高时效性下的轨迹局部相似性查询算法的设计与实现

目录

[第一章 绪论 3](#_Toc527491328)

[1.1研究背景及意义 3](#_Toc527491329)

[1.2本文研究内容 6](#_Toc527491330)

[1.3本文组织结构 7](#_Toc527491331)

[第二章 相关理论与关键技术 8](#_Toc527491332)

[2.1轨迹表示方法 8](#_Toc527491333)

[2.1.1仅基于位置信息的轨迹表示方法 8](#_Toc527491334)

[2.1.2基于时间信息的轨迹表示方法 8](#_Toc527491335)

[2.1.3基于文本信息的轨迹表示方法 9](#_Toc527491336)

[2.1.4基于网格的轨迹表示方法 9](#_Toc527491337)

[2.2现有轨迹相似性计算方法 9](#_Toc527491338)

[2.2.1欧氏距离（EU） 9](#_Toc527491339)

[2.2.2动态时间弯曲距离（DTW） 10](#_Toc527491340)

[2.2.3最长公共子序列距离（LCSS） 11](#_Toc527491341)

[2.2.4编辑距离（EDR） 12](#_Toc527491342)

[2.2.5带真实惩罚的编辑距离（ERP） 12](#_Toc527491343)

[2.2.6路网上的轨迹相似性查询 13](#_Toc527491344)

[2.2.7城市运输系统中基于段的轨迹相似性计算 15](#_Toc527491345)

[2.2.8简化的轨迹相似性计算 18](#_Toc527491346)

[2.2.9基于签名的轨迹相似性计算 19](#_Toc527491347)

[2.3 高效的轨迹索引技术 20](#_Toc527491348)

[2.3.1 R树索引 20](#_Toc527491349)

[2.3.2 Voronoi图 21](#_Toc527491350)

[2.4本章小结 21](#_Toc527491351)

[第三章 标准空间下距离的算法 23](#_Toc527491352)

[3.1 数值不均衡问题的提出 23](#_Toc527491353)

[3.1.1 轨迹的表示形式 23](#_Toc527491354)

[3.1.1 数值不均衡问题的定义 23](#_Toc527491355)

[3.1.2 数值不均衡问题的影响 23](#_Toc527491356)

[3.2 标准空间 25](#_Toc527491357)

[3.2.1 标准空间的定义 25](#_Toc527491358)

[3.2.2 标准空间的构建方法 25](#_Toc527491359)

[3.3点和线段的时空距离定义 26](#_Toc527491360)

[3.2.1点距离的定义 26](#_Toc527491361)

[3.2.2 标准空间下的轨迹段的距离 26](#_Toc527491362)

[第四章 基于标准空间的局部轨迹相似性算法 28](#_Toc527491363)

[4.1局部轨迹相似性问题的提出 28](#_Toc527491364)

[4.2轨迹段相似度及代价定义 28](#_Toc527491365)

[4.2.2 轨迹段的相似度 28](#_Toc527491366)

[4.2.2 获取轨迹段相似度的代价 28](#_Toc527491367)

[4.3基于代价的局部轨迹相似性算法 28](#_Toc527491368)

[第五章 实验设计与分析 29](#_Toc527491369)

[第六章 总结与展望 29](#_Toc527491370)

# 第一章 绪论

## 1.1研究背景及意义

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，那么一些手机客户端可以根据用户喜好，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。

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之外，还有对野生动物行为轨迹的采集，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查询技术的发展。

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。

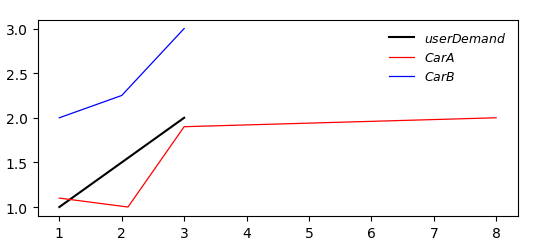
**定义2**（轨迹相似性计算函数）给定一个定义在轨迹数据上的空间T，以及任意两条T中的轨迹x与y，则T上的相似性计算函数d定义为：

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，并获得相对而言更重要的轨迹相似性信息，因为相似的那段比不相似的那段更有价值。因此在面对不同场景时我们需要根据具体情况采用合适的相似性计算方法。

