

# 时空轨迹聚类方法研究进展

龚 玺<sup>1,2</sup>, 裴 韬<sup>1</sup>, 孙 嘉<sup>2,3</sup>, 罗 明<sup>4</sup>

(1. 中国科学院地理科学与资源研究所 资源与环境信息系统国家重点实验室, 北京 100101; 2. 中国科学院烟台海岸带研究所, 烟台 264003; 3. 中国科学院研究生院, 北京 100049; 4. 香港中文大学地理与资源管理系, 香港)

**摘 要:** 时空轨迹(Trajectory)是移动对象的位置和时间的记录序列。作为一种重要的时空对象数据类型和信息源,时空轨迹的应用范围涵盖了人类行为、交通物流、应急疏散管理、动物习性和市场营销等诸多方面。通过对各种时空轨迹数据进行聚类分析,可以提取时空轨迹数据中的相似性与异常特征,并有助于发现其中有意义的模式。本文根据时空轨迹数据的特点,系统综述了时空轨迹聚类方法的研究进展。首先,从理论、可行性和应用的角度分析了时空轨迹数据及其聚类方法研究的重要性,并论述了时空轨迹的定义、模型与表达;然后,按照相似性度量所涉及的不同时间区间将现有的时空轨迹聚类方法划分为6类,并对每一类方法的原理及特点进行了评述;最后,讨论了现有方法面临的主要问题和挑战,并对时空轨迹聚类研究的发展进行了展望。

**关 键 词:** 时空轨迹;时空数据挖掘;聚类;相似性度量;研究进展

## 1 引言

传统的GIS研究中,人们常常只关注于某一时刻对地理空间中的属性与空间信息的分析,这实际上只是描述了研究对象的一个快照,没有对连续的时态数据作专门处理,但时间、空间和属性作为地理实体及地理现象本身固有的3个基本特征,是反映地理实体的状态和演变过程重要组成部分<sup>[1]</sup>。随着卫星定位技术、无线通信、跟踪检测设备及视频实时采集技术的快速发展,人们能够方便地以低廉的价格获得时空轨迹数据。例如,通过传感器遥测野生动物或者鱼类的活动,通过旅行日志记录交通工具的运动状况,通过条形码的检入检出了解物流的状况,通过信用卡刷卡记录或者电话通话记录来跟踪用户的位置,甚至通过互联网搜索某对象的相关事件来确定该对象的运动轨迹等。空间对象的位置、属性都可能随着时间的推移而发生变化,人们不仅需要知道某一对象的属性和空间信息,更要了解该对象的来龙去脉,以便对其形成原因作出评估,对未来情况进行预测。时空轨迹数据恰能有效地表达时空对象的这些特性,通过分析各种不同对象的时空轨迹数据,有助于对人类行为模式、交通

物流、应急疏散管理、动物习性、市场营销、计算几何以及模拟仿真等各个领域进行研究。综上所述,无论从理论、可行性还是应用的角度来看,时空轨迹数据的研究都非常必要。

为了能够从大量时空轨迹数据中发现有趣的、隐藏的、未知的知识,需要使用空间数据挖掘作为分析方法。空间数据挖掘为研究者们提供了很多有效的数据分析工具<sup>[2]</sup>。在数据驱动的空间数据挖掘方法中,聚类分析和关联规则挖掘是两种重要的手段,其区别在于关联规则挖掘是一个异中求同的过程,而聚类分析则是同中求异的过程。通过聚类能够识别对象空间中稠密和稀疏的区域,将数据中的相似性与异常特征提取出来,从而发现全局分布模式和数据属性之间有趣的相关<sup>[3]</sup>。这正符合人们对时空轨迹数据分析的要求,即在没有任何先验知识的情况下,先将数据聚合成不同的类,再对各类所代表的模式进行解读从而获得知识。

本文根据时空轨迹数据的特点,系统综述了时空轨迹聚类方法的研究进展。首先,主要阐述时空轨迹数据的定义、模型及其表达;然后,分类介绍了各种时空轨迹聚类方法的原理并对其特点进行分析和评述;最后,讨论了现有方法面临的主要问题

收稿日期:2010-10; 修订日期:2011-02.

基金项目:中国科学院知识创新工程重要方向项目(KZCX2-YW-QN303);中国科学院地理资源所自主部署创新项目(200905004);863项目(2009AA12Z227)。

作者简介:龚玺(1986-),男,硕士研究生,主要研究方向为空间数据挖掘。E-mail: gongx@lreis.ac.cn

通讯作者:裴韬(1972-),男,副研究员,主要从事空间数据挖掘和空间信息统计等的研究。E-mail: peit@lreis.ac.cn

和挑战,并展望了轨迹聚类研究的发展趋势。

## 2 时空轨迹数据

时空轨迹(Trajectory)数据具有与其他数据不同的重要特征,主要体现在定义、模型和表达3个方面。它既是一种重要的时空对象数据类型,又是一种重要的信息源,因此其应用范围也非常广泛。

### 2.1 时空轨迹的定义

时空轨迹是移动对象的位置和时间的记录序列<sup>[4]</sup>。抽象地来看,如式(1)所示,时空轨迹是时间到空间的映射,由一个以时间为自变量的连续函数 $o$ 表示的,当给定某一个时刻 $t(t \in R^+)$ 时,通过该函数可以得到 $t$ 时刻该对象所处的 $d$ 维空间 $R^d$ (一般是二维或者三维空间)中的位置。 $o: R^+ \rightarrow R^d$  (1)

### 2.2 时空轨迹的模型

从定义中我们可以看出,时空轨迹是连续的,但通常用一组时空记录点序列,以离散的方式表示。例如,对时空对象的实际轨迹曲线进行采样,用得到的集合来代表时空轨迹<sup>[5]</sup>。因此,时空轨迹的模型如式(2)所示:

$$T = \{(x_1^1, \dots, x_1^d, t_1), (x_2^1, \dots, x_2^d, t_2), \dots, (x_n^1, \dots, x_n^d, t_n)\} \quad (2)$$

式中:  $T$  代表一条轨迹,序列中每一个 $(d+1)$ 元组 $(x_n^1, \dots, x_n^d, t_n)$ 代表轨迹对象 $t_n$ 时刻在 $d$ 维空间中的一个记录点,其空间位置是 $(x_n^1, \dots, x_n^d)$ (例如,二维空间位置通常以 $(x_n, y_n)$ 表示,三维空间位置则通常以 $(x_n, y_n, z_n)$ 表示)。

### 2.3 时空轨迹数据的表达

为了对时空轨迹进行比较,常常需要通过其模型重构时空轨迹,这就是时空轨迹数据的表达。轨迹表达的方法有很多种,本节将结合Nanni对轨迹重构方法的分类方式<sup>[5]</sup>,按照对轨迹记录点间对象运动过程的不同认识,分3部分阐述时空轨迹数据的表达。

#### 2.3.1 基于全局回归模型的时空轨迹数据表达

如果时空对象的运动方式整体上服从某一规则,那么可对该对象的所有记录点进行全局回归,用关于时间 $t$ 的回归方程代表时空对象的轨迹<sup>[5]</sup>。如图1所示,黑点和白点分别代表两条不同轨迹的记录点,两条直线是采用线性回归所得到的轨迹。由于这种模型过于简化,重构的时空轨迹也不与所

有采样点重合,往往不能满足实际的需要。

#### 2.3.2 基于局部插值模型的时空轨迹数据表达

有时时空对象的运动方式并非全局一致,但可以假设在相邻记录点间的局部运动是服从特定规则的,不同的规则可以用不同的局部插值方法来表达。最常见的规则是相邻记录点间对象作匀速直线运动,该规则可以用线性插值方法表达(图2a),这种模型在时空轨迹模拟<sup>[6-9]</sup>和分析<sup>[10-12]</sup>中均被广泛使用,并且可以采用时空路径(Space-time Path)的方式来可视化表达<sup>[13]</sup>(图2b)。这种表达方式将二维的空间和一维的时间整合到一个三维坐标系中表示,每个记录点的 $x, y$ 坐标对应记录点的空间坐标,第三维坐标则对应记录点的时间值,图中实线表示的是时空路径,虚线为时空路径在空间维上的投影。

