以往的相似性算法大多使用的查询轨迹Q和数据轨迹T中对应采样点的距离来计算得到轨迹距离，经典的算法如DTW、EDR、ERP算法等。这种方法存在的缺陷就是，采样策略的细微变化会导致最终相似性结果的变化。同样一个对象相同方式移动产生的轨迹，其采样策略的变化，会导致与数据轨迹T产生成倍增长的距离。比如以5s为采样频率产生Q1，以10s为频率产生Q2，数据轨迹为T，那么Q1的采样点个数是Q2的两倍，会导致基于对应采样点距离的相似度计算方法产生的轨迹距离相差很大，若Q1与T的距离为d（Q1，T），那么d（Q2，T）约为2\* d（Q1，T）。

移动对象轨迹数据的分析挖掘需要关心轨迹间的距离，比如空间距离、时间距离等，还要关心轨迹的形状是否相似。最后数据分析的质量好坏与取决于距离的计算函数和形状相似的度量方法。

问题：找出给定数据轨迹T中，和查询轨迹Q最相似的轨迹段。（假设T比Q长）

**第三章 利用正态空间解决多维空间下数值不均衡问题**

**问题定义：**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。

**前人工作：**2014年VLDB的一篇文章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。









**存在的问题：**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。

**数值不均衡问题：**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。

**正态空间：**为了解决上述问题，我使用统计学中的Standardization方法将所有维度分别进行标准化处理，构造一个正态空间。构造方法是首先获取两条轨迹所有样本点Q={q1, q2, q3…,qm}，R={r1,r2,r3…rn}。其中样本点格式为（，，），计算三个维度各自的均值和方差，，，，，。然后针对所有点pi使用下面公式做标准化。

得到正态空间下的轨迹点集合，。其中每个点=(,,)。

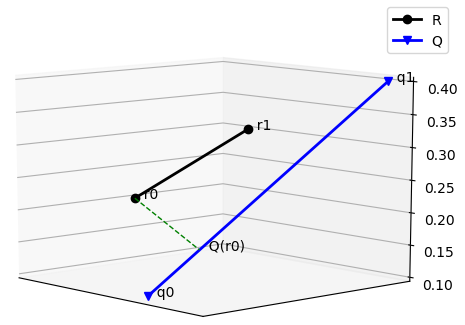
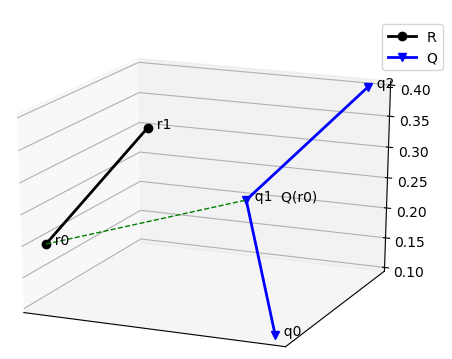
**对应样本点：**

我们计轨迹Q上的样本点qi在轨迹R上的对应样本点为R（q）。这里有可能会出现R（qi）和R（q（i+1））是同一个样本点的情况，即qi和qi+1对应到了轨迹R上的同一个样本点。

## 3.3对应轨迹段相似性

在介绍对应轨迹段的相似性之前，需要介绍到对应点和对应轨迹段的概念。

轨迹上的对应点指的是使用一个样本点对齐算法，将数据轨迹R上的所有样本点对齐到查询轨迹Q上去，那么R上的某个样本点会获得轨迹Q上的一个对应点，我们将在轨迹Q上的对应点记作，如图所示，的对应点是轨迹Q上的样本点。这里需要明确的是，并不一定是数据轨迹Q上的一个样本点。因为考虑减少到采样策略的不同对轨迹相似性计算造成的影响，本文中的样本点对齐算法的设计将有可能使对齐到查询轨迹Q两个样本点之间的连线上，如图所示，的对应点在轨迹Q的样本点和之间。这里需要用到的样本点对齐算法将在后文中进行介绍。



在获得对应样本点的基础上，我们可以得到对应轨迹段的概念。数据轨迹R上的连续两个样本点和，会在查询轨迹Q上分别获得他们的对应点和，那么轨迹段的对应轨迹段就是。

对应轨迹段可能会出现多种情况，我们对每种情况的处理方式不尽相同。第一种情况是和之间的轨迹段是一条直线段。其中和有可能是轨迹Q的样本点，也有可能是轨迹Q中某两个样本点连线上的点。这个情况在讲对应点已经讨论过了，如图所示，的对应点是轨迹Q的样本点，的对应点在样本点和之间。