大多数对轨迹相似性的查询研究和常用的一些轨迹相似性计算函数，比如动态时间规整算法，最长公共子序列算法和编辑距离算法一般针对的是完整轨迹，最后计算结果得到的是两条完整轨迹的距离或者表达轨迹相似程度的数值。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，比如三分之一的长度和Q是很重合的，有的时候这三分之一的轨迹包含的信息远大于另外三分之二的轨迹，但是之前整段轨迹比较相似性的一个缺点就是其余三分之二的不相似的轨迹段容易掩盖掉这三分之一的特征，因此我们需要额外使用一个方法，将这最相似的三分之一的轨迹段找出来。

举一个例子来说明这个问题。比如为用户推荐顺风车，用户需要的路线是顺风车CarA预计行驶轨迹的三分之一，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，导致掩盖了相似的也是用户最需要的那部分结果，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，而计算完整轨迹的相似性计算没法解决这个问题。此外，顺风车的推荐还需要考虑一个时间因素，如果下图中CarA和CarB的轨迹是一小时之前，而用户发出呼叫顺风车的请求是在现在，那么任何一辆车都不能推荐给用户。针对这种时效性的问题，还必须考虑轨迹之间的时间距离。之前的许多研究都涉及到了时间方面的因素，比如将两点之间时间戳的差值作为时间距离，最后使用一个系数和空间距离结合起来，得到一个轨迹间的相似程度。但是大多数情况是根据算法研究人员经验值去确定参数值，并且不同规模的轨迹可能还需要调整参数的大小，参数的确定很难给出一个明确的意义。



国内外很多专家学者对轨迹相似性进行了深入的研究，使用了不同的空间网络、不同的轨迹格式表示以及不同的维度企图去找到一种更好地方法去表示出轨迹之间的相似程度。但是由于研究的问题会涉及到具体的场景，由于大家研究的问题不尽相同，所以研究出了很多的相似性表示方法。首先空间网络有欧式空间下和路网下的轨迹相似性，这两种研究场景最大的区别在于欧式空间不考虑道路对移动的限制，可以使用一条直线距离去衡量任意两个地点的距离，而由于路网空间对人员和车辆的限制，两个地点之间的距离必须使用真实的道路距离去衡量。还有轨迹格式会根据研究的问题不同而选取不同的格式，比如在欧式空间下采用网格表示轨迹，还有不考虑时间因素的情况下，轨迹点信息中仅包含经纬度的信息，如果考虑时间、空间、速度和运动方向的话，轨迹点的格式就会同时包含经纬度、时间戳、瞬时运动速度以及运动的方向。而本文考虑到欧式空间对于研究的便捷性以及计算的高效性，所以不考虑道路交通情况，而是研究欧式空间下包含时间、空间、运动方向等信息的轨迹相似性。

## 1.2本文研究内容

尽管在轨迹相似性方向上已经有很多研究成果，但是上文中提到的两个问题，几乎没有一个很好的解决方案。第一个问题是轨迹的时间距离在与空间距离结合的时候，普遍使用的参数结合的方法不能赋予参数一个明确的含义。第二个问题是以往的相似性函数忽略了局部相似的情况。为解决以上两个问题，本文采用了标准空间去结合包括时间和空间在内的各个维度，并提出基于标准空间的相似性计算方法。具体包括以下几个方面：

（1）标准空间的定义。本文基于统计学知识和欧式空间构造出了标准空间，并阐述了标准空间的特点，以及如何解决之前的研究工作中存在的问题，可以更好地研究轨迹的相似程度。

（2）在标准空间的基础上，给出各种距离的定义。包括点与点之间的标准距离，轨迹段与轨迹段之间的间隔距离，以及轨迹段在形状上的改进余弦距离。标准空间中的距离可以包含构造标准空间中的每一个维度表达的信息，更好地表示空间中的轨迹段的相似程度。

（3）基于标准空间的局部轨迹相似性算法的研究。研究的核心问题是如何利用标准空间的特性，在很长的数据轨迹中找出与查询轨迹最相似的部分。设计了局部轨迹相似性查询算法，找出最相似部分轨迹并表示数据轨迹与查询轨迹的相似程度。

