电子商务数据分析

第3章 轨迹大数据挖掘

朱桂祥 (9120201070@nufe.edu.cn)
南京财经大学信息工程学院
江苏省电子商务重点实验室
电子商务信息处理国家级国际联合研究中心
电子商务交易技术国家地方联合工程实验室

https://github.com/zgx881205/Analysis-of-E-commerce-Data

南京财经大学 NANJING UNIVERSITY OF FINANCE & ECONOMICS

苏州大学校训

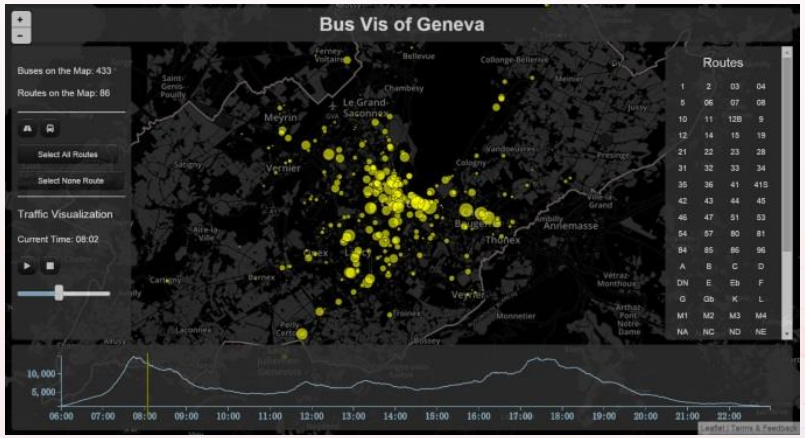
“养天地正气，法古今完人”原文出自1923年元月国父孙中山先生的墨宝，溯其渊源可至《孟子·公孙丑上》：“我善养吾浩然之气”，后经宋朝文天祥《正气歌》的引用进一步拓展了其内涵：“况浩然者乃天地之正气也”，“天地有正气，杂然赋流形。下则为河岳，上则为日星。于人曰浩然，沛乎塞苍冥……”培养天地间坚毅不屈的气节，师法古今完美道德的圣贤，苏州大学校训蕴含着“仰以察古，俯以观今”的气度，又渗透着“观乎天文，以察时变;观乎人文，以化成天下”的传统文化底蕴，如同《正气歌》一样引人进入一种至高至上的境界。

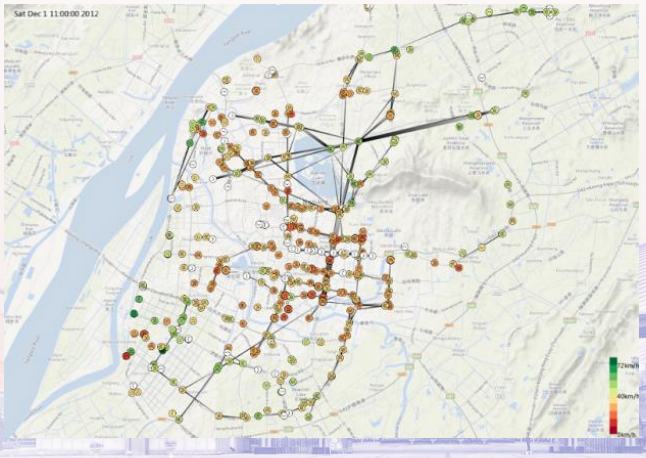
https://tv.cctv.com/2014/10/27/VIDE1414372321925275.shtml

轨迹大数据分析

1.1 轨迹数据概念和特征

轨迹数据就是时空环境下，通过对一个或多个移动对象运动过程的采样所获得的数据信息，包括采样点位置、采样时间、速度等，这些采样点数据信息根据采样先后顺序构成了轨迹数据。





轨迹大数据分析

1.1 轨迹数据概念和特征

■ 代表性轨迹数据：

轨迹数据分类

- 人类活动轨迹
- 交通工具活动轨迹
- 动物活动轨迹
- 自然现象活动轨迹

表 1　　代表性轨迹数据

数据种类	采集方式	采样频率	日均数据量（采样点）	数据总量
车辆轨迹	车载 GPS	秒级、分钟级	千万–亿级	TB 级
移动轨迹	地图 APP	秒级、分钟级	千万–百亿级	TB、PB 级
手机轨迹	蜂窝基站	分钟级	十亿–百亿级	TB、PB 级
公交轨迹	公交卡	小时级	百万–千万级	TB、PB 级
卡口数据	卡口抓拍	分钟级	千万级别	TB 级
行为轨迹	社交媒体	分钟、小时级	百万–千万级	PB 级

轨迹大数据分析

1.1 轨迹数据概念和特征

轨迹数据的4V:
①大规模(Volume)
②实时高速(Velocity)
③多样性(Variance)
④高价值(Value)

移动对象轨迹数据库的特征：

时空采样性：轨迹是时空环境下的采样序列。
异频采样性：轨迹的采样间隔差异显著。导航服务秒级，社交媒体行为的小时级采样，
数据质量差：连续运动的轨迹被离散化表示。
路网相关性：交通轨迹数据做GPS空间向路网空间的映射。

轨迹大数据分析

1.2 轨迹数据应用

轨迹数据能够刻画人类的活动和行为历史，蕴含了群体性的移动模式和规律。

表 2　　代表性轨迹分析应用

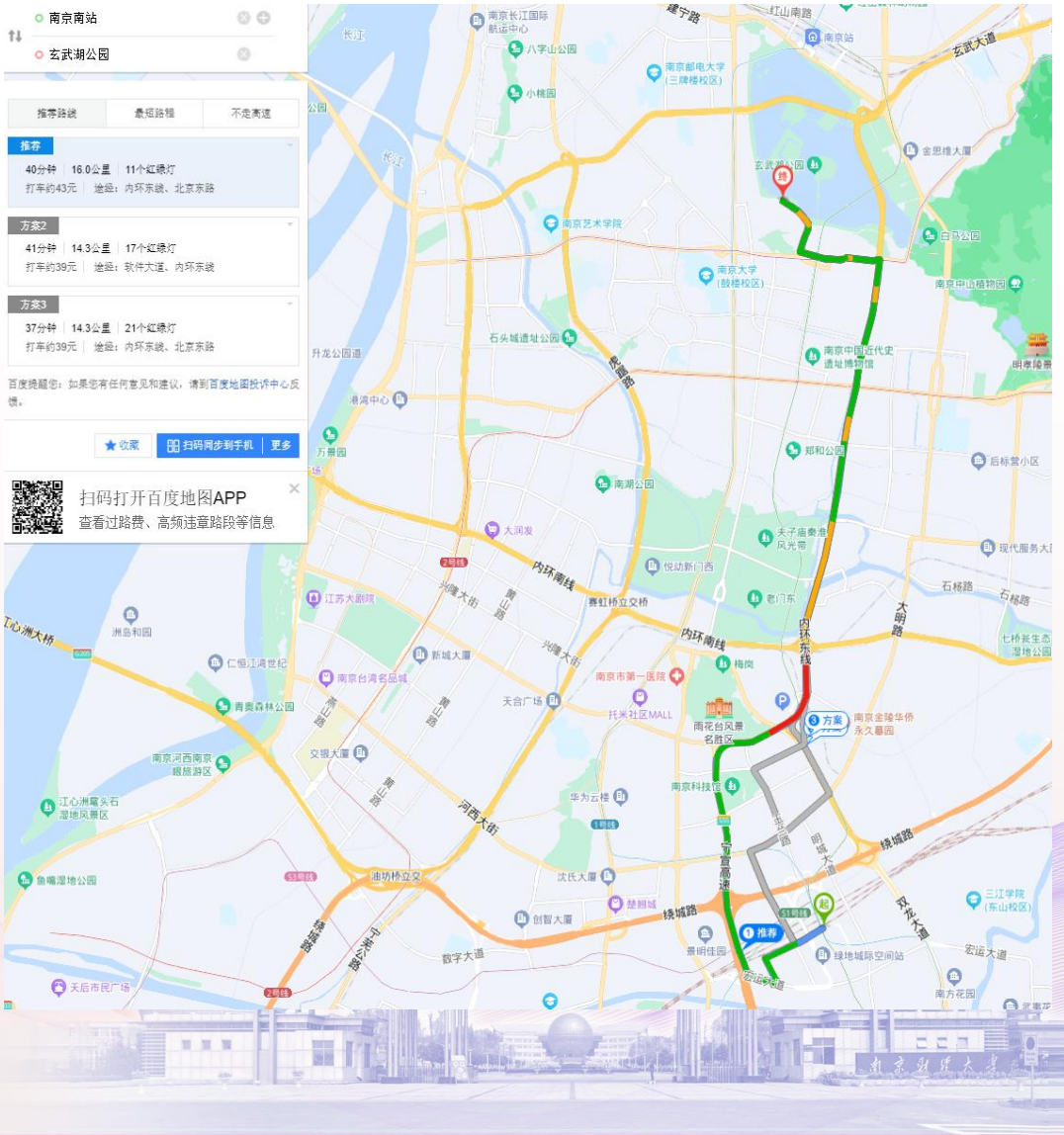
应用	所用数据	应用现状
大众化经验路径推荐	出租车 GPS 轨迹、私家车移动轨迹数据、气象数据、交通路网数据、历史事故数据等	广泛应用在地图服务公司，显著提升服务水平
交通路况精准预测	GPS 数据（流）、路网路况数据、气象数据、大型活动记录、重大事故数据等	用于地图服务和交通指挥系统，但精度尚需提高
城市规划智能决策	轨迹数据、地图数据、兴趣点数据、消费数据、价格数据、公交线路、历史事故等数据	用于数据驱动的规划决策，多源数据集成与融合是难点
个性化服务与活动推荐	车辆与手机轨迹、社交网络与社交媒体数据、兴趣点和签到、评论数据等	用于基于位置的服务推荐，需提高语义理解和推荐算法
出租车服务	出租车 GPS 轨迹、私家车移动轨迹、公交线路与轨迹等数据	应用于相关业务优化，有进一步提升空间