第二种情况是由于数据轨迹Q移动速度和移动距离等因素，和可能会对齐到Q上的同一个点，即和有可能是轨迹Q中的同一个样本点，如图所示，和的对应点和为同一个点，并且该点还是轨迹Q的样本点。

第三种情况是虽然和是两个连续的样本点，但是和中间可能隔着多个样本点，如图所示，和的对应点和中间夹着轨迹Q的样本点和。







由于后面计算轨迹相似性的需要，本文提出标准空间下的对应轨迹段相似性的概念。在标准空间中，数据轨迹段和对应的查询轨迹段之间的相似性由两部分组成，第一部分是其形状上的相似性，第二部分是轨迹段之间标准空间距离。下面依次对这两部分进行讲解。

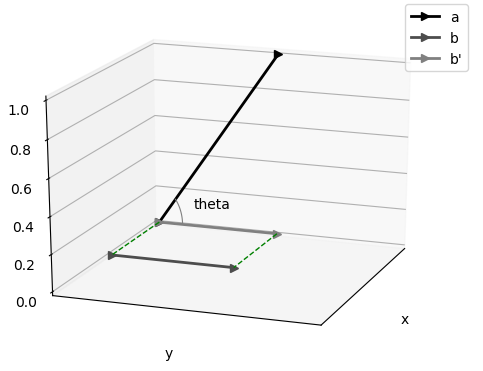
### 3.3.1轨迹段形状相似性

轨迹段形状相似性和轨迹段之间的夹角以及轨迹段的长度相关，轨迹段之间夹角越小的情况下，两条轨迹段长度越长，那么轨迹段形状越相似。下面将从夹角和轨迹段长度两方面讨论轨迹段形状上的相似性。

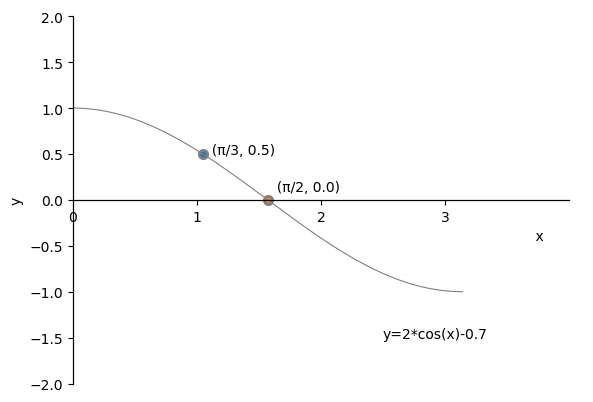
#### 3.3.1.1改进余弦距离

在我们建立的三维标准空间中，两条轨迹段可能存在异面的情况，如图所示。我们在这一小节采用向量的方式去讨论轨迹段的夹角，那么这里需要利用三维空间中的向量去求两个异面线段的夹角的余弦距离。三维空间中向量和向量的余弦距离如公式所示。在轨迹段余弦距离中，的取值范围为[0, ]，的取值范围为[-1,1]。

，其中。



但是使用余弦距离衡量一条轨迹段与其对应轨迹段之间的形状相似程度有两个问题，第一个问题是余弦距离对角度差介于和之间的轨迹段不敏感，即不能有效地对较大的锐角的轨迹段做出惩罚，如图所示，因为在轨迹的相似性里面，如果两条轨迹朝着一个方向走的时候，夹角过大，会导致轨迹形状的不相似，同时物体距离会随着时间推移越来越大，这是需要抑制的一种情况。第二个问题是两条轨迹段的相似程度不仅和角度相关，还和轨迹段长度相关。当两条轨迹段夹角为0时，在一定范围内，轨迹段长度越长，表示这两条轨迹段越相似。当两条轨迹段夹角为时，轨迹段长度越长，表示两条轨迹段不相似的程度越高。