（4）算法实现及实验设计。实现局部轨迹相似性查询算法，并设计实验，使用微软提供的真实轨迹数据集去验证算法的高效性和准确性。

## 1.3本文组织结构

# 第二章 相关理论与关键技术

本章主要介绍一些轨迹相似性相关理论与关键技术，包括轨迹的表示方法、现有的一些轨迹相似性查询算法以及一些加速轨迹相似性查询的索引技术。

## 2.1轨迹表示方法

移动对象在移动过程中，我们对其按既定的规则进行采样，会获取一系列采样点，这些采样点再按照时间先后进行排序，就会大致还原移动对象的移动过程。实际情况中，轨迹是移动对象的连续移动过程中采样得到的一系列离散的位置序列。并且在不同的场景下，为了达到不同的目的，我们需要采用不用的轨迹表示形式。下面将介绍一些常用的轨迹表示方法。

### 2.1.1仅基于位置信息的轨迹表示方法

一般情况下，轨迹可以表示为一系列包含信息的点组成的有序集合，即T = <p1, p2, …, pn>。在欧式空间下，点pi是一个坐标的形式，pi = <loni, lati>，loni表示经度，lati表示纬度。在路网空间下，我们可以将路网模型化为图的数据结构的形式，即G = (V, E)，此时点pi表示的就是图G的顶点集合V中的一个点。其实在真实的路网环境下，采样得到的点也是用坐标形式表示的，我们会使用map-matching算法将采样点映射到路网模型的顶点中。单纯地记录轨迹位置信息的优点是简单方便，没有太多的数据冗余，并且可以使用多种简单的相似性度量方法来计算轨迹相似性，比如DTW、LCSS和EDR，这些相似性度量方法针对的都是仅包含空间信息的轨迹，简单高效。

### 2.1.2基于时间信息的轨迹表示方法

我们都知道，轨迹信息有空间尺度和时间尺度，但是上面对轨迹的表述方法仅仅考虑了轨迹的空间尺度，没有考虑到移动物体在采样点空间位置下的时间信息。在研究用户的移动模型或者预测用户下一时刻的位置等情况下，为了更准确地得到研究结果，我们一般会在记录用户位置信息的同时，记录下用户处于该位置的时间信息。我们可以将含有时间信息的轨迹表示为T = <(p1, t1), (p2, t2), …,(pn, tn)>，pi表示坐标或者顶点集合中的点，ti表示的是用户处于pi位置的时刻。记录下轨迹的时间信息比单纯记录空间位置的轨迹应用地更广泛，不但可以更加详细的描述原始轨迹，还可以使用一些对时间信息敏感的轨迹相似性度量方法对轨迹进行相似性度量，获得更为准确和有效的相似性结果。

### 2.1.3基于文本信息的轨迹表示方法

在某种特殊情况下，我们可能无需考虑轨迹的时间信息，甚至空间信息也不是首要考虑的，但是我们需要好好利用轨迹的文本信息。比如我们要开发一个推荐系统，通过研究用户的个人偏好和个性化的要求，然后给出合理的推荐方案。在生成推荐方案时，为了结合用户的偏好和要求，我们需要考虑轨迹的文本特性，由此产生了基于文本特性的轨迹信息。基于文本特性的轨迹可以表示为T = <(p1, k1), (p2, k2), …,(pn, kn)>，pi表示坐标或者顶点集合中的点，ki表示pi位置的文本描述。

### 2.1.4基于网格的轨迹表示方法

网格表示法也是一种常见的轨迹表示方法。将图平面按照一定规则划分成网格，采样点的id用所在网格的id表示，同时记录下该采样点进入所在网格和离开所在网格的时刻。网格表示法的一般表示形式为：TR={(c, I)|cC, I=(tin, tout)}，其中c代表整个网格集合C中的一个网格，tin和tout分别表示进入和离开网格的时刻。