#### 2.3.3 基于领域知识模型的时空轨迹数据表达

如果没有内插函数作为重构轨迹的依据,那么在任意相邻的记录时刻间,时空对象理论上可能在空间中的任何位置出现,但多数情况下各种领域知识会限制该对象出现的位置<sup>[13]</sup>。例如,由于存在移动速度的限制,在某个记录时刻后,该时空对象只能存在于以该记录点为顶点的一个圆锥体内<sup>[14]</sup>(图3a);或者由于道路的限制,对象只能沿交通网络运动<sup>[15-16]</sup>;或者用户在运动过程中需要使用信息通讯技术,故受到网络覆盖区域的限制<sup>[17-18]</sup>等等。这些情况下,时空棱镜(Space-time Prism)是一种很好的可视化表达方式<sup>[14]</sup>(图3b),两相邻记录点的空间位置分别是 $L_1$ 和 $L_2$ ,记录时间分别为 $t_1$ 和 $t_2$ ,坐标表示方法与时空路径相同,记录点间的棱镜部分表示对象可能出现的时空范围,而该棱镜在空间维平面上的投影则表示对象的潜在活动区域(Potential Activity Area)。

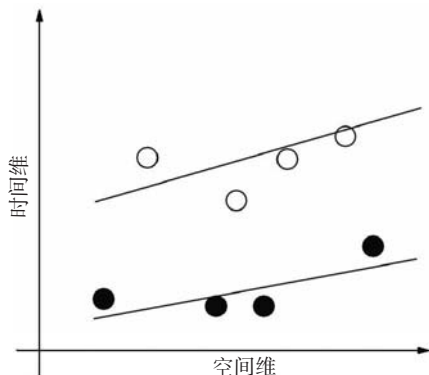


图1 基于全局回归模型的时空轨迹数据表达<sup>[5]</sup>

Fig.1 Trajectory data expression based on global regression model<sup>[5]</sup>

2.4 时空轨迹数据研究的应用范围

时空轨迹数据在许多应用领域中都是重要的研究对象,除了时间地理学(Time Geography)中人的活动轨迹之外<sup>[13]</sup>,还有生物学领域中精子运动的轨迹<sup>[19]</sup>和人的笔迹轨迹<sup>[20]</sup>、化学领域中分子的运动轨迹<sup>[21]</sup>、气象学领域中气团<sup>[22]</sup>和飓风的轨迹<sup>[23-24]</sup>以及体育领域中球员的运动轨迹<sup>[25-26]</sup>等。

3 时空轨迹聚类方法

为了从时空轨迹数据中提取其相似性与异常,并发现其中有意义的模式,时空轨迹聚类分析方法被广泛采用。该方法的主要目的是试图将具有相似行为的时空对象划分到一起,而将具有相异行为的时空对象划分开来。其关键是根据时空轨迹数据的特点,设计与定义不同轨迹间的相似性度量。因为要将数据集划分成不同的类别,必须定义一种

相似性的测度来度量同一类样本间的类似性和非同类样本间的差异性<sup>[27]</sup>,而各种时空轨迹聚类方法间的主要区别也正是在于其相似性度量的不同。

两个对象之间的相似度(Similarity)是这两个对象相似程度的数值度量,相异度(Dissimilarity)是这两个对象差异程度的数值度量,距离(Distance)常被看作是相异度的同义词<sup>[28]</sup>。因而,两个对象越类似,它们的相似度就越高,相异度就越低,距离越小。通常,相似度的取值范围是 $[0,1]$ (0代表完全不相似,1代表完全相似),而相异度(距离)的取值范围是 $[0,\infty)$ (0代表完全相似, $\infty$ 代表完全不相似),它们通常是可以互相转化的,所以使用“相似性度量”作为相似度和相异度(距离)的统称。

依照相似性度量所涉及的不同时间区间,可将现有的时空轨迹聚类方法划分为6类(表1)。从表1第二列的相似时间区间示意图可以看出,这6类方法对于相似时间区间的要求是逐渐放松的,从要求

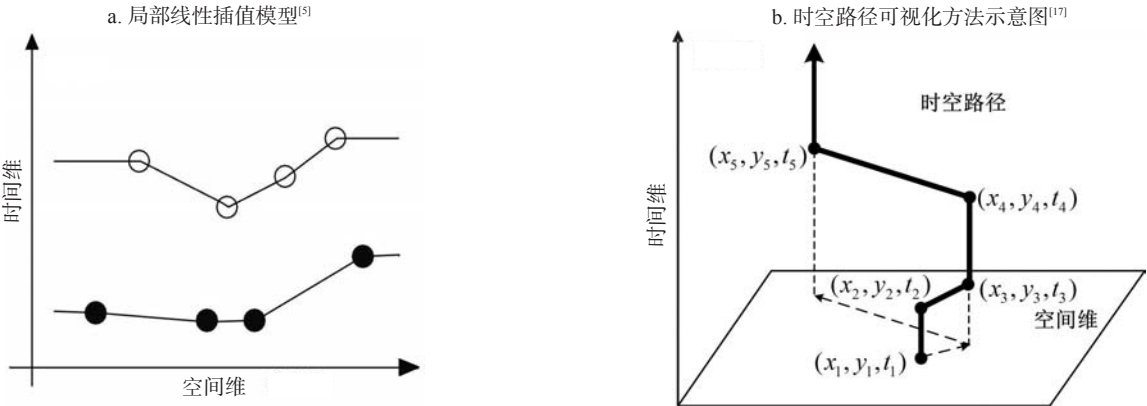


图2 基于局部插值模型的时空轨迹数据表达  
Fig.2 Trajectory data expression based on local interpolation model

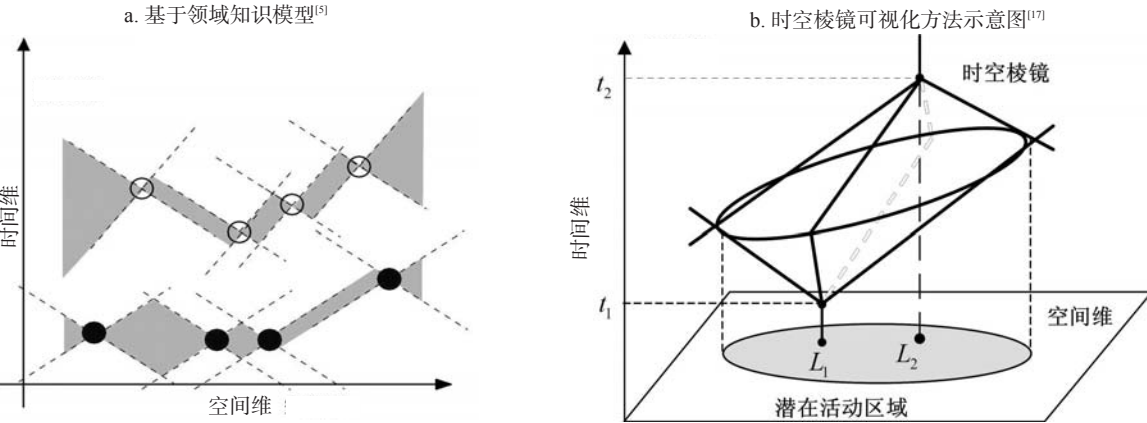


图3 基于领域知识模型的时空轨迹  
Fig.3 Trajectory data expression based on domain knowledge-based model

时间全区间相似,到局部时间区间相似,最后到无时间区间对应相似。这种分类方式既能体现人们对时空轨迹相似性认知的多样性,又能反映时空轨迹相似性度量的发展过程。下面本文将对各类时空轨迹聚类方法进行介绍和评述。

3.1 时间全区间相似的聚类方法

时间全区间相似的聚类方法将时空轨迹看作一个整体,并要求同一聚类中的轨迹在各个时刻都对应相似。这类方法所使用的相似性度量主要有轨迹间欧氏距离和最小外包矩形距离等。

3.1.1 轨迹间欧氏距离

轨迹间欧氏距离和点与点的欧氏距离有所不同。它首先将轨迹用相同维度的坐标向量表示,然后计算每一个时刻上对应两点的欧式距离,再对这些距离进行综合(如求和,求平均值、最大值或者最小值),就可以得到轨迹间欧式距离。例如,式(3)就是二维空间中,以求和方式综合的轨迹间欧式距离公式<sup>[29-30]</sup>:

$$Eu(R,S)=\sum_{i=1}^n dist(r_i,s_i)$$
$$dist(r_i,s_i)=\sqrt{(r_{i,x}-s_{i,x})^2+(r_{i,y}-s_{i,y})^2}$$

