基于标准空间的局部轨迹相似性算法的设计与实现

目录

[第一章 绪论 3](#_Toc527386771)

[1.1 研究背景 3](#_Toc527386772)

[1.2 本文研究内容 4](#_Toc527386773)

[1.3 本文组织结构 4](#_Toc527386774)

[第二章 相关理论与关键技术 5](#_Toc527386775)

[2.1 现有轨迹相似性计算方法 5](#_Toc527386776)

[2.2 现有距离计算方法 5](#_Toc527386777)

[第三章 标准空间定义 6](#_Toc527386778)

[3.1 数值不均衡问题的提出 6](#_Toc527386779)

[3.1.1 数值不均衡问题的定义 6](#_Toc527386780)

[3.1.2 数值不均衡问题的影响 6](#_Toc527386781)

[3.2 标准空间 8](#_Toc527386782)

[3.2.1 标准空间的定义 8](#_Toc527386783)

[3.2.2 标准空间的构建方法 8](#_Toc527386784)

[3.3 标准空间下点和线段的距离计算 9](#_Toc527386785)

[3.2.1 标准空间下的点距离 9](#_Toc527386786)

[3.2.2 对应样本点的获取 9](#_Toc527386787)

[3.2.3 标准空间下的线段距离 9](#_Toc527386788)

[第四章 基于标准空间的局部轨迹相似性算法 10](#_Toc527386789)

[4.1局部轨迹相似性问题的提出 10](#_Toc527386790)

[4.2轨迹段相似度及代价定义 10](#_Toc527386791)

[4.2.2 轨迹段的相似度 10](#_Toc527386792)

[4.2.2 获取轨迹段相似度的代价 10](#_Toc527386793)

[4.3基于代价的局部轨迹相似性算法 10](#_Toc527386794)

[第五章 实验设计与分析 11](#_Toc527386795)

[第六章 总结与展望 11](#_Toc527386796)

# 第一章 绪论

## 1.1 研究背景

近年来，随着移动设备和 GPS 的不断发展，人们已经可以很轻松地获取移动物体的地理位置信息，为了可以更好地利用这些信息，将需要使用一些技术手段去对这些信息进行处理，而更好地处理前提是需要更多的数据去支持算法的运行，从而又带动了地理位置的采集，形成一个良性循环。

数据采集中获取到的数据有很多种类。比如用户手持移动电话，通信公司就会根据手机信号和信号发射基站的位置去确定用户的具体位置，根据一个制定好的采样策略，收集用户所在位置的经纬度、当前时间，并根据多次采集数据去计算得到用户的平均移动速度和移动方向等。根据用户的移动数据，可以获取到大量用户的个性化信息，比如该用户经常去某家餐厅就餐，那么一些app可以根据用户喜好，为用户推荐类似口味的餐厅。或者根据用户一周内频繁出现的场所，为用户推荐周边的美食或娱乐场所。还可以根据用户的移动速度的变化，判断用户在某段路程里打了出租车，可以为用户推荐上车周边更好打车的地点。让用户可以不刻意得去记录自己的日常行为，仅仅被记录下行为轨迹，便可以获得个性化的推荐。

除了用户轨迹，现在很多出租车和私家车上都安装了车载GPS，通过车载GPS的工作可以将汽车每日的移动轨迹上传到服务器，然后通过对一个城市大量出租车、私家车的轨迹分析，可以得到很多信息。比如通过分析一天的轨迹信息中道路上车辆行驶速度，可以得到该城市每日早高峰晚高峰大约会出现在什么时间段，建议不赶时间的司机错峰行驶。还可以通过实时轨迹数据得到当前时间道路的车流量，判断该条道路在该时刻的拥堵状况，道路拥堵信息可以在广播频道里司机进行实时指导路线，或者在手机的出行app里动态展示，为司机挑选相对通畅的道路。

在野生动物保护领域，一些动物保护组织或者野生动物纪录片的导演需要了解动物生活习性并摆放摄影机对动物日常行为进行摄制，这里就需要用到无线传感器和GSM/GPS等远程跟踪动物的设备，通过在野生动物的身上安装一个追踪节点，就可以记录并跟踪动物所在的位置信息。然后可以使用带有红外的相机去探测野生动物的具体位置，保证了不管是白天还是黑夜，都可以记录动物的影像。此外还可以通过无人机在空中跟踪拍摄一些警觉性较高的动物，获取宝贵的野生动物轨迹数据和影像资料。通过其中的轨迹数据我们可以得到动物经常活动的区域及其栖息地，通过研究栖息地的气候条件比如采光，温度和湿度去判断该种群的生存条件。还可以根据轨迹数据获得一些有迁徙习性的动物的迁徙路线，研究气候变化对其迁徙路线的影响。

