一种支持轨迹大数据潜在语义相关性挖掘的谱聚类方法

廖律超[1,2],蒋新华[1,2],邹复民[2],贺文武[2],邱 淮[3]
(1.中南大学信息科学与工程学院,湖南长沙 410075; 2.福建工程学院福建省汽车电子与电驱动技术重点实验室,福建福州 350108; 3.福建省交通运输厅福建省交通信息通信中心,福建福州 350001)

摘 要: 针对交通管理优化和轨迹大数据挖掘的实际应用需求,本文提出了一种支持交通轨迹大数据潜在语义相关性挖掘的交通路网谱聚类方法(TSSC).首先研究了交通轨迹数据的向量空间建模方法,其次通过随机投影法快速提取大规模轨迹数据矩阵的特征信息并构建其低维语义子空间,然后基于语义子空间挖掘轨迹数据的潜在语义相关特性,在此基础上通过谱聚类方法实现了交通路网的快速聚类.通过本文提出的方法对总里程1400多万公里的实际交通轨迹数据进行实验分析表明,本方法可根据交通轨迹大数据的潜在语义相关性对交通路网进行快速的谱聚类处理,从而在复杂的交通路网间快速挖掘其潜在特性,为交通规划及其管理优化提供决策支持信息,同时也为时空大数据的聚类挖掘提供了一种新的解决方案.

关键词: 交通轨迹; 大数据; 数据挖掘; 语义空间; 谱聚类
中图分类号: TN911.23 **文献标识码:** A **文章编号:** 0372-2112 (2015)05-0956-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.05.019

A Spectral Clustering Method for Big Trajectory Data Mining with Latent Semantic Correlation

LIAO Lü-chao[1,2], JIANG Xin-hua[1,2], ZOU Fu-min[2], HE Wen-wu[2], QIU Huai[3]
(1. *School of Information Science and Engineering, Central-South University, Changsha, Hunan* 410075, *China*;
2. *Fujian Key Laboratory for Automotive Electronics and Electric Drive, Fujian University of Technology, Fuzhou, Fujian* 350108, *China*;
3. *Fujian Transport Information &Telecommunications Center, Fujian Communication Department, Fuzhou, Fujian* 350001, *China*)

Abstract: To facilitate traffic understanding, planning and management optimization, we present a new spectral clustering method(TSSC) for big trajectory data mining based on latent semantic correlation. First, a matrix model is proposed to represent vehicle trajectories and the underlying road network with a grid-vehicle matrix, which is then transformed to a low-dimensional semantic subspace with random projection. Second, through matrix decomposition we extract hidden characteristics of the mass trajectory data and construct a similarity matrix for road network cells. Third, we adopt and implement a fast spectral clustering method to discover road network clusters based on the similarity matrix in the semantic space. Finally, we evaluate our approach with a large trajectory data set collected by the Fujian Communications Department, which has 19,719 vehicles and a total mileage of more than 14 million kilometers. Experiment results show that the approach can efficiently cluster the road network with traffic context semantic information derived from massive trajectory data. The approach is capable to discover inherent characteristics of complex road networks and provide insights for traffic planning and management optimization.

Key words: traffic trajectory; big data; data mining; semantic space; spectral clustering

1 引言

随着卫星定位技术与基于位置服务(LBS)的迅速发展,已有越来越多的车辆安装了GPS/北斗等具备定位功能的装置,这些车辆周期性上报其位置、方向及速度等信息,形成海量的行车轨迹数据.以福州市为例,目前仅用于行业监管的浮动车数据每天就超过了2000万条,行车轨迹的总里程达到350万公里以上.

收稿日期:2014-04-14;修回日期:2015-01-07;责任编辑:孙瑶
基金项目:国家自然科学基金(No.61304199,No.41471333,No.61101139);福建省高校杰出青年科研人才计划(No.JA14209);福建省自然科学基金(No.2013J01214,No.2012J01247);福建省科技重大专项专题项目(No.2011HZ0002-1,No.2013HZ0002-1);福建省交通科技计划(No.201318)

我们知道,这些海量的轨迹数据中往往蕴藏了交通模式和驾驶行为习惯等丰富的潜在信息,通过充分挖掘这些潜在的、隐藏的信息,可以为城市道路规划与交通管理优化等提供非常有价值的决策支持信息.因此,通过轨迹数据挖掘研究驾驶行为与交通路网之间的相关性,进而优化交通路网规划,已成为国内外的研究热点之一[1].

目前,相关的研究包括城市基础设施的空间聚集特性[2]、城市交通流分析[3],动态路径寻优[4],热门地理区域发现[5],驾驶出行 OD 估计[6]、驾驶员的社会角色发现[7]、道路的几何特征分析[8]以及交通地图的自动更新[9]等多个方面.但是,现在这些研究主要还是针对原始轨迹数据直接进行分析处理的,然而对于实际的轨迹数据,由于受路网的地理特性及轨迹数据的内在特性影响,轨迹数据的空间分布往往具有非平稳性,即轨迹数据的属性值在空间上的变化并不均匀,甚至在同一道路的各局部空间也可能呈现完全不同的统计特性.因此,对原始数据直接进行统计分析度量的方法容易产生较大误差,尤其在具有较高的空间非平稳性时,往往导致那些具有特征意义的轨迹差别被大量的并没有特征意义的小差别所淹没,难以在海量轨迹数据中挖掘出其潜在的共性行为.

为此,本文提出了一种可支持交通轨迹大数据潜在语义相关性挖掘的谱聚类算法(Trajectory Semantic Spectral Clustering, TSSC).首先研究了交通轨迹大数据的向量空间建模、并通过随机投影方法实现大规模轨迹数据矩阵的快速语义计算,进而结合谱聚类方法实现了基于潜在语义相关性挖掘的交通路网谱聚类,并进行了相应的实验验证与探讨.