轨迹大数据分析

1.1 轨迹数据概念和特征

轨迹数据应用领域：

1. 大众化经验路径推荐：
从南京南站到玄武湖的导航地址：
在不同的时间，不同的交通条件下，结果也是不同的。基于大众的轨迹数据，寻找最优的导航路径。



轨迹大数据分析

1.1 轨迹数据概念和特征

2.交通情况精准预测：

通过轨迹流的统计，评估不同区域的进出流量，获取实时的交通态势。
通过轨迹分析，综合运用大数据的外部性，做到指挥决策的先知先觉。

3.城市规划智能决策：

通过轨迹，分析不同城市不同区域的社会功能，对城市不同区域的发展和规划进行辅助决策。

4.个性化服务与活动推荐：

社交媒体的轨迹记录了用户的位置行为，通过对轨迹的行为理解，为用户推荐个性化的景点。基于位置的服务（Location Based Services，LBS），广告和推荐。

5.出租车服务：

监控出租车的行驶路线，通过海量的历史数据，找到出租车的最优路线，对绕路等行为进行欺诈检测。

轨迹大数据分析

1.1 轨迹数据概念和特征

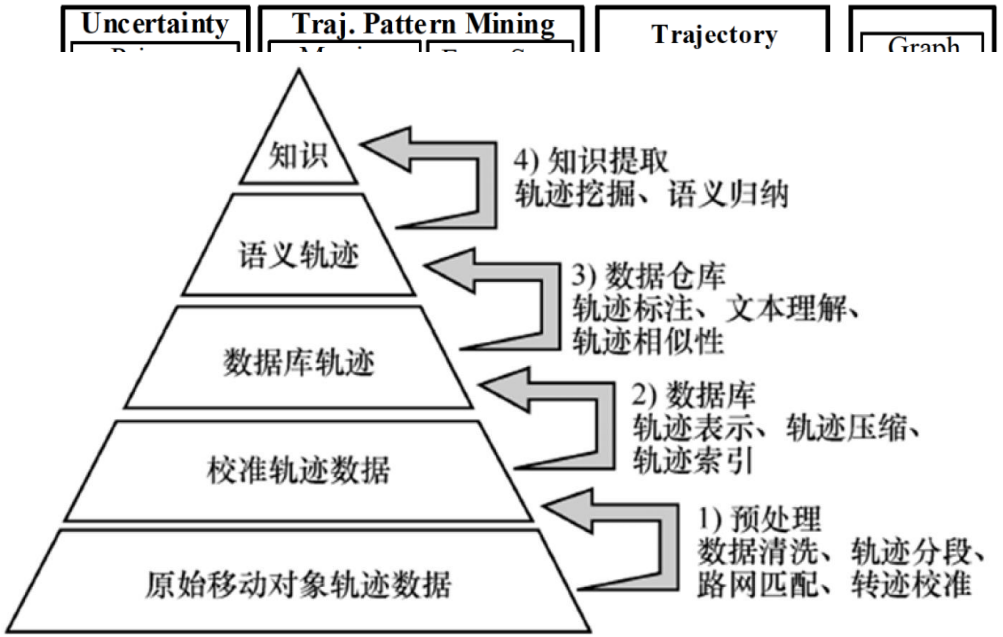


图 3-1. 轨迹数据金字塔

Figure 1 Paradigm of trajectory data mining

轨迹数据挖掘的综述文章：

[1] Zheng Y. Trajectory data mining: an overview[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2015, 6(3): 29.

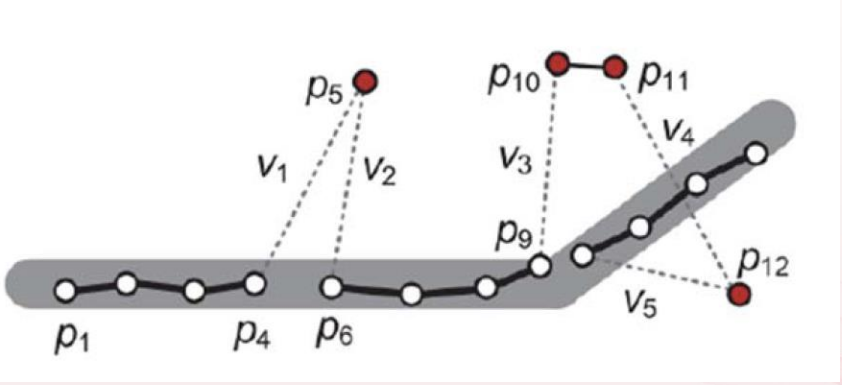
轨迹大数据分析

1.2 轨迹数据预处理技术

轨迹数据作为轨迹大数据处理对象,其预处理效果将直接影响轨迹数据挖掘的效果 轨迹数据存在着一系列的数据质量问题：数据缺失、数据冗余、数据不一致等。

■ 噪声过滤

轨迹噪声数据：



注：红色为噪声点

产生的原因：

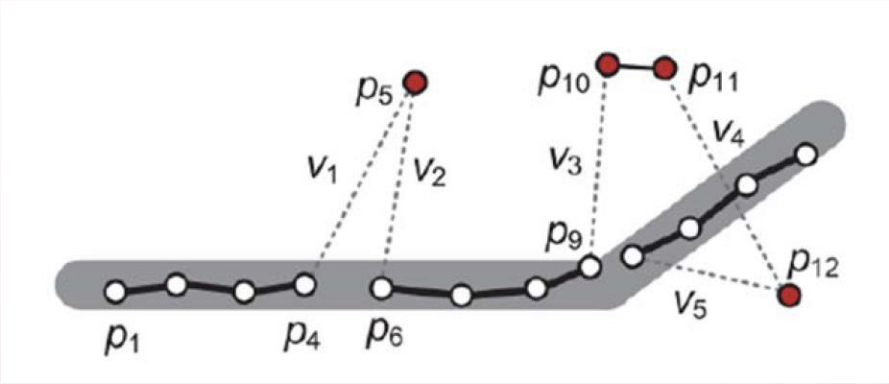
由于数据误差，导致位置在路网之外。

轨迹大数据分析
1.2 轨迹数据预处理技术

如何解决问题？
分别对横坐标和纵坐标进行计算

滑动窗口(sliding window)

轨迹噪声数据：



均值过滤(Mean filter) 对较大的错误值比较敏感
$(1,3,4,7,10^{10})$ -> $2*10^9$

中值过滤(Median filter) 对较大的错误值不敏感
$(1,3,4,7,10^{10})$ -> 4.

