# 基于轨迹的身份合一方案

# 基本定义

定义1：用户轨迹tr。用户轨迹是一系列三维时空点的集合，记为, 其中表示轨迹tr在时刻的位置为，<<<，n是轨迹tr的采样点数目。

定义2：匿名轨迹。手机号的电围网轨迹。

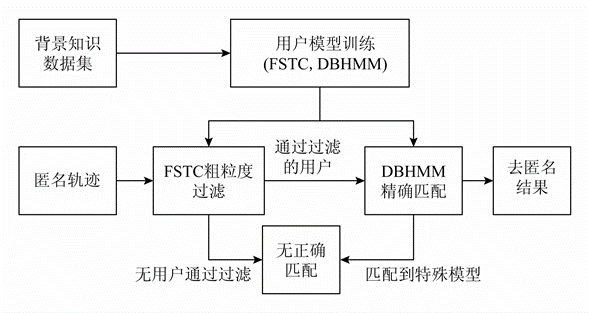
定义3：轨迹数据集TD。一个轨迹数据集包含多个用户，每个用户有多条轨迹，，表示用户i的第j条轨迹。

定义4：背景知识数据集。身份证在各种场景中的使用轨迹数据集。

定义5:开放场景．假定不一定包含的全部用户，即敌手并不知道自己是否掌握匿名轨迹对应的用户模型．

定义6:非开放场景．假定包含的全部用户，即敌手明确自己掌握了匿名轨迹对应的用户信息，只是不知道是哪个用户．

# 方案框架



根据现有背景知识数据建立用户的行为模型。每个用户的模型分为2层：第1层是作为粗粒度特征的频繁时空块（FSTC）。第2层是进行精准去匿名的基于密度隐马尔可夫模型（DBHMM）。匿名轨迹首先与用户的特征频繁位置进行粗粒度匹配，如果不够相关，则不进行进一步精确匹配，直接输出为无正确匹配项；反之则可以得到1个粗匹配概率，随后进行DBHMM的精确匹配得到一个精确匹配概率，最终匹配概率最高的用户就是匿名轨迹对应的用户身份。

# 基于密度的隐马尔可夫模型

主要对隐马尔可夫模型的隐含态进行了优化，现有的方案有些直接指定隐含态，有些隐含态只考虑地理维度或者时间维度，有些是指定隐含态的数目后采用EM算法获取每个用户隐含态，但是由于行为的差异性很大，上述方案都无法很好的兼顾用户的个体差异性。

本方案采用密度连通聚类算法（density-joinable cluster, DJ-Cluster）找出代表隐含态的密集时空点集合，其主要思想是通过参数 Minpts和生成初始簇，然后逐步合并密度连通的初始簇直到无法进一步合并，得到最终的密度聚类。Minpts是构成初始簇需要的最小点数目，是初始簇范围的最大半径。这样就能够发现任意形状的聚类和聚类的数目完全由数据本身决定，使它能够更准确地发现每个用户轨迹中的个性化隐含态。

### 密度隐含态的生成

采用欧氏距离来度量2个时空点和之间的距离．为了保证距离的一致性，纬度和经度以m为单位，时间以s为单位:

其中a=111000，

和传统的隐含态生成方式不同，密度隐含态是根据以下两个条件生成：

1)}。是隐含态s中包含的时空点数目，。

2) 任意和距它最近的。

### 模型训练和优化

定义基于密度的隐马尔可夫模型为。每个参数的按如下方式获取：

隐含态S=: 由基于密度的聚类从用户轨迹中获得。

观察态：每个元素都是用户访问过的一个时空点。

初始概率分布：每个隐含态的初始概率分布二如下所示：

，

其中。

转移概率矩阵:隐含态之间的转移概率，隐含态转移到隐含态的转移概率如下所示：

其中，表示状态转移到状态的次数，表示状态发生转移的次数。

发生概率分布：每一个隐含态都以不同的概率表现为一系列观察态，一个观察态在隐含态的映射概率为

,

其中是观察态在隐含态中出现的次数。

### 流程图

开始

一个用户的轨迹数据集

计算

使用DJ-Cluster得到隐含态S

统计

得到初始概率分布

计算

得到转移概率矩阵

得到

计算

得到发生概率分布

隐马尔可夫模型M={O,S,Π,A,E}

结束

# 身份合一过程

有2个重要部分：频繁时空块过滤和无正确匹配判断。

### 频繁时空块过滤

为了在确保准确率的同时提高效率，需使用一种粗粒度的过滤机制将无关的用户模型过滤掉，只留下少数候选用户模型与匿名轨迹进行精确匹配，先给出一些基本的解释和定义：

定义7: 时空块（spatio-temporal cube)。将三维（时间、纬度和经度）空间按一定的间隔进行划分，三维空间则表现为大量的时空块。每天按小时分为24个间隔，纬度和经度按1km进行划分。最多可以得到个时空块，每个时空块的体积约为。其中，是用户轨迹纬度范围内划分后的数量，是用户轨迹经度范围内划分后的数量。

