射到语义位置)

定位样本序列的数据规约方法

(pp. 64-67)



(a) raw sequence $|P| = 32$

(b) after intra-merge $|P| = 16$

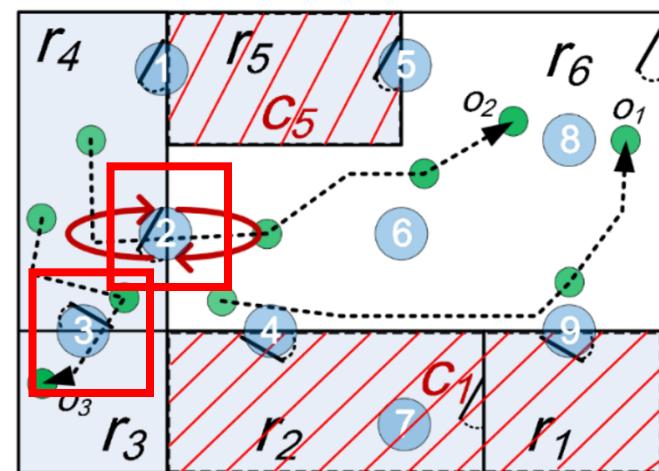
(c) after inter-merge $|P| = 8$

对象样本序列能够生成的最大可能路径数量为 $\prod_{1 \leq i \leq n} |\pi_l(X_i)|$

内合并操作 (inner-merge)：在每个时刻，合并在室内空间位置图上同一条边上的等价P-location；（原理：等价P-location能够搜索一组完全相同的室内单元）

间合并操作 (inter-merge)：多个连续的样本集合，若它们的P-location集合完全相同，可以进行近似合并；（原理：对象可能在一个位置停留很久，产生重复定位观测）

无关对象筛除：若与P-location相关 (overlap) 的所有可能语义位置都不在查询集合中，则对应的对象和其采样集合序列可以快速排除。



单个S-location的流量计算算法

(pp. 67-68)

算法 4.2 Flow(Indoor semantic location q , 1DR-tree $tree$, Query time interval $[t_s, t_e]$)

```
1 LeafEntrySet  $les \leftarrow tree.RangeQuery([t_s, t_e])$ 
2 initialize a hash table  $H_O : \{oid\} \rightarrow \{\mathcal{X}\}$ 
3 for each leaf entry  $le \in les$  do
4     append  $le.X$  to  $H_O[le.oid]$ 
5  $flow \leftarrow 0$ 
6 for each key  $oid \in H_O.keys$  do
7      $\langle (X_1, \dots, X_n), psls \rangle \leftarrow ReduceData(H_O[oid], \{q\})$ 
8     if  $psls$  is null then continue
9     path set  $P \leftarrow \{\langle (loc, prob) \rangle \mid (loc, prob) \in X_1\}$ 
10    for  $i$  from 1 to  $n$  do
11        for each path  $\phi \in P$  do
12            remove  $\phi$  from  $P$ 
13            for each sample  $e \in X_i$  do
14                if  $M_{IL}[\phi.tail.loc, e.loc] \neq \emptyset$  then
15                     $\phi' \leftarrow append(\phi, e); add \phi' to P$ 
16     $pr \leftarrow 0; pr_{sum} \leftarrow 0$ 
17    for each path  $\phi \in P$  do
18         $pr_\phi \leftarrow \prod_{1 \leq j \leq |\phi|} \phi[j].prob; pr_{sum} \leftarrow pr_{sum} + pr_\phi$ 
19        if  $pr_{\phi \sim q} > 0$  then
20             $pr \leftarrow pr + (pr_{\phi \sim q} \cdot pr_\phi)$ 
21     $flow \leftarrow flow + \frac{pr}{pr_{sum}}$ 
22 return  $flow$ 
```

1DR-tree 对样本序列进行时间属性索引

对每个对象的样本序列进行数据规约

每个时刻往下迭代，通过笛卡尔积计算候选路径集合

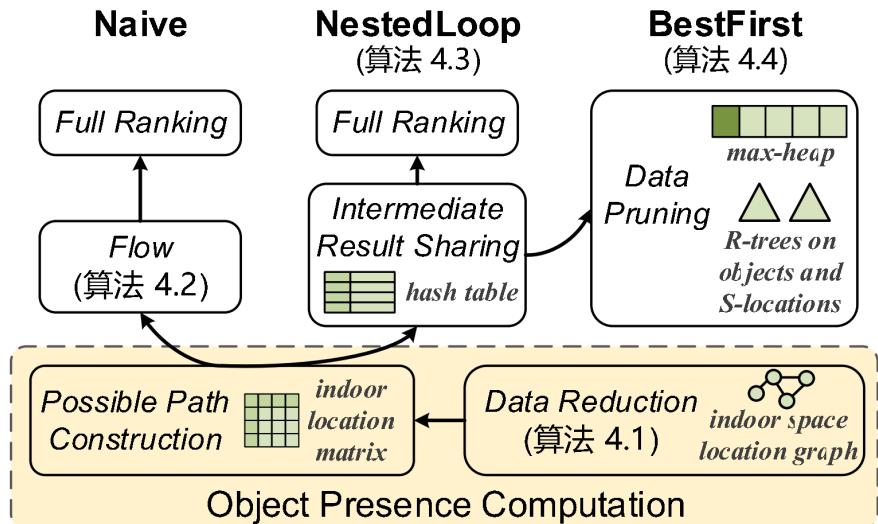
通过检查位置矩阵来验证候选路径的有效性；
避免生成许多无效的路径分支，存入硬盘

$\triangleright \phi$ has chance to pass $Cell(q)$

计算每条可能路径对语义位置的流量贡献

Top-k室内热点语义位置挖掘算法

不同算法的搜索策略 (pp.68)



Naïve方法: 为每个查询语义位置调用Flow算法;

NestedLoop方法: 通过哈希表，实现了中间计算结果的共享，每个对象处理后，其对所有相关语义位置的出现度被缓存下来；

BestFirst方法: 加入空间索引剪枝和最佳优先搜索策略，对无希望的候选位置和移动对象进行排除。

循环嵌套算法 (pp.69-70)

```

10 initialize a hash table  $H_\phi : \{path\} \rightarrow 2^Q$ 
11 for  $i$  from 1 to  $n$  do
12     for each path  $\phi \in P$  do
13         remove  $\phi$  from  $P$ 
14          $list_Q \leftarrow$  remove  $H_\phi[\phi]$  from  $H_\phi$ 
15         for each sample  $e \in X_i$  do
16             if  $M_{IL}[\phi.tail.loc, e.loc] \neq \emptyset$  then
17                  $\phi' \leftarrow$  append( $\phi, e$ ); add  $\phi'$  to  $P$ 
18                  $list'_Q \leftarrow (C2S(M_{IL}[\phi.tail.loc, e.loc]) \cap Q)$ 
19                  $H_\phi[\phi'] \leftarrow list_Q \cup list'_Q$ 
20 initialize a hash table  $H_{ls} : Q \rightarrow \{score\}$ 
21  $pr_{sum} \leftarrow 0$ 
22 for each path  $\phi \in P$  do
23      $pr = \prod_{1 \leq j \leq |\phi|} \phi[j].prob$ ;  $pr_{sum} \leftarrow pr_{sum} + pr$ 
24     for each query S-location  $q \in H_\phi[\phi]$  do
25          $H_{ls}[q] \leftarrow H_{ls}[q] + (pr_{\phi \leadsto q} \cdot pr)$ 
26 for each query S-location  $q \in H_{ls}.keys$  do
27      $H_Q[q] \leftarrow H_Q[q] + \frac{H_{ls}[q]}{pr_{sum}}$ 
28 return the top- $k$  from  $H_Q.keys$  with the highest scores