除了对用户位置的信息采集，对出租车移动路线信息的采集，对野生动物行为轨迹的采集之外，还有军事领域对地方目标轨迹的实时监测以实现精准打击，对飓风移动路径数据的采集来预报和预防自然灾害等等。随着数据采集设备的改良和采集方式的优化，各个领域都产生了海量的轨迹数据。所以对轨迹数据的分析利用变得十分重要。人们为了发掘海量数据中隐藏的价值，得到丰富的数据特征空间以及用户轨迹的规律性信息，开发了聚类分析、隐私保护和行为预测等一系列的应用技术。而这些技术的实现都得益于移动对象轨迹的相似性度量技术的发展。

轨迹数据展示了移动对象的时空动态，以数据的形式存储了空气、动物、车辆和人类的运动信息，在预测风暴移动、研究动物迁徙、规划城市建设和提供出行路线等方面有着重要的应用。而这些应用都需要轨迹数据库提供一个高效的轨迹查询功能，其中就包括轨迹的相似性查询，其定义如下。

**定义1**（轨迹相似性查询）给定一个查询轨迹Q，轨迹数据库S和一个阈值，返回S的一个子集，使得

d(Q,),

其中d(·)是轨迹数据之间的相似性计算函数。

在移动对象的轨迹相似性查询中，相似性计算函数是核心。我们给出相似性计算函数的形式化定义。

**定义2**（轨迹相似性计算函数）给定一个定义在轨迹数据上的空间T，以及任意两条T中的轨迹x与y，则T上的相似性计算函数d定义为：

其中R是实数空间。

当前该领域主要围绕两个问题进行研究，一是研究合适的相似性计算函数，二是研究高效的检索机制。选择一个合适的相似性计算函数和利用函数制定高效的检索机制至关重要，这些因素同时决定了查询方法的好坏。比如有时候我们无需对采样得到的整段轨迹计算与其他轨迹的相似度，只需要对一小段子轨迹选取合适的函数进行相似性计算即可，这样就可以在一定程度上减少运算时间。因此，在面对不同场景时我们需要根据具体情况采用不同的相似性查询方法。

以上对轨迹相似性的查询和已有的一些轨迹相似性计算函数一般针对的是完整轨迹。比如给定一条查询轨迹Q，一条数据库中的数据轨迹R，通过相似性计算函数得到的是轨迹Q和轨迹R从起始位置到终止位置所有样本点按照时间顺序连成轨迹段的相似性。但是实际情况下，轨迹数据库中的单条轨迹在很大部分轨迹段上与查询轨迹Q并不相似，但是有一小部分，比如四分之一的长度和Q是很重合的，有的时候这四分之一的长度包含的信息也会很有用，但是之前的方法没有把这些信息挖掘出来。比如为用户推荐顺风车，用户需要的路线是顺风车CarA预计行驶轨迹的四分之一，另外一辆顺风车CarB的预计行驶轨迹与用户需要的路线很接近，但是大于CarA与用户接近的那部分轨迹距离，那么最好的情况应该是为用户推荐CarA。但是如果使用之前轨迹相似性算法计算轨迹全段的相似程度，可能出现由于CarA的轨迹很大部分与用户轨迹不相似，而为用户推荐了CarB的结果。这并不是使用相似性算法的本意，于全部轨迹的相似性计算没法解决这个问题。

那么如果其中有一小部分很相似的轨迹段变会被很长的不相似的轨迹段给掩藏起来，最后会找到一条

## 1.2 本文研究内容

## 1.3 本文组织结构

# 第二章 相关理论与关键技术

## 2.1 现有轨迹相似性计算方法

## 2.2 现有距离计算方法

# 第三章 标准空间定义

## 3.1 移动用户轨迹的表示方法

## 3.1 数值不均衡问题的提出

### 3.1.1 数值不均衡问题的定义

数值不均衡问题，即时空数据某一维度数值跨度很大，而相比而言另一维度的跨度很小。例如轨迹开始与终止时间跨越为十二个小时，即43200秒，空间距离跨度为4000米，从数值上看，时间维度的数值远大于空间维度的数值。或者汽车从北京（116.408598, 39.923079）开往南昌（115.700302, 28.934003），在经度上跨度很小，约为120千米，而纬度上跨度很大，约为1320千米，纬度的跨度远大于经度跨度。

### 3.1.2 数值不均衡问题的影响

Personalized trajectory matching in spatial networks（2014年VLDB）使用以下的方法计算轨迹间距离，假设查询轨迹q和数据轨迹t上对应点为v1和v2。论文使用空间距离为v1和v2的路网距离，时间距离为v1和v2的时间戳之差，然后将空间距离和时间距离作为指数函数的指数部分，获取空间影响因子Is（v1,v2）和时间影响因子It（v1，v2），如公式（2）和（3）所示，该影响因子代表样本点之间的时空相似程度。在相似度函数的计算中，利用LCSS的方法，找出所有样本点对以获得最大空间和时间相似度相似度，如（4）和（5）所示，其中权重q.head.w代表点q在相似度计算中的重要程度，也是文章的创新点之一。最后使用参数将相似度结合起来，得到轨迹相似度，如公式（6）所示。