## 2.2现有轨迹相似性计算方法

### 2.2.1欧氏距离（EU）

欧氏距离的计算首先要得到两条轨迹的对应点，按照时间先后，一一对应，如图2.1所示。



图2.1 欧氏距离一对一的对应关系

然后将所有对应点的欧氏距离进行综合处理，可以求和、求均值、取中值等，下面以求和的方式举例。给定两个一维的时间序列，，序列A和序列B的欧氏距离表达式如式(1)所示。欧式距离实际上是Lp-norms在p=2情况下的一个特例。Lp-norms定义如式（2）所示。当p=1时，L1-norms叫做曼哈顿距离。

(1)

(2)

欧氏距离有很多优点，比如计算简单，长度为n的两条轨迹，可以在时间内计算出它们的相似度，而且它满足三角不等式，如式（3）所示，运用三角不等式可以计算两条轨迹之间的距离下限，从而可以进行高效的轨迹查询。

(3)

虽然欧氏距离计算十分简单，时间复杂度低，但是缺点也是显而易见的。第一，使用欧氏距离的前提就是两条轨迹必须要拥有相等的长度，因为欧氏距离的公式决定了两条轨迹必须使用相对应的点来进行计算二维距离。第二，欧式距离不能处理局部时间偏移，局部时间偏移是指由于采样策略或对象移动速度的不同，轨迹上的样本点不能在时间上一一对应，在另一条轨迹上的对应点可能是一段时间之前或者一段时间之后的。第三，使用欧氏距离进行相似性计算容易受到噪声的影响，因为在欧氏距离的计算中，轨迹中的每个点对应到另一条轨迹上的点，如果有噪声点，那么噪声点对最后结果会产生一定的影响，带来更大的距离。随着数据量的变大和研究的深入，我们现在一般用此函数对轨迹数据进行预处理，利用其时间代价低的优点，起到一个初步筛选的作用。

### 2.2.2动态时间弯曲距离（DTW）

由于样本点采集设备的误差等原因，两条轨迹数据的样本点在时间上不能一一对应，会产生局部时间偏移的问题，只有将轨迹在时间维度上进行拉伸之后才能进行有效的相似性计算。动态时间弯曲距离将计算两条轨迹中最小对应距离之和，而不是按照时间关系一一对应，如图2.2所示。



图2.2 DTW一对多的对应关系

DTW的二维空间上的计算公式如式(4)所示。其中m和n分别表示轨迹A和轨迹B的采样点的个数，即轨迹长度。Head()函数表示轨迹的第一个采样点。d(Head(A),Head(B))表示轨迹A和轨迹B的第一个采样点之间的欧氏距离。而Rest()函数表示轨迹除去第一个点剩余的部分。动态时间弯曲距离的公式是用递归定义的，公式的含义是轨迹A和B的第一个采样点之间的欧氏距离加上轨迹剩余部分的最小的一个DTW值，直到轨迹剩余部分长度为零。下面所有公式中涉及到的Head()和Rest()函数和动态时间弯曲距离中的Head()和Rest()函数意义相同。

(4)

由于动态时间弯曲距离可以通过复制某些点来解决局部时间偏移的问题，弥补了欧氏距离只能处理等长的轨迹数据的缺点，所以应用范围比欧氏距离更广。但是动态时间弯曲的时间复杂度是，计算代价比欧氏距离大。此外，与欧氏距离一样，计算动态时间弯曲距离时，每一个点都会被强制性找出其对应点，所以也会产生噪声干扰的问题。

### 2.2.3最长公共子序列距离（LCSS）

顾名思义，最长公共子序列距离计算的是两条轨迹中最长的公共子序列的长度，以此来表示两条轨迹的相似度，计算公式如公式(5)所示。

(5)

其中：

实际上，最长公共子序列距离表示的并不是空间距离，而是“得分”，两条轨迹的得分越高，表示它们相似度就越高。由计算公式可知，在递归过程中，每当两条子轨迹的第一个采样点的欧氏距离小于一个阈值，并且两段子轨迹的长度在一定的阈值以内，就认为这两个点是匹配的，可以给当前结果加一分，继续取二者的子轨迹进行递归，否则就取子轨迹组合中最大的得分，直到子轨迹的长度为零。

相比较前面介绍的两种函数而言，最长公共子序列距离可以有效地避免噪声的干扰。因为噪声点对应到另一条轨迹上时，距离会大于阈值，噪声点将不会匹配上另一条轨迹上的点，从而排除了噪声点的干扰。在时间复杂度上，最长公共子序列距离和动态时间弯曲距离一样，也需要的时间开销。