```

- 阶段一（行1--10）：
 - 为每个对象进行定位样本序列的连接操作；存入哈希表（objectID为主键）；
 - 对移动对象（其可能语义位置的MBR）进行R-tree索引，记为 R_C ；
 - R_C 索引的每个非叶结点记录了其包含所有条目的个数（用于上界估计）；
- 阶段二（行11--18）：
 - 将移动对象R-tree R_C 和查询语义位置R-tree R_Q 进行Join；
 - 对每个查询语义位置，其相交的 R_C 的非叶结点记录的子条目个数可以作为其流量值的上估（overestimate）-因为对象在任何语义位置的出现度都不可能大于1；
 - 将流量值和对象的语义位置q存入最大堆中进行处理；
- 阶段三（行19--43）：
 - 按照最大堆顺序优先处理更高流量估计值的语义位置；
 - 在两个join的R-tree上进行搜索，直到 R_Q 到达叶子（查询语义位置）或者 R_C 到达叶子（移动对象）；
 - 如果一个查询语义位置q已经处理完毕，加入到top-k结果中，如果已经有k个，返回；
 - 如果查询语义位置（或非叶结点）碰到一组移动对象，就计算进行这些移动对象对多个语义位置的出现度贡献；
 - 流量估计值很低的语义位置在top-k结果返回前可能无需计算。



对比方法

- **Naive**、**NestedLoop (NL)** 、**BestFirst (BF)**；
- **SC**: 简单计数，选取定位记录中最高概率的（第一个）样本，并丢弃其它样本。如果选取样本相应的P-location被某个语义位置q包含，则q的流量值加1；
- **SC-p**: 选取所有概率超过给定阈值p的样本。SC和SC-p都允许一个P-location被多个S-location所计入，但SC-p可能考虑更多的样本和P-location；
- **MC**：进行一定轮次的蒙特卡洛模拟，每轮生成一个IUPUT的确定性实例。随后，它通过在确定性定位记录上构建有效路径来计算每个候选位置的流量值；
- 使用原始序列不进行数据规约的方法：**Naive-ORG**、**NL-ORG**、**BF-ORG**。

Java实现；PC主机运行。

度量模型

- 计算效率 (Efficiency)
 - 多次运行的平均时间开销；
 - 剪枝率：无需计算耗时的对象出现度的对象的比例；
- 结果有效性 (Effectiveness)
 - 召回率：top-k挖掘结果中为真实top-k密集区域的比例；
 - Kendall系数：挖掘结果和真值top-k结果的排序一致性度量；1代表完全吻合，-1代表完全反序；



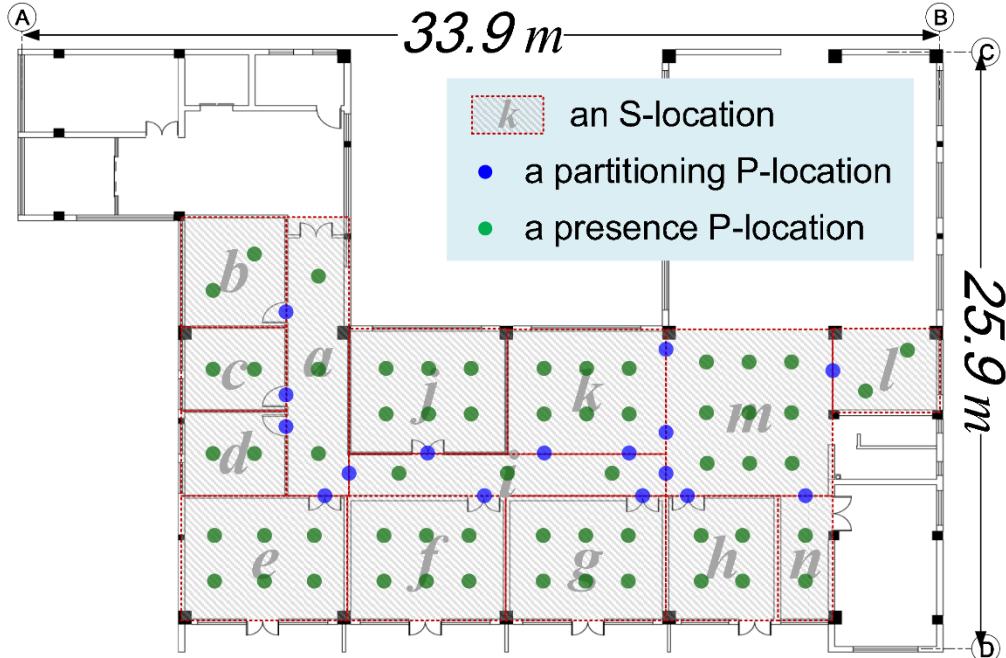
土木科技楼5F 14个语义位置：9个办公室和5个过道；

75个P-location中16个为分割P-location；

截取最高流量的150分钟，共计35个用户的64,846条不确定定位记录；

最大样本容量 (mss, maximum sample-set size) 为4，即一个样本集合最多包含4个P-location采样；最大定位间隔T为3s。

选取数据的平均定位精度约为2.1m。



	参数	设定
查询区域占比	k	1, 2, 3 , ..., 8
最大样本容量	$ Q $ (% of S-locations)	20%, 40%, 60 %, 80%, 100%
查询时间段	mss	1, 2, 3, 4
	Δt (minute)	30 , 60, 90

方法	运行时间 (sec.)	剪枝率 (%)	Kendall 系数 τ	召回率 (%)
SC	0.6	-	0.007	62.2
SC- ρ ($\rho = 0.25$)	1.1	-	0.382	75.6
MC, 900 rounds	1.7×10^4	-	0.712	86.7
BF	4.4	59.4	0.859	93.3
NL	9.5	19.2	same as above.	
Naive	59.1	19.2	same as above.	
BF-ORG	1.4×10^4	50.3	0.893	95.6
NL-ORG	2.3×10^4	0	same as above.	
Naive-ORG	1.6×10^5	0	same as above.	

▷ 默认参数下计算效率和结果有效性对比

结论：

- SC和SC-p通过统计观测样本，无需构建候选路径，因此计算效率高；但有效性度量明显很低；
- 采用不确定流量计算模型，BF和NL的有效性度量明显很高（返回同样的top-k结果）；
- BF在计算效率和结果有效性方面取得了良好的平衡；
- 不采用数据规约的方法，如BF-ORG, NL-ORG和Naive-ORG比其对应的数据规约版本要慢多个数量级；这展示了数据规约方法的重要性；
- BF算法，包括BF-ORG算法，其剪枝效率都很高，而NL和Naive的19.2%的剪枝率是通过数据规约方法得到的；
- MC算法采用多轮模型，虽然每次模拟生成的确定性IUPPT实例计算代价低，但需要进行多次计算，故要达到同样的有效性结果，需要多几个数量级的计算时间。

高效性和有效性验证

不确定性带来的影响 (pp.76)

- 当位置报告中包含更多概率样本时，各算法的运行时间增加，MC算法要慢几个数量级，当最大包含4个位置样本时，BF算法仍可在4.4s内完成一次查询时长为半小时的热点语义位置挖掘；
- 更多概率样本能提升BF的结果有效性，不确定数据模型比确定性数据模型在解决流量分析问题时更为有效。

高效性验证 (pp.77-78)

