基于移动对象轨迹相似性的局部轨迹相似性算法的设计与实现

目录

[第一章 绪论 2](#_Toc527318942)

[1.1 研究背景 2](#_Toc527318943)

[1.2 本文研究内容 3](#_Toc527318944)

[1.3 本文组织结构 3](#_Toc527318945)

[第二章 相关理论与关键技术 4](#_Toc527318946)

[2.1 现有轨迹相似性计算方法 4](#_Toc527318947)

[2.2 现有距离计算方法 4](#_Toc527318948)

[第三章 标准空间解决数值不均衡问题 5](#_Toc527318949)

[3.1 数值不均衡问题 5](#_Toc527318950)

[3.1.1 数值不均衡问题的定义 5](#_Toc527318951)

[3.2.2 未考虑数值不均衡存在的问题 5](#_Toc527318952)

[3.2 标准空间 7](#_Toc527318953)

[3.2.1 标准空间的定义及构建方法 7](#_Toc527318954)

[3.2.2 路网空间向标准空间的转换及距离计算 8](#_Toc527318955)

[第四章 相似性算法设计 9](#_Toc527318956)

# 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，比如用户当前所在位置的经纬度、时间、速度和移动方向等，对路网中的移动车辆和行人的追踪越来普遍，随之产生的是海量的轨迹数据。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）给定一个轨迹T，轨迹数据库S和一个阈值，返回S的一个子集，使得

d(T,),

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[1]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。本文重点介绍当前主流的一些相似性计算函数。

## 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 标准空间下的距离

## 3.1 标准空间解决数值不均衡问题

### 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。

### 3.2.2 未考虑数值不均衡存在的问题

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。

### 3.1.3 标准空间的定义及构建方法

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。

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=(,,)。

## 3.2 标准空间下的距离

### 3.2.1 点距离

标准空间下

### 3.2.2 基于点距离的轨迹段距离

# 第四章 相似性算法设计