基于稀疏化最小生成树聚类的个性化轨迹隐私保护算法

王 超,杨 静,张健沛
(哈尔滨工程大学计算机科学与技术学院,黑龙江哈尔滨 150001)

摘 要: 现有的轨迹匿名算法没有充分考虑轨迹内外在特征信息以及移动对象个性化的隐私需求.为此,本文提出个性化轨迹 k-匿名的概念,并提出轨迹结构相似性度量模型,综合考虑轨迹方向、速度、转角和位置等内外在特征信息;然后,提出基于稀疏化最小生成树聚类的个性化隐私保护算法,通过稀疏化的方法降低最小生成树聚类的执行时间,通过贪婪策略生成近似最优的轨迹 k-匿名集合.实验结果表明,本文的轨迹结构相似性度量模型能更加准确地度量轨迹间的相似性,所提算法花费了更少的时间代价,具有更高的数据可用性.

关键词: 轨迹相似性; 个性化轨迹 k-匿名; 稀疏化; 最小生成树聚类; k-节点划分
中图分类号: TP309.2 **文献标识码:** A **文章编号:** 0372-2112 (2015)11-2338-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.11.029

The Sparse Minimum Spanning Tree Clustering Based Personalized Trajectory Privacy Protection Algorithm

WANG Chao, YANG Jing, ZHANG Jian-pei
(*College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang* 150001, *China*)

Abstract: The existing trajectory anonymity methods can not reflect the trajectory internal and external characteristics information well, and ignore personalized privacy requirements of moving objects. To solve these problems, we propose a new similarity measure model of trajectory structure, which considers the trajectory internal and external characteristics information of direction、speed、angle and location. On this basis, we propose the sparse minimum spanning tree clustering based personalized trajectory privacy protection algorithm. It reduces runtime by sparse methods, and generates an approximate optimal trajectory k-anonymity set by greedy strategy. Finally, the results showed our new similarity measure model of trajectory structure can calculate distance of trajectories more accurately, and our method offers better utility and costs less time than previous proposals in the literature.

Key words: trajectory similarity; personalized trajectory k-anonymity; sparse methods; minimum spanning tree clustering; k-node partition

1 引言

移动社会网络的兴起及移动智能终端的发展带来了海量新数据[1],尤其是个人位置和轨迹数据.位置和轨迹数据含有丰富的时空信息,对其进行分析和挖掘可以支持多种与移动对象相关的应用[2].因此,发布并分析这样的时空数据非常必要.然而,个人的隐私信息很可能会随着轨迹数据的发布而受到威胁.在轨迹数据中,最大的隐私威胁就是"敏感位置泄露",如果攻击者确定某人在哪些时间访问了哪些位置,那么攻击者就能确定此人在发布数据库中的真实记录,并获取此人的其他轨迹信息,进而推理得到此人的行为模式、社会习惯等隐私,造成个人隐私信息泄露.因此,面向移动社会网络轨迹发布的隐私保护方法是一个亟待解决的问题.

对于关系型数据的隐私保护,学术界进行了广泛研究[3~7],但关系型数据的隐私保护不能直接应用到轨迹隐私保护中,为此,研究人员提出了多种轨迹匿名算法[8~13],使发布的轨迹数据集满足轨迹 k-匿名,从而实现用户轨迹的隐私保护.然而,不合适的轨迹相似性度量导致匿名过程中不必要的信息损失,降低数据可用性.因此,如何准确地衡量轨迹相似性是一个关键问题.

在计算轨迹相似性时,O.Abul[8]、Huo[10]使用欧式距离作为度量标准,O.Abul[9]、Chen[11]用编辑距离作为度量标准,Tiakas[12]用线性时空距离作为度量标准,Gao

收稿日期:2014-02-19;修回日期:2014-04-11;责任编辑:蓝红杰
基金项目:国家自然科学基金(No.61370083, No.61073041, No.61073043);高等学校博士学科点专项科研基金(No.20112304110011, No.20122304110012);哈尔滨市科技创新人才研究专项资金(优秀学科带头人 No.2011RFXXG015)

等[13]用轨迹角度来衡量轨迹的相似性和方向,Yuan等[14]用轨迹的特征信息进行聚类,但是忽略了轨迹的时间维度对轨迹相似性的影响.以上方法都不能很好地反映轨迹的内外在特征信息对轨迹相似性的影响;而且忽略了移动对象个性化的隐私需求.

针对以上问题,本文提出轨迹结构相似性度量模型,综合考虑轨迹的方向、速度、转角和位置等内外在特征信息;然后提出个性化轨迹 k-匿名的概念,并提出基于稀疏化最小生成树聚类的个性化隐私保护算法,将轨迹 k-匿名集合的选择问题,转换成稀疏化图的个性化 k-节点划分问题,通过稀疏化的方法降低聚类时间,通过贪婪的策略生成近似最优的轨迹 k-匿名集合;最后,通过实验验证了方法的有效性.