- 随着k值增大，BF和NL的性能差异减小，当k等于 $|Q|$ 时，二者几乎等价；
- 随着 $|Q|$ 值增大，BF和NL的性能差异加大，BF的优势更为明显；
- 随着查询时间段的增大，BF的优势提升越发显著。

有效性验证 (pp.78-79)

- SC和SC-p的两项有效性度量都很差；MC方法尽管需要极大的时间开销，其最佳性能也差于BF方法；BF方法在各项参数调节下表现都最好；
- k值增大，有效性度量降低，但BF方法的召回率始终高于0.88；
- $|Q|$ 值增大，有效性度量降低，但即使对所有语义位置进行流量排名，Kendall系数仍高于0.75；召回率总体高于0.85；
- 查询时间段增大，有效性度量仅略有下降，这是由于考虑了室内拓扑结构，可以去除更多无效路径，从而更加毕竟真实的对象运动情况；

合成数据集实验



实验设置 (pp.79-80)

- Vita模拟器生成的**大规模**数据集；
- 5450个P-location; 649个S-location;
- 10K的移动对象数量；
- 模拟最大定位周期：1, 3, 5, 7s；
- 模拟定位误差：3,5,7 m；

有效性验证 (pp.82-83)

- BF的有效性度量始终优于MC、SC和SC-p算法；
- 随着k值增大，其有效性优势更加明显；
- 随着 $|Q|$ 增大，其有效性优势更加明显；BF方法的Kendall系数始终高于0.83，而简单计数方法则为负数；
- **有效性度量对移动对象数量不敏感**；
- 随着查询时间段增大，BF的有效性度量仅略为下降，优势更为明显。

不确定性带来的影响 (pp.81-82)

- 加大定位周期，观测数据越来越稀疏，计算效率提升，而BF算法的结果有效性仅略为下降；相比其他算法的优势更大；这是因为考虑拓扑结构来生成有效路径受到采样稀疏的影响更小；
- 加大定位误差，BF和NL算法的运行时间反而下降，这是因为排除了更多无效的候选路径；同时，结果有效性下降，但当定位误差高达7m时，BF方法Kendall系数仍高于0.77，召回率高于0.87；
- 在**移动数据具有较高不确定性时**，提出的BF算法仍可高效、有效地工作。



本项工作小结

- 使用了具有时空不确定性的历史室内定位数据，每个对象在过去某时刻的位置被描述为一组概率样本。
- 充分考虑了移动数据的时空不确定性特点和室内拓扑结构对对象移动的约束，并建立了有效的室内流量定义及计算模型。
- 提升流量计算的效率：
 - 提出了加速相关数据访问的数据结构；
 - 减少中间处理数据量的数据规约方法；
 - 总体的室内流量计算算法。
- 利用流量计算技术，我们设计了高效的top-k热点语义位置挖掘算法。
- 利用合成和真实数据集进行了全面实验评估：
 - 提出的数据规约方法可明显降低中间计算的数据量；
 - 提出的室内热点语义位置挖掘算法是高效、可扩展及有效的；
 - 数据具有较高不确定性时，仍能返回与真实情况高度一致的一组热点语义位置。
- 相关长文在TKDE审稿中。



目录

1 研究背景和动机

2 研究基础和现状

3 室内密度分析挖掘

4 室内流量分析挖掘

5 室内移动语义挖掘

6 结论和展望

• 研究动机：

- 移动行为分析 - 热点位置^[15,18]或路线^[12,14]发现、移动模式挖掘^[8,11,60]、店内营销^[25,30]；
- 一类行为分析侧重于语义层面 - 商场那些店铺组合是购物者最常购买的？；
- 采用的移动数据是几何坐标，需要翻译为带有语义信息的序列：

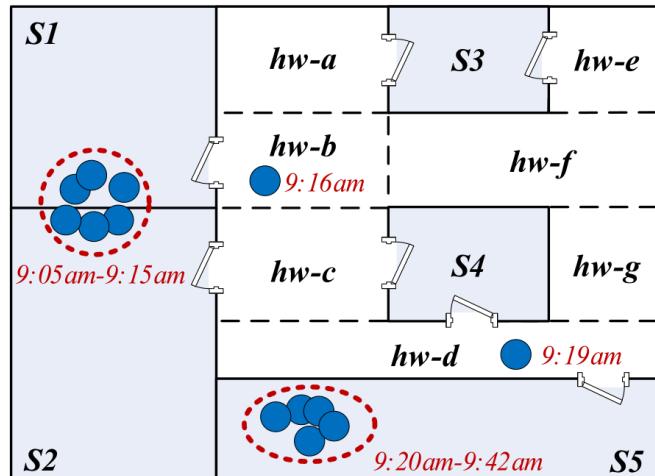
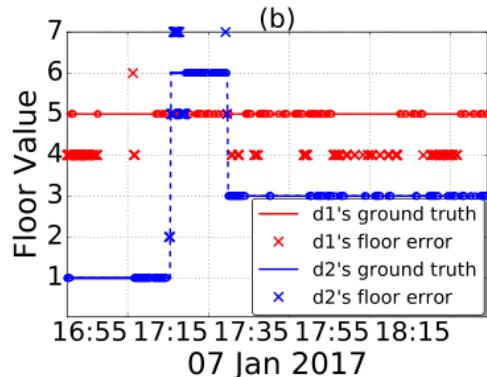
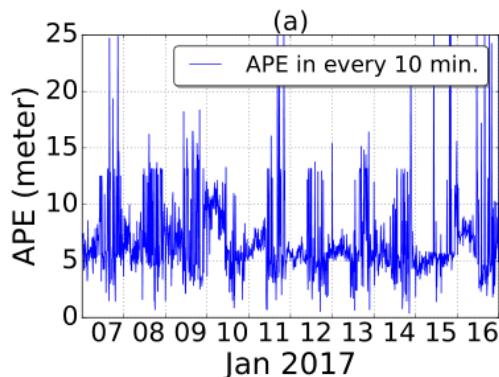
$o_1 : (\text{stay}, \text{Nike}, 1:02pm-1:18pm) \rightarrow (\text{pass-by}, \text{Adidas}, 1:19pm-1:20pm)$

$\rightarrow (\text{stay}, \text{Cashier}, 1:21pm-1:24pm)$

- (空间、时间、移动事件) 三元标注，利于理解、方便查询/存储、可视化。

• 研究挑战：

- 无线定位获得的原始移动数据非常“脏”：



- 室内空间中容纳了大量密集的实体，使得对象的运动变得复杂而难以进行标注；
- 移动设备倾向关闭无线开关，室内定位记录常常是稀疏和离散的。从稀疏的定位数据中获得完整的移动语义序列是很难的，具有多种可能性。



- 输入 室内定位结果表 (IPT) :
 - 每个移动对象的原始定位序列被称为 p-sequence;
- 移动语义 (mobility semantics) - 三元组:
 - 时间标注为一个时间段;
 - 空间标注为一个室内区域 (分析人员根据语义预先定义) ;
 - 事件标注为 pass-by (无意图直接路过) 和 stay (为满足特定需求而产生的停留) 。
- 输出 移动语义序列 (ms-sequence) 。
- 问题定义:

问题 5.1 (室内移动语义挖掘) 给定 IPT 、时间段 \mathcal{T} , 和一组室内区域 \mathcal{R} , 对 IPT 中任一对象 o_i , 室内移动语义挖掘将 o_i 的 p -sequence $\Theta_{o_i, \mathcal{T}} = \langle \theta_{i1}, \dots, \theta_{in} \rangle$ 翻译为对应的 ms -sequence $\Lambda_{o_i, \mathcal{T}} = \langle \lambda_{i1}, \dots, \lambda_{im} \rangle$ 。

o	$l(\mathbf{x}, \mathbf{y}, \mathbf{f})$	t
o_1	(2.5, 10.7, 1)	t_1
o_2	(5.1, 38.5, 4)	t_1
o_1	(2.3, 11.2, 2)	t_4