定义8：频繁时空块（FSTC)。频繁时空块是针对一个用户的定义，满足下列条件的时空块就是该用户的频繁时空块：

1)将三维空间划分为时空块后，某个用户轨迹中的每个时空点都有对应的时空块存在；

2)时空块中包含的用户轨迹时空点数目为 N，且，其中K是自定义阂值。

在进行过滤时，一条匿名轨迹将与每个用户的频繁时空块进行匹配，先计算p:

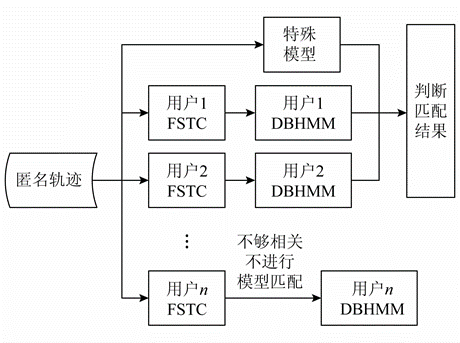
如果，则该用户为潜在用户，进行下一步的模型匹配。

### 无正确匹配判断

在非开放场景下，将身份合一看作多模型匹配问题，匿名轨迹与已有的多个用户模型相匹配，找出匹配概率最大的作为身份合一结果。在开放场景下，比非开放场景多出一种无正确匹配的情况，可以看作是需要增加一个特殊背景模型的多模型匹配问题。因此，问题的重点就转变为如何构建一个可以与大多数用户的轨迹匹配的特殊背景模型。

本方案选择选择混合多个用户的数据训练一个模型作为特殊背景模型。由于这个模型在构建时混杂了多个用户的数据，普适性要远远地高于根据一个用户的数据训练出的模型．因此，无正确匹配的匿名轨迹与已有用户模型的匹配概率会低于与背景模型的匹配概率（相似用户情况除外）。

### 流程图



身份合一过程可以进一步分为3个子阶段：

1)过滤阶段。通过频繁时空块过滤机制选择出可以进人下一个阶段的用户模型。

2)匹配阶段。利用Viterbi算法计算匿名轨迹数据集中的每条轨迹和背景模型及通过过滤的用户模型的匹配概率。

3）判断阶段。假定数据集中有N条轨迹，首先找出每条轨迹的Viterbi匹配概率最大的top-k用户模型，然后从个结果中选出出现次数最多的 k个用户作为结果集。如果背景模型在结果集中排第一，则该匿名轨迹集对应的用户不存在；否则，结果集去除背景模型后作为最终的去匿名结果。

# 性能分析

### 时间复杂度

从时间复杂度的角度来说，基于频繁时空块过滤的时间复杂度是，基于隐马尔可夫模型方法通过Viterbi算法进行匹配的时间复杂度是，n是待匹配的轨迹点的数目，m是隐马尔可夫模型的隐含态数目。此外，直接对比需要的基本操作要少于Viterbi算法每次计算的基本操作，所以2种方法的基准消耗时间也不同。因此，过滤后再使用模型匹配消耗的时间要远远少于直接通过模型匹配消耗的时间。

### 准确性

1)数据描述

第1个数据集是Geo1ife，包含微软亚洲研究院在2007年4月到2011年10月期间收集的 182个北京用户的高频率GPS数据．每条记录包括经度、纬度、海拔、日期和时间．为了保证模型的准确性，我们选择涉及日期超过15 d的共105个用户的数据进行实验．

第2个数据集是Gowa11a，包含California 和Nevada的10162个用户从2009年2月到2010 年10月期间的456 988条签到数据．每条签到数据都包括用户ID、签到时间、签到坐标（经纬度）信息．选择了签到数据多于100条的共1 072个用户的数据进行实验．

在非开放场景下，我们选择每个用户的80% 数据作为训练集进行用户行为模型训练，剩余的 20％数据作为测试集进行模型的准确率测试，训练集和测试集没有任何交集。

在开放场景下，我们分别在2个数据集的测试集中额外添加了新的测试数据，这些测试数据对应的用户模型并不存在。无匹配模型的测试用户与最初的测试用户的比例约为2:3。

2)数据预处理

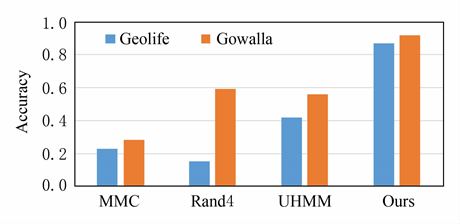
原始数据集并不适合直接用来作用户建模，直接使用初始的经纬度时空信息可能引发2个问题。首先，原始数据精度过高容易导致过拟合。其次，数据精度过高也会导致计算量过大。所以实验中对地理位置数据进行了一定的泛化，经纬度精度为小数点后3位，时间精度以0.5h为单位。

#### 非开放场景下结果分析

在非开放场景下与以下3种去匿名方法进行准确率对比：

1)MMC。对训练数据和测试数据数据分别提取POT后构建马尔可夫链并对比2个马尔可夫链的相似度；

2)Rand4。从训练轨迹中随机选择 4个点作为一个用户的唯一标识，以此进行匹配； 3)UHMM。同样选择了隐马尔可夫模型，但指定一天的24 h为24个隐含态来对用户建模。



#### 开放场景下结果分析

对比的是本方案模型在开放场景和非开放场景的准确率（accuracy)，查准率 (precision）和查全率（recall）。