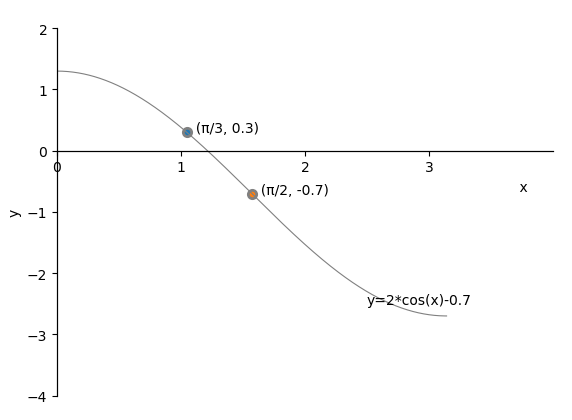


为了解决对较大锐角不能做出惩罚的问题，可以将余弦距离减去一个惩罚系数，的取值范围为，如公式所示，使得两向量夹角为较大锐角的情况下，他们的余弦距离为负值，达到了对大锐角惩罚的目的。但是这个方法一个隐藏的问题，因为时，所以的取值范围为[-1, 1]，则的取值范围为，在取0的时候，原本余弦距离为1，现在为，大于0，所以小于1。因此这个方法不但加大了对大锐角的惩罚，同时也减少了对小锐角的激励，存在一定的缺陷。

，其中

基于以上讨论，为了既可以加大对大锐角的惩罚程度，同时不降低对小锐角的激励程度，本文提出了改进余弦距离，如公式所示。改进余弦距离通过前面的倍数2加大了对小锐角情况的激励，同时使用惩罚系数也加大了大锐角以及钝角情况的惩罚。假设取，从改进余弦距离的函数图像中，我们可以看出当夹角在时，改进余弦距离为正数，当夹角在时，改进余弦距离为负数，且改进余弦距离在定义域内会随着角度的增大而减小，是一个单调递减函数，值域为[-1.7, 1.3]。

，其中



#### 3.3.3轨迹段形状相似距离

上一节介绍的改进余弦距离可以用来将对应轨迹段之间的夹角的相似程度，但是两条轨迹段的形状相似除了与夹角有关，还和轨迹段长度相关。

我们首先仍然假设两条轨迹段均为直线段，即如图所示的第一种情况。因此在获得可以获得对应轨迹段之间的夹角和轨迹段的长度后，我们得到了轨迹段与其对应的直线段的轨迹段的形状相似距离，如公式所示。

在轨迹段之间夹角很小的情况下，在一定程度上，数据轨迹段的长度越长，代表两条轨迹越相似。但是如果超过了这个程度，即比的长度大很多倍时，对应轨迹段之间的相似性如果再随着轨迹段长度的增长而增加，就不符合我们对轨迹形状上的相似性的要求了，因为在轨迹段形状相似性的要求是，在小锐角的情况下，两条轨迹段的长度越长越相似。因此不能仅根据一条轨迹段的长度边长去延伸。也就是说对数据轨迹段长度的激励需要有一个限制。

而依据余弦距离的几何意义，为映射到方向上的距离，这个距离带有方向性，如果夹角为钝角，距离就是负数。那么如果将余弦距离替换为改进余弦距离，也是映射的一种方式，那么的值就是由映射到方向后的值，使用改进余弦距离投影后，如果距离特别长而对应轨迹段的距离特别短，投影结果不能反映轨迹段之间长度的差异，为了抑制这种情况，在轨迹段形状相似距离计算时，我们将的长度规定为轨迹段形状相似距离的上限，即形状相似距离允许的最大激励不得超过轨迹段的长度，再大的部分我们认为是无效部分。

基于以上讨论，我们给出与对应轨迹段的轨迹段形状相似距离的计算公式，如公式所示。由公式可以看出，轨迹段之间夹角为一个小锐角时，两条轨迹段长度越长，二者相似距离越大。

其中，

上面讨论了轨迹段的对应轨迹段是直线段的情况，我们还需要对另外两种情况进行讨论。

当出现第二种情况，即和是轨迹Q中的同一个样本点时，由于此时为0，所以这种情况对应的轨迹段形状相似距离为0。

当出现第三种情况，即和轨迹段中间经过一个或多个样本点，此时不是一条直线段，不能使用上面的公式直接进行计算。

#### //fixme: 形状相似性与论文 A segment based trajectory similarity measure in the urban transportation system 中的形状相似性参数做对比。

在这篇论文中，也使用了一个描述形状相似的函数作为轨迹段距离的参数，该函数可以描述对应轨迹段之间形状上的相似程度，如公式所示，是两条对应轨迹段中间夹的锐角。该函数与本论文中提到的形状相似性距离最本质的区别就是的计算不会考虑两条轨迹段的方向是否同向，因为只获取两条轨迹段所夹的锐角，所以只要两条有方向的轨迹段夹角越接近0或越接近，那么作为轨迹段距离系数的就越小，最终计算出的轨迹间的距离越小，轨迹越相似。