(3)

式中:  $R$ 、 $S$  分别表示两条轨迹,记录点数均为  $n$ ;  $Eu(R,S)$  为轨迹  $R$ 、 $S$  间的欧式距离;  $r_i$ 、 $s_i$  分别表示轨迹  $R$ 、 $S$  上第  $i$  个记录点;  $r_{i,x}$ 、 $r_{i,y}$ 、 $s_{i,x}$ 、 $s_{i,y}$  分别表示记录点  $r_i$ 、 $s_i$  的  $x$  坐标和  $y$  坐标;  $dist(r_i,s_i)$  表示记录点  $r_i$  和  $s_i$  间的欧式距离。

Agrawal 等在 1993 年就提出了该方法并用于解决序列的相似性问题<sup>[29]</sup>。为了提高这种方法的

效率,Faloutsos 等和 Chan 提出了一些通过离散傅里叶变换(Discrete Fourier Transform, DFT)和离散小波变换(Discrete Wavelet Transform, DWT)来降维的近似办法<sup>[45-46]</sup>。Chakrabarti 也提出了一种名为 APCA(Adaptive Piecewise Constant Approximation) 的近似方法<sup>[47]</sup>。但是这些方法都不能应用于采样率不同或者尺度不同的轨迹数据。Yanagisawa 等先将轨迹分段线性表示,然后内插重采样,再计算轨迹间欧氏距离,这样处理能够将采样率不同的轨迹数据进行比较<sup>[48]</sup>。而 Keogh 等则认为可以先对轨迹进行全局缩放再计算轨迹间欧式距离,该方法能够有效解决尺度不同的问题<sup>[49]</sup>。但是由于轨迹间欧氏距离的基本思想是严格计算轨迹在每个时刻的对应距离,因此这类方法对噪声较敏感。

3.1.2 最小外包矩形距离

该方法可以看作一种简化时空轨迹的方法。它首先将整条轨迹划分成一些相对平滑的轨迹区间,再将每条子轨迹用其最小外包矩形(Minimum Boundary Rectangle, MBR)表示,这样每条轨迹就变成了一个最小外包矩形的序列<sup>[31]</sup>(图 4),图中虚线矩形框和实线矩形框分别代表虚线轨迹和实线轨迹的最小外包矩形序列,通过比较最小外包矩形序列即可度量时空轨迹间的相似性。

Lee 等定义了最小外包矩形间的距离计算规则,并将各对外包矩形间的距离加权平均作为整体轨迹间的距离<sup>[31]</sup>。而 Elnekave 等则将最小外包矩形重叠部分的大小作为整条轨迹相似性度量<sup>[32]</sup>。由于使用最小外包矩形代替了轨迹区间,这种方法平滑了轨迹的细节,并在一定程度上缓解了噪声的影响,但是如何有效地将轨迹划分成平滑轨迹区间仍有待研究。这类时间全区间相似聚类方法的优点在于非常直观,易于理解,但那些不在一一对应时刻上完全相似的轨迹,则可能被遗漏。

3.2 全区间变换对应相似的聚类方法

该类方法在全区间相似聚类方法的基础上,放松了对时间维的限制,即时空轨迹的时间维可以局

表 1 时空轨迹聚类方法分类

Tab.1 Methods of the trajectory clustering		
相似性度量类别	相似时间区间示意图	代表聚类方法
时间全区间相似		轨迹间欧氏距离 <sup>[29]</sup>
		最小外包矩形距离 <sup>[30-32]</sup>
全区间变换对应相似		DTW <sup>[33]</sup>
多子区间对应相似		最长公共子序列距离 <sup>[34]</sup>
		编辑距离 <sup>[35]</sup>
单子区间对应相似		子轨迹聚类 <sup>[36]</sup>
		时间聚焦聚类 <sup>[37]</sup>
		移动微聚类 <sup>[38]</sup>
		移动聚类 <sup>[39]</sup>
单点对应相似		历史最近距离 <sup>[40]</sup>
		Fréchet 距离 <sup>[40]</sup>
无时间区间对应相似	无	单向距离 <sup>[41]</sup>
		特征提取方法 <sup>[42-44]</sup>

注:相似时间区间示意图中的实线部分为相似区间,虚线部分为不相似区间。

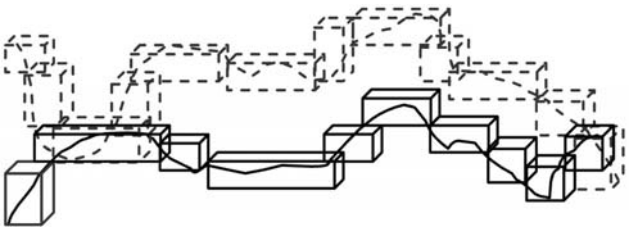


图 4 时空轨迹最小外包矩形<sup>[32]</sup>

Fig.4 Minimum boundary rectangle of trajectories<sup>[32]</sup>

部拉伸和缩放,只需要保证轨迹记录点的时间顺序,而不需要在——对应的时刻上进行比较,其中基于DTW(Dynamic Time Warping)距离的方法就是典型代表。

$$DTW(R,S)=\begin{cases} 0 & \text{if } m=n=0 \\ \infty & \text{if } m=0 \text{ or } n=0 \\ dist(r_1,s_1)+\min\begin{cases} DTW(Res\ t(R),Res\ t(S)) \\ DTW(Res\ t(R),S) \\ DTW(R,Res\ t(S)) \end{cases} & \text{o t h e r w i s e} \end{cases}$$

(4)

式中:  $DTW(R,S)$  表示时空轨迹  $R$  与  $S$  间的  $DTW$  距离;  $m$  和  $n$  分别代表时空轨迹  $R$  与  $S$  的记录点个数;  $dist(r_i,s_i)$  表示两个记录点  $r_i$  和  $s_i$  之间的欧式距离;  $Res\ t(R)$  和  $Res\ t(S)$  分别表示轨迹  $R$  与  $S$  去掉第一个记录点所得的轨迹区间,其他各项意义与前述相同。从式(4)可以看出:如果两条轨迹都无记录点,那么  $DTW$  距离为0;如果只有一条轨迹无记录点,则  $DTW$  距离为无穷大;如果两条轨迹均存在记录点,则采用递归的方式求取最小的距离作为  $DTW$  距离,在求取最小距离的过程中会产生记录点的最优对应关系。如图 5a 所示,某些点(如  $r_i$ 、 $s_n$ )在计算  $DTW$  距离时多次使用,实际上是对时间维的局部拉伸;图 5b 是记录点对应关系的矩阵表示,黑色方块表示的是最优对应关系形成的  $DTW$  路径,其中每个方块的权值为相应记录点间的欧式距离, $DTW$  距离就是整条路径的权值之和。

Sankoff 和 Kruskal 最早使用 DTW 来度量不等长序列的相似性<sup>[33]</sup>,但该方法存在计算量过大的问题<sup>[50]</sup>,而通过建立索引则能提高计算效率<sup>[51-52]</sup>。也有学者在该方法基础上做出一些其他的改进,例如, Little 和 Gu 先用路径和速度曲线来表示轨迹,再用 DTW 度量距离<sup>[53]</sup>;而 Vlachos 等则将轨迹引入极坐标空间,通过角度与长度来表示轨迹,再计算轨迹间的 DTW 距离<sup>[54]</sup>。

DTW 方法可较好地发现时间维局部缩放后才相似的时空轨迹,解决了采样率不同和时间尺度不一的问题。但计算 DTW 距离时,轨迹间的记录点映射需要具有连续性,因此对于噪声很敏感。此外,如果两条轨迹在小部分区间内完全不相似,该方法将无法识别。