当窗口较小的时候，
对于连续的错误不再适用。
如(p_{10}, p_{11}, p_{12})

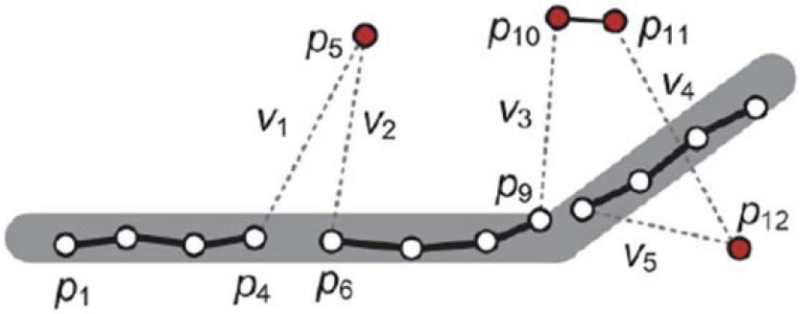
轨迹大数据分析

1.2 轨迹数据预处理技术

卡尔曼滤波(Kalman filtering):
线性模型假设符合高斯噪声分布。
使用当前的状态，预测或更正下一个状态。
下一个状态线性独立于当前的状态。

粒子滤波(Particle filtering)
模拟测量噪声和轨迹。

轨迹噪声数据：



基于启发式的滤波
本质上是适用估计值替代噪声。
P_4->p_5, p_5 -> $P_{6,}$ p_9 -> $P_{10,}$ p_{11} -> P_{12}

这些速度超过了异常的速度，需要过滤。

轨迹大数据分析

1.2 轨迹数据预处理技术

驻留点检测：

什么是驻留点？

用户在某个位置停留了一段时间。

(p_3)->(stay point 1)

用户在某个位置停留了一段时间。

驻留点2，

用户围绕某个位置驻留(p_5,p_6,p_7,p_8) -> (stay point 2)。

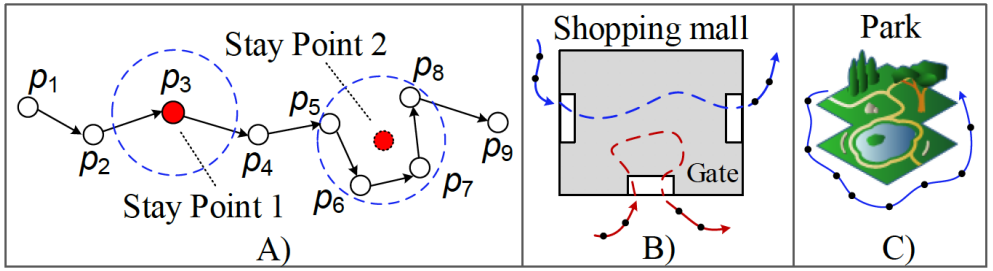


Figure 3. Stay points in a trajectory

图(B) 用户的轨迹在某个地点范围内。

图(c) 用户的轨迹围绕着某个地点进行。 从空间的点序列，变成了有意义的地点序列

$$\boldsymbol{P} = p_1 \rightarrow p_2 \rightarrow \cdots \rightarrow p_n, \Rightarrow \boldsymbol{S} = s_1 \xrightarrow{\Delta t_1} s_2 \xrightarrow{\Delta t_2}, \ldots, \xrightarrow{\Delta t_{n-1}} s_n,$$

轨迹大数据分析

1.2 轨迹数据预处理技术

轨迹压缩：

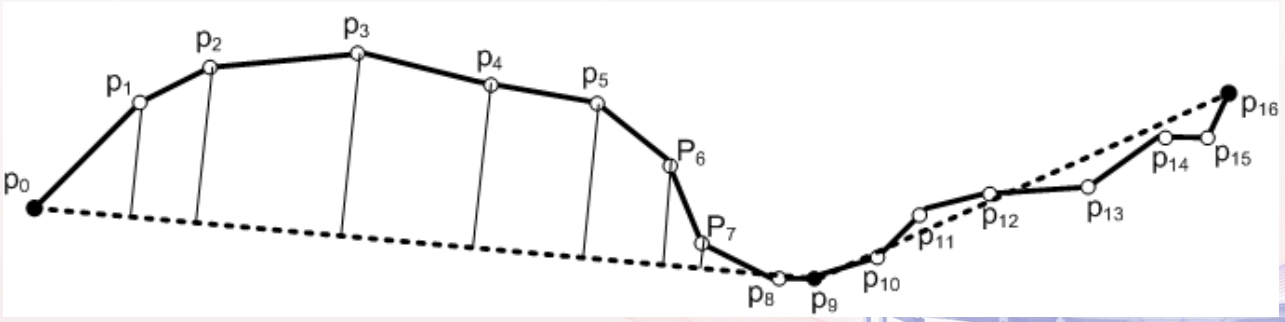
为什么做轨迹压缩？

原始轨迹数据的空间位置以每秒的精度进行保存。

真实的应用并不需要这样高精度的数据。

线下压缩：对全部的轨迹数据进行压缩，时间范围已知。

在线压缩：对实时的轨迹数据进行压缩。



研究问题：如何把这个轨迹用更少的点表示呢？

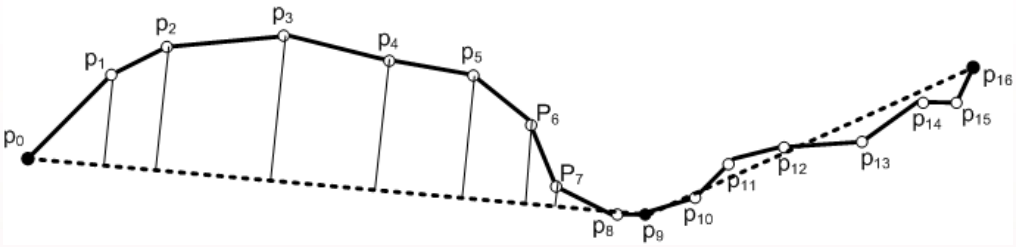
轨迹大数据分析

1.2 轨迹数据预处理技术

轨迹压缩

目标：

减少轨迹数据的大小，
保留精度。



性能指标：

处理时间
压缩率
错误率

原轨迹上某点的位置和对应于压缩轨迹的某点位置的距离。

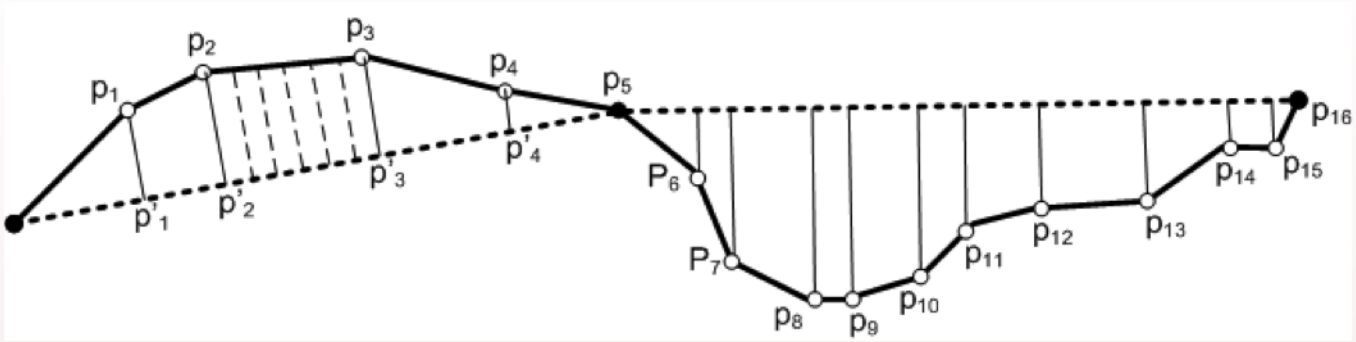
轨迹大数据分析

1.2 轨迹数据预处理技术

轨迹压缩

PED (Perpendicular Euclidean Distance)