端到端的室内移动轨迹翻译套件 – TRIPS (<http://longaspire.github.io/trips/>)

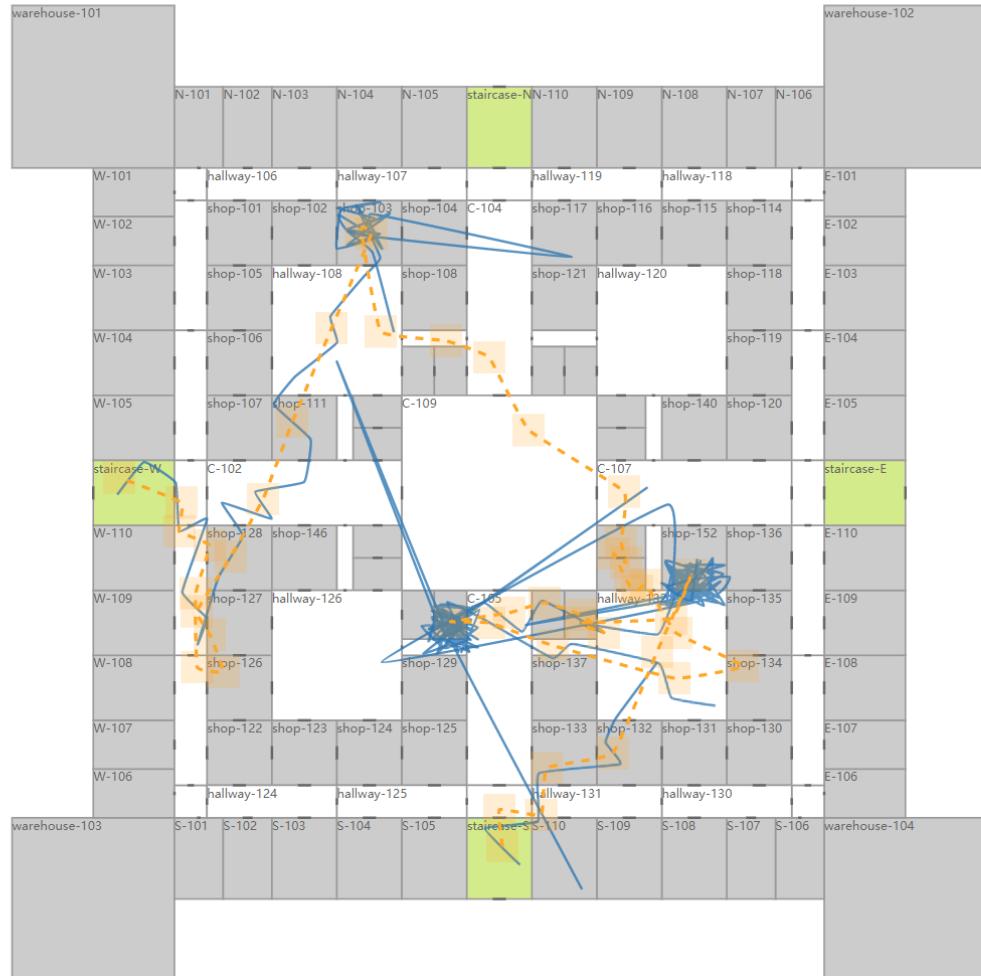
Current Filename: test-device

Legend

- Ground Truth
- Raw Data
- Cleaned Raw Data
- Mobility Semantics
- Tooltip
- Room
- Hallway
- Staircase
- Door
- False Floor Values

Floor Chooser

Floor Name	Number of Raw Records
floor-0	0
floor-1	351
floor-2	311
floor-3	329
floor-4	0

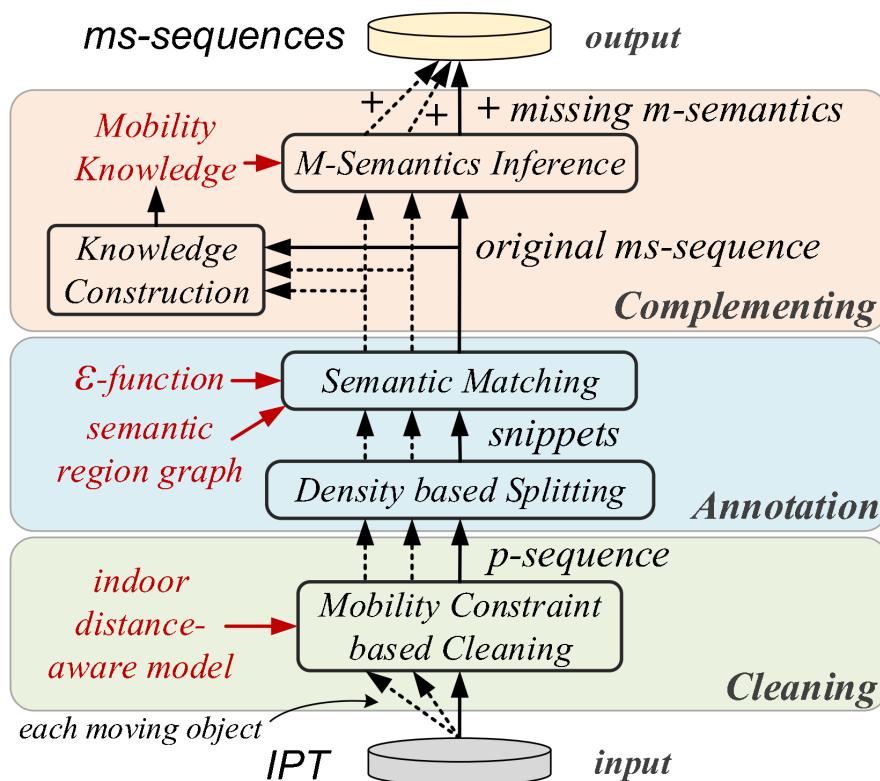


Mobility Semantics Timeline

stay pass-by

- 3F N-301 08:00:00--08:09:47
- 3F hallway-303 08:09:48--08:09:52
- 3F hallway-306 08:09:53--08:09:55
- 3F hallway-303 08:09:56--08:09:59
- 3F W-301 08:10:00
- 3F hallway-303 08:10:01--08:10:04
- 3F hallway-304 08:10:05--08:10:12
- 3F W-303 08:10:13--08:10:15
- 3F hallway-304 08:10:16--08:10:19
- 3F hallway-305 08:10:20--08:10:32
- 3F C-301 08:10:33--08:10:42
- 3F staircase-W 08:10:43--08:10:47
- 3F C-301 08:10:48--08:10:52
- 3F staircase-W 08:10:53--08:11:12
- 2F staircase-W 08:11:13--08:11:32
- 2F C-201 08:11:33--08:11:37
- 2F C-202 08:11:38--08:11:57
- 2F hallway-213 08:11:58--08:12:00
- 2F C-202 08:12:01--08:12:04

[VLDB18] 管理多源输入数据（定位源/事件训练数据/语义区域）、后台翻译、可视化展示。构建移动语义序列提供了一种直观、简洁的方式来理解室内移动对象的一般行为，因此也是进行上层的室内行为分析的必要基础。