3.3 多子区间对应相似的聚类方法

为解决上一类方法无法识别小部分不相似区间的问题,多子区间对应相似的聚类方法在定义相

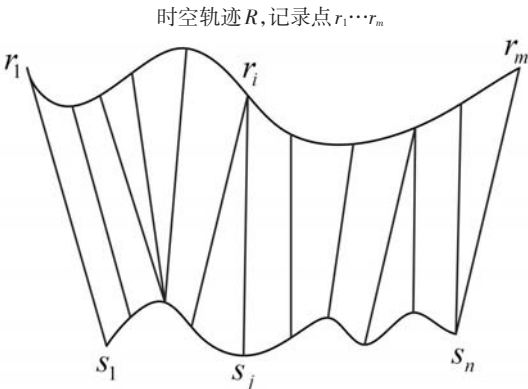
基于DTW距离的方法在保证时空轨迹对象记录点顺序不变的前提下,通过重复之前的记录点来完成时间维的局部缩放,以此求出轨迹间的最小距离作为相似性度量。具体计算公式为<sup>[30,33]</sup>:

似性度量时,不要求整条轨迹相似,而是寻找不重叠的多个相似子区间,并将所有子区间的相似性汇总成轨迹间的相似性度量。其中最长公共子序列距离和编辑距离是比较常见的方法。

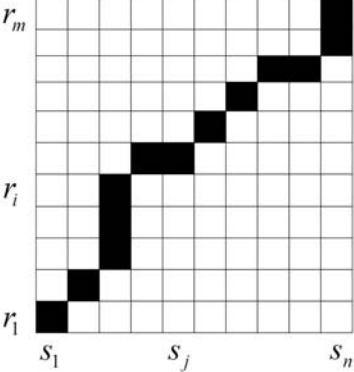
3.3.1 最长公共子序列距离

最长公共子序列 (Longest common sub-sequence, LCSS)是指两个或者多个序列中存在的最长的共同子序列。对于时空轨迹来说,计算其最长公共子序列并转化为 LCSS 距离可以衡量轨迹间的

a. 轨迹间计算DTW距离时的记录点的对应关系



b. 记录点的对应关系映射图



注: 黑色方块代表的是最优的DTW路径。

图5 DTW距离示意图<sup>[51]</sup>

Fig.5 Illustration image of DTW distance<sup>[51]</sup>

相似程度<sup>[30,34]</sup>。LCSS的计算一般通过递归方式求

$$LCSS(R,S)=\begin{cases} 0 & \text{if } m=n=0 \\ LCSS(Res t(R), Res t(S))+1 & \text{if } |r_{1,x}-s_{1,x}|\leq\delta \text{ and } |r_{1,y}-s_{1,y}|\leq\varepsilon \\ \max\{LCSS(Res t(R),S), LCSS(R, Res t(S))\} & \text{o t h e r w i s e} \end{cases} \quad (5)$$

式中:  $LCSS(R,S)$  表示时空轨迹  $R$  与  $S$  间的  $LCSS$  长度,  $\delta$  和  $\varepsilon$  分别表示  $x$  轴和  $y$  轴上的相似阈值, 也就是说, 当横坐标差小于  $\delta$  且纵坐标差小于  $\varepsilon$  时, 认为这对记录点相似,  $LCSS$  值加1, 其他各项意义与前述相同。当轨迹记录点数都为0时,  $LCSS(R,S)$  为0; 若记录点个数不为0, 则用递归的方式判断共有子序列长度的最大值。如图6所示, 实线和虚线分别为一维空间中的两条时空轨迹, 横轴为时间, 纵轴为一维空间坐标, 区间1、2、3内是  $LCSS$  公式所定义的共有子序列, 两条轨迹在这3个子区间内是对应相似的。

在应用中, 通常使用轨迹间的距离作为相似性度量, 因此研究者们将  $LCSS$  转换为距离的形式来进行聚类, 其转换方式是<sup>[52]</sup>:

$$D_{LCSS}(R,S)=1-\frac{LCSS(R,S)}{\min(m,n)} \quad (6)$$

式中:  $D_{LCSS}(R,S)$  表示时空轨迹  $R$  与  $S$  间的  $LCSS$  距离;  $\min(m,n)$  表示  $R$  与  $S$  的记录点个数的较小值。这个过程使得轨迹间的  $LCSS$  转换为  $[0,1]$  间的距离。Agrawal 最早用这种方法来计算一维时间序列的相似性, 他认为如果两个时间序列的  $LCSS$  超过某一阈值, 则可以认为它们相似<sup>[34]</sup>。该方法可以很方便地应用于高维时间序列和时空轨迹的相

得, 如式(5)所示:

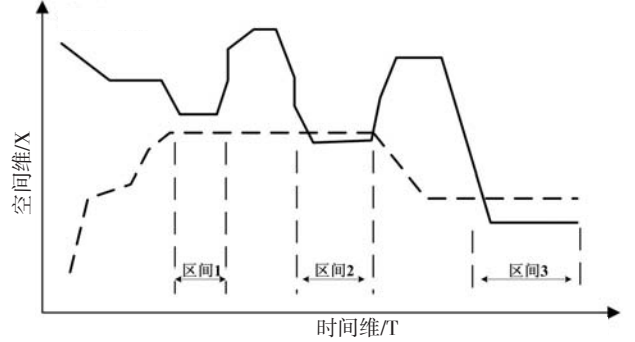


图6 LCSS示意图

Fig.6 Illustration image of LCSS

似性度量中, 例如 Vlachos 就提出可以先对轨迹进行平移变换再使用  $LCSS$  方法<sup>[55]</sup>。

$LCSS$  方法不需要所有的记录点全部匹配, 因此不相似的区间会被剔除。此外, 因为点与点的距离被概化为0和1, 所以即使噪声点参与到  $LCSS$  的计算中, 其影响也会被减弱。

### 3.3.2 编辑距离

编辑距离(Edit Distance)是指两个序列(文本或者模式等)进行比较时, 若只进行增、删、改操作, 一个序列完全变成另一个序列所需最小操作次数<sup>[35]</sup>, 也可以很容易地将其扩展为时空轨迹间的编辑距离, 即:

$$ED(R,S)=\begin{cases} n & \text{if } m=0 \\ m & \text{if } n=0 \\ ED(Res t(R), Res t(S)) & \text{if } m,n>0 \text{ and } r_1 \text{ is equal to } s_1 \\ \min\begin{cases} ED(Res t(R), Res t(S))+1 \\ ED(Res t(R), S)+1 \\ ED(R, Res t(S))+1 \end{cases} & \text{o t h e r w i s e} \end{cases} \quad (7)$$

式中:  $ED(R,S)$  表示轨迹  $R$  和  $S$  的记录点序列  $(r_1, \dots, r_m)$  和  $(s_1, \dots, s_n)$  间的编辑距离,  $m$  和  $n$  分别代表时空轨迹  $R$  与  $S$  的记录点个数, 其他各项意义与前述相同。如果其中一条轨迹的记录点个数为0时, 编辑距离为另一条轨迹的记录点个数; 如果两条轨迹均存在记录点, 且首个记录点坐标相同, 则编辑距离不变; 否则编辑距离增加, 并采用递归的

方式求取最小值作为编辑距离。例如, 两条一维时空轨迹的空间坐标序列为  $\{1,3,4,5\}$  和  $\{1,2,3,4,6\}$ , 那么它们的编辑距离就为2, 因为序列  $\{1,3,4,5\}$  经过第二位增加2和最后一位改成6两次操作就可以变成序列  $\{1,2,3,4,6\}$ 。

在编辑距离的计算过程中, 要求首个记录点坐标相同的判断条件往往过于严格, 影响了计算效率

和精度,因此很多学者对其进行了改进。Bozkaya 提出的改进方法是:不要求序列在增、删、改操作后完全相同,而只要在一定阈值内相似即可<sup>[56]</sup>。Chen 和 Ng 则提出 ERP(Edit distance with Real Penalty)距离并用于度量时间序列的相似性<sup>[57]</sup>。ERP 与 DTW 类似,也可对时间维进行局部缩放,因此能够处理不同尺度的数据;但是 ERP 用真实值来度量距离,而不是将距离概化为 0 和 1,所以该方法同样对噪声敏感。此外,Chen 等提出 EDR(Edit Distance on Real sequence)距离也是编辑距离的扩展<sup>[30]</sup>,该距离通过正态化处理解决了空间维缩放的问题;与 LC-SS 和 EDR 相比,EDR 不仅对噪声不敏感,而且对于轨迹间不同尺寸的缺口(即两段相似区间中的不相似区间)指定不同的惩罚距离,这使得该度量方法更加精确。