### 2.2.4编辑距离（EDR）

EDR的核心思想是从字符串领域借鉴来的。为了判断两个字符串之间的相似程度，根据对其中一个字符串做增加、删除和修改操作，其中删除一个字符串中的字符可看做是在另一个字符串的增加字符。增加字符的操作是为了使两个字符串序列长度相等，我们把增加的字符叫做间隙元素（gap）。两个字符串之间的距离如式（6）所示。

(6)

然而时间序列中的元素是实数，有时候不会像字符那样完全相等，所以当两个实数之差小于阈值时，我们就认为这两个实数相等，因此时间序列中元素之间的距离如式（7）所示。EDR是基于时间序列中元素的距离distedr得到的，如式（8）所示，其中序列R和S的长度分别是m和n，Rest（R）和Rest(S)是序列R和S除第一个元素以外的剩余元素。EDR能够处理时间序列偏移的能力就是由于当r1和s1不相等时，取值为R、S和其剩余部分相结合EDR的最小值，从而匹配了最合适的点对。

(7)

(8)

### 2.2.5带真实惩罚的编辑距离（ERP）

ERP[11]是L1-norms和EDR的一个结合，在计算两个元素之间距离的时候，当遇到两个非间隙元素时采用元素间真实的L1-norms距离而不是0，当其中有一个元素是间隙元素时，利用一个常数g来参与L1-norms距离计算，因此ERP的计算结果中包含了两条轨迹之间真实的距离。ERP中两个序列中元素的距离表示如式（9）所示。基于disterp的ERP计算公式如式（10）所示，类似于EDR的计算方法，当序列R和S长度均不为0时，ERP将计算R、S与其剩余部分结合的ERP最小值，因此ERP同样可以处理局部时间偏移。

(9)

(10)

### 2.2.6路网上的轨迹相似性查询

当将一段轨迹展示在路网中时，考虑到实际道路情况，点与点之间不一定有直线道路相连，使用二维空间的Lp-norms即欧氏距离来计算两条轨迹的相似度和实际相似情况可能相差较大。因此我们需要重新定义一个适用于路网的相似性函数。

在计算相似性之前，需要将轨迹映射到路网上去，如图2.5所示。给定移动对象a和b的移动轨迹Ta和T b，轨迹格式为T={(l1, v1, t1), (l2, v2, t2),…(lm, vm, tm)}，其中li=(lgi, lai)表示样本点坐标，vi表示对象在ti时刻的移动速度。我们用da(lai,Tb)表示从样本点lai到轨迹Tb的路网距离。lai到Tb路网距离指，Tb上距离lai最近的点到lai的路径距离。用DG表示路网G的直径。在不同条件下，对图G中的轨迹进行相似性查询，我们有不同的距离函数。

图2.5 将道路转化为路网

路网上Ta和Tb的距离dN(Ta, Tb)的计算公式如式（14）所示。其中，轨迹Ta叫做查询轨迹，轨迹Tb叫做目标轨迹，m代表轨迹Ta的样本点个数。

(14)

当查询用户有对兴趣点的查询需求时，式中的样本点将会拥有权值wai，wai的大小代表样本点在相似性查询操作时不同的重要性。带权值的轨迹距离公式如式（15）所示。

(15)

当需要对路况信息或者交通拥堵信息进行分析时，轨迹的实时速度信息就十分重要了。令SG代表当前道路最高限速，ds(lai,lbi)代表轨迹A上的第i个样本点lai与轨迹B上的第i个样本点lbi的速度之差，其中lbi是轨迹B上到lai最近的样本点。带速度信息的轨迹距离公式如式（16）所示。

(16)

分析路况信息时，时间信息也是一个很重要的因素。我们令代表轨迹A上的第i个样本点lai与轨迹B上的第i个样本点lbi的时间之差，其中lbi是轨迹B上到lai最近的样本点。la1是轨迹Ta的第一个点，lam是轨迹Ta的最后一个点，lb1是轨迹B上距la1最近的点，lbm是轨迹B上距lam最近的点。带时间信息的轨迹距离公式如式（17）所示。

(17)