2　轨迹数据及其矩阵建模

2.1　相关定义

定义 1(计算空间)　将整个待计算路网的经纬度区间$[Lng_{\min}, Lng_{\max}]$与$[Lat_{\min}, Lat_{\max}]$所围成的区域称为计算空间 S,即:

$$S = \int_{Lng_{\min}}^{Lng_{\max}} (Lat_{\max} - Lat_{\min}) * \mathrm{d}_{Lng} \tag{1}$$

定义 2(空间网格)　将计算空间 S 按经度优先原则划分为等分的 M 列和 N 行,形成的$(M \times N)$个网格称为空间网格集 G,其中每个网格称为空间网格 g_i,简称网格,即:

$$G = \bigcup_{i=1}^{M \times N} g_i \tag{2}$$

其中,

$$i = n \times M + m \tag{3}$$

m 和 n 为网格 g_i 所在的列号和行号,i 为空间索引号,即 $g_{m,n}$可用 g_i 来表示.

显然,对于计算空间 S 和网格空间 G 具有 $S \equiv G$,且对于任意 $i, j(i \neq j)$,具有 $g_i \cap g_j = \varnothing$,同时给定计算空间 S 中的任意位置 $p(x, y)$,则在网格空间 G 中有 $\exists\, g_i \vdash p \in g_i$.

定义 3(路网空间)　空间网格集 G 中与交通路网 R 交集非空的空间网格集合称为路网空间 $\hat{R}$,即:

$$\hat{R} = \{ g_i \mid g_i \wedge R \neq \varnothing \} \tag{4}$$

轨迹数据的空间覆盖往往受路网可达性限制,因此路网空间是轨迹数据可能出现的空间网格集合,即计算空间 S 的一个子集,通过定义路网空间 $\hat{R}$,可大幅压缩空间维度,有效提升轨迹大数据挖掘的时间性能.

定义 4(交通轨迹)　周期性采集得到的某辆车在行驶过程中的空间、时间及其他信息组成的时间序列称为交通轨迹,即:

$$T_i = \langle p_1, p_2, p_3, \cdots, p_j \rangle \tag{5}$$

其中,$p_i = \langle x_i, y_i, t_i \rangle$为至少包含经度、纬度和时间信息的时空数据,$j = len(T_i)$为交通轨迹时间序列的长度.

由于在计算空间 S 中,任意位置 $p(x, y)$都有唯一的空间网格 g_i 与之相对应,因此轨迹数据可用空间网格序列等效表示.为避免空间网格序列中连续多个空间网格造成的处理难度,本文对序列中连续相同网格进行了进一步的删冗处理,即对于式(5)有等效的表示方式:

$$T_i = \langle g_1, g_2, g_3, \cdots, g_k \rangle \tag{6}$$

其中 $k \leqslant len(T_i)$.

2.2　轨迹数据向量空间建模

面对海量的等效空间网格序列,本文引入文本语义挖掘的思想,探索轨迹数据的语义特性挖掘.文本语义信息挖掘的核心思想是通过将所有的文本数据按固定顺序进行向量空间建模,并进而计算词组之间的语义相关性,虽然其建模过程中使数据产生了失序现象,但通过潜在语义分析仍可以反映较高的语义匹配准确度[10],同时还可覆盖到其缩写、等效表达、同义词等不同的等效关系[11].与文本语义信息类似,通过提取轨迹数据的语义信息,有望发现地理空间位置之间的语义相似性,这种语义相似性体现在当一辆车经过一个位置时,往往也会经过另外一个地理位置.为此,本文首先通过空间网格方法,将高精度的原始轨迹数据,转化为有限精度的等效空间网格系列,既保持了数据间的差异信息,也有利于提取其共性信息,并对其进行向量空间建模,从而将不规则、非结构化的交通轨迹数据转化为统一的结构化数据矩阵,为进一步结合潜在语义分析及谱聚类算法进行轨迹数据的内在特性分析提供了重要的结构化数据支持.

根据定义 3 与定义 4,交通轨迹数据可以通过路网空间 $\hat{R}$ 中的网格序列来描述,而不同的交通轨迹则可

通过车辆 ID 号来进行标识.因此,可构建“网格-车辆”矩阵 $\boldsymbol{X}$,其中矩阵的行对应于路网空间的网格编号(g_i),矩阵的列对应于交通轨迹的车辆编号(v_j),以实现大规模交通轨迹数据的结构化处理.

假定接入轨迹数据平台的车辆有 n 辆,每辆车为矩阵的一列,路网空间的网格数为 m 个,每个网格为矩阵的一行,即构成了一个 $m\times n$ 的“网格-车辆”矩阵 $\boldsymbol{X}$.在矩阵 $\boldsymbol{X}$ 中,根据矩阵元素 $x(i,j)$的不同赋值及其物理含义,可以对交通轨迹大数据进行不同方面的数据挖掘.

$$\boldsymbol{X}=\begin{pmatrix} & v_1 & v_2 & v_3 & \cdots & v_n \\ g_1 & x(1,1) & x(1,2) & x(1,3) & \cdots & x(1,n) \\ g_2 & x(2,1) & x(2,2) & x(2,3) & \cdots & x(2,n) \\ g_3 & x(3,1) & x(3,2) & x(3,3) & \cdots & x(3,n) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ g_m & x(m,1) & x(m,2) & x(m,3) & \cdots & x(m,n) \end{pmatrix} \tag{7}$$

定义 5(网格热度矩阵) 不妨令矩阵 $\boldsymbol{X}$ 的元素赋值为该车辆在该网格的行经次数,则可得到网格热度矩阵 $\boldsymbol{H}$(简称为“$\boldsymbol{H}$ 矩阵”),即:

$$\boldsymbol{H}=\begin{pmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{m1} & \cdots & h_{mn} \end{pmatrix} \tag{8}$$

其中,h_{ij}为车辆 j 在网格 g_i 中的行经次数.由于交通拥堵等原因可能导致车辆在同一网格内连续多次采集形成的重复数据,显然不能认为是多次经过,因此对连续在同一网格内的多个数据统计为一次.

根据 $\boldsymbol{H}$ 矩阵的定义可知,通过对 $\boldsymbol{H}$ 矩阵的语义空间相关性分析,有望实现空间网格基于道路驾驶热度的自动聚类,并进一步反应道路间的相关程度,为评估城市路网间的流量特性与通行模式[1]提供基础数据支撑.为此,本文将基于 $\boldsymbol{H}$ 矩阵研究可支持交通轨迹大数据潜在语义相关性挖掘的谱聚类算法,并进行相关实验研究与验证.