最长公共子序列距离和编辑距离都考虑的是多个子区间对应相似的情况,这一大类方法能发现非整体相似的时空轨迹,但是所发现的相似时间区间是离散且不确定的,这种区间在数学上比较清晰,但是并不直观,不易被人们观察和理解。

3.4 单子区间对应相似的聚类方法

为解决多子区间对应相似聚类方法中相似区间不直观的问题,单子区间对应相似的聚类方法在其基础上,对时间区间的要求进一步放松:只需获得一个最大的相似子区间,就能衡量轨迹间的相似性。这类方法主要有子轨迹聚类、时间聚焦聚类、移动微聚类和移动聚类。

3.4.1 子轨迹聚类

子轨迹聚类方法由 Lee 等在 2007 年提出的,它采用的是先划分再聚合的思路(Partition-and-group Framework)<sup>[36]</sup>,其流程如图 7a 所示:首先将时空轨迹看作一组点序列,然后按照最小描述长度(Minimum Description Length, MDL)原则<sup>[36]</sup>将轨迹划分为一些子轨迹,再用基于密度的聚类方法对这些子轨迹聚类,最终可以得到子轨迹的运动模式和整条轨迹的相似子区间<sup>[36,58-59]</sup>。

该方法的相似性度量由 3 种距离的加权和表示,分别是其垂直距离  $d_{\perp}$ 、平行距离  $d_{\parallel}$  和角度距离  $d_{\theta}$ 。轨迹  $L_i$  和  $L_j$  间的 3 种距离如图 7b 所示,其中:  $s_i$ 、 $s_j$ 、 $e_i$ 、 $e_j$  分别代表轨迹  $L_i$  和  $L_j$  的起点和终点;  $p_s$  和  $p_e$  分别表示  $s_j$  和  $e_j$  在轨迹  $L_i$  上的投影;  $l_{\perp 1}$ 、 $l_{\perp 2}$ 、 $l_{\parallel 1}$ 、 $l_{\parallel 2}$  则分别表示图中对应端点间的欧

氏距离,  $\|L_j\|$  表示轨迹  $L_j$  的长度;  $\theta$  是两条子轨迹的夹角 ( $0^{\circ} \leq \theta \leq 180^{\circ}$ )。

虽然子轨迹聚类方法能发现具有相似性的单个最大时间区间,但是由于该方法预先将轨迹划分成子轨迹,并以子轨迹为基本单位进行聚类,因此相似时间区间会受到子轨迹时间区间的限制,具有一定的局限性。

3.4.2 时间聚焦聚类

时间聚焦聚类(Time-focused Clustering)方法可以较好地解决子轨迹聚类存在的问题。该方法先定义了一个聚类过程,该过程是将某一时间区间内轨迹间的欧氏距离作为相似性度量,并采用基于密度的聚类方法 OPTICS<sup>[60]</sup>对轨迹进行聚类;然后对每一个不同的时间区间均进行一次上述聚类过程;最终目标是发现使轨迹聚类结果最优(即类内相似度大、类间相似度小)的时间区间,并记录这个区间和相应的聚类结果<sup>[37]</sup>。如图 8 所示,轨迹集合由三类轨迹添加一些噪声轨迹组成,每条虚线代表一条时空轨迹,实线代表最优轨迹聚类结果,即在实线所示的时间区间内,聚类结果中的各个类别类内相

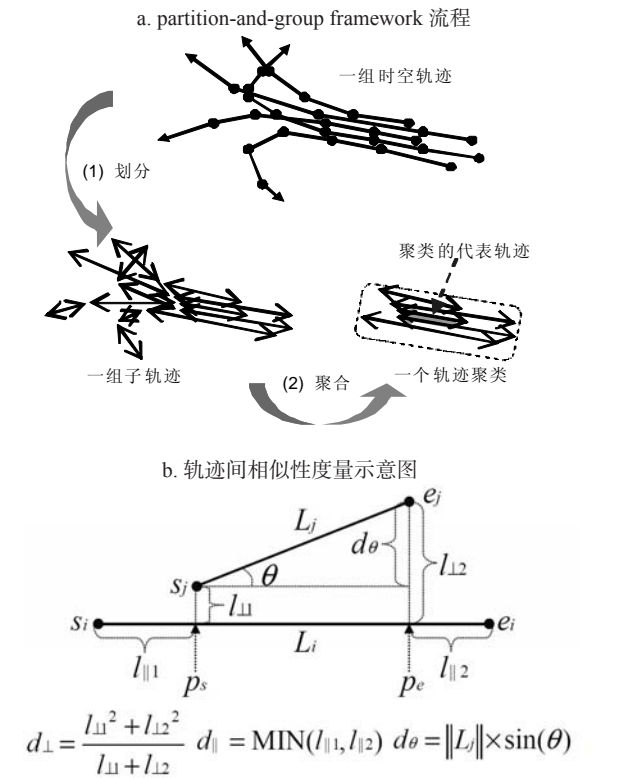


图 7 子轨迹聚类示意图<sup>[36]</sup>

Fig.7 Illustration image of sub-trajectory clustering<sup>[36]</sup>

似度大,类间相似度小,能够被清晰地区分。

### 3.4.3 移动微聚类

移动微聚类(Moving Micro Clustering, MMC)方法将数据挖掘中经典的BIRCH方法<sup>[61]</sup>应用于移动对象轨迹数据,在原有的微聚类(Micro Cluster)基础上增加了时间维信息,也就是说在同一聚类中的对象不仅在当前时刻位置靠近,还需要保持相近(共同运动)一段时间<sup>[12]</sup>。该方法把每个微聚类的中心看作一个对象继续聚类,即不予区分微聚类和对象,最终形成层次结构完成聚类。在这个过程中,移动微聚类需要根据一个预先定义的外包矩形来判断聚类是否应该分裂或合并,但这个矩形的参数需要人为定义,这就是这个方法的不足之处。

### 3.4.4 移动聚类

前面提到的方法中,聚类都可以被认为是“内涵固定”的,即相似时间区间内,类内的个体集合不变。移动聚类(Moving Cluster)改变了这种对聚类的看法,将聚类视为一个动物群落,群落中时有个体迁入,时有个体迁出,但是处于群落中的个体始终保持聚集。这种方法认为移动对象是由多个时间片(Time Slices)上的空间位置组成,首先分别对每个时间片上的点进行聚类,然后计算连续时间片中聚类所包含点的重合程度,如果大于一定阈值,那么这个移动聚类成立。如图9所示(图中每条折线代表一条轨迹,而每一个时间片上的聚类用深色圆形表示),这个聚类中成员的生命周期(Lifetime)不需要一致,只需在某一时间段内保持聚集即可成为聚类。这种方法更关注聚类所涉及的区域而非其中的轨迹对象,所以可以认为它是一种介于聚类与频繁模式挖掘之间的方法<sup>[19, 38]</sup>。

## 3.5 单点对应相似的聚类方法

这类方法是将轨迹间的相似性概括为某一对记录点间的相似性,即以一个时间点来代表整个时间区间。其中历史最近距离和Fréchet距离是最主要的两种方法。

### 3.5.1 历史最近距离

历史最近距离是任意两条轨迹在给定时间范围内同一时刻的最近距离<sup>[39]</sup>,其计算如式(8)所示:

$$\text{MinDist}(R, S, T) = \min\{\text{dist}(R(t), S(t)) | \forall t \in T\} \quad (8)$$

式中:  $\text{MinDist}(R, S, T)$  表示时空轨迹  $R$  与  $S$  在时间区间  $T$  内的历史最近距离,  $R(t)$  和  $S(t)$  分别表示轨

迹  $R$  与  $S$  在  $t$  时刻的空间位置,  $\text{dist}(R(t), S(t))$  表示  $R(t)$  和  $S(t)$  间的欧氏距离。如图10所示,时空轨迹  $R$  和  $S$  的历史最近距离为  $d_{\min}$ 。

### 3.5.2 Fréchet距离

Fréchet距离可以这样形象地理解:在遛狗过程中,假设狗沿一条轨迹连续运动,它的主人沿另一条轨迹连续运动,他们在各自轨迹上任意一点速度都可以变化,甚至可以停止,但是不能折返,用遛狗绳将他们相连,Fréchet距离就是在保证两者运动不