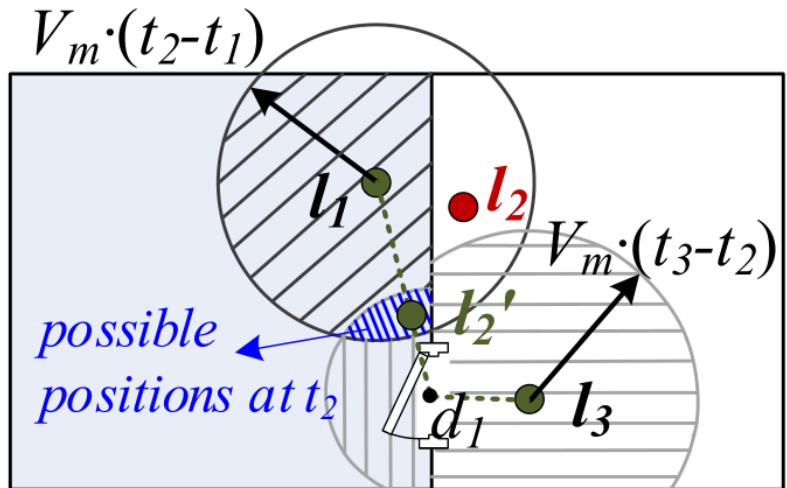


三层的挖掘模型，逐步提升数据质量并最终构建出有效的室内移动语义：

- **清洗层** 对每条p-sequence中的数据错误进行处理，充分考虑记录在距离敏感模型（distance-aware model）中的室内移动性约束来进行数据清洗；
- **标注层** 首先使用时空密度聚类将每条p-sequence分割为一组数据片段，并在事件识别模型和语义区域图模型的基础上，利用语义匹配将每个分割片段转换为一组语义元组；
- **补全层** 对标注层获得的每条原始ms-sequence进行数据补全。利用已标注的移动语义，构建移动知识。随后，利用移动知识进行概率推断，每条ms-sequence缺失的语义元组被生成并插入到原始序列中。



- 同时处理三种室内移动数据错误:
 - 随机错误（跳跃到其他室内分区）；
 - 位置异常值（重大偏差）；
 - 楼层错误值。
- 室内移动性约束识别错误:
 - 室内拓扑结构下的运动约束条件；
 - 基于室内距离敏感模型进行速度检查。
- 两阶段的修复过程:
 - 楼层值上平滑，修复存在的潜在错误；
 - 基于前后的有效位置进行室内平面位置的插值：



$$dist_I(\theta_p.l, l) = \frac{\theta_i.t - \theta_p.t}{\theta_s.t - \theta_p.t} \cdot dist_I(\theta_p.l, \theta_s.l)$$

- 错误的识别和修复可以从全局检查中找到的任一有效位置开始前向/后向处理。



算法 5.2 SplitMatchAnnotation(P-sequence Θ_o , Event identification function \mathcal{E} , Semantic region graph G_R)

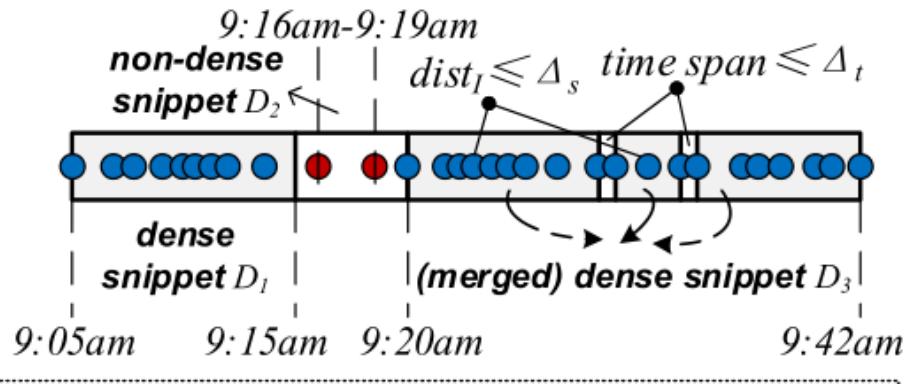
```

1 time-ordered sequence  $\Lambda_o \leftarrow \langle \rangle$ 
2  $\mathcal{A}_{snpt} \leftarrow DensityBasedSplitting(\Theta_o)$ 
3 for each snippet  $\Theta_o^*$  in  $\mathcal{A}_{snpt}$  do
4      $\Lambda_o^* \leftarrow SemanticMatching(\Theta_o^*, \mathcal{E}, G_R)$ 
5      $\Lambda_o \leftarrow \Lambda_o \cup \Lambda_o^*$ 
6 return  $\Lambda_o$ 
```

split-and-match

对每个分割片段进行三元组（时间、事件、空间）的语义标注

基于时空聚类
的序列分割



- 改进的ST-DBSCAN，找出空间和时间属性同时紧凑的记录片段，可能对应于停留事件：
 - 根据局部采样频度的自适应聚类参数调节；
 - 采用最小室内移动距离度量；
 - 加入时空容忍阈值，避免过度分割；

分割片段的语义匹配

事件标注 (pp.94)

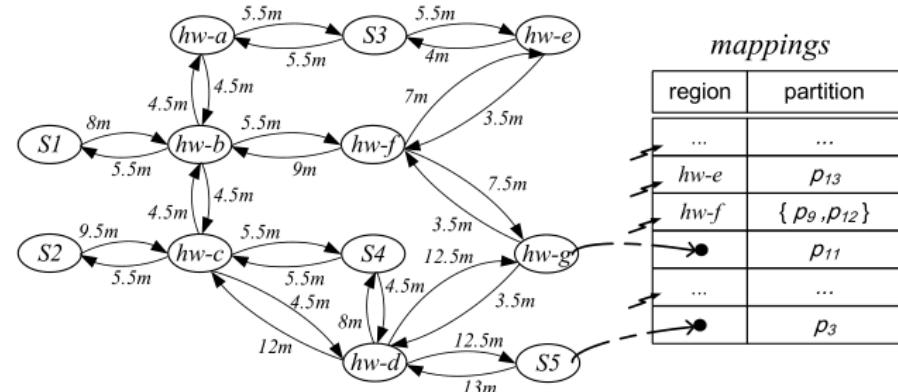
- 基于监督学习模型，设计了移动事件识别函数；
- 对每个分割片段抽取移动特征向量：
 - 密集程度；位置估计方差；采样条件；覆盖范围；相关区域；行走距离/速度；转向次数；
- 逻辑斯蒂回归模型进行停留/路过事件的分类；
- 引入co-training机制^[143]，以迭代形式将具有高置信度的预测结果作为新的训练数据加入到模型强化中。

$$dist_{gr}(r_i, r_j) = \max_{l \in r_i, d \in P2D_{\square}(R2P(r_j))} dist_I(l, d)$$

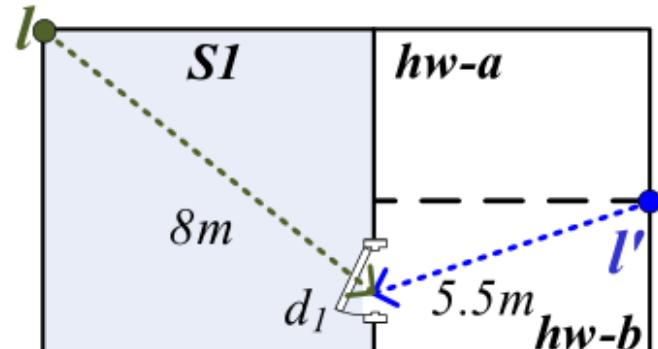
面积越大的语义区域，到其他区域的GRD越大，意味着通过这个区域需要的时间通常越多。

空间标注 (pp.95-97)