3 基于轨迹大数据的交通路网语义谱聚类

基于语义谱分析的路网空间聚类就是在矩阵分解构建的语义空间中,构造路网网格之间的带权无向完全图 $\bar{G}(V,E)$,各个路网网格看成无向图 $\bar{G}$ 的顶点 V,带权边集合 $E=\{w_{i,j}\mid i,j\in V\}$表示图中两顶点间的相似度,也就是地理空间中两个网格之间的语义相似度,从而把路网空间的聚类问题,转化为无向图的子图划分问题.显然,子图划分的关键就是要设计一种划分准则,使得划分后的子图间具有最小的相似度,而子图内部则实现相似度的最大化[12].

3.1 轨迹数据语义相似性度量

轨迹数据聚类的本质是按照轨迹数据的相似性进行对象的划分,而聚类划分的结果往往是使某种表示聚类质量的评价函数最优.因此,如何评价轨迹数据间的距离或相似度是聚类处理的关键问题之一.

轨迹数据是由空间位置信息组成的时间系列数据,同时包含时间与空间信息,轨迹数据的相似度是它们在时间与空间维度的相互邻近程度.目前的度量轨迹数据相似度的方法主要以独立坐标点为元素来进行处理的,并没有考虑轨迹间的语义相关性,难以发现海量轨迹之间的潜在共性信息,而且算法时间复杂度达 $O(n^2)$[13],难以满足大规模轨迹数据的处理需求.为此,本文提出一种结合语义空间欧氏距离及其高斯核函数的轨迹数据相似性度量方法,即首先将通过奇异值分解(SVD)构建大规模轨迹数据的语义空间,然后根据轨迹数据潜在语义信息的欧氏距离,通过高斯核函数进行其相似性度量.

奇异值分解是一种有效的上下文语义信息提取方法.通过对网格热度矩阵 $\boldsymbol{H}$ 进行奇异值分解,生成一个由若干左奇异正交向量构成的降秩空间,就是交通路网的语义空间.通常,语义空间隐含了交通驾驶人员对各道路节点的兴趣情况,其感兴趣程度由语义空间的特征向量与对应的奇异值的内积得到[14].

根据奇异值分解定理,已知矩阵 $\boldsymbol{H}$ 是 $M\times N$ 的实矩阵,不管其行列是否相关,必定存在 N 阶正交矩阵 $\boldsymbol{U}=(u_1,u_2,u_3,\cdots,u_r)$和 $\boldsymbol{V}=(v_1,v_2,v_3,\cdots,v_r)$,使得

$$\boldsymbol{H}=\boldsymbol{U}*\mathrm{diag}(\sigma_1,\sigma_2,\cdots,\sigma_r)*\boldsymbol{V}^{\mathrm{T}} \tag{9}$$

其中,对角矩阵 $\mathrm{diag}(\sigma_1,\sigma_2,\cdots,\sigma_r)$保存的是矩阵 $\boldsymbol{H}$ 的奇异值,这些奇异值按照从大到小的顺序排列为 $\sigma_i(i=1,2,\cdots,r)$,$\boldsymbol{U}$ 的列向量是 $\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}$ 的特征向量,$\boldsymbol{V}$ 的列向量是$\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}$ 的特征向量,r 为矩阵 $\boldsymbol{H}$ 的秩.

因此,将网格热度矩阵 $\boldsymbol{H}$ 进行奇异值分解,即可得到反应交通路网网格间上下文信息的语义空间,即左奇异向量矩阵,并且其语义空间前面的 k 个分量信号即可反映原始信号的概貌(其余的信号则反映了原始信号的细节变化)[15].

$$\boldsymbol{H}_k=\boldsymbol{U}_k*\mathrm{diag}(\sigma_1,\sigma_2,\cdots,\sigma_k)*\boldsymbol{V}_k^{\mathrm{T}} \tag{10}$$

其中降秩左奇异向量 $\boldsymbol{U}_k$ 称为降维语义子空间,根据该语义子空间可有效进行交通路网间共性特征模式提取及其异常模式发现.维数 k 的选取是影响算法性能的一个关键性指标,通常跟前 k 个最小奇异值的大小相关,若奇异值 σ_{k+1}相对于 $\sigma_1,\sigma_2,\cdots,\sigma_k$ 来说其下降速度明显减小,则这个 k 值是合适的[16].因此,本文在后续的算法验证过程中将遵循本原则进行 k 值的选取.

在降维语义子空间中,由于路网语义空间中低维

度部分的语义向量主要代表着交通路网网格间的共性信息,而高维度部分的语义向量则主要代表着各条道路的特性,且语义信息的重要程度与对应的奇异值成正比[14],因此将降维语义子空间的列向量与其对应的奇异值相乘,即可构成反应了交通路网中潜在的上下文语义信息的加权语义子空间 $\hat{\boldsymbol{U}}_k$.

$$\hat{\boldsymbol{U}}_k = \boldsymbol{U}_k * \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_k) \tag{11}$$

在路网加权语义子空间中,任意路网网格均具有了等维的潜在语义序列($u_1, u_2, \cdots, u_k$),并包含了网格间绝大部分的潜在语义信息,因此可采用欧氏距离来评价任意两个网格间的语义相似程度,即:

$$sem\_dist(A,B) = \left(\sum_{i=1}^{k} | A[i] - B[i] |^2\right)^{1/2} \tag{12}$$

其中,$sem_dist(A,B)$为路网网格 A 与 B 在语义空间的距离,$A[i](i=1,2,\cdots,k)$为 A 的潜在语义序列.

3.2 轨迹大数据语义计算

大数据是不能集中存储、难以在可接受时间内分析处理且数据整体呈现高价值的海量复杂数据集[17].交通轨迹数据挖掘正是个典型的大数据问题,上节的"空间-车辆"矩阵模型可有效解决轨迹大数据的结构化问题,但大矩阵的奇异值分解问题是本文对交通轨迹大数据挖掘的最关键的难题.由于奇异值分解的计算时间复杂度达到 $O(\min\{m^2n, mn^2\})$,随着矩阵空间维数的增加,其计算量将急剧增加,显然,难以适应大规模数据矩阵的应用需求[18].

为提升轨迹大数据挖掘的可行性,本文拟通过路网空间训练进行空间网格降维,并采用随机投影方法将大规模原始矩阵投影到低维随机矩阵中,从而将大数据集转化为便于进行奇异值分解和谱聚类处理的小数据集,实现交通轨迹大数据内在相关特性的快速挖掘与分析.

