

MTSAX: 一种新的多元轨迹索引方法

王 飞¹, 庞 悦¹, 周向东¹, 陈海波²

(1. 复旦大学 计算机科学技术学院, 上海 200433; 2. 国网上海市电力公司, 上海 200122)

摘 要: 轨迹数据的分析与查询在移动数据管理、位置服务等领域具有十分重要的应用价值。为提高海量多元轨迹数据的分析和挖掘效率, 提出一种新的多元轨迹数据索引方法 MTSAX, 并给出一种多维空间编码方法: GeoWord 编码, 在 iSAX 索引框架的基础上, 设计移动对象轨迹索引方法。在真实轨迹数据集上的实验结果表明, 与传统基准方法相比, MTSAX 具有更好的轨迹查询性能。

关键词: 轨迹索引; 空间编码; 时间序列; 移动对象; 相似性查询

中文引用格式: 王 飞, 庞 悦, 周向东, 等. MTSAX: 一种新的多元轨迹索引方法[J]. 计算机工程, 2018, 44(5): 1-6.

英文引用格式: WANG Fei, PANG Yue, ZHOU Xiangdong, et al. MTSAX: A Novel Method of Multivariate Trajectory Index[J]. Computer Engineering, 2018, 44(5): 1-6.

MTSAX: A Novel Method of Multivariate Trajectory Index

WANG Fei¹, PANG Yue¹, ZHOU Xiangdong¹, CHEN Haibo²

(1. School of Computer Science, Fudan University, Shanghai 200433, China;

2. State Grid Shanghai Electric Company, Shanghai 200122, China)

[Abstract] The analysis and query of trajectory data have very important application value in the fields of mobile data management, location services and so on. In order to improve the analysis and mining efficiency of massive multivariate trajectory data, a new multiplex trajectory data index method MTSAX is proposed. A multi-dimensional spatial coding method is given: GeoWord coding, based on the iSAX index framework, designing a moving object track index method. Experimental results on real trajectory datasets show that MTSAX can achieve better trajectory query performance compared with the traditional benchmark methods.

[Key words] trajectory index; spatial coding; time series; moving object; similarity query

DOI: 10.19678/j.issn.1000-3428.0046289

0 概述

随着传感器、无线通信网络以及 GPS 定位等技术的飞速发展, 各种基于位置的应用产生了海量轨迹数据, 如导航地图上的出行记录、跑步软件上的历史线路等。轨迹是一种位置相关的时间序列^[1]。由于实际应用中移动对象往往同时包含多种信息, 如速度和运动方向等, 因此多元轨迹数据在导航规划、位置服务等领域^[2-3]具有非常重要的应用价值, 得到了越来越多的研究与关注。

为了提高海量多元轨迹数据的分析和挖掘效率, 轨迹数据的索引技术是关键。学术界对轨迹索引技术进行了大量研究, 以往工作主要在面向点查询和范围查询^[4]的轨迹索引研究^[5-8]。近年来出现的面向相似轨迹查询的方法中, 基于空间划分的索

引^[9]通常对原始空间采取固定划分的方式; 基于特征压缩的索引通常会丢失大量原始空间信息^[10]; 基于度量空间的索引^[11]运用三角不等式等下界过滤技术提高查询效率, 但该类索引仍难以克服高维数据索引面临的“维灾问题”。

本文提出一种新的多元轨迹索引方法 MTSAX。对单个轨迹点设计基于 Geohash 的 GeoWord 编码, 根据经度、纬度、速度等数据的密集程度和数据的重要性进行动态编码, 保留指定精度下的轨迹信息。将每条轨迹分成若干等长子段来压缩原始轨迹, 每个子段采用 GeoWord 编码, 所有子段的 GeoWord 编码组成的符号串代表一条轨迹, 即用 MTSAX 表示。对于 MTSAX 表示的整条轨迹, 设计了基于一元时间序列索引 iSAX 技术的多元轨迹索引结构, 实现一种新的轨迹索引 MTSAX。

基金项目: 国家高科技研究发展计划项目(2015AA050203); 国家自然科学基金(61370157); 上海市科委项目(14511107400); 国家电网公司上海市电力公司科技项目(52094016001Z)。