垂直欧拉距离

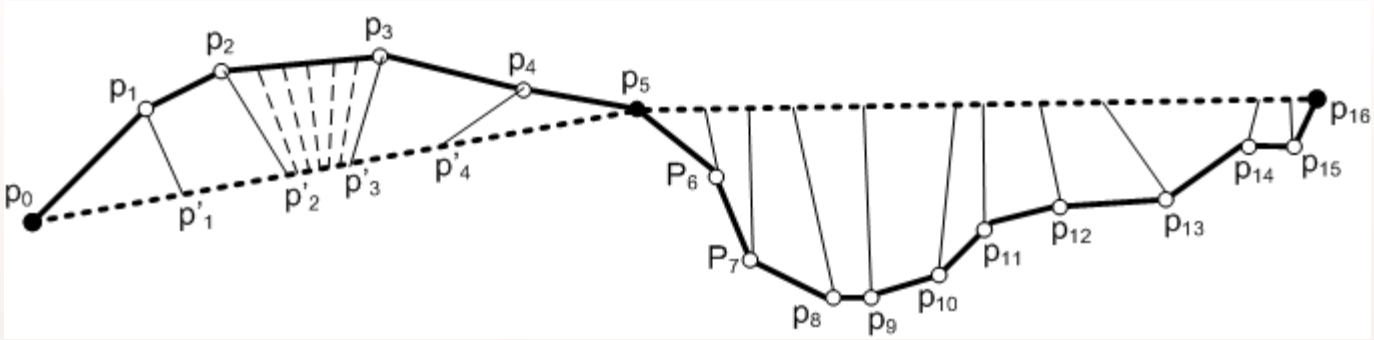


轨迹大数据分析

1.2 轨迹数据预处理技术

轨迹压缩

TSED (Time Synchronized Euclidean Distance)时间同步的欧拉距离



轨迹大数据分析

1.2 轨迹数据预处理技术

轨迹压缩的分类.

批处理压缩:

收集所有轨迹点的数据集合，压缩后，传送给服务器

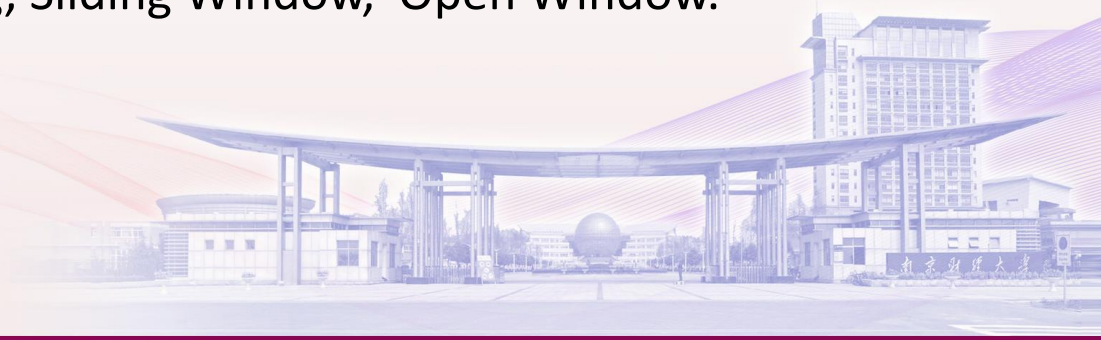
主要技术: Douglas-Peucker 算法, top-down time-ratio (TD-TR), Bellman‘s 算法.

在线数据压缩：

根据精度的要求，选择在线的数据更新。

应用: 交通监控等。

主要技术 Reservoir Sampling, Sliding Window, Open Window.



轨迹大数据分析

1.2 轨迹数据预处理技术

Douglas-Peucker (DP) 算法
(perpendicular Euclidean distance)

(1) 把原轨迹用近似的线段替代。

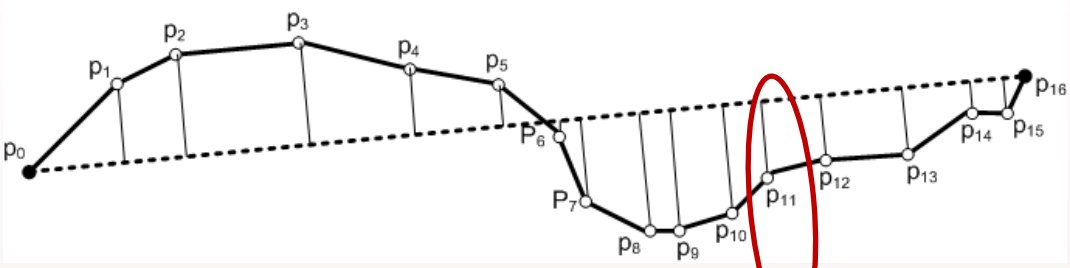
(2)计算替代后的错误率,看看是否满足要求。如果满足要求，则停止。如果没有满足要求，寻找分裂点(splitting point)，该分裂点会导致最大的错误率。

(3)递推的过程继续，一直到所有的替代后的轨迹和原轨迹的错误都满足阈值的要求。

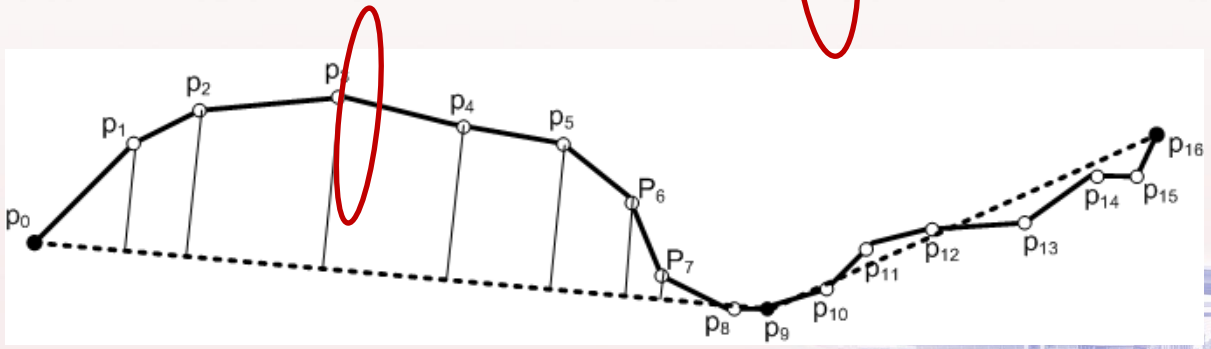
轨迹大数据分析

1.2 轨迹数据预处理技术

Douglas-Peucker (DP) 算法
(perpendicular Euclidean distance)