当结合权值、时间和空间来计算轨迹距离，我们有公式（18），其中和分别代表对应的子计算方法的权值参数，并且两个。如果将速度信息加入公式，公式如式（19）所示。

(18)

(19)

上述五种情况的基本思想是首先找出对应点，给定查询轨迹上的点，它的对应点是目标轨迹上到该点路网距离最近的点。然后考虑每一组点对之间距离、权值、速度和时间占总体的比例，最终得出以上公式。优点是并没有采取按照时间来匹配对应点的想法，计算结果可能更符合原始相似情况，并且计算思路简单、清晰，扩展方便。缺点是样本点权值的选取没有给出一个具体的方法，可能会使计算效果不好。

### 2.2.7城市运输系统中基于段的轨迹相似性计算

轨迹点是根据一个给定的采样方法获取的，不同的采样方法给轨迹相似性计算带来了很大的影响。比如两条完全相同的轨迹，但是由于采样开始时间不同，就造成了样本点的错位，如图2.6所示。



图2.6 不同采样方法造成的样本点错位

传统的相似性计算方法，比如DTW就没有考虑这个问题。LCSS忽略了轨迹的空间距离，EDR没有考虑到轨迹的形状因素。由于传统轨迹相似度计算方法计算结果的不准确，所以Mao等人提出了基于段的轨迹相似性计算方法[14]。

首先介绍点段距离的概念。点段距离是两条轨迹对应点之间的特殊距离，表示为图2.7中的阴影面积，由样本点R、S以及各自前后样本点的中点连接而成。由于不规则阴影面积的计算比较复杂，而阴影部分正比于图2.8中两个虚线三角形的面积之和，所以将阴影部分面积的计算转换为图2.8中三角形面积的计算。但是当seg1和seg2很长时，效果并不好，所以用三角形的高，p1到seg2的距离和p2到seg1的距离来代表p1与p2之间的距离，如图2.9所示，其中为p1到轨迹S的点段距离。对应的距离公式如式（20）所示。

图2.7 使用轨迹段面积代替点点距离 图2.8三角形面积正比于阴影面积



图2.9 使用三角形的高代替三角形面积

(20)

其中，

然后介绍预测距离。给定一组对应点和，时间戳分别为ti和tj，ti和tj不相等，令时间戳大的点等于A，时间戳小的点等于B，如图2.10所示。



图2.10 预测B在与A相同时刻的位置

然后利用轨迹R在ti-1时刻的位置和ti-1与ti间的平均速度，来预测tj时刻的位置，得到，的预测距离计算公式如式（21）所示。

(21)

其中，

那么这两个点的时间距离可以转化为tj时刻，轨迹R的位置到的距离，如图2.10所示。的预测距离计算公式如式（22）所示。

(22)

图2.11中，融合点段距离和预测距离，我们可以得到对应样本点和之间的时空距离公式，如式（23）所示，其中t是时间距离的敏感参数，值越大表示时间距离越重要。

(23)



图2.11 利用夹角计算形状相似性

利用样本点之间的时空距离，可以得到该样本点和后一个样本点形成的轨迹段之间的距离，即段段距离，来计算轨迹段和轨迹之间的形状相似程度，如式（24）所示，其中是与之间的轨迹段，是与之间的轨迹段。

(24)

在判断两条轨迹的相似程度的时候，形状上的相似也十分重要。结合形状因素的段段距离如式（25）所示，其中是两条轨迹段之间的夹角，和的计算公式如式（26）和式（27）所示。

(25)

(26)

(27)

前面得到了结合形状因素的轨迹段之间的时空距离，我们将其当做一个距离计算因子，代替DTW函数中使用的对应点之间距离，可以得到式（28），其中指的是轨迹R的第一个样本点和第二个样本点之间的轨迹段，指的是出掉第一个轨迹段剩下的所有轨迹段。

(28)