3.2.1 空间网格降维

将交通轨迹转化为空间网格序列表示时,网格数量与轨迹精度成反比.考虑到除交叉路口外,道路之间的间距一般都是大于 100m 的,因此在本文后续的系统实验中,网格大小设置为 $100 \times 100\mathrm{m}^2$.此时,以福建省福州市的市区及周边区域为例,总共的空间网格数多达 14.3 万个,显然这时矩阵 $\boldsymbol{H}$ 的奇异值分解计算量将非常巨大.

基于交通轨迹覆盖实际上受路网可达性限制的事实及路网空间的稀疏性,本文通过将离散的、非连续的路网空间映射到连续的数值空间,并剔除非路网空间网格以显著压缩矩阵 $\boldsymbol{H}$ 的行数,而通过该数值空间的逆映射,则可快速重构交通驾驶轨迹.同时,由于大规模交通轨迹数据的空间覆盖实际上反映了路网结构[19],因此路网网格集也可以通过大规模的轨迹数据进行训练得到,从而实现路网空间的动态提取,使矩阵 $\boldsymbol{H}$ 的空间网格降维与实际的路网变化保持一致.

3.2.2 随机投影降维

通过空间网格降维,对矩阵 $\boldsymbol{H}$ 的行数得到了显著压缩,但随着轨迹数据车辆数的不断增加,矩阵 $\boldsymbol{H}$ 的列数将越来越大,其奇异值分解仍然是个难题.

由于轨迹大数据语义空间的谱聚类主要是基于网格间的欧氏距离进行评价,而随机投影(Random Projections,RP)是一种高效的距离保持的数据压缩方法,因此本文拟通过随机投影方法进一步进行矩阵 $\boldsymbol{H}$ 的投影降维.随机投影是通过一组随机向量将高维数据投影到低维空间实现数据的压缩,其基本思想来自于 JL 定理.目前,该方法已成为大数据处理的重要技术之一[18].

定理 1　(Johnson-Lindenstrauss(JL)定理)[20],给定任意整数 n 及 $\varepsilon \in (0,1)$,设 k 为满足以下不等式的正整数:

$$k \geq 4(\varepsilon^2/2 - \varepsilon^3/3)^{-k} \ln n \tag{13}$$

则在 d 维欧氏空间中的任意 n 个数据点构成的集合 W,存在映射关系 $f: R^d \rightarrow R^k$,且使得集合 W 中的任意两个数据点 u, v 满足以下不等式:

$$(1-\varepsilon) \| u - v \|^2 \leqslant \| f(u) - f(v) \|^2 \leqslant (1+\varepsilon) \| u - v \|^2 \tag{14}$$

JL 定理基于矩阵扰动理论,定量分析了随机投影矩阵与原始矩阵的逼近情况,表明了高维欧氏空间中的 n 个数据点可以映射到 $O(\log n/\varepsilon^2)$ 维子空间中,并且这些点之间的距离仍得到近似保持,距离变化幅度不超过扰动界$(1 \pm \varepsilon)$,而文献[21]则进一步分析了基于随机投影的大规模矩阵奇异值分解结果的有效性.

通过随机投影方法,可将大规模的矩阵 $\boldsymbol{H}$ 投影到一个更低维的随机矩阵 $\boldsymbol{\Omega}(\boldsymbol{\Omega} \in \mathbf{R}^{n \times l})$,其中 l 为指定的低维随机投影空间维数.

$$\boldsymbol{P} = \boldsymbol{S} * \boldsymbol{\Omega} \tag{15}$$

其中,低维随机矩阵 $\boldsymbol{\Omega}$ 一般采用高斯分布随机生成.但由于高斯随机矩阵对于构造大规模随机矩阵及其对数据的投影操作计算复杂度较高,根据文献[22],本文选用更为简单的随机分布模式如式(16)所示,该分布模式中随机值有 2/3 的概率为 0,因此可节省 2/3 的投影计算开销.

$$\omega_{i,j} = \sqrt{3} \cdot \begin{cases} +1, & (p = 1/6) \\ 0, & (p = 2/3) \\ -1, & (p = 1/6) \end{cases} \tag{16}$$

通过以上随机投影方法,将奇异值分解时间复杂度由 $O(\min\{m^2n, mn^2\})$大幅降为 $O(mnl)$[18],从而显著提高大规模"网格-车辆"矩阵的奇异值分解效率.

3.3 基于轨迹语义相关的交通路网谱聚类

路网网格之间的相关性可以通过其语义相似图来

表征.给定路网空间数据集$\{\boldsymbol{g}_1, g_2, \cdots, g_n\}$以及任意两个网格 g_i 与 g_j 的语义相似关系 $W_{i,j}(W_{i,j}\geq 0)$,即可构建一个全连通的语义相似图 $\bar{G}(V,E)$,其中每个顶点 v_i 代表路网空间中的一个网格 g_i,边 $e_{i,j}$的权值表示网格间的语义相似度,从而可通过构建图的拉普拉斯矩阵进行相关性分析,并根据其相关特性进行聚类处理,这种聚类算法可以根据大规模轨迹数据进行路网特性分析,同时也可以支持任意形状的路网聚类处理[23].

以下给出交通路网谱聚类算法的具体流程:

(1)构造交通轨迹的语义空间.根据前面章节的分析,首先需要建立大规模轨迹数据矩阵 $\boldsymbol{H}$,然后通过网格压缩进行空间降维,再通过随机投影降维后进行矩阵奇异值快速分解,即可得到交通路网的语义空间.进一步通过选取交通轨迹语义空间中的前 k 个向量以表征原空间中的绝大部分共性信息,并通过与对应奇异值的内积运算,构建反应交通路网潜在上下文语义信息的加权语义子空间 $\hat{\boldsymbol{U}}_k$.