寻找分裂点



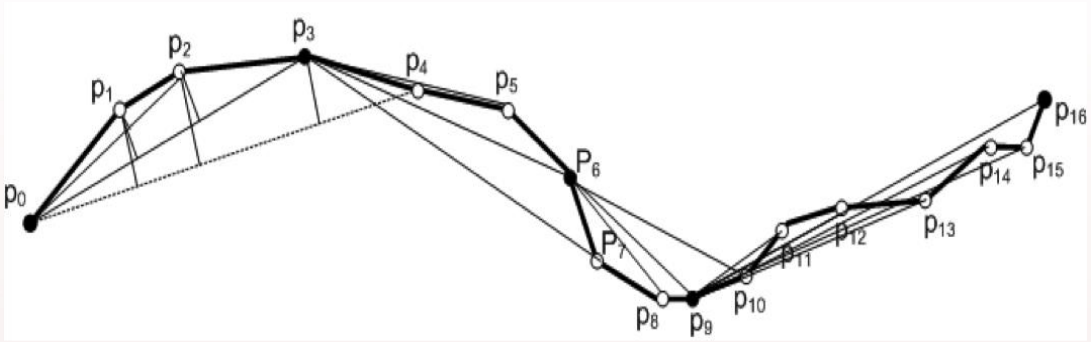
递归直到所有分割满足要求

轨迹大数据分析

1.2 轨迹数据预处理技术

sliding window算法

当滑动窗口从{p_0}增长到{p_0, p_1, p_2, p_3}, 在原轨迹和拟合轨迹的所有错误，并没有超过阈值。

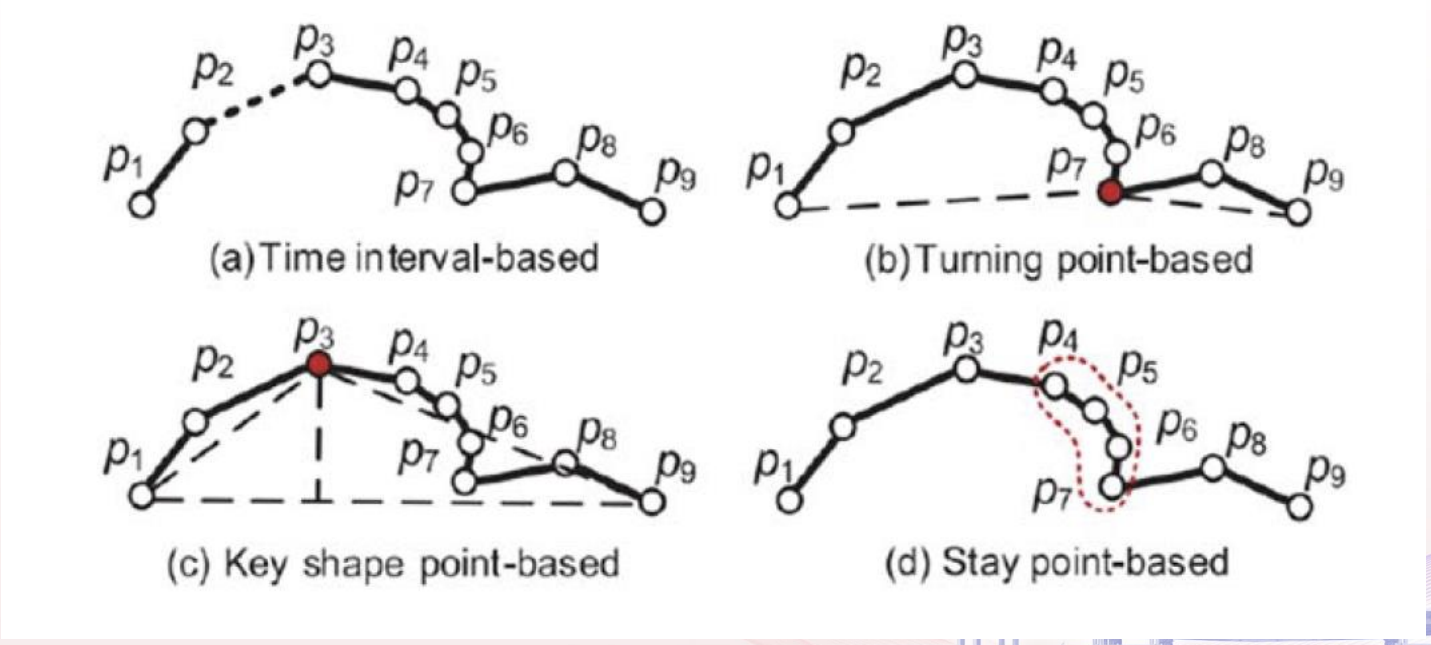


当p_4加入的时候，错误超过阈值了，这个时候p_4作为分割点，进行下一步的计算

轨迹大数据分析

1.2 轨迹数据预处理技术

轨迹分割



轨迹大数据分析

1.2 轨迹数据预处理技术

地图匹配算法

把初始的经纬度坐标数据 转化为 路段的序列数据

几何的地图匹配算法

拓扑的地图匹配算法

概率的地图匹配算法

其他技术。

轨迹大数据分析

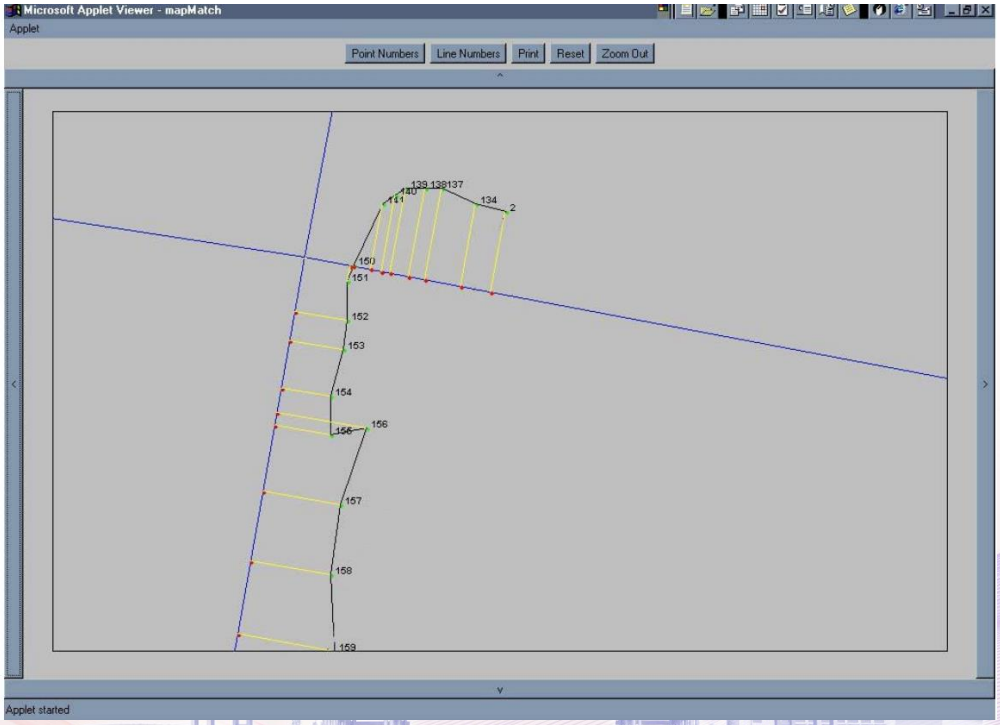
1.2 轨迹数据预处理技术

地图匹配算法

几何地图匹配算法。

把GPS的位置点与最近的道路匹配

如右图的公交车路径选择问题



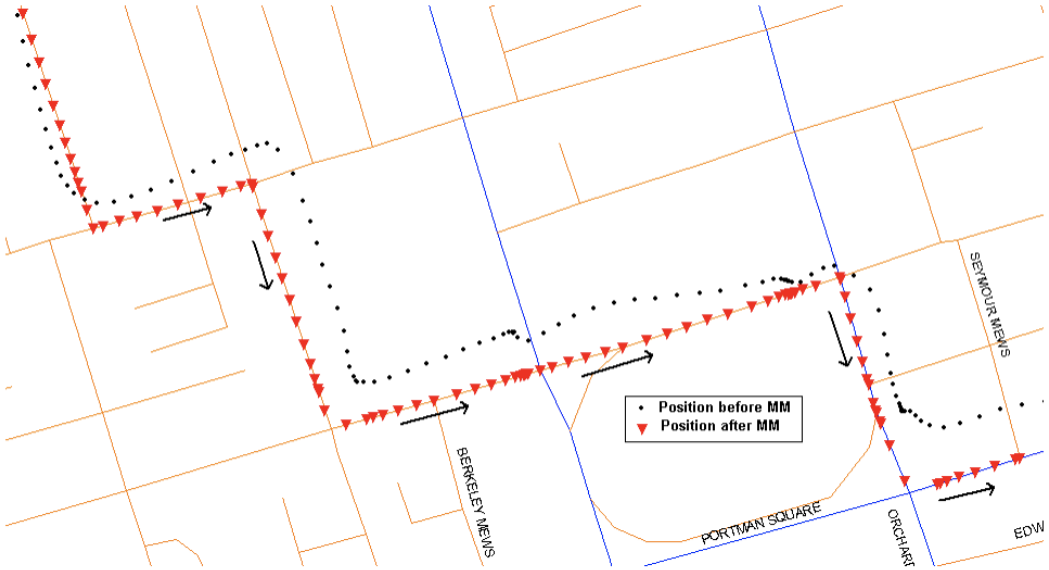
轨迹大数据分析

1.2 轨迹数据预处理技术

地图匹配算法

拓扑地图匹配算法。

地图匹配算法的实例



(Presentation map scale 1 cm : 20 m)

Fig. 13: Map matching results on complex urban road network.