- 基于语义区域图（底层为空间索引）



- 连接语义区域的边：确保达到距离 (GRD, guaranteed reaching distance)





- 语义序列可能在时间上是不完全的，对应的定位记录间可能还存在着未观测到的信息。
- 提出了概率推理方法来恢复缺失的移动语义：
 - 人们常在较小范围内的两个室内目的地间进行非常相似的移动；
 - 序列中的每个停留区域可视为室内目的地，而其间构建的多个路过语义可聚合在一起，来捕获两个目的地间的相似移动；
 - 进一步考虑室内移动性约束，可以根据已经标注的移动语义序列来推断补全未观察到的移动情况；

已标注数据中得到先验的相似移动知识

算法 5.3 InferenceBasedComplementing(Set of ms-sequences \mathcal{S}_Λ , Semantic region graph G_R)

```

1 set  $\mathcal{S}'_\Lambda \leftarrow \emptyset$ 
2 hash table  $\mathcal{MK} \leftarrow \boxed{\text{ConstructMobilityKnowledge}(G_R, \mathcal{S}_\Lambda)}$  ▷ Construct candidate path sets
3 for each original ms-sequence  $\Lambda_o$  in  $\mathcal{S}_\Lambda$  do                                ▷ Infer missing mobility semantics
4    $\Lambda_o \leftarrow \boxed{MSemanticsInference(\Lambda_o, \mathcal{MK}, G_R)}$ 
5   add  $\Lambda_o$  to  $\mathcal{S}'_\Lambda$ 
6 return  $\mathcal{S}'_\Lambda$ 

```

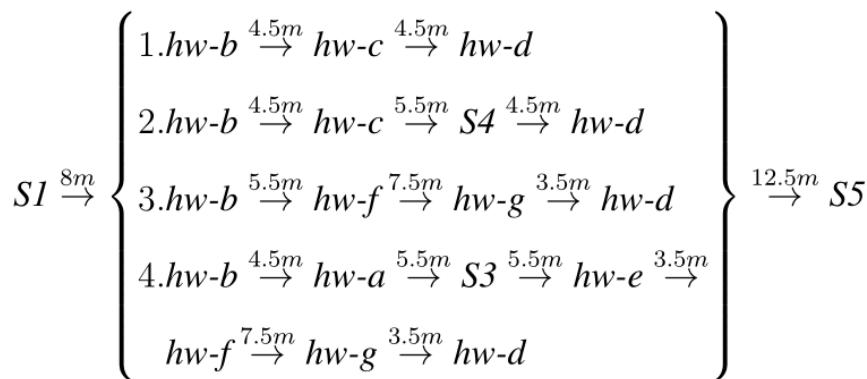
利用拓扑结构和先验知识对不完整观测序列进行三元组推断

移动知识构建

给定两个目的地S1和S5，其二者之间的相似运动可以由候选路径集合（一个语义区域图的子图）和各相关语义区域间的转移概率进行表示。（可以通过历史观测数据进行计算）

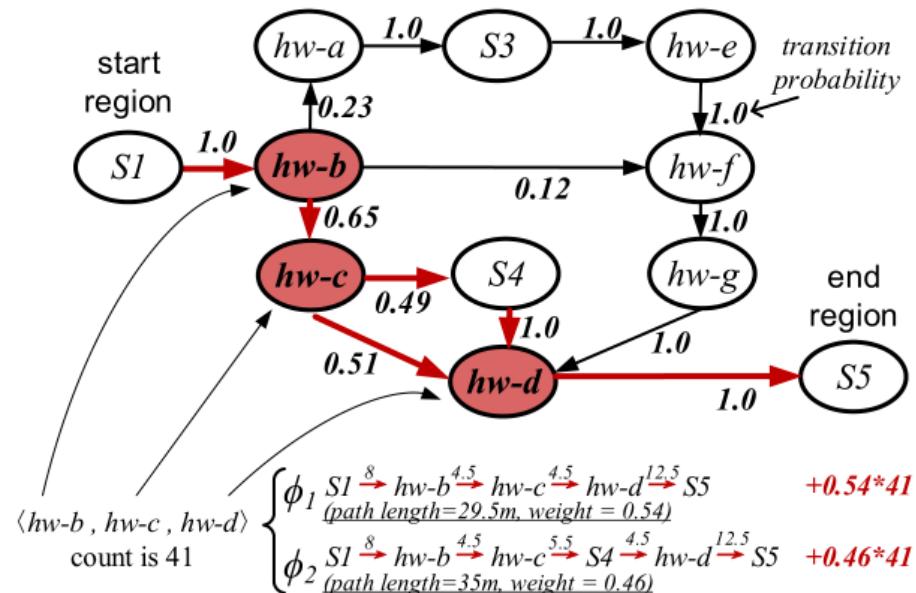
候选路径集合 (pp.98-99)

- 基于语义区域图，两个观测到的停留区域间的候选路径（语义区域序列）可以通过A*算法进行搜索；



转移概率 (pp.99-100)

- 将历史观测序列的频度赋值到语义区域图的子图的连接边上。



缺失语义元组推断

最可能路径推断 (pp.101-102)

- 根据先验的转移概率，从每两个连续的观测区域中找到最优（具有最大连乘概率）的语义区域序列 – 最大后验估计问题 (max-product 算法^[146])；

$$\begin{aligned} \arg \max_{\phi} P(\phi | PT^{(o)}) &= \arg \max_{r_a^{\triangleright} \rightarrow \dots \rightarrow r_b^{\triangleright} \subseteq \phi} P_t(r_s^{\shortparallel}, r_a^{\triangleright}) \prod_{x=a}^{b-1} P_t(r_x^{\triangleright}, r_{x+1}^{\triangleright}) P_t(r_b^{\triangleright}, r_q^{\triangleright}) \\ &\quad \arg \max_{r_c^{\triangleright} \rightarrow \dots \rightarrow r_d^{\triangleright} \subseteq \phi} P_t(r_q^{\triangleright}, r_c^{\triangleright}) \prod_{y=c}^{d-1} P_t(r_y^{\triangleright}, r_{y+1}^{\triangleright}) P_t(r_d^{\triangleright}, r_e^{\shortparallel}) \end{aligned}$$

时间段推断 (pp.102-103)

- 根据语义区域间的确保到达距离，对未观测到的空白时间段进行分配；

$$t_s^{(x)} = \lambda_p \cdot \tau \cdot t_e + \Delta t \cdot \frac{\sum_{i=p}^x dist_{gr}(r_p, r_i)}{\sum_{i=p}^q dist_{gr}(r_p, r_i)}; \quad t_e^{(x)} = \lambda_p \cdot \tau \cdot t_e + \Delta t \cdot \frac{\sum_{i=p}^{x+1} dist_{gr}(r_p, r_i)}{\sum_{i=p}^q dist_{gr}(r_p, r_i)}$$

真实数据集实验



实验设置 (pp.103-104)

- 杭州某购物中心Wi-Fi三边测距定位系统；截取2017年1月1日-31日运营时间数据；
- 平均每天：MAC地址7,647个，定位记录数量2,907,904；共计237,057p-sequence；
- 基于室内距离计算的平均误差2m-25m；
- 202家店铺作为语义区域；
- 采用TRIPS[VLDB18]系统进行可视化地移动语义标注；9,687条ms-sequence作为评估的真值，剩下1,004来自于1月1日的序列作为事件识别模型的训练数据；
- 移动知识：共生成10,682个有向区域对之间的候选路径集合和转移概率信息。
- 每天对识别模型和移动知识进行更新；

度量模型 (pp.104)