(2)生成路网空间语义距离矩阵.通过构造加权语义子空间 $\hat{\boldsymbol{U}}_k$,不仅通过提取少量低维语义向量便可估计原有超高维数据之间的距离关系,从而节省大量的相似度评估等计算开销,同时将路网属性特征的表征由原始的异构属性集,转化为等维的潜在语义向量,因此可方便地采用欧氏距离来评估各个路网间的语义距离关系,并生成对称的路网空间语义距离矩阵 $\boldsymbol{D}$.

$$\boldsymbol{D}=\begin{pmatrix} & g_1 & g_2 & g_3 & \cdots & g_n \\ g_1 & 0 & & & & \\ g_2 & d(2,1) & 0 & & & \\ g_3 & d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \\ g_n & d(n,1) & d(n,2) & d(n,3) & \cdots & 0 \end{pmatrix} \tag{17}$$

(3)构造路网空间语义相似度矩阵.根据语义距离矩阵 $\boldsymbol{D}$,有多种方式可以计算的路网空间节点的相似度,其中基于高斯核函数(Gaussian Kernel)相似性度量的谱聚类已证明具有较好的收敛性[24],为此,本文选用高斯核函数(也称为径向基核函数 Radial Basis Function Kernel)进行路网网格间的语义相似相似度计算,并构造对称的路网空间语义相似度矩阵 $\boldsymbol{W}$:

$$W_{i,j}=\exp\left(\frac{-(d_{i,j})^2}{2\sigma^2}\right) \tag{18}$$

其中,$d_{i,j}$为语义空间欧氏距离,其值越大则表示路网网格之间的相似度越小,σ 为核函数的带宽参数,用于描述路网网格语义相似度减小的速度,σ 越大则核函数的频带越宽,函数曲线越平滑,相似度下降速度越慢.但是,过大的 σ 容易出现欠学习(过平滑),难以有效评价路网网格的相似度,而过小的 σ 则容易出现过学习(欠平滑),使得路网相似度评价容易受个别不希望的突变量所影响.因此,在不同的应用环境中,需要通过调节 σ 值,以取得最佳的相似度评价效果.

(4)构造拉普拉斯矩阵以表征路网空间的图谱特性.根据图谱论[25],当两个图的拉普拉斯矩阵具有相同的特征值集时,它们被称为是谱相似的,并且谱相似的图不必同构,但同构的图必然谱相似.因此,对于语义相似图 $\bar{G}(V,E)$的分析,可从图的拉普拉斯矩阵出发,通过分析特征多项式、特征值以及特征向量来研究图的性质,为此本文通过语义相似度矩阵 $\boldsymbol{W}$,构造无向图的拉普拉斯矩阵 $\boldsymbol{L}$ 作为一个图 $\bar{G}(V,E)$的矩阵表示.

$$\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{Z}^{-1/2}\boldsymbol{W}\boldsymbol{Z}^{-1/2} \tag{19}$$

其中,$\boldsymbol{I}$ 为单位矩阵,$\boldsymbol{Z}$ 为对角阵,表示图中每个顶点及其所连接顶点间的关联程度,且有:

$$z_{i,j}=\sum_{j=1}^{n}W_{i,j} \tag{20}$$

(5)根据拉普拉斯矩阵特性进行路网谱聚类.根据谱聚类算法计算拉普拉斯矩阵 $\boldsymbol{L}$ 的特征值及其特征向量,并抽取前 p 个最小特征值对应的特征向量,构建矩阵 $\boldsymbol{Q}$,通过 k 均值聚类算法,对矩阵 $\boldsymbol{Q}$ 行向量构成的数据集聚合成 n 类,并根据聚类结果将原始数据矩阵 $\boldsymbol{H}$ 分类成 $\boldsymbol{H}_1, \boldsymbol{H}_2, \cdots, \boldsymbol{H}_n$.

4 实验结果与分析

4.1 实验环境

本文的实验运行环境是 Windows 7(64bit)操作系统,实验工作站的硬件配置是 Intel i7-3630QM 2.4G 双核 CPU,内存为 32GB,算法采用 Matlab 语言编写.

4.2 数据集与预处理

本文采用福建省福州市的浮动车轨迹数据作为算法实验测试数据集,整个数据集由 2013 年 12 月 1 ~ 4 日(共 4 天)的浮动车轨迹数据组成,数据集覆盖地理空间的经度范围为[119.113, 119.684],纬度范围为[25.904,26.155],区域面积约为 1430km^2,覆盖了福州市主城区及周边区域,数据集包括出租车、两客一危,重型货车、半挂牵引车等车辆类型 10 类共 19719 辆,包含了 8429 万条浮动车位置信息,交通轨迹总里程约 1416×10^4km.实验将覆盖的地理空间区域进行网格划分,网格的长宽均为 100m,整个地理空间划分为约 14.3 万个网格.通过大量轨迹数据进行网格训练,将空间压缩为 36879 个网格,图 1 为压缩后的路网空间图,其中每个黑色像素点为路网网格,白色像素点为网格训练所滤除的网格,空间降维压缩率达到 26%.

本文首先基于以上实验数据构造网格热度矩阵 $\boldsymbol{H}$,矩阵的行列数分别为 36 879 和 19 719,分别对应路

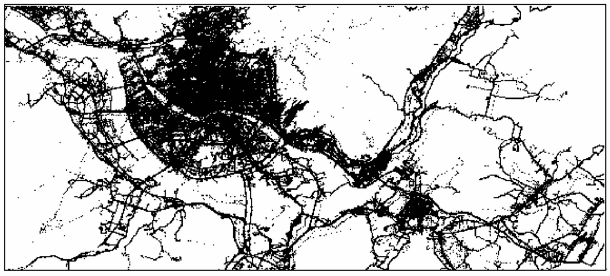


图1　福州市道路路网空间图

网空间的网格数和采集数据的车辆数,矩阵的元素值为该车辆经过该网格的行经次数.由于车辆驾驶区域覆盖率及浮动车轨迹数据的稀疏性的影响,矩阵 $\boldsymbol{H}$ 是个稀疏矩阵,矩阵密度为 0.0297.

4.3　实验结果

为快速提取轨迹数据间的上下文语义信息以实现路网潜在的相关性分析,对矩阵 $\boldsymbol{H}$ 直接进行奇异值分解的时间和空间复杂性都是难以接受的,本文实验工作站的硬件环境也不足以支撑其计算,故本文未能做具体的对比实验.而通过本文方法对矩阵进行奇异值分解则具有不错的时间空间复杂度表现,表 1 为选定不同的随机矩阵维度进行奇异值分解的结果情况.实验结果表明,当随机矩阵维度大于 200 时,随机矩阵分解结果趋于稳定.