该方法主要优点有三个：

（1）改进DTW函数，采用轨迹段到轨迹段的距离来代替点到点的距离计算，可以减少对轨迹采样方法的敏感程度。

（2）利用预测的方法，对于一个点对，预测时间戳靠前的点在下一刻的位置，使得两个点时间戳相同，将时间距离转换为空间距离，考虑到相似性计算中去。

（3）将形状因素加入到相似性计算中，提高形状相似性方面的精度。

### 2.2.8简化的轨迹相似性计算

由于利用轨迹相似性可以进行未来某一时刻的位置预测，Liu等人首先提出了基于社会传染理论的位置预测算法，然后提出了一个简化的轨迹相似性计算方法来找出带预测用户的相似用户组，以此支撑社会传染理论的运行，

根据用户在某地活动消耗的的时间比仅仅路过该地的时间长，并减少相似性计算的复杂度，我们将完整的轨迹分解成小轨迹。分解条件就是用户的时间跨度，是一个时间阈值。分解后得到用户的小轨迹集。我们可以得到所有用户的小轨迹集。

然后两个小轨迹集之间的重合程度来计算他们的相似程度。令代表用户的布尔型向量，的值如式（29）所示。

(29)

当比较用户和用户之间的相似程度时，将和做与运算（），取1的个数作为轨迹之间的相似度值。

这个方法的优点是相似度计算过程简单快速，没有过多地求轨迹之间的空间距离以及考虑时间、速度等因素。缺点就是计算结果可能随阈值设置的好坏而变化，若设置过大，导致小轨迹过长，会让两条相似的轨迹重合的小轨迹变少，从而不能反映真实的相似程度，精度不够。

### 2.2.9基于签名的轨迹相似性计算

当前的轨迹相似性函数很大程度依赖两条轨迹的对应样本点，然后计算对应样本点之间包含各种信息的距离，但是由于采样频率或者物体移动速度不同，样本点很可能不能一一对应，如图2.12所示。因此Ta等人提出了BDS[16]计算两条轨迹之间的相似程度。与找轨迹的对应点的方法不同，BDS在计算轨迹Ti和Tj的相似度时，通过累加Ti的每个样本点到Tj的最短距离，如图2.13所示。

图2.12按时间匹配点对 图2.13按点到轨迹段最小距离匹配

这个最短距离的定义是将Tj上的所有样本点按照时间顺序连线，形成Tj-1条轨迹段，Ti上的第k个样本点到这Tj-1条轨迹段的最短距离就是Ti上样本点到Tj的距离，其计算公式如式（30）所示。

(30)

而Ti上样本点到轨迹段的最短距离分为两种情况，如果点到轨迹段的垂线与轨迹段相交，距离就是垂线的长度，否则就是样本点到轨迹段里自己最近的端点的距离，如图2.14所示。



图2.14点到轨迹段的距离

BDS的计算公式如式（31）所示，其中是一个归一化的距离，最终结果是一个介于0到1之间的值，表示轨迹和轨迹之间的相似程度，值越大，表示两条轨迹越相似。

(31)

其中，

该相似性函数的计算量很大，对于两条轨迹，要计算出每一个样本点到另一条轨迹所有段的最短距离。为了弥补这个缺点，使用该方法之前需要利用网格作为轨迹签名，先进行一次筛选，然后在计算轨迹上第k个点到轨迹的距离时，只需要计算点以该点为中心的一定范围内的网格中存在的轨迹段的距离，极大程度地减少了计算次数。

这个相似性计算方法方法的优点是不用硬性地将两条轨迹中的点进行配对，一个样本点的对应点可能在另一条轨迹段中两个样本点之间，将两条相似的轨迹更好地进行吻合。缺点就是由于仅考虑了两条轨迹的空间位置，没有考虑其他信息，比如时间和速度信息，因此仅适用于比较两条道路的相似性，不能完整地反映移动对象的详细信息。

## 2.3 高效的轨迹索引技术

### 2.3.1 R树索引

R树是一种针对空间数据建立的树形索引，其核心思想是将每片需要建立索引的区域使用一个外接矩形框起来，作为树的叶子节点，即叶子节点包含着指向该片区域的指针。然后在建树的过程中，矩形之间允许多层嵌套，每个矩形都是一个节点，外层的矩形是被嵌套的矩形的父节点，最终构造成一颗高度平衡树。解决了传统数据库对空间数据检索效率低下的问题。

由于R树需要构建成一颗高度的平衡树，所以有一个构建规则。假设M是树中所有节点允许包含的最大的条目数，m是节点包含的最小条目数，通常情况下m=M/2，此时R树的构建规则如下：