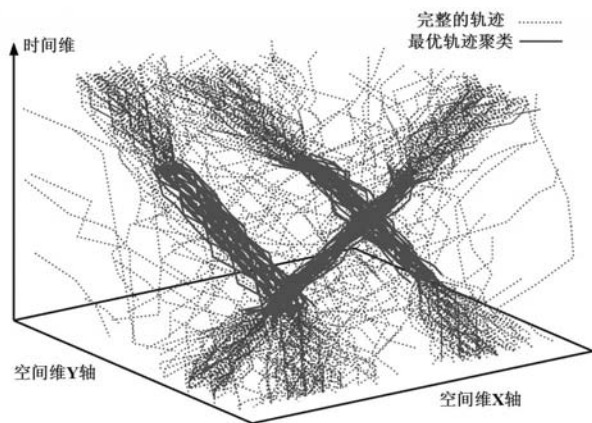


图8 Time-focused 聚类方法示例<sup>[37]</sup>

Fig.8 Example of time-focused clustering over a time interval<sup>[37]</sup>

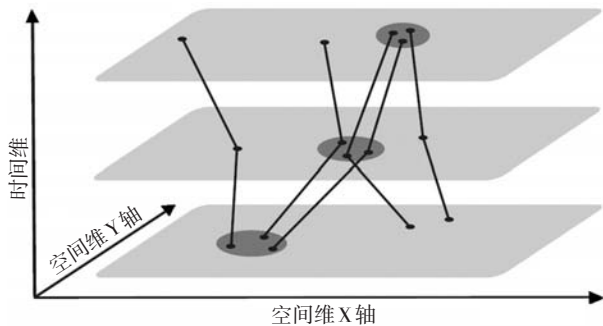


图9 移动聚类示例<sup>[62]</sup>

Fig.9 Example of moving cluster<sup>[62]</sup>

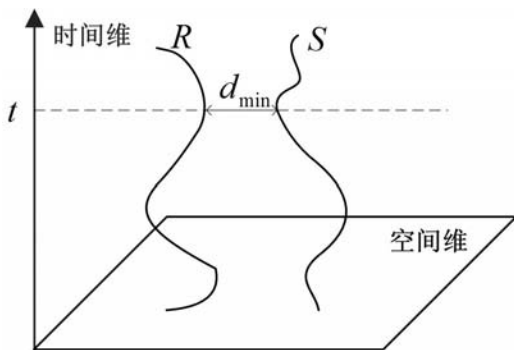


图10 历史最近距离示意图

Fig.10 Illustration image of the nearest historical distance

互相干扰的前提下,所需最短的遛狗绳的距离。Fréchet距离原本是用来度量两条曲线间距离的,由于时空轨迹可以看作时空路径曲线,所以Fréchet距离也可以用来度量轨迹间的相似性<sup>[40,63]</sup>。

历史最近距离和Fréchet距离这两种方法,都是将轨迹间的距离抽象为某一个时刻上点与点的距离,以此来代表整条轨迹间的距离。不同之处在于,历史最近距离是一种乐观的相似性度量,只要两条轨迹在某一时刻曾经靠近过就认为它们相似,而Fréchet距离是一种悲观的相似性度量,它认为必须在每一时刻都靠近才能说两条轨迹相似。

3.6 无时间区间对应相似的聚类方法

无时间区间对应相似的聚类方法在度量轨迹间相似性时,将时间维的限制进一步放宽,通常只考虑空间位置的相似性,或者将时空轨迹转换到属性空间进行比较,比如单向距离方法和特征提取方法等。

3.6.1 单向距离

单向距离(One-way Distance, OWD)是一种将时间相似性弱化的相似性度量方法。该方法源自巴士线路设计问题,在这类问题中,速度和方向信息不重要,时间顺序的意义也不大,只有空间形状的相似性比较重要<sup>[41]</sup>,为了定义OWD,首先需要定义点  $p$  到轨迹  $T$  的距离:

$$D_{point}(p,T)=\min_{q\in T} D_{Euclid}(p,q) \tag{9}$$

式中:点  $q$  为轨迹  $T$  上一点,  $D_{Euclid}(p,q)$  表示点  $p$  和点  $q$  间的欧式距离。而轨迹  $T_1$  到  $T_2$  的OWD定义为  $T_1$  上的点到  $T_2$  被  $T_1$  所截取部分的距离积分:

$$D_{owd}(T_1,T_2)=\frac{1}{|T_1|}\left(\int_{p\in T_1} D_{point}(p,T_2)dp\right) \tag{10}$$

由式(10)可以看出,OWD描述的是一种形状上的相似性,并且是一条轨迹到另一条轨迹的有向距离,这种距离不具有对称性。通常,将  $D_{owd}(T_1,T_2)$  和  $D_{owd}(T_2,T_1)$  的均值作为轨迹  $T_1$  和  $T_2$  间的距离:

$$D(T_1,T_2)=\frac{1}{2}(D_{owd}(T_1,T_2)+D_{owd}(T_2,T_1)) \tag{11}$$

例如,两条一维轨迹的空间坐标序列为  $T_1:\{1,2,3,4\}$  和  $T_2:\{8,6,4,2\}$ ,那么  $T_1$  上每个点到  $T_2$  的距离分别为1、0、1、0,  $T_1$  上每个点到  $T_2$  的距离分别为4、2、0、0,那么轨迹  $T_1$  到  $T_2$  的OWD为0.5,轨迹  $T_2$  到  $T_1$  的OWD距离为1.5,  $T_1$  与  $T_2$  间的距离则为1。单向距离在定义轨迹间相似性时,尽管同样

是抽象成点与点的距离,但是不考虑时间上的顺序关系,这点与历史最近距离和Fréchet距离不同。

3.6.2 特征提取方法

该方法不对轨迹本身直接进行比较,而是先从轨迹中提取特征,再通过特征来定义相似性度量。例如,Perng等<sup>[42]</sup>和Faloutsos等<sup>[43]</sup>分别从轨迹中提取Landmark和Signature特征,然后对这些特征定义运算规则来计算轨迹间的距离;Pelekis则将轨迹分别抽象为速度特征和方向特征,并定义了对应的相似性度量进行聚类<sup>[44]</sup>。

3.7 其他聚类方法

除了以上6类方法外,还有一些方法也可以用来定义轨迹间的相似性度量并用于聚类分析,如人机交互方法、基于模型的方法、递增层次方法、基于图形的方法等。其中,人机交互方法是根据用户提出的一些限制来调整参数获得距离度量函数<sup>[64-65]</sup>;基于模型的方法尝试针对轨迹数据特点建模,然后将从同一模型构造出来的对象聚成一类<sup>[66-68]</sup>;递增层次方法认为轨迹是由不同元件组成的序列,不同轨迹按泛化程度形成层次结构从而完成聚类<sup>[69]</sup>;基于图形的方法则是将轨迹曲线看作一些线状图形的组合,依此为轨迹编码,通过这种编码可以有效地查询指定形状的数据,从而将形状相似的轨迹聚成一类<sup>[70-71]</sup>。

4 结语与展望

随着数据获取手段的快速发展,人们不仅仅对事物在某一时刻的空间特性感兴趣,更渴望了解事物随时间的发展演变过程,时空轨迹恰能有效地记录和表达这一时间与空间相结合的过程。而时空轨迹聚类方法则是发现其中所蕴涵知识的重要手段,时空轨迹聚类方法的建立与完善不仅可以拓展空间数据挖掘理论,而且具有广泛的应用前景。时空轨迹聚类方法将点状对象的相似性度量扩展到线状对象的比较中,在传统数据挖掘方法中融入了对时间语义的不同认识,通过对时空轨迹进行聚类分析,能使一些原本只能用于可视化或定性观察的轨迹数据转换为多类定量的时空对象的行为模式,进而结合对象属性转换成知识,帮助人们更好地理解个体在时间与空间维度上的特性,并对各类对象

的行为作出相应的预测,从而指导人们的生产生活。例如,通过罪犯的行动轨迹了解其作案模式,以便进行预警和抓捕;通过台风的轨迹认识其形成和运动模式,以便对人群进行疏散;通过人口迁移轨迹数据发现其中不同人群的迁移习惯,以便制定相关政策,帮助人们搬迁定居等。