作者简介: 王 飞(1991—), 男, 硕士研究生, 主研方向为数据索引; 庞 悦, 博士研究生; 周向东, 教授; 陈海波, 高级工程师。

收稿日期: 2017-03-09

修回日期: 2017-05-04

E-mail: 14210240024@fudan.edu.cn

1 相关工作

文献[4]提出轨迹查询的3种类型:点查询,范围查询和轨迹相似查询。面向点查询和范围查询的轨迹索引又可划分为移动物体的历史位置索引^[5]、当前位置索引^[6]、未来位置索引^[7]和全时空位置索引^[8]。轨迹相似查询有多种方式,常见的面向相似轨迹查询的索引方法如下:

1) 基于空间划分的轨迹索引。该方法通常采用划分空间单元格的方式对原始空间进行编码,借助空间编码建立索引。文献[9]提出轨迹索引方法 TRSTJ,首先使用 PAA^[12] 方法将原始轨迹分段表示,然后将 PAA 表示的轨迹空间切分成相同大小的单元格,并为每个单元格分配一个符号,最终一条轨迹被表示成一个字符串。

2) 基于特征压缩的轨迹索引。该方法提取轨迹的特征并编码,借助特征编码建立索引。文献[10]提出一种多元时间序列索引方法,该方法将多元时间序列以多变量求和的方式转化为一元时间序列,使用 PAA 方法把一元时间序列变成 N 维向量,最后使用 R 树来索引该 N 维向量。

3) 基于度量空间的轨迹索引。该方法先选择若干参考点,定义某种距离,再计算所有轨迹相对于参考点的距离,最后在查询时通过这些参考点过滤掉不符合要求的轨迹。文献[11]提出一种基于参考点距离的方法,建立了多元时间序列索引 LBS,在飞行数据集上进行相似性查询。

2 多元轨迹符号化索引 MTSAX

2.1 一元时间序列索引方法 iSAX

文献[12]在分段聚集近似 PAA^[12] 和符号化聚集近似 SAX^[13] 的基础之上,提出一种可索引的符号化聚集近似方法 iSAX^[14],可以索引海量一元时间序列,支持时间序列相似性查询。iSAX 假设原始时间序列服从高斯分布,并将原始时间序列转换成均值为 0 标准差为 1 的标准时间序列。iSAX 采用 PAA 方法将时间序列分成若干等长子段,然后根据正态分布的概率区间将均值表示的序列离散化为符号串,最后将符号化表示的时间序列进行树状索引。iSAX 可以根据时间序列的分布情况采用不同的离散化基数来动态编码。

2.2 多维空间编码 GeoWord

Geohash 是一种二维空间编码,对经纬度交替编码,缺乏灵活性。多元轨迹不仅包含移动物体在运动过程中的经纬度位置,还包含速度、方向等信息;不同变量的重要性一般也不同,例如在轨迹中经纬度位置通常比速度、方向更加重要。因此,本文提出一种基于 Geohash 的多维空间编码 GeoWord,可以对经纬度位置和速度、方向等附加信息同时编码,并考虑了不同变量数据的密集程度及重要性。

GeoWord 算法需要给定单个轨迹点的多元信息和对应的离散化基数。离散化基数用于将 PAA 得到的连续值离散化为指定精度的编码,对于经度、纬度这样的空间位置可以看做是对空间切分的精度。GeoWord 可以采用等切分获得分割点,也可以通过对数据进行采样估计获得分割点。图 1 以等切分为例展示了 GeoWord 编码,初始切分中的 GeoWord 编码为 $0^21^21^2$,其中, 0^2 、 1^2 、 1^2 分别由经度、纬度和速度 3 个变量的编码组成,上标表示该变量的离散化基数。