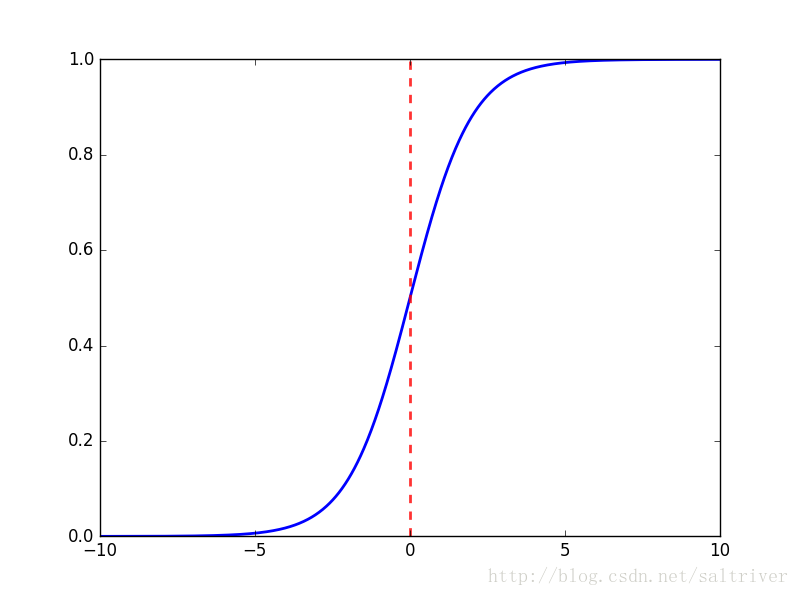
#### // fixme : 完成此处的逻辑解释

而本片论文为了获取最相似的轨迹段，需要

#### 3.3.4轨迹段形状相似性因子

轨迹段形状相似性距离描述了二者形状上的相似程度，由数据轨迹段和其对应轨迹段夹角以及二者长度决定，值域为全体实数，为了得到轨迹段形状相似性距离对轨迹段整体的影响程度，需要使用一个函数将轨迹段形状相似距离转化为一个影响因子。

数学中有一个函数叫sigmoid函数，它的数学公式如公式所示。它是一个S形的函数，定义域为，值域为(0,1)，当x趋近于负无穷时，y趋近于0，当x趋近于正无穷时，y趋近于1，x等于0时，y等于0.5，并且该函数是一个中心对称函数，对称中心为(0,0.5)，是一个良好的阈值函数，可以将全体实数映射到0和1之间。



我们在这里采用sigmoid函数作为将轨迹段形状相思距离映射为形状相似因子的函数，下面给出形状相似性因子的公式，如公式所示。

其中，是轨迹段和其对应轨迹段的形状相似性距离。

因为刚才讨论的仅仅是向量之间余弦距离的情况，没有对轨迹段之间的余弦距离做出讨论，由于对应轨迹段会有三种空间分布，所以需要分三种情况来计算对应轨迹段的余弦距离。因为第一种情况中两条轨迹段均为直线段，所以可以将轨迹段看做向量，直接使用余弦距离。

针对第二种空间分布，即

#### // fixme:还没有说完几个特殊情况

#### 3.3.5对应轨迹段的空间距离

对应轨迹段间的空间距离是轨迹距离的重要组成部分，对应轨迹段在空间上距离越远，那么对应的轨迹段相似程度越小，越近相似程度越大。这里空间距离指的是标准空间距离，包括已经被标准化的时间维度距离。最终轨迹之间的距离由对应轨迹段间的距离构成。我们记和到其对应点距离的平均值作为轨迹段之间的距离，计算方法如公式所示。

#### 3.3.6对应轨迹段相似性

对应轨迹段的相似性由轨迹段形状相似性和轨迹段的空间距离决定