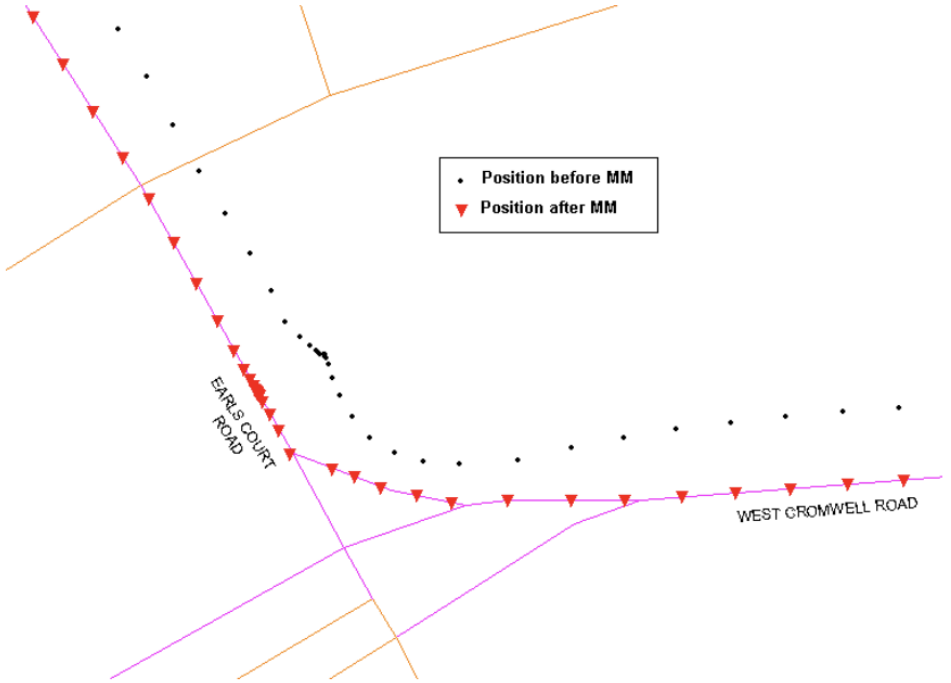
轨迹大数据分析

1.2 轨迹数据预处理技术

地图匹配算法

拓扑地图匹配算法[1]。

地图匹配算法，
在一个复杂的路口拐弯的例



(Presentation map scale 1 cm : 10 m)

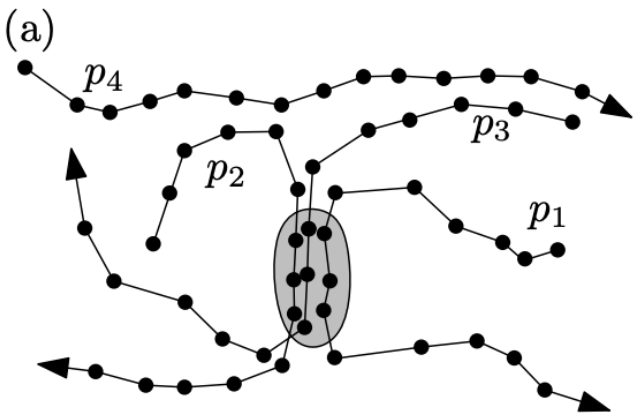
Fig. 14: Map matching results on left-turn maneuvering at complex junction.

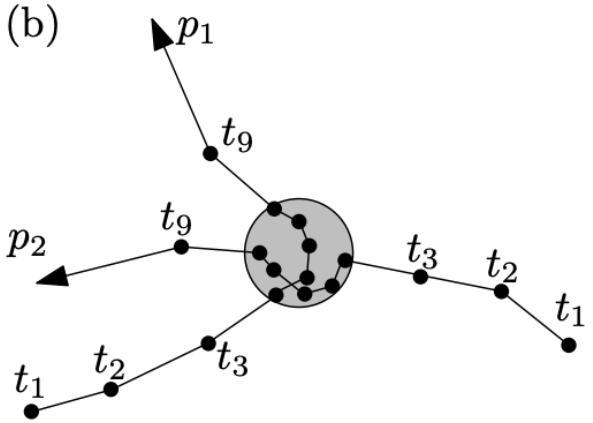
轨迹大数据分析

1.3 轨迹数据挖掘技术

伴行模式(flock)

(a) flock(m, k, r)
一组至少m个个体，
一起运动至少k个连续
的时间点，并且彼此动态
在一个半径为r的范围内。





(b)stay(m, k, r)
一组至少m个个体，
一起至少k个连续
的时间点，并且静态彼此
在一个半径为r的范围内。

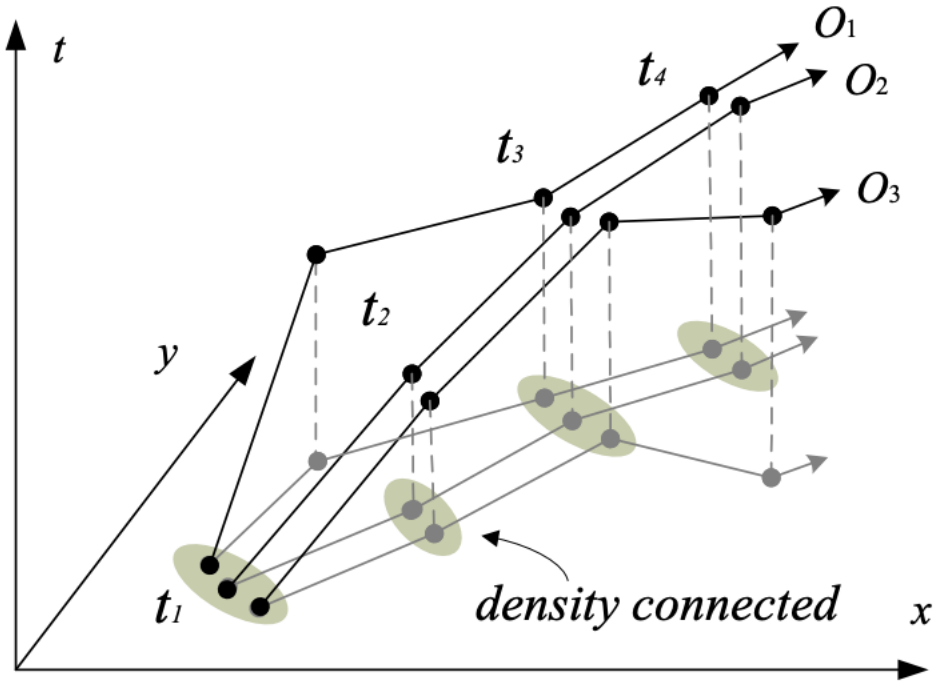
轨迹大数据分析

1.3 轨迹数据挖掘技术

伴行模式(convoy)
convoy(m, k, e)

一组m个个体，在至少k个连续的时间点内，任意两个个体的距离小于e.

比如一个车队的汽车一起行驶，或一群候鸟在天空飞过。



轨迹大数据分析

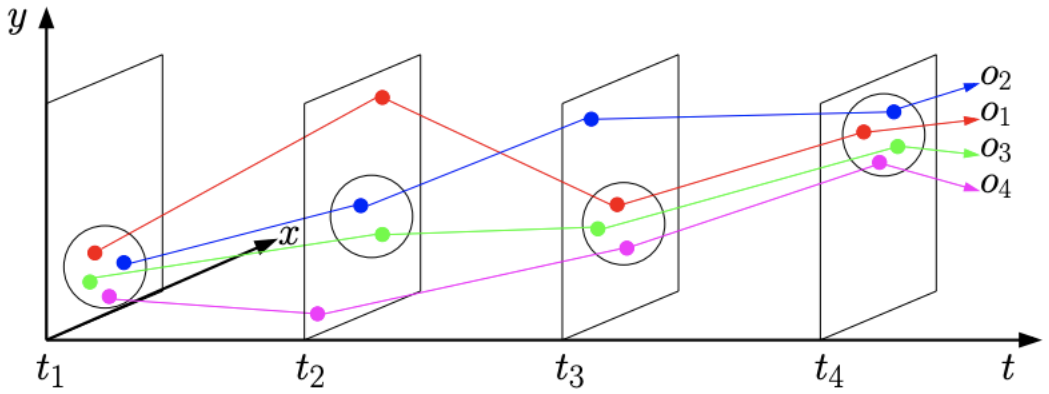
1.3 轨迹数据挖掘技术

伴行模式(swarm)
swarm(m, t0)