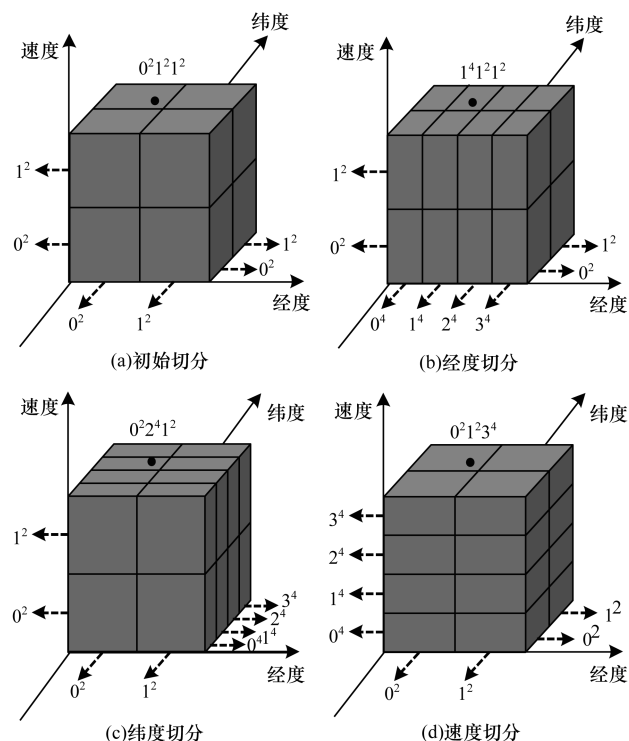


图 1 GeoWord 编码示意图

GeoWord 可以选择不同变量进一步离散化,如图 1 中在初始切分的基础上可以对经度、纬度和速度进行切分。切分变量的选择依赖于不同变量数据的密集程度和变量本身的重要性,切分的目的是为了更好索引轨迹,关于切分变量的选择策略见 2.3 节 MTSAX 索引构建方法中的节点分裂介绍。

GeoWord 编码可以得到单个多元轨迹点在指定精度下的压缩表示,从而得到整条轨迹的压缩表示;对于压缩表示,执行 GeoWord 的逆过程可以得到单个多元轨迹点在指定精度下的信息,从而得到整条轨迹在指定精度下的信息。

算法 1 GeoWord 算法

输入 第 i 个轨迹点 $(p_{i1}, p_{i2}, \dots, p_{ij}, \dots, p_{im})$, 及其对应的离散化基数 $\{a_{i1}, a_{i2}, \dots, a_{ij}, \dots, a_{im}\}$, p_{ij} 表示第 i 个轨迹点的第 j 个变量, a_{ij} 是 p_{ij} 的离散化基数

输出 第 i 个轨迹点 GeoWord 编码

1. $g_i = \text{null}$ // 保存 GeoWord 编码

其中, g_i 为 $(p_{i1}, p_{i2}, \dots, p_{ij}, \dots, p_{im})$ 经过 GeoWord 编码得到的表示, $1 \leq i \leq w$, w 个 GeoWord 编码 $\{g_1, g_2, \dots, g_i, \dots, g_w\}$ 组成一个 MTSAX 表达。

2.4 MTSAX 索引构建方法

MTSAX 索引是树结构,第一层可以看作是叉树,且第一层所有节点基数相同,即对原始空间采用同样的划分精度。从第二层开始,由于叶子节点根据数据的密集程度进行二分裂,则以第一层节点为根节点的子树是二叉树。MTSAX 索引构建因此可以看作是对叉树和二叉树混合在一起的树的构建,MTSAX 索引建立详细过程如算法 2 所示。

算法 2 MTSAX 索引建立

输入 需要插入的多元轨迹 ts

输出 MTSAX 索引插入多元轨迹 ts

1. 获取 ts 的 MTSAX 表示 G
2. if 当前索引节点存在 MTSAX 表示为 G 的后继节点
3. 获取 MTSAX 表示为 G 的后继节点 $node$
4. if $node$ 为叶子节点
5. if $node$ 存储的轨迹数量小于节点分裂阈值 th
6. $node$ 直接插入 ts
7. else//叶子节点索引的轨迹已满,需要分裂
8. 新建中间节点 $newnode$
9. $newnode$ 递归插入 ts 和 $node$ 索引的所有轨迹
10. 删除 $node$ 节点
11. 将 $newnode$ 作为当前节点的后继节点
12. else// $node$ 为中间节点
13. $node$ 递归插入 ts
14. else
15. 新建 MTSAX 表示为 G 的叶子节点 $newLeaf$
16. 叶子节点 $newLeaf$ 直接插入 ts

MTSAX 索引采用二分裂的方式,因此,分裂过程中可以升高某子段的某个变量的基数,将原先叶子节点索引的轨迹分配到 2 个新的叶子节点。节点分裂后还可能新的数据插入,所以从概率上选择最可能均分数据的某子段的某个变量升高基数。依次计算要分裂的叶子节点索引的所有轨迹数据在 w 个子段的 m 个变量的均值 μ 和标准差 σ ,若某子段的某个变量的均值越靠近升高基数后得到的新分割点 a ,同时对应的标准差越大,那么从概率上就越可能新的分割点均匀地分配数据;考虑到各变量的量纲不同,需要除以其最大值 max 归一化;考虑不同变量本身的重要性 $weight$,这样在相似性查询中可以更加倾向重要性更大的变量。通过求解式(5)可以寻找最可能均分数据的分裂子段和分裂变量。图 4 所示为每段包含 2 个变量的三段轨迹数据从基数 2 升高到基数 4 的情况,求解得到在第 2 段的第 2 个变量分裂能够从概率上最均匀的分配数据。

$$\operatorname{argmin}_i \frac{|\mu_i - a|}{3\sigma_i \times \max_i \times \text{weight}_i} \quad (5)$$