（1）所有叶子节点包含m至M个记录索引。如果一个叶子节点是树的根节点，那么它包含的记录数可以小于m。

（2）叶子中存储的指针指向的记录，这些记录均为可以覆盖其代表空间中区域的最小外包矩形。

（3）一个非根且非叶子的节点，其包含的孩子数目在m到M之间。

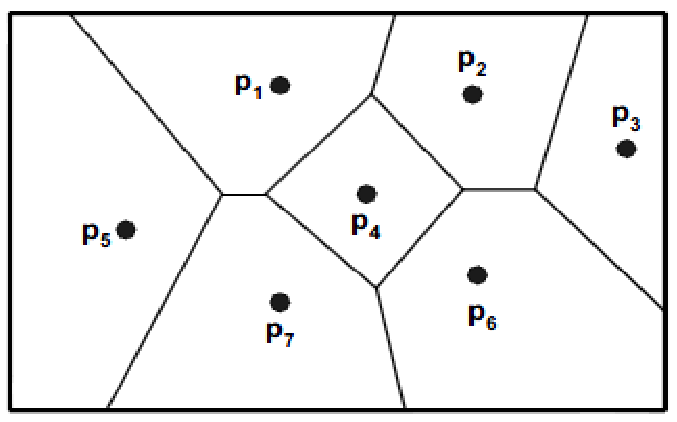
（4）对于一个非叶子节点，节点中的记录是节点所有孩子节点代表区域的最小外包矩形。

（5）R树的所有叶子节点均在同一层。

由于以上特性，相比传统数据库，R树可以很快检索空间数据。但是R树允许外包范围重叠，所以使得一个区域可以对应多条搜索路径，严重影响了R树性能，使空间搜索效率降低。因此R+树和R\*树和三维R树被提出来解决搜索区域的重叠问题。

### 2.3.2 Voronoi图

Voronoi图根据空间中的点集的分布，将一个欧式空间划分为多个凸多边形，凸多边形的每一条边都是一个点与周围点的垂直平分线，该点叫图多边形的质心。每个凸多边形都有一个质心，任意一个凸多边形内的点到其他凸多边形质心的欧式距离都小于到本凸多边形质心的欧氏距离，如图所示。后来有学者对Voronoi图进行了进一步研究，根据发生点和生成面的不同，提出了广义Voronoi图，其中包括有序k阶Voronoi图，最远点Voronoi图，权重Voronoi图等。Voronoi图作为空间数据索引中的一类重要索引结构，解决了很多传统数据库不能解决的问题。



## 2.4本章小结

本章主要介绍了不同的轨迹的表示方法以及为解决不同问题而提出来的一些轨迹相似性计算方法。有些相似性使用范围较为广泛，比如DTW、LCSS和EDR，但是这些算法不针对某个具体场景或具体问题，并没有研究具体情况下该引入哪些特征去描述轨迹或者该怎么去优化轨迹间的相似性表示。因此现在轨迹相似性的研究更加偏向于研究某个具体场景下的轨迹相似性，比如城市运输系统中基于段的相似性查询，和路网上的轨迹相似性查询等。此外，本章还介绍了加速空间数据搜索的几个应用比较广泛的索引结构，

# 第三章 标准空间下距离的算法

## 3.1 数值不均衡问题的提出

### 3.1.1 轨迹的表示形式

### 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点和线段的时空距离定义

### 3.2.1点距离的定义

由于定义了标准空间，统一了时间维度和空间维度，允许了时间维度和空间维度进行运算，所以点与点之间的时空距离计算在这里变得很方便，标准空间下所有维度的计量单位都是相同的，可以直接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。样本点和之间距离的计算公式如式所示。其中和是时间维度转换过来的数值，在三维的时空标准空间中使用了与三维空间欧氏距离计算相同格式的公式，体现了标准空间对时间和空间的融合。

### 3.2.2 标准空间下的轨迹段的距离

#### 3.2.2.1 利用改进余弦距离计算轨迹段的形状距离

#### 3.2.2.2 轨迹段的位置距离

# 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 对应样本点的获取

4.3.2

# 第五章 实验设计与分析

# 第六章 总结与展望