有关时空轨迹聚类方法的研究在国际上起步不久,却已经成为相关领域研究的热点之一,并取得了一定的研究进展。尽管如此,时空轨迹聚类方法仍面临许多困难与挑战,有待进一步研究和解决,主要体现在以下5个方面:

(1) 当前大部分时空轨迹数据聚类方法仍然是将时间看作原空间对象的附加维,这种处理方式难免使时间与空间有所分隔,与人们对于事物的直观认识有出入;

(2) 现有聚类方法对于某些轨迹数据类型并不完全适用,例如人口迁移轨迹数据,该类数据的时间维是不等长的,但是由于具有年龄的语义,时间维不能拉伸,这种情况目前还没有很好的方法可以处理;

(3) 聚类结果在转换成知识的过程中存在一些问題,例如所发现的知识或者过于简单,近乎于常识,或者过于复杂,让人们无法直观理解;

(4) 海量的轨迹数据的不断产生,一方面为研究者提供了丰富的数据源,但另一方面,也要求研究者从中选择有效的数据并提高算法效率;

(5) 在处理与人有关的时空轨迹时,如何保护对象的隐私等也成为研究者应当考虑的问题。

## 参考文献

- [1] 王家耀,魏海平,成毅,等. 时空GIS的研究与进展. 海洋测绘, 2004, 24(5): 1-4.
- [2] Han J, Kamber M, Tung A K H. Spatial clustering methods in data mining//Miller H J, Han J. Geographic Data Mining and Knowledge Discovery. London: Taylor & Francis, 2001: 188 - 217.
- [3] Han J, Kamber M. Data Mining: Concepts and Techniques. 2<sup>nd</sup> ed. San Francisco: Morgan Kaufmann, 2006: 383.
- [4] Li X, Han J, Kim S, et al. Roam: Rule- and motif-based anomaly detection in massive moving object data sets//Proceedings of the Seventh SIAM International Conference on Data Mining. Philadelphia: SIAM, 2007.
- [5] Nanni M. Clustering methods for spatio-temporal data. Pi-

- sa, Italy: University of Pisa[D], 2002.
- [6] Theodoridis Y, Silva J R O, Nascimento M A. On the generation of spatiotemporal datasets. Advances in Spatial Databases, 1999, 1651/1999: 147-164.
- [7] Giannotti F, Mazzoni A, Puntoni S, et al. Synthetic generation of cellular network positioning data//Proceedings of the 13th annual ACM International Workshop on Geographic Information Systems. New York, NY, USA: ACM, 2005: 12-20.
- [8] Tzouramanis T, Vassilakopoulos M, Manolopoulos Y. On the generation of time-evolving regional data. Geoinformatica, 2002, 6(3): 207-231.
- [9] Saglio J M, Moreira J. Oporto: A realistic scenario generator for moving objects. Geoinformatica, 2001, 5(1): 71-93.
- [10] Saltenis S, Jensen C S, Leutenegger S T, et al. Indexing the positions of continuously moving objects. Sigmod Record, 2000, 29(2): 331-342.
- [11] Tao Y, Papadias D. Time-parameterized queries in spatio-temporal databases//Proceedings of the 2002 ACM SIGMOD International Conference on Management of data. New York, NY, USA: ACM, 2002: 334-345.
- [12] Li Y, Han J, Yang J. Clustering moving objects//Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2004: 617-622.
- [13] Hägerstrand T. What about people in regional science? Papers in Regional Science, 1970, 24(1): 6-21.
- [14] Lenntorp B. Paths in space-time environments: A time-geographic study of movement possibilities of individuals. Environment and Planning A, 1977, 9: 961-972.
- [15] Miller H J. Modelling accessibility using space-time prism concepts within geographical information systems. International Journal of Geographical Information Systems, 1991, 5(3): 287-301.
- [16] Kwan M, Hong X. Network-based constraints-oriented choice set formation using GIS. Geographical Systems, 1998, 5: 139-162.
- [17] Yu H, Shaw S L. Exploring potential human activities in physical and virtual spaces: A spatio-temporal GIS approach. International Journal of Geographical Information Science, 2008, 22(4): 409-430.
- [18] Shaw S L, Yu H B. A GIS-based time-geographic approach of studying individual activities and interactions in a hybrid physical-virtual space. Journal of Transport Geography, 2009, 17(2): 141-149.
- [19] Berezensky M, Greenspan H, Cohen-Or D, et al. Segmentation and tracking of human sperm cells using spa-

- tio-temporal representation and clustering. *Medical Imaging 2007: Image Processing*, 2007, 6512: 65122M.1-65122M.12.
- [20] Erez K, Goldberger J, Sosnik R, et al. Analyzing movement trajectories using a markov bi-clustering method. *Journal of Computational Neuroscience*, 2009, 27(3): 543-552.
- [21] Gabarro-Arpa J, Revilla R. Clustering of a molecular dynamics trajectory with a hamming distance. *Computers and Chemistry*, 2000, 24(6): 693-698.
- [22] Cape J N, Methven J, Hudson L E. The use of trajectory cluster analysis to interpret trace gas measurements at Mace Head, Ireland. *Atmospheric Environment*, 2000, 34(22): 3651-3663.
- [23] Camargo S J, Robertson A W, Gaffney S J, et al. Cluster Analysis of typhoon tracks. Part I: General properties. *Journal of Climate*, 2007, 20(14): 3635-3653.
- [24] Camargo S J, Robertson A W, Gaffney S J, et al. Cluster analysis of typhoon tracks. Part II: Large-scale circulation and Enso. *Journal of Climate*, 2007, 20(14): 3654-3676.
- [25] Kang C H, Hwang J R, Li K J. Trajectory analysis for soccer players//*Proceedings of the Sixth IEEE International Conference on Data Mining Workshops*. Washington, DC, USA: IEEE Computer Society, 2006: 377-381.
- [26] Laube P, Imfeld S, Weibel R. Discovering relative motion patterns in groups of moving point objects. *International Journal of Geographical Information Science*, 2005, 19(6): 639-668.
- [27] 蔡元龙. 模式识别. 西安: 西北电讯工程学院出版社, 1986: 18.
- [28] Tan P N, Steinbach M, Kumar V. *Introduction to Data Mining*. Boston: Pearson Addison-Wesley, 2006: 60-65.
- [29] Agrawal R, Faloutsos C, Swami A. Efficient similarity search in sequence databases. *Foundations of Data Organization and Algorithms*, 1993, 730: 69-84.
- [30] Chen L, Özsu M, Oria V. Robust and fast similarity search for moving object trajectories//*Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA: ACM, 2005: 491-502.
- [31] Lee S, Chun S, Kim D, et al. Similarity search for multidimensional data sequences//*Proceedings of the 16th International Conference on Data Engineering*. Washington D. C. USA: IEEE Computer Society, 2000: 599-608.
- [32] Elnekave S, Last M, Maitnon O. Incremental clustering of mobile objects//*Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop*. Washington, DC, USA: IEEE Computer Society, 2007: 585-592.
- [33] Sankoff D, Kruskal J. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. MA, USA: Addison-Wesley, 1983.
- [34] Agrawal R, Lin K I, Sawhney H S, et al. Fast similarity Ssearch in the presence of noise, scaling, and translation in time-series databases//*Proceedings of the 21th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995: 490-501.
- [35] Crochemore M, Rytter W. *Text Algorithms*. New York, NY, USA: Oxford University Press, 1994.
- [36] Lee J G, Han J, Whang K Y. Trajectory clustering: A partition-and-group framework//*Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA: ACM, 2007: 593-604.
- [37] Nanni M, Pedreschi D. Time-focused clustering of trajectories of moving objects. *Journal of Intelligent Information Systems*, 2006, 27(3): 267-289.
- [38] Kalnis P, Mamoulis N, Bakiras S. On discovering moving clusters in spatio-temporal data//*Medeiros C B, Egenhofer M, Bertino E. Proceedings of the 9th International Symposium on Advances in Spatial and Temporal Databases*. Berlin: Springer-Verlag, 2005: 364-381.
- [39] Gao Y, Zheng B, Chen G, et al. Algorithms for constrained K-nearest neighbor queries over moving object trajectories. *Geoinformatica*, 2010, 14(2): 241-276.
- [40] Alt H, Godau M. Computing the fr chet distance between two polygonal curves. *International Journal of Computational Geometry and Applications*, 1995, 5(1): 75-91.
- [41] Lin B, Su J. One way distance: For shape based similarity search of moving object trajectories. *Geoinformatica*, 2008, 12(2): 117-142.
- [42] Perng C S, Wang H, Zhang S R, et al. Landmarks: A new model for similarity-based pattern querying in time series databases//*Proceedings of the 16th International Conference on Data Engineering*, 2000: 33-42.
- [43] Faloutsos C, Jagadish H, Mendelzon A, et al. A signature technique for similarity-based queries//*Proceedings of the Compression and Complexity of Sequences 1997*. Washington, DC, USA: IEEE Computer Society, 1997: 2-20.
- [44] Pelekis N, Kopanakis I, Ntoutsis I, et al. Mining trajectory databases via a suite of distance operators//*Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop*. Washington, DC, USA: IEEE Computer Society, 2007: 575-584.
- [45] Faloutsos C, Ranganathan M, Manolopoulos Y. Fast sub-