表 1　随机投影矩阵分解结果

测试序号	随机矩阵维度	分解结果(前 5 个奇异值)					计算时间(s)
		S_1	S_2	S_3	S_4	S_5	
1	50	6.7265	2.7625	1.7097	1.3473	1.3141	22.8
2	100	6.8105	3.0714	2.0330	1.7345	1.5677	43.0
3	200	6.8553	3.0714	2.0330	1.7345	1.5677	14.7
4	300	6.8691	3.1086	2.0829	1.7804	1.6103	20.9
5	400	6.8757	3.1239	2.1024	1.8068	1.6347	32.5
6	500	6.8791	3.1312	2.1126	1.8177	1.6501	34.7

根据实验结果,本文选择随机矩阵维度为 500 进行 $\boldsymbol{H}$ 矩阵的降维处理,此时 $\boldsymbol{H}$ 矩阵奇异值分解用时仅为 34.7s.

根据奇异值的排列特性可知,前若干个奇异值占了绝大部分的比重,表明语义空间中的若干维信息代表了主要的语义信息.为了避免语义空间降维中造成关键信息丢失,本文以整个奇异值曲线(图 2)中的下降加速度为语义空间降维值的选取依据,即将奇异值下降速度明显减小的位置点作为语义空间维度.

由图 2 可知,第 13 个以后的奇异值变化幅度明显降低,并逐步趋于平稳.为此,本文把交通轨迹语义子空间维度设为 12 维,并根据语义信息构造路网相似度矩阵,进而通过拉普拉斯矩阵分析图谱间的相似关系,实现交通路网的语义谱聚类.通过以上方法,可以将交通路网聚合成若干类.当类别数越大,则粒度越小,可更好地分析路网的局部特性,而当分类数越小,则粒度越大,可更好地分析路网的全局特性,具体类别数的选择可根据实际应用场景的分类粒度要求进行灵活调整.本文以 10 类为例,将福州市路网网格按驾驶热度进行谱聚类分析,实验结果表明(图 3 用不同灰度标示分类结果),算法除了个别类别因数据量少(如类别 5、类别 7 及类别 8),难以进行定性分析外,绝大部分的类别均具有实际的物理意义,并实现了城区高速路、城市快速路、城区主干道、城区次干道、城区支路、郊区道路、国省道及其他道路的挖掘提取.

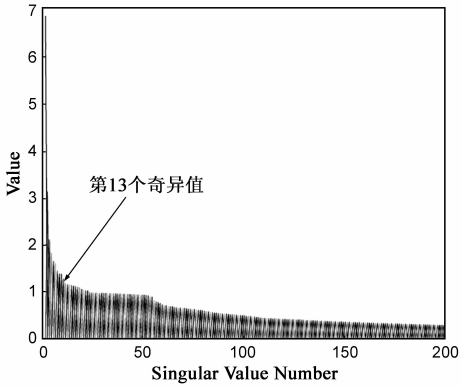


图2　语义空间奇异值曲线图

图3　路网空间聚类图

首先根据人工经验判断可发现,福州市路网网格根据车辆经过热度的聚类结果与实际的道路属性基本一致.为了进一步评估以上语义谱聚类的物理意义,本文将空间网格按聚类进行排序聚合,同时用类别编码乘以 1000 的线段作为该类别的覆盖区间,并统计各类别区间中每个网格经过的车辆数(如图 4).

图 4 表明,本文的聚类方法将顺序和相关性已完全独立的路网网格,根据车辆行经热度合理地划分成了 10 类.同时,不同类别之间(如类别 3 与类别 2 及类别 10)的行经车辆统计存在较大的交叉重叠,说明本文的语义谱聚类方法还结合了其他内在特性.

为了进一步分析各类别内部网格间的内在相关性,本文从各类别的网格数量、经过的总车辆数、车辆

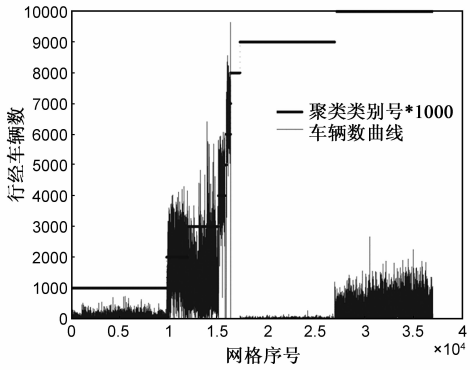


图4 各类别网格行经车辆数

经过类别内网格的总次数、每网格经过车辆的最大次数、最少次数、方差以及每类道路对应的实际物理意义等方面进行分析(如表 2),其中第 7、第 8 类网络因没有车辆经过,故未列出.

表 2 道路聚类结果统计分析

道路聚类 ID	道路网格数量	车辆经过次数情况					实际物理意义
		网格集经过总车辆数	网格集经过总车次	单网格最大次数	单网格最少次数	方差	
1	9740	14556	673464	732	1	66.0	城区支道、郊区道路
2	2197	15343	5583077	4305	4	860.9	城区主干道
3	3075	15922	7498529	6413	3	781.2	高速路
4	682	14043	5340441	6071	39	769.2	城区次干道
5	126	89	200	15	1	1.7	
6	445	13630	8254374	9641	3303	788.9	城市快速路
9	9639	11053	97507	263	0	12.0	其他道路
10	10021	15689	4486538	2670	1	301.1	国省道

表 2 的统计数据进一步表明,聚类结果总体反应了路网中的行驶热度,但并不完全根据热度的统计特性来进行聚类处理,而是充分结合了交通轨迹各网格间的强相关性.同样以类别 3 为例,其方差系数为 781.2,表明类别中各个网格之间行经车辆数统计差距较大,说明与直接基于网格热度统计分析的分类方式不同,在本文的语义谱聚类中,某些热度差距很大的网格因为交通轨迹的相关性,也自动地聚合到同一个类别中.

为了验证聚类内部路网网格间的相关性,本文提取了类别 3 的道路集(如图 5),及其螺洲大桥附近区域(经度方向网格序号范围[200,300]和维度方向网格序号范围为[150,200]所覆盖的的路网区域)的子道路集(如图 6).