- 标注的计算效率 (Efficiency)
 - 每个对象序列一个线程，标注一条p-sequence的平均运行时间；
- 标注结果的有效性 (Effectiveness)
 - 很难在时间和空间上找到几乎完全匹配的语义元组；
 - 给定真值的三元组，若标注元组的空间和事件标注和其相同，且

$$\frac{|\lambda \cdot \tau \cap \lambda_g \cdot \tau|}{|\lambda_g \cdot \tau|} \geq \eta$$

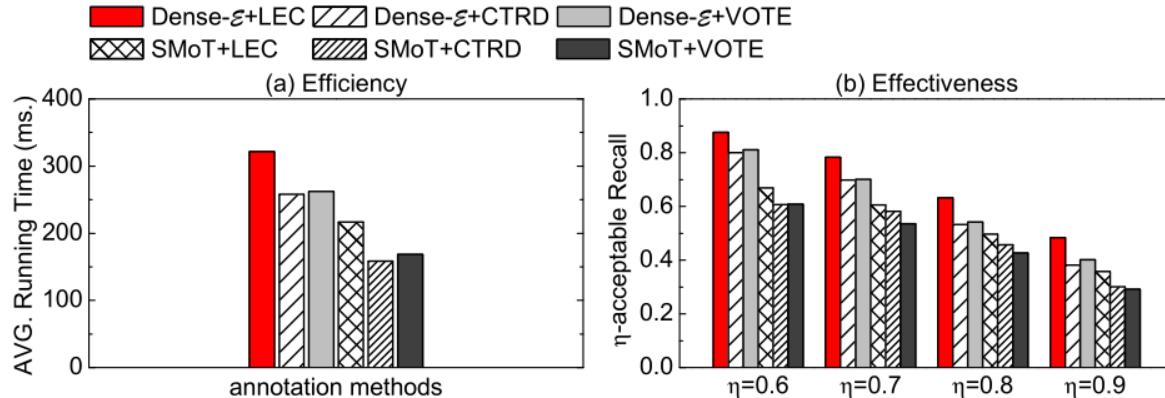
认为它是 η 可接受的；

- 越大的 η 表示对标注结果要求越高；
- 真值中可以找到 η 可接受的标注元组的比例，就是 η 可接受的召回率；

Java实现；Xeon服务器；多线程运行。

实验评估结果

标注方法比较 (pp.104-105)



- 对比方法：组合不同的已有事件和空间标注技术；
事件标注方法：SMoT^[131]；
空间标注方法：Centroid；
Multi-Voting；
- 提出的语义标注方法，在计算效率和结果有效性取得不错的平衡，具有良好性能表现；

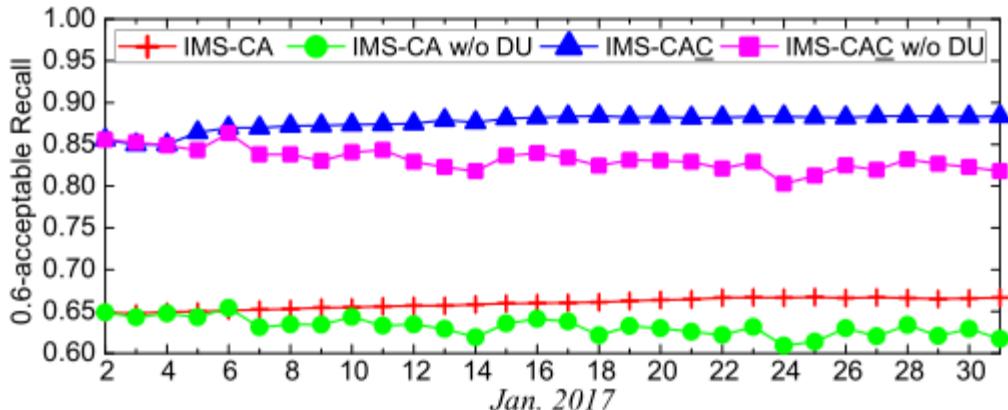
数据清洗和数据补全的效果 (pp.105-106)

方法	序列平均移动语义个数	η -可接受召回率			
		$\eta=0.6$	$\eta=0.7$	$\eta=0.8$	$\eta=0.9$
IMS-A	11.94	0.3555	0.2926	0.2187	0.1642
IMS-CA	10.23	0.6615	0.5577	0.3825	0.2825
IMS-AC	14.51	0.4645	0.3858	0.2638	0.2155
IMS-CAC	14.12	0.8756	0.7828	0.6318	0.4834

- 对应于三层框架IMS-CAC，在保留标注算法的基础上，在前后分别添加清洗和/或补全方法；
清洗层是必要的，对未清洗的定位序列进行清洗的有效性很差；
补全层的添加，会增加标注语义元组的个数，提高了有效性；
三层的处理框架表现最佳；

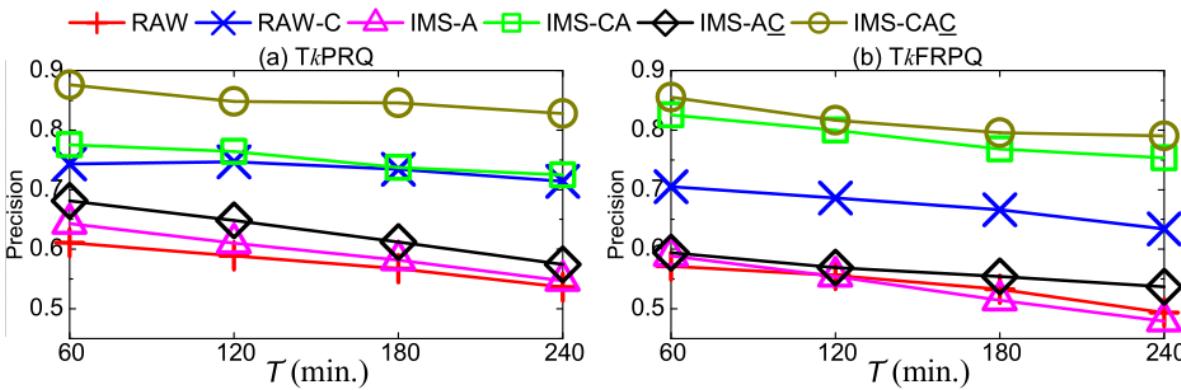
实验评估结果

连续更新机制 (pp.106-107)



- IMS-CAC: 同时利用加入前一天标注的数据来强化事件识别模型和移动知识；
- IMS-CA: 仅强化事件识别模型；
- w/o DU: 不进行每日更新；
- 当原始定位数据不断流入系统时，进行连续更新能够提高标注的有效性。

语义元组在查询应答中的表现 (pp.107-109)



- Top-k 热点区域查询；
- Top-k 频繁区域对查询；
- RAW 和 RAW-C，利用阈值规则对原始移动序列进行语义的提取并回答查询，1个月数据约 3.44GB；
- 标注的语义元组直接进行查询应答，大小约 220MB；
- 挖掘的语义元组在查询应答中表现良好。

合成数据集实验



实验设置 (pp.109-110)