该方法的问题是没有考虑到相同差距值在不同维度具有不同的信息，忽视了空间上的单位米和时间上的单位秒的概念，将距离和时间简单作为一个数值带入公式中进行影响因子的计算。然后论文认为可以仅根据用户的时间相似度和空间相似度的重要程度确定参数。但是数值上大小相同的空间和时间影响因子，可能描述了轨迹不一样的相似程度。

假设两条轨迹的一对对应点v1和v2，空间距离是5米，时间戳差距是0分钟，另外一对对应点v3和v4空间距离是0米，时间戳差距是5分钟。此外两条轨迹在距离上首尾跨度1200米，时间跨度是10分钟。如果的确定不考虑轨迹时空跨度，即时空数据的不均衡，那么假设用户认为时间和空间因素同等重要的情况下为0.5，通过计算得到的Is(v1,v2)和It(v3,v4)数值相等，It(v1,v2)和Is(v3,v4)数值相等。但是1200米的轨迹上对应点之间5米的差距是很小的，而相比而言10分钟的跨度内，有5分钟的时间差距是很大的，计算结果很明显与用户的需求不符合。

与之类似的另一个问题，在欧式空间下，很多方法直接使用欧氏距离作为点与点之间的空间距离。而欧式空间中的x轴和y轴是两个不同的维度。如果轨迹在x轴上跨度很大，而在y轴上跨度很小，那么对应样本点在x轴上的距离差距和在y轴上的距离差距也不可以简单使用各自数值去计算。

## 3.2 标准空间

### 3.2.1 标准空间的定义

出现上述问题的关键在于之前的方法忽视了不同维度的数值传达信息的不同。上述论文中的方法在下面的情况下，影响不是很大，以二维空间举例，当轨迹沿着与经线和纬线夹角为45度的方向延伸时，轨迹在两个维度有相同的计量单位，并且有着相差不大的数值范围，此时使用欧氏距离可以描述轨迹的相似程度。但是下面两种情况，之前的方法不能很好的适用。第一种情况是轨迹与经线或纬线夹角很小，会使得轨迹在经纬度上的距离值相差很多倍。第二种情况是加入了时间维度，首先与空间维度的计量单位不同，那么相同的差值，表示的相似度信息也是不同的。

为了解决上述问题，本文提出一个标准空间的概念，在标准空间中，所有维度的计量单位均相同，不同维度可以在点与点的距离计算中直接进行运算，包括空间和时间。目的是消除量纲和数值不均衡对距离计算的影响。

### 3.2.2 标准空间的构建方法

假设有两条轨迹Q={q1, q2, q3…,qm}，R={r1,r2,r3…rn}。其中样本点格式为（，，），有两个方法可以将Q和R向标准空间转化。

第一个方法是Standardization，其构造方法是首先获取两条轨迹所有样本点，计算三个维度各自的均值和方差，，，，，。然后针对点pi使用下面公式做标准化。该方法得到的标准空间中，每个维度均服从均值为0，方差为1的正态分布。

第二个方法是Min-Max scaling，其构造方法是先获取三个维度的最大和最小值，，，，，，。然后针对所有点pi使用下面公式做标准化。该方法得到的标准空间的特点是所有数值均分布在0到1之间。

使用上述任意一种标准化方法均可得到属于轨迹Q和轨迹R的标准空间，标准空间下的，。其中每个点=(,,)。

## 3.3 标准空间下点和线段的距离计算

### 3.2.1 标准空间下的点距离

由于标准空间下所有维度的计量单位都是相同的，所以可以将时间维度纳入点对的距离计算之中，样本点之间的距离计算可以使用Lp-norm距离，其中本文以p=2，即欧氏距离来举例。

### 3.2.2 对应样本点的获取

### 3.2.3 标准空间下的线段距离

# 第四章 基于标准空间的局部轨迹相似性算法

## 4.1局部轨迹相似性问题的提出

现在一般的轨迹相似性算法计算的都是两条完整轨迹的相似性，使用一个数值去衡量轨迹之间的相似程度，数值越大代表轨迹越相似。本文解决的问题是给定查询轨迹R，求数据轨迹Q中与查询轨迹最相似的部分。

求两条轨迹之间最相似的部分。

## 4.2轨迹段相似度及代价定义

### 4.2.2 轨迹段的相似度

### 4.2.2 获取轨迹段相似度的代价

## 4.3基于代价的局部轨迹相似性算法

4.3.1

# 第五章 实验设计与分析

# 第六章 总结与展望