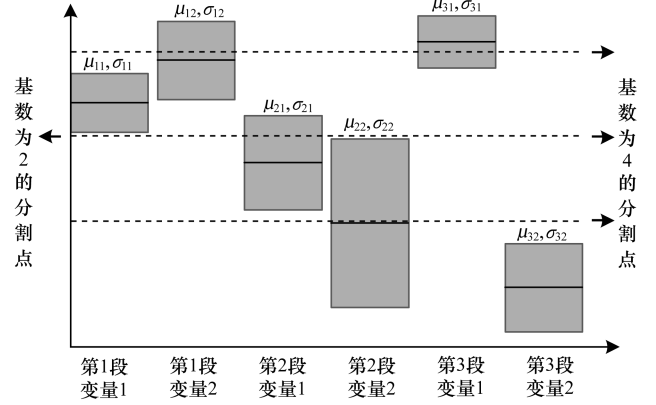


图 4 GeoWord 对切分变量的选择

2.5 基于 MTSAX 的多元轨迹相似轨迹查询

MTSAX 索引相似查询假设相似的两条多元轨迹具有相同的 MTSAX 表示,查询的结果是与查询轨迹距离最近的轨迹。MTSAX 索引是层次且没有重叠的,可以遍历索引树找到对应的索引节点,获取其索引的所有轨迹,分别计算这些轨迹与查询轨迹之间的距离,返回距离最小的轨迹。使用式(6)计算轨迹 s 和 t 之间的距离,计算过程中考虑了不同变量的重要性,距离越小越相似。

$$D(s, t, w) = \sqrt{\sum_{i=1}^m w_i \sum_{j=1}^n (s_{ij} - t_{ij})^2} \quad (6)$$

其中, m 表示轨迹的变量数目, n 表示轨迹中包含的轨迹点数目, s_{ij} 和 t_{ij} 分别表示轨迹 s 和 t 在第 i 个变量的第 j 个轨迹点, w_i 表示第 i 个变量的权重。

3 实验结果与分析

3.1 数据集

航运数据:从全球免费在线航运船舶跟踪网站 www.vesselfinder.com 爬取 40 000 条船的轨迹,每条轨迹有 200 个轨迹点,每个轨迹点包含经纬度坐标和航速 3 项信息。

GeoLife 数据集:GeoLife 数据集^[15]是一个关于被广泛使用的轨迹数据集,有 18 670 条轨迹,24 876 978 个轨迹点,轨迹总距离为 1 292 951 km。轨迹由用户步行、骑自行车、乘坐公交车、出租车等多种出行方式产生,每个轨迹点包含经度、纬度和海拔 3 项信息。

3.2 对比方法

对比方法如下:

1) 在基于空间划分的轨迹索引中,选择文献[9]提出采用固定空间划分的 TRSTJ 作为对比方法。

2) 在基于空间划分的轨迹索引中,选择文献[10]提出的基于多变量加和的轨迹索引方法,将其简记为 ADD。ADD 的处理过程与本文方法处理

步骤类似,主要区别在于 ADD 对原始的多元轨迹采用多变量加和的方式,而 MTSAX 采用 GeoWord 编码。

3) 在基于度量空间的轨迹索引中,选择文献[11]提出的基于参考点距离的 LBS 作为对比方法。

3.3 结果分析

为验证 MTSAX 索引方法的有效性,本文在相同初始基数下分别对 MTSAX、TRSTJ、ADD、LBS 和原始数据顺序扫描 5 种方法随机查询 100 次,通过平均查询时间比较这些方法的查询性能。

3.3.1 航运数据上的查询性能对比

查询性能对比过程如下:

1) MTSAX 和顺序扫描的对比

从图 5 和表 1 中可以看出,顺序扫描与空间划分无关,基数对顺序扫描没有影响,不同基数下顺序扫描的时间相同。MTSAX、TRSTJ、ADD 和 LBS 建立了相应的索引机制,因而均比顺序扫描要快。