对照图 5 和福州市交通路网图可知,该自动聚类形成的网格集合与福州市各高速道路及三环高速路的集合高度吻合,表明本文方法实现的路网网格聚类正确地反应了实际交通路网的相关特性.同时,由图 6 可知,尽管本文方法没有对各路网网格按照道理属性进行特别处理,但是本文方法有效地挖掘了这种语义空间的潜在相关性,实现了同一道路绝大多数网格的自动聚类,保持了较好的路网完整性.

进一步对图 6 路网区域的驾驶热度进行统计分析表明(如图 7),区域内通过本文方法聚合在同一类别的道路,各网格间的驾驶热度本身具有较大的非平稳性.因此,若直接使用全数据集的统计方法,显然将造成交通轨迹语义空间上下文信息的严重丢失,难以实现同一道路的聚合.

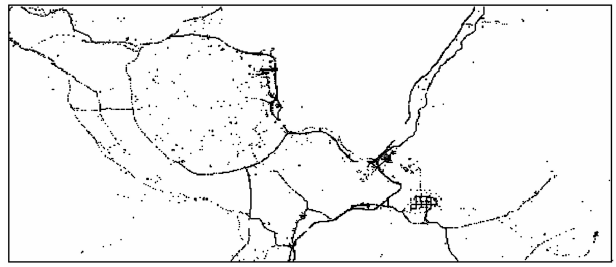


图5 路网聚类结果图(第3类)

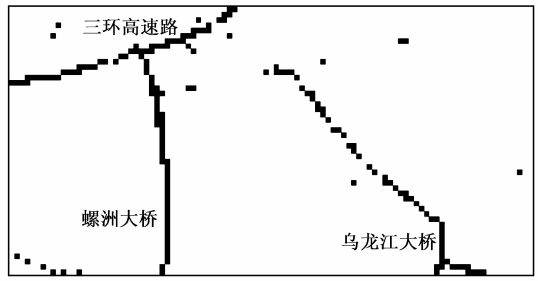


图6 局部路网图(第3类)

另一方面,为了评估算法的时间性能,本文给定数据的维度(矩阵的行数)为 36 879,通过调整样本量大小(矩阵的列数)生成实验样本数据,对比本文方法与传统谱聚类方法的聚类时间,图 8 为其时间曲线对比图,实验数据表明,传统谱聚类方法随着样本量的大小,系统计算时间呈线性增长,而本文方法则接近平稳,其主要原因在于本文方法将高维数据投影到低维语义空间中,实现了其快速聚类处理,从而为时空轨迹大数据的快速处理提供了有效的方法.

5 结论

通过交通轨迹大数据的向量空间矩阵建模、快速语义计算以及谱聚类等理论研究,提出了一种可支持交通轨迹大数据潜在语义相关性挖掘的交通路网谱聚类方法(TSSC).实验结果表明,该方法不仅具有良好的处理性能,以支持轨迹大数据的处理,还可有效轨迹大数据潜在的、隐藏的语义信息,为交通规划及其管理优化提供各种决策支持信息.同时,该方法也为时空轨迹

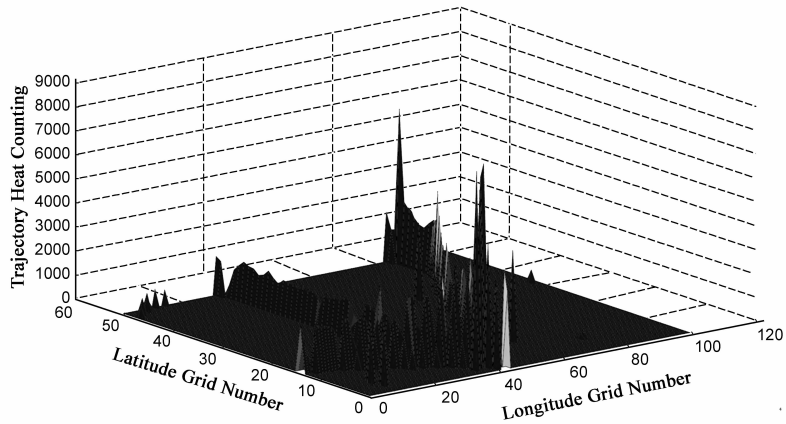


图7　局部路网热度分布图(第3类)

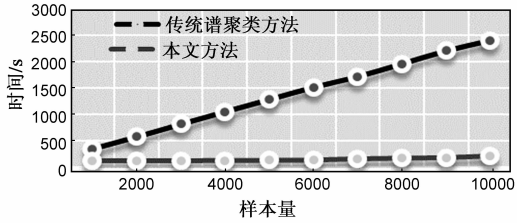


图8　时间性能曲线

大数据的深度挖掘提供了一种新的数据建模及其快速挖掘方法,并可进一步用于轨迹聚类及交通拥堵点发现等交通流共性行为特征识别等应用场景.

参考文献

[1] Fang Z, Shaw S-L, Tu W, et al. Spatiotemporal analysis of critical transportation links based on time geographic concepts: a case study of critical bridges in Wuhan, China[J]. Journal of Transport Geography, 2012, 23(3): 44 – 59.

[2] Bell M G. Policy issues for the future intelligent road transport infrastructure[A]. IEE Proceedings-Intelligent Transport Systems[C]. USA: IEE, 2006. 147 – 155.

[3] Shi W, Kong Q-J, Liu Y. A GPS/GIS integrated system for urban traffic flow analysis[A]. Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems (ITSC)[C]. USA: IEEE, 2008. 844 – 849.

[4] 曹政才,韩丁富,等.面向城市交通网络的一种新型动态路径寻优方法[J].电子学报,2012,40(10):2062 – 2067.
Cao Zheng-cai, Han Ding-fu, et al. Anovel dynamic path optimization method for urban traffic networks[J]. Acta Electronica Sinica, 2012, 40(10): 2062 – 2067. (in Chinese)

[5] Masutani O, Iwasaki H, Tei K, et al. Real-time POI detection andrating using floating car data[A]. Proceedings of 14th World Congress on Intelligent Transport Systems[C]. Beijing: ITS, 2007. 1 – 5.

[6] Yang Y, Lu H-P, Hu Q. A bi-level programming model for origin destination estimation based on FCD[A]. Proceedings of the 10th International Conference of Chinese Transportation Professionals[C]. USA: American Society of Civil Engineers, 2010. 117 – 124.