- sequence matching in time-series databases. *ACM SIGMOD Record*, 1994, 23(2): 419-429.
- [46] Chan K. Efficient time series matching by wavelets//*Proceedings of the 15th International Conference on Data Engineering*. Washington, DC, USA: IEEE Computer Society, 1999: 126-133.
- [47] Chakrabarti K, Keogh E, Mehrotra S, et al. Locally adaptive dimensionality reduction for indexing large time series databases. *ACM Transactions on Database Systems (TODS)*, 2002, 27(2): 188-228.
- [48] Yanagisawa Y, Akahani J, Satoh T. Shape-based similarity query for trajectory of mobile objects//*Proceedings of the 4th International Conference on Mobile Data Management*. London, UK: Springer-Verlag, 2003: 63-77.
- [49] Keogh E, Palpanas T, Zordan V, et al. Indexing large human-motion databases//*Proceedings of the Thirtieth International Conference on Very Large Data Bases*. VLDB Endowment, 2004: 780-791.
- [50] Berndt D, Clifford J. Using dynamic time warping to find patterns in time series//*Proceedings of KDD-94 Workshop*, 1994: 359-370.
- [51] Sakurai Y, Yoshikawa M, Faloutsos C. FTW: Fast similarity search under the time warping distance//*Proceedings of the Twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. New York, NY, USA: ACM, 2005: 326-337.
- [52] Vlachos M, Hadjieleftheriou M, Gunopulos D, et al. Indexing multi-dimensional time-series with support for multiple distance measures//*Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2003: 216-225.
- [53] Little J, Gu Z. Video retrieval by spatial and temporal structure of trajectories//*Proceedings of SPIE, the International Society for Optical Engineering* SPIE, 2001: 545-552.
- [54] Vlachos M, Gunopulos D, Das G. Rotation invariant distance measures for trajectories//*Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2004: 707-712.
- [55] Vlachos M, Kollios G, Gunopulos D. Discovering similar multidimensional trajectories//Agrawal R, Dittrich K, Ngu A H H. *Proceedings of the 18<sup>th</sup> International Conference on Data Engineering*. Washington, DC, USA: IEEE Computer Society, 2002: 673-684.
- [56] Bozkaya T, Yazdani N, Özsoyoğlu M. Matching and indexing sequences of different lengths//*Proceedings of the Sixth International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 1997: 128-135.
- [57] Chen L, Ng R. On the marriage of lp-norms and edit distance//*Proceedings of the Thirtieth International Conference on Very Large Data Bases*. VLDB Endowment, 2004: 792-803.
- [58] Liu J P, Zhang Y L, Liu G. Partition and density-based clustering for moving objects Trajectories//*Proceedings of the Third International Conference on Computer Science & Education*. Xiamen: Xiamen University Press, 2008: 182-187.
- [59] Lee J G, Han J, Li X, et al. Traclass: Trajectory classification using hierarchical region-based and trajectory-Based clustering//*Proceedings of International Conference on Very Large Data Base(VLDB'08)*. VLDB Endowment, 2008: 1081-1094.
- [60] Ankerst M, Breunig M M, Kriegel H P, et al. Optics: Ordering points to identify the clustering structure//*Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA: ACM, 1999: 49-60.
- [61] Zhang T, Ramakrishnan R, Livny M. Birch: An efficient data clustering method for very large databases. *ACM SIGMOD Record*, 1996, 25(2): 103-114.
- [62] Nanni M, Kuijpers B, Körner C, et al. Spatiotemporal data mining//Giannotti F, Pedreschi D. *Mobility, Data Mining, and Privacy: Geographic Knowledge Discovery*. Berlin: Springer-Verlag, 2008: 267-296.
- [63] Brakatsoulas S, Pfoser D, Salas R, et al. On map-matching vehicle tracking data//*Proceedings of the 31st International Conference on Very Large Data Bases*. VLDB Endowment, 2005: 853-864.
- [64] Schreck T, Bernard J, Tekusova T, et al. Visual cluster analysis of trajectory data with interactive kohonen maps. *Information Visualization*, 2009, 8: 14-29.
- [65] Yu W, Gertz M. Constraint-based learning of distance functions for object trajectories//*Proceedings of the 21st International Conference on Scientific and Statistical Database Management*. Berlin, Heidelberg: Springer-Verlag, 2009: 627-645.
- [66] Gaffney S, Smyth P. Trajectory clustering with mixtures of regression models//*Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 1999: 63-72.
- [67] Chudova D, Gaffney S, Mjolsness E, et al. Translation-invariant mixture models for curve clustering//*Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY,

USA: ACM, 2003: 79-88.

- [68] Alon J, Sclaroff S, Kollios G, et al. Discovering clusters in motion time-series data//Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03). Washington, DC, USA: IEEE Computer Society, 2003: 375-381.
- [69] Ketterlin A. Clustering sequences of complex objects//Proceeding of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD-97). AAAI Press, 1997: 215-218.
- [70] Agrawal R, Psaila G, Wimmers E L, et al. Querying shapes of histories//Proceedings of the 21th International Conference on Very Large Data Bases. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995: 502-514.
- [71] Kim S W, Yoon J, Park S, et al. Shape-based retrieval of similar subsequences in time-series databases//Proceedings of the 2002 ACM Symposium on Applied Computing. New York, NY, USA: ACM, 2002: 438-445.

## Review of the Research Progresses in Trajectory Clustering Methods

GONG Xi<sup>1,2</sup>, PEI Tao<sup>1</sup>, SUN Jia<sup>2,3</sup>, LUO Ming<sup>4</sup>

- (1. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China; 2. Yantai Institute of Coastal Zone Research, CAS, Yantai 264003, Shandong, China; 3. Graduate University of Chinese Academy of Sciences, Beijing 100049, China;
- 4. Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong, China)

**Abstract:** A trajectory is a sequence of the location and timestamp of a moving object. It is not only an important type of spatio-temporal data, but also a critical source of information. Extracting patterns from different trajectory data can help people understand the drives and outcomes of individual and collective spatial dynamics, such as human behavior patterns, transport and logistics, emergency evacuation management, animal behavior, and marketing. Recently, a larger number of trajectory data are available for analyzing the temporal and spatial pattern, as the result of the improvements of tracking facilities and sensor networks. Therefore, clustering analysis needs to be used to find the implicit patterns in it. Based on the characteristics and the similarity measurements of trajectory data, this paper reviewed the research progresses in trajectory clustering methods. Firstly, the significance of research on trajectory data and its clustering methods was presented. Then the definition, models as well as several visualization methods of trajectories were summarized. After that, the authors classified the existing trajectory clustering methods into 6 main categories according to the similarity measurement of them, and analyzed each of the trajectory clustering methods, along with their respective pros and cons by category. Finally, some research challenges and future directions were discussed.

**Key words:** trajectory; spatio-temporal data mining; clustering; similarity measurement; research progress

本文引用格式:

龚玺, 裴韬, 孙嘉, 等. 时空轨迹聚类方法研究进展. 地理科学进展, 2011, 30(5): 522-534.