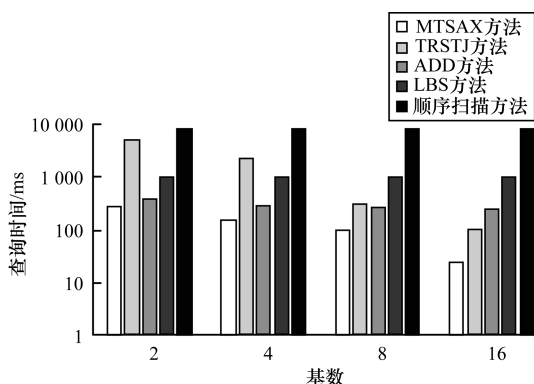


图5 航运数据上的查询性能对比

表1 航运数据上的查询性能

方法	基数为 2	基数为 4	基数为 8	基数为 16
MTSAX	257.27	148.47	96.40	22.77
TRSTJ	4 534.68	2 063.50	279.60	91.93
ADD	346.60	267.30	246.40	233.50
LBS	916.35	916.35	916.35	916.35
顺序扫描	7 693.73	7 693.73	7 693.73	7 693.73

2) MTSAX 和 TRSTJ 的对比

从图 5 和表 1 中可以看出,相同基数下 MTSAX 的查询性能均比 TRSTJ 要好,如基数为 2 时 MTSAX 查询速度是 TRSTJ 的 17.6 倍,基数为 4 时 MTSAX 查询速度是 TRSTJ 的 13.9 倍。这是 TRSTJ 对空间采取固定划分方式,导致 TRSTJ 索引中数据分布很不均匀,极少数的索引节点索引了大多数的轨迹,大多数的索引节点只索引了少部分的轨迹。因此,TRSTJ 大部分的查询发生在极少数的

索引节点上,而这些索引节点又索引了大量轨迹,查询需要扫描节点上索引的所有轨迹,该查询已经退化为顺序扫描,性能大大降低。

与 TRSTJ 相比,MTSAX 索引节点索引轨迹的数量分布相对均匀。MTSAX 可以随着数据量的大小动态调整索引结构,当单个节点包含的轨迹数量超过指定阈值节点则分裂,虽然增加了查询的深度,但是在每个索引节点上查询时间大大缩短,从而提高了整体查询的性能。

3) MTSAX 和 ADD 的对比

从图 5 和表 1 中可以发现,在相同基数下 MTSAX 效果均比 ADD 方法要好。这是因为 ADD 方法将多个变量加和,不同变量数据的变化被抵消,索引节点存放了很多不相似的轨迹,查询时间变长。

4) MTSAX 和 LBS 的对比

LBS 采用基于度量空间的方法,与空间划分无关,因此在不同初始空间基数下结果相同。由于时间序列的高维特性,基于度量空间的方法会面临维灾问题,当维度超过 15 时索引性能会随着维度的膨胀明显下降,而实验中每条轨迹包含 200 个轨迹点,远远大于 15。此外,LBS 将整条轨迹分成多个子段,查询整条轨迹需要将多个子段拼接在一起,降低了查询效率。

3.3.2 GeoLife 数据集上的查询性能对比

图 6 与表 2 是在 GeoLife 数据集上进行查询得到的性能对比,可以看出,MTSAX、TRSTJ、ADD、LBS 和顺序扫描 5 种方法的查询表现与航运数据是相似的。顺序扫描与空间划分无关,MTSAX、TRSTJ、ADD、LBS 均建立了相应的索引机制,均比顺序扫描要快。在相同基数下,MTSAX 的查询性能表现最佳。如基数为 2 时 MTSAX 查询速度是 TRSTJ 的 13.9 倍、ADD 方法的 1.3 倍、LBS 方法的 11.2 倍、顺序扫描的 26.2 倍;基数为 4 时 MTSAX 查询速度是 TRSTJ 的 8.5 倍、ADD 方法的 1.3 倍、LBS 方法的 11.9 倍、顺序扫描的 27.8 倍。