[7] 马宇驰,杨宁,谢琳,等.基于轨迹时空关联语义和时态熵的移动对象社会角色发现[J].计算机研究与发展,2012,49(10):2153 – 2160.
Ma Yu-chi, Yang Ning, Xie Lin, et al. Socialroles discovery of moving objects based on spatial-temporal associated semantics and temporal entropy of trajectories[J]. Journal of Computer Research and Development, 2012, 49(10): 2153 – 2160. (in Chinese)

[8] Liu C, Jian Z, Meng X. Combining float car data and multi-spectral satellite images to extract road features and networks [A]. Progress in Location-Based Services[C]. Berlin Heidelberg: Springer, 2013. 29 – 43.

[9] Li J, Qin Q, Xie C, et al. Integrated use of spatial and semantic relationships for extracting road networks from floating car data [J]. International Journal of Applied Earth Observation and Geoinformation, 2012, 19(5): 238 – 247.

[10] Platzer C, Dustdar S. A vector space search engine for web services[A]. Proceedings of the Third IEEE European Conference on Web Services[C]. USA: IEEE, 2005. 1 – 9.

[11] Kanerva P, Kristofersson J, Holst A. Random indexing of text samples for latent semantic analysis[A]. Proceedings of the 22nd Annual Conference of the Cognitive Science Society [C]. USA: Cognitive Science Society, 2000. 1036 – 1037.

[12] Von Luxburg U. A tutorial on spectral clustering[J]. Statistics and computing, 2007, 17(4): 395 – 416.

[13] Chen L, zsu M T, Oria V. Robust and fast similarity search for moving object trajectories[A]. Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data [C]. USA: ACM, 2005. 491 – 502.

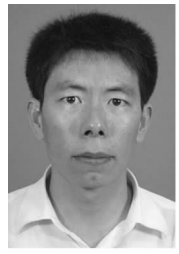
[14] 刘云峰,齐欢,Hu Xiangen,等.基于潜在语义空间维度特性的多层文档聚类[J].清华大学学报:自然科学版,

2005,45(S1):1783 – 1786.
[15] 赵学智,叶邦彦.多分辨 SVD 包理论及其在信号处理中的应用[J].电子学报,2012,40(10):2039 – 2046.
Zhao Xue-zhi, Ye Bang-yan. Multi-resolution SVD packet theory and its application to signal processing[J]. Acta Electronica Sinica,2012,40(10):2039 – 2046.(in Chinese)
[16] 秦洋,王立宏,等.基于拉普拉斯矩阵的 DNA 序列集相似性分析[J].北京交通大学学报:自然科学版,2009,33(6):137 – 140.
Qin Yang,Wang Li-hong,et al. Analysis of similarity between DNA sequence sets basedon laplace matrix[J]. Journal of Beijing Jiaotong University,2009,33(6):137 – 140.(in Chinese)
[17] Zou B, Li L, Xu Z, et al. Generalizationperformance of fisher linear discriminant based on markov sampling [J]. IEEE Transactions on Neural Networks and Learning Systems, 2013,24(2):288 – 300.
[18] Mahoney M W. Randomizedalgorithms for matrices and data [A]. Advances in Machine Learning and Data Mining for Astronomy[C]. USA: Chapman and Hall/CRC, 2012. 647 – 672.
[19] Li Q, Zhang T, Yu Y. Using cloud computing to process intensive floating car data for urban traffic surveillance[J]. International Journal of Geographical Information Science,2011, 25(8):1303 – 1322.
[20] Johnson W B, Lindenstrauss J. Extensions of Lipschitz mappings into a Hilbert space[J]. Contemporary mathematics, 1984,26(1):189 – 206.
[21] Papadimitriou C H, Tamaki H, Raghavan P, et al. Latent semantic indexing: A probabilistic analysis[A]. Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems[C]. USA: ACM, 1998.159 – 168.
[22] Achlioptas D. Database-friendly random projections: Johnson-Lindenstrauss with binary coins[J]. Journal of Computer and System Sciences,2003,66(4):671 – 687.
[23] 徐森,周天,于化龙,等.一种基于矩阵低秩近似的聚类集成算法[J].电子学报,2013,41(6):1219 – 1224.
Xu Sen, Zhou Tian, Yu Hua-long, et al. Matrixlow rank approximation-based cluster ensemble algorithm[J]. Acta Electronica Sinica,2013,41(6):1219 – 1224.(in Chinese)
[24] 高炜,周定轩.与一般相似度函数相关的谱聚类的收敛性[J].中国科学:数学,2012,42(10):985 – 994.
Gao Wei,Zhou Ding-Xuan. Convergence of spectral clustering with a general similarity function[J]. Scientia Sinica Mathematica,2012,42(10):985 – 994.(in Chinese)
[25] Merris R. Laplacian matricesof graphs: a survey[J]. Linear Algebra and Its Applications,1994,197(3795):143 – 176.

作者简介

廖律超 男,1980 年生于福建长汀,高工,现为中南大学信息科学与工程学院博士研究生.主要研究领域为交通数据挖掘与知识发现、交通大数据处理技术等,主持了国家自然科学基金、福建省自然科学基金、福建省高校杰出青年科研人才计划等项目 8 项.
E-mail:achao@fjut.edu.cn

蒋新华 男,1956 年生于湖南长沙,博士生导师,中南大学交通信息工程及控制学科教授,福建工程学院教授.主要研究方向为交通大数据处理关键技术、智能控制理论与先进 PID 控制方法、下一代移动互联网关键技术.
E-mail:xhjiang@fjut.edu.cn



邹复民 男,1976 年生于湖南隆回,博士,现为福建工程学院信息科学与工程学院教授,主要研究方向为车联网、交通信息工程与云计算技术.
E-mail:fmzou@fjut.edu.cn

贺文武 男,1972 年生于湖南常德,理学博士,现为福建工程学院数理系副教授,美国密歇根大学访问学者.主要研究方向为统计机器学习与数据挖掘等.
E-mail:hwwhbb@163.com

邱 淮 男,1963 年生于福建上杭,教授级高工,现就职于福建省交通运输厅,主要研究方向:交通信息采集,异构交通信息融合及交通运输云计算平台关键技术等.