一组m个个体在空间上相互靠近。

不要求在连续k个时间彼此靠近，但至少在一个时间段t0内相互靠近。

允许有短暂的分开，再汇聚。



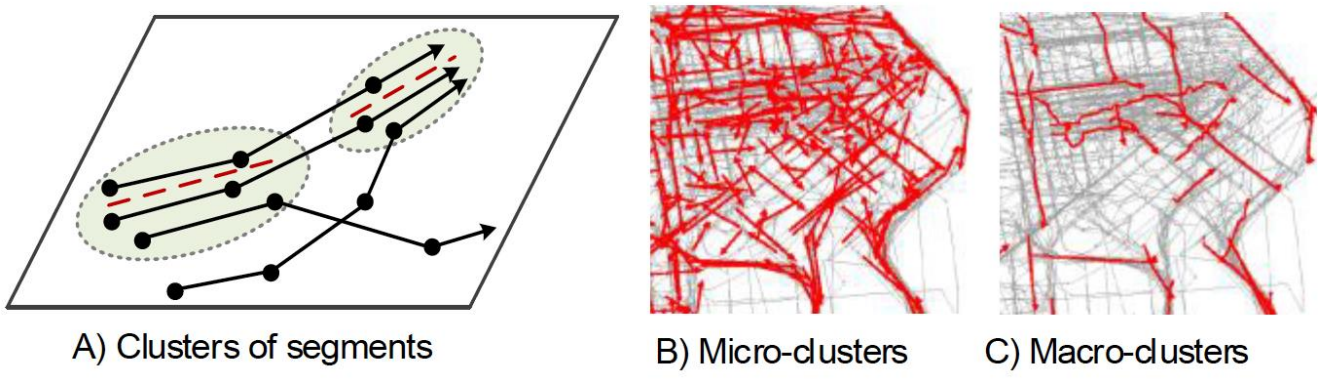
轨迹大数据分析

1.3 轨迹数据挖掘技术

轨迹聚类

轨迹分段，
轨迹分段后聚类，
如图(A)所示。

轨迹微观的聚类(B)
轨迹宏观的聚类(C)



轨迹大数据分析

1.3 轨迹数据挖掘技术

序列模式：一组个体在一个相似的时间间隔内，经过了一些公共的位置。
比如A,B 两个物体经历了l_1, l_2, l_4 3个公共的地点。

自由空间的序列挖掘
路网的序列挖掘

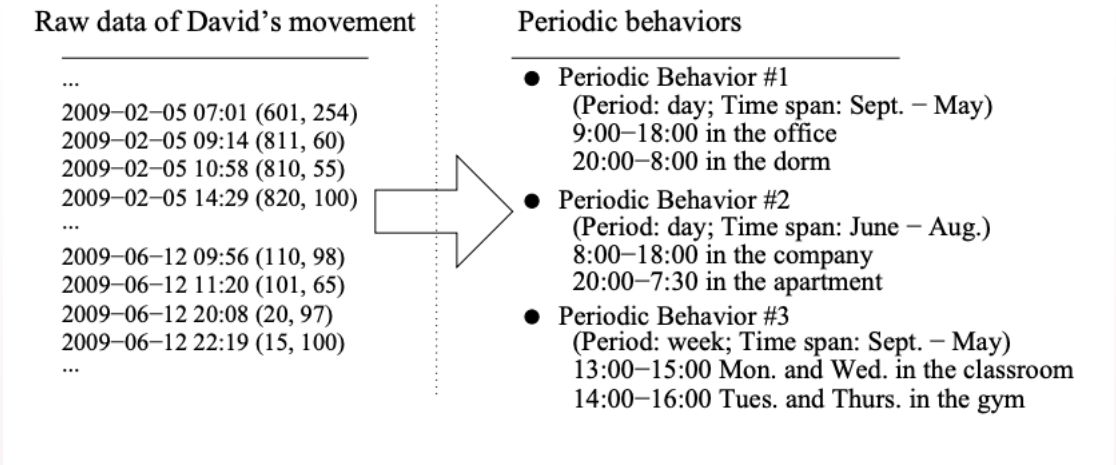
$$A: l_1 \xrightarrow{1.5h} l_2 \xrightarrow{1h} l_7 \xrightarrow{1.2h} l_4. \qquad B: l_1 \xrightarrow{1.2h} l_2 \xrightarrow{2h} l_4,$$

轨迹大数据分析

1.3 轨迹数据挖掘技术

周期模式：

通过David的移动位置，发现David在不同时间的行为模式。



轨迹大数据分析

1.3 轨迹数据挖掘技术

轨迹的语义转化

移动性理解：理解移动对象的共性趋势和汇总性趋势。
(比如一群候鸟的迁徙路径(给一群候鸟带上传感器，定时传回位置信息.)

行为理解：将轨迹的数据与兴趣点和签到相结合。理解特定用户的行为
(在哪里坐地铁，在哪里上班，喜欢在哪里看电影，喜欢去哪里吃饭等。)

轨迹的相似性：时空轨迹的相似性，语义轨迹的相似性。进行轨迹的搜索挖掘。
(判断两个人的轨迹是否相似，是否参加了同一个活动。)

轨迹大数据分析

1.3 轨迹数据挖掘技术

轨迹的语义标注

传统的轨迹，仅仅包含时间和地理位置坐标的信息。

语义标注：包含了地理位置的语义信息。

轨迹的行为理解：把轨迹数据和兴趣点(Points Of Interests POI),以及签到数据相结合(check in).

区域标注：轨迹和兴趣点相结合，计算公交站点之间的移动规律。

路段标注：地图匹配算法，识别车辆行驶路段，及车辆在路段中的位置。

位置标注：识别轨迹兴趣点(POI)，基于预先定义的热点集合，推测移动对象的行为，挖掘轨迹数据中的周期行为。

轨迹大数据分析

1.4 轨迹数据实例-苏宁云商

实验目标：
研究电商大数据线上和线下融合(O2O)，并为电子商务提供决策支持。

实验内容:

(1) 基于室内的用户轨迹数据，研究空间布局和建模，用户驻留点定位，行走推测等。

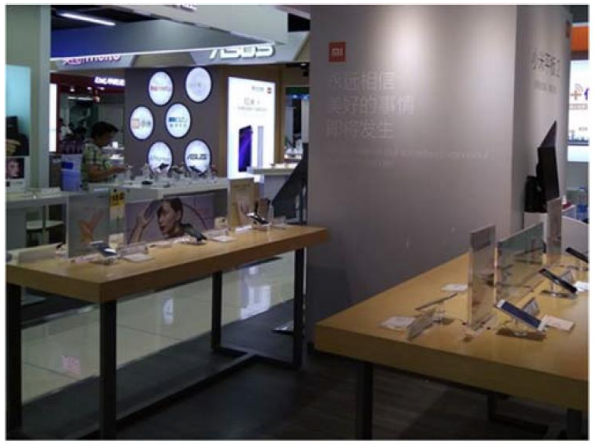
(2) 结合用户的轨迹数据，研究位置与线下购买行为和活动策划的相关性。

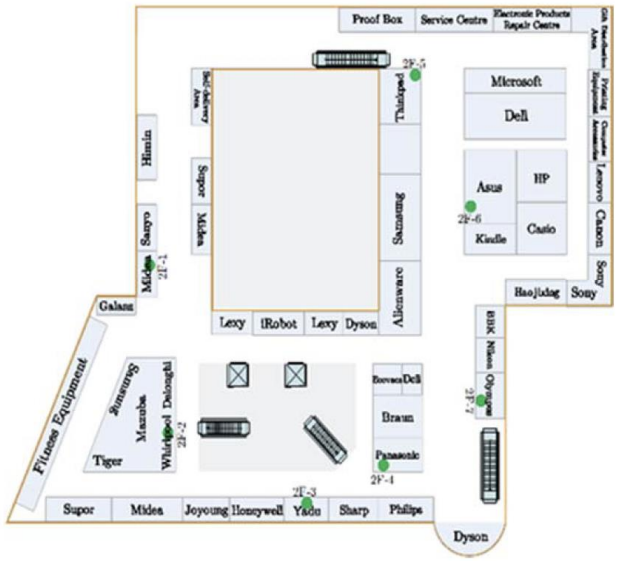
(3) 通过线上和线下的数据融合，构建室内综合导航和推荐。

轨迹大数据分析

1.4 轨迹数据实例-苏宁云商

数据采集：
苏宁云商线上数据：服务器网页的访问日志，APP的访问日志。
线下数据：23个Wi-Fi传感器。





轨迹大数据分析

1.4 轨迹数据实例-苏宁云商

数据采集：

线下数据：23个Wi-Fi传感器，

MAC (手机的物理地址)
Sensor_ID (传感器的标识)
Timestamp (时间戳信息)
Phone_brand (手机品牌)
RSSI (接收信号的强度,[0,99],
可估计手机的位置范围)