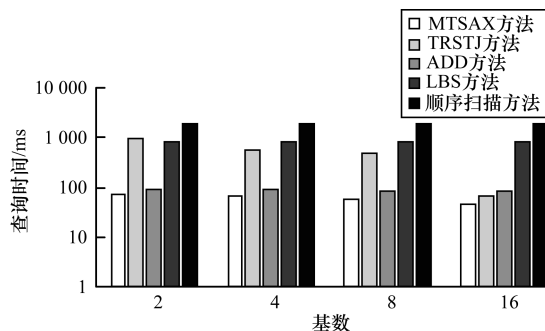


图6 GeoLife 数据集上的查询性能对比

表 2 GeoLife 数据集上的查询性能

方法	基数为 2	基数为 4	基数为 8	基数为 16
MTSAX	67.38	63.34	55.62	44.26
TRSTJ	935.78	539.93	468.66	62.66
ADD	89.05	83.77	82.90	78.93
LBS	752.36	752.36	752.36	752.36
顺序扫描	1 762.90	1 762.90	1 762.90	1 762.90

4 结束语

针对现有轨迹索引方法存在的问题,本文提出一种多元轨迹符号化索引 MTSAX。该方法针对单个多元轨迹点设计了一种基于 Geohash 的新编码 GeoWord,在 iSAX 索引框架的基础上,给出了移动对象历史轨迹索引方法 MTSAX。实验结果表明,在相同基数下 MTSAX 搜索性能均优于已知的基准索引方法,在海量数据下 MTSAX 对近似查询可以快速响应。下一步将侧重于建立分布式轨迹索引,提升索引建立的速度。

参考文献

- [1] 龚旭东. 轨迹数据相似性查询及其应用研究[D]. 合肥:中国科学技术大学,2015.
- [2] HECTOR G, HAN Jiawei, LI Xiaolei, et al. Adaptive fastest path computation on a road network: a traffic mining approach [C]//Proceedings of the 33rd International Conference on Very Large Data Bases. Washington D. C., USA: IEEE Press, 2007: 794-805.
- [3] GORJI S M, MOHAMMAD S, ABDOLLAH H. Non-stationary time series clustering with application to climate systems[M]. Berlin, Germany: Springer, 2014.
- [4] ZHENG Yu, ZHOU Xiaofang. Computing with spatial trajectories[M]. Berlin, Germany: Springer, 2011.
- [5] ELIAS F. Indexing objects moving on fixed networks[C]//Proceedings of International Symposium on Spatial and Temporal Databases. Washington D. C., USA: IEEE Press, 2003: 289-305.
- [6] KIM K S, KIM S W, KIM T W, et al. Fast indexing and updating method for moving objects on road networks[C]//Proceedings of Web Information Systems Engineering Workshops. Washington D. C., USA: IEEE Press, 2003: 34-42.
- [7] SALTENIS S, JENSEN C S, LEUTENEGGER S T. Indexing the positions of continuously moving objects [C]//Proceedings of 2000 ACM SIGMOD International Conference on Management of Data. New York, USA: ACM Press, 2000: 331-342.
- [8] LIN Dan, JENSEN C S, OOI B C, et al. Efficient indexing of the historical, present, and future positions of moving objects [C]//Proceedings of the 6th International Conference on Mobile Data Management. Washington D. C., USA: IEEE Press, 2005: 59-66.
- [9] PETKO B, MARIOS H, TSOTRAS V J. Time relaxed spatiotemporal trajectory joins [C]//Proceedings of the 13th Annual ACM International Workshop on Geographic Information Systems. New York, USA: ACM Press, 2005: 182-191.
- [10] 李正欣, 张凤鸣, 李克武, 等. 一种支持 DTW 距离的多元时间序列索引结构[J]. 软件学报, 2014, 25(3): 560-575.
- [11] KANISHKA B, ZHU Qiang, OZA N C, et al. Fast and flexible multivariate time series subsequence search [C]//Proceedings of 2010 IEEE International Conference on Data Mining. Washington D. C., USA: IEEE Press, 2010: 48-57.
- [12] EAMONN K, KAUSHIK C, MICHAEL P, et al. Dimensionality reduction for fast similarity search in large time series databases [J]. Knowledge and Information Systems, 2001, 3(3): 263-286.
- [13] JUNEJO I N, AL A Z. Using SAX representation for human action recognition [J]. Journal of Visual Communication and Image Representation, 2012, 23(6): 853-861.
- [14] JIN S, EAMONN K. iSAX: disk-aware mining and indexing of massive time series datasets [J]. Data Mining and Knowledge Discovery, 2009, 19(1): 24-57.
- [15] ZHENG Yu, ZHANG Lizhu, XIE Xing, et al. Mining interesting locations and travel sequences from GPS trajectories [C]//Proceedings of the 18th International Conference on World Wide Web. Washington D. C., USA: IEEE Press, 2009: 791-800.

编辑 索书志