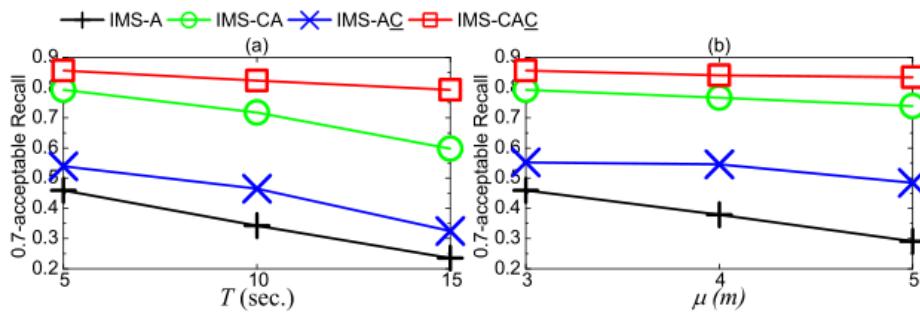
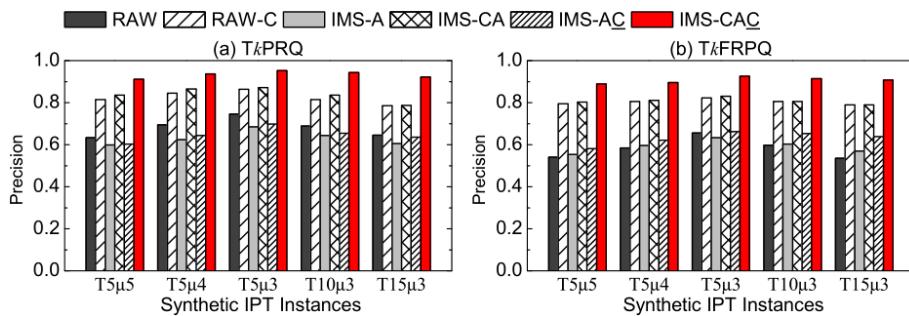
- Vita模拟器生成的大规模数据集；
- 1410个分区、2200扇门的10层建筑；
- 423个语义区域作为标注的标签；
- 模拟最大定位周期：5, 10, 15s；
- 模拟定位误差：3,4,5 m；

挖掘的有效性验证 (pp.110)

- 定位周期加大，定位数据越稀疏，不采用补全的框架表现越差，而三层框架表现良好，仅略有下降；
- 定位误差提升，不采用数据清洗的框架的有效性迅速下降，三层框架表现稳定，仅略有下降；
- 在原始数据中时空不确定性较大时，三层的挖掘框架的有效性仍然很高；

查询应答 (pp.111-112)

- 加大定位周期，构建的语义元组在查询应答中的准确度仅略有下降；
- 加大定位误差，查询应答的准确度也下降，但在5m的定位误差上仍然可在两个查询上分别达到91%和88%以上的准确度。





本项工作小结

- 解决一个从原始的、具有时空和语义不确定性的室内定位序列中挖掘用户移动语义的问题。
- 设计了一个三层移动语义挖掘模型及对应的数据处理方法：
 - 提出了基于室内移动性约束的原始数据清洗方法，通过识别典型的室内定位错误并根据移动约束进行两段式修正，来减少原始定位序列中的数据错误；
 - 提出了基于分割匹配的移动语义标注方法，首先根据序列中位置报告的时空密度将序列进行分割，随后利用学习得到的移动事件识别函数和构建的语义区域图对移动语义进行匹配和构建；
 - 提出了基于概率推断的数据补全方法，利用历史数据构建出两个室内目的地间的候选路径集合及区域间的转移概率，并通过最大后验概率估计找出当前观测序列的最可能路径及其中每一区域的时间标注，完成缺失移动语义的恢复。
- 利用合成和真实数据集进行了全面实验评估：
 - 模型可高效地对原始定位数据进行处理，得到准确的移动语义；
 - 挖掘的语义元组也能有效、高效地对典型的数据查询进行应答。
- 短文录用在VLDB18，长文仍在投。



目录

1 研究背景和动机

2 研究基础和现状

3 室内密度分析挖掘

4 室内流量分析挖掘

5 室内移动语义挖掘

6 结论和展望

工作总结



- 针对室内**空间结构**、室内**定位机制**和室内**移动对象**的一般性特点，对室内移动数据中常见的**时空不确定性和语义不确定性**进行了有效的建模分析，对具有实际应用价值的分析挖掘问题进行了有效高效的求解。

研究问题	移动数据设定	不确定性分析	挖掘目标
室内密度分析挖掘	在线快照数据	时空不确定性	室内密集区域
室内流量分析挖掘	历史范围数据	时空不确定性	室内热点语义位置
室内移动语义挖掘	历史范围数据	时空、语义不确定性	用户移动语义元组

- 提出了面向时空不确定性的**室内区域密度分析计算模型**及相应的**密集区域挖掘方法**。
- 提出了面向时空不确定性的**室内语义位置流量分析计算模型**及相应的**热点语义位置挖掘方法**。
- 提出了面向时空和语义不确定性的**室内用户移动语义挖掘方法**。



- 扩展和加强室内密度分析挖掘方法:
 - 利用已提出的室内缓冲/核心区域和距离衰减不确定区域模型，对连续室内密度查询和室内移动对象聚类等相关问题进行探究；
 - 通过离线学习的方法从历史数据中构建对象不确定运动模型，以适应更普适情况下室内区域密度的计算分析。
- 进一步扩展和完善室内流量分析挖掘方法:
 - 室内对象行为的进阶建模，以加强提出的室内流量的计算衡量方法；
 - 考虑定位位置的密集程度对流量计算准确度的影响，找到流量分析和定位系统部署的最优平衡点；
 - 考虑连续版本的室内热点语义位置挖掘方法，包括对室内空间位置图和位置矩阵的动态维护，及挖掘算法中搜索策略的改良优化等。
- 改进和丰富室内移动语义挖掘方法，扩展具体应用范畴:
 - 在缺失语义元组的推断过程中，加入对历史数据中时间标注的建模过程，以更有效地预测缺失语义的标注信息；
 - 对常见的符号定位数据进行扩展，以形成通用的移动语义抽取和计算分析平台；
 - 利用具有更高置信度的数据源，如用户事务日志、签到数据或移动社交媒体数据，来修正和增强现有移动层面的语义序列。

攻读博士学位期间的研究成果



- **Huan Li**, Pai Peng, Hua Lu, Lidan Shou, Gang Chen, and Ke Chen. E²C²: Efficient and Effective Camera Calibration in Indoor Environments. UbiComp, 2015: 9-12. (CCF-A类会议, 短文)
- **Huan Li**, Hua Lu, Xin Chen, Gang Chen, Ke Chen, and Lidan Shou. Vita: A Versatile Toolkit for Generating Indoor Mobility Data for Real-World Buildings. PVLDB, 2016, 9(13): 1453--1456. (CCF-A类会议, 短文)
- **Huan Li**, Hua Lu, Lidan Shou, Gang Chen, and Ke Chen. In Search of Indoor Dense Regions: An Approach Using Indoor Positioning Data. TKDE, 2018, 15 pages. (CCF-A类期刊, 长文)
- **Huan Li**, Feichao Shi, Hua Lu, Gang Chen, Ke Chen, and Lidan Shou. TRIPS: A System for Translating Raw Indoor Positioning Data into Visual Mobility Semantics. PVLDB, 2018, 4 pages. (CCF-A类会议, 短文)
- **Huan Li**, Hua Lu, Lidan Shou, Gang Chen, and Ke Chen. Finding Most Popular Indoor Semantic Locations Using Uncertain Mobility Data. TKDE, 2018, 14 pages. (CCF-A类期刊, 在投)
- **Huan Li**, Hua Lu, Gang Chen, Ke Chen, Qinkuang Chen, and Lidan Shou. Towards Translating Raw Indoor Positioning Data into Mobility Semantics. PVLDB, 2018, 13+4 pages. (CCF-A类会议, 在投)



THANKS

不确定室内移动数据的分析挖掘方法研究



- 答辩人：李环
- 2018年6月7日