探测点	MAC	品牌	最早探测时间	最后探测时间
ZYM-1	60f...	三星	2016/10/10 8:56	2016/10/10 8:57

通过 WiFi，将线下采集到的数据与线上移动端相关联

MAC	品牌	商品/页面ID	页面描述	访问时间
60f...	三星	437060	K7自动折叠车	2016/10/10 15:28

轨迹大数据分析

1.4 轨迹数据实例-苏宁云商

数据预处理：
Wi-Fi传感器的覆盖范围是圆形的区域，因此会检测到路过商场的无效用户。

(1)去除非营业时间的记录，只保留[10am,9pm]的记录。
(2) 去除距离传感器很远的手机记录，仅保留RSSI<=60的记录。
(3) 去除卖场中短时间停留的用户记录。 用户至少被检测到2次，而且时间间隔大于10分钟。

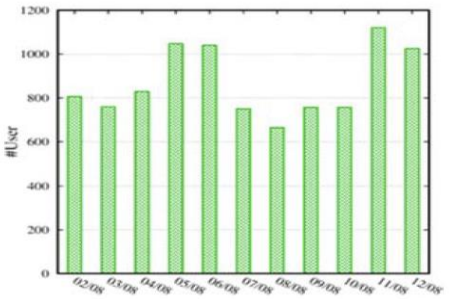
表 3-3. 数据预处理的结果

	原始数据	过滤器 1	过滤器 2	过滤器 3
用户数	201，621	171，620	14，602	5，587
记录数	6510307	5504721	529579	501427

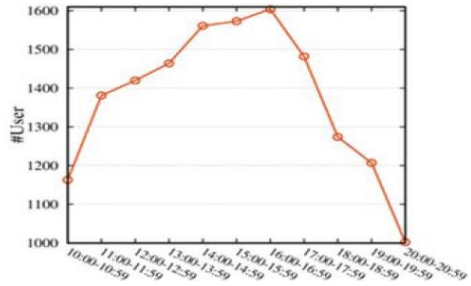
轨迹大数据分析

1.4 轨迹数据实例-苏宁云商

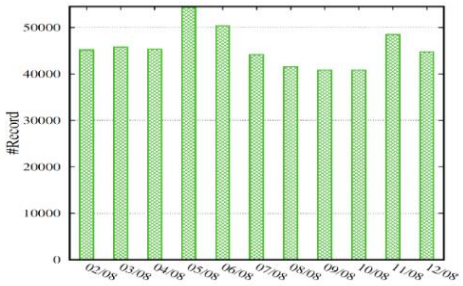
用户轨迹的基本统计信息：



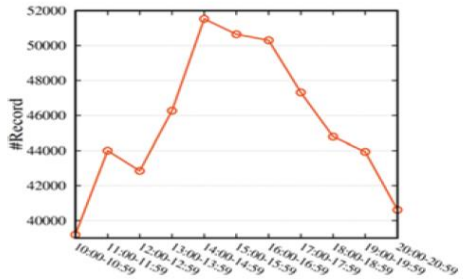
(a)每日顾客数



(b)每个时段的顾客数



(c)每日记录数



(d) 每个时段的记录数

图 3-11. 基本统计信息

轨迹大数据分析

1.4 轨迹数据实例-苏宁云商

用户行为分析

(1) 用户访问天数 (NoVD):
用户在2017年8月2日到12日之间的访问天数 (0, 11)
(2) 访问时间段 (NoVT): 上午10点到下午9点， 每个小时一个间隔, (0, 11)
(3) 记录数(NoR), 从传感器收集的用户记录编号。(0, 500)
(4) 传感器数量：部署了23个传感器 [0, 23]

通过以上4个特征, 使用Min-Max预处理技术进行归一化，
然后用Kmeans聚类的方法，把用户分为5类。

轨迹大数据分析

1.4 轨迹数据实例-苏宁云商

用户行为分析

表 3-4. 五个集群的规模和重心

	C#1	C#2	C#3	C#4	C#5
#User	4015	135	1116	210	111
NoVD	0.008	0.961	0.03	0.62	0.629
NoVT	0.008	0.795	0.017	0.237	0.381
NoR	0.021	0.936	0.062	0.229	0.915
NoS	0.084	0.192	0.309	0.256	0.388

注：2-5 行是每个集群重心的特征值

轨迹大数据分析

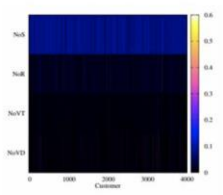
1.4 轨迹数据实例-苏宁云商

群体用户行为的解释：

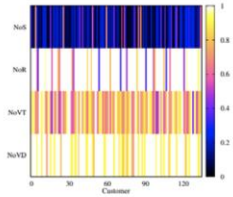
(1) 对于Cluster #1和Cluster#3,由于NoVD, NoVT, NoR的数据都比较少，说明这2个人群很少访问。Cluster#3比Cluster#1相对忠诚，因为NoS 和和NoVD的数字较高。

(2) 对于Cluster#2,由于NoVD，NoVT和NoR很大，但是NoS很小，所以是展览用的固定位置的手机。

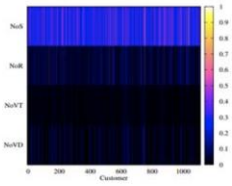
(3) Cluster#4，Cluster#5,他们的NoVD，NoVT 和NoR值远远高于Cluster＃1 和Cluster＃3，但略低于Cluster＃2。此外，它的NoS 值显然高于Cluster＃2。所以是工作人员。Cluster#4 的 NoS值比Cluster#5低，所以是固定员工。而Cluster#5是移动频繁的员工。



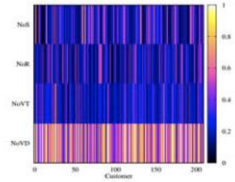
(a)Cluster #1



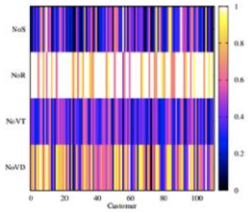
(b)Cluster #2



(c)Cluster #3



(d)Cluster #4



(e)Cluster #5

轨迹大数据分析

1.4 轨迹数据实例-苏宁云商

用户个体行为分析：

识别热点区域：利用回归模型，找出热点区域。
识别热点路径：利用马尔科夫模型，找出热点路径。

用户个体行为分析：

识别热点区域：利用回归模型，找出热点区域。

表 3-5. 顾客停留的热点区域

Sensor_ID	Coefficient	Descriptions of Region
1F-1	0.11	On the entrance and near Starbucks coffee
1F-3	0.088	On the entrance and near China Mobile
1F-7	0.073	Selling Chinese mobile phones, e.g., Huawei
1F-2	0.068	Near elevators and also Chinese mobile phones
2F-5	0.098	Near escalators and laptops marketing
2F-2	0.059	Selling small appliance with broad brands
3F-1	0.095	Selling TVs with broad brands, e.g., LG
3F-3	0.084	Near elevators and selling Skyworth TVs
4F-2	0.075	Near elevators and selling small appliance
4F-1	0.07	Selling air-conditioners, e.g., Haier and Daikin

轨迹大数据分析

1.4 轨迹数据实例-苏宁云商

用户个体行为分析：

识别热点区域：利用回归模型，找出热点区域。
识别热点路径：利用马尔科夫模型，找出热点路径。

用户个体行为分析：

识别热点路径：利用马尔科夫模型，找出热点路径。

表 3-6. 热点路径

Source	Intermediate Points	Destination
1F-1	1F-2	1F-4
	1F-2→1F-4	1F-5
	1F-3→1F-7	1F-6
	1F-3	1F-7
1F-7	1F-5→1F-4→1F-2	3F-1
	1F-5→1F-4→1F-2→3F-4	3F-2
	1F-5→1F-4→1F-2	3F-3
	1F-5→1F-4→1F-2	3F-4
	1F-5→1F-4→1F-2→3F-3	3F-5
4F-4	1F-2	1F-4
	1F-2→1F-4	1F-5
	1F-2→1F-4 →1F-5	1F-6
	1F-2→1F-4	1F-7