2　相关工作

为了保护移动对象的轨迹隐私,研究人员做了大量研究.You[15],Gao[16]提出生成假轨迹的方法,得到较高的服务质量,但不能确保轨迹间良好的相似性.Terrovitis[17],Chen[18]提出使用抑制技术来实现轨迹数据的匿名,然而,太多的轨迹片段被抑制会造成巨大的信息损失.Nergiz[19]提出基于泛化的算法,首先通过轨迹点匹配的方式进行轨迹聚类,然后将轨迹的对应点泛化为最小边界矩形,最后重构匿名的轨迹数据,并发布重构后的原子轨迹;Domingo-Ferrer[20]提出基于微聚集和排列的匿名算法,使用微聚集算法对轨迹进行聚类,最后使用位置排列算法,对轨迹数据进行重构.Mahdavifar[21]认为不同的移动对象有不同的隐私级别,并提出个性化的轨迹隐私保护算法,使用匹配点算法进行轨迹匿名,以此平衡信息损失和隐私保护程度.但是,上述方法都不能很好地反映轨迹的内外在特征信息对轨迹相似性的影响,而且大部分忽略了移动对象个性化的隐私需求,认为所有移动对象的隐私保护级别是相同的.

3　研究基础

限于篇幅,下文中涉及到的轨迹、子轨迹和轨迹 k-匿名的概念请参考文献[20].

定义 1　轨迹隐私级别.轨迹隐私级别是保证该轨迹不能被攻击者从匿名集合中识别出来,所需要的最小轨迹数量.

本文使用 $L=\{l_1,l_2,\cdots,l_{\max}\}$ 表示轨迹集合中轨迹隐私级别的有序集合,对于任意整数 $i,j(1\leqslant i<j\leqslant \max)$,有 $l_i<l_j$.对于任意轨迹 T,$l(T)$ 表示该轨迹的隐私级别.

轨迹数据发布时个人隐私泄露的风险与攻击者掌握的背景知识有关,攻击者掌握的背景知识越多,个人隐私泄露风险越大,反之亦然.本文假设攻击者掌握以下背景知识:

(1)发布的匿名轨迹集合 D^*;(2)原始的目标轨迹 $T(T\in D,D$ 表示原始轨迹集合)的子轨迹 S.

定义 2　攻击模型.对于已知的子轨迹 $S(S\leqslant T)$和匿名轨迹集合 D^*,如果攻击者能确定匿名轨迹 T^* $(T^*\in D^*)$是 T 的匿名轨迹,则认为用户的个人隐私遭到攻击.

定义 3　个性化轨迹 k-匿名.给定轨迹集合 D,子轨迹 S,如果对于任意轨迹 $T(T\in D)$,都满足轨迹 k-匿名(k 为轨迹 T 的轨迹隐私级别),那么轨迹集合 D 满足个性化轨迹 k-匿名.

4　轨迹相似性度量

4.1　轨迹结构

定义 4　相交轨迹.两条轨迹 $T_i=\{(t_1^i,x_1^i,y_1^i),(t_2^i,x_2^i,y_2^i),\cdots,(t_n^i,x_n^i,y_n^i)\}$ 和 $T_j=\{(t_1^j,x_1^j,y_1^j),(t_2^j,x_2^j,y_2^j),\cdots,(t_m^j,x_m^j,y_m^j)\}$,$I=\max(\min(t_n^i,t_m^j)-\max(t_1^i,t_1^j),0)$.如果 $I>0$,那么轨迹 T_i 和 T_j 是相交轨迹;否则,轨迹 T_i 和 T_j 不是相交轨迹.

下文涉及到的轨迹 $p\%$-相交、同步轨迹和同步轨迹集合的概念请参考文献[20].本文的轨迹结构相似性度量以同步轨迹为基础,为计算轨迹相似性,必须将非同步轨迹转换为同步轨迹.具体算法见文献[20]算法1.

定义 5　轨迹结构.轨迹结构是轨迹内部特征属性集合,包括轨迹方向、速度、转角和位置.

定义 6　轨迹段.同步轨迹 T_i,T_j 的采样时间为 $(t_1,t_2,\cdots,t_n)$,$t_k(1\leqslant k<n)$ 为任意采样时间,则 t_k 和 t_{k+1} 间的轨迹片段即为轨迹段,用 L_k^i,L_k^j 表示.

4.2　轨迹方向距离

图1、2描述了轨迹段的方向距离.轨迹段方向距离表示轨迹段在运动趋势上的偏转程度,用 DirDist 表示.

$$\begin{aligned}\mathrm{DirDist}(L_k^i,L_k^j)&=\begin{cases}\min(|L_k^i|,|L_k^j|)\cdot\sin\alpha, & 0\leqslant\alpha\leqslant\pi/2\\ \min(|L_k^i|,|L_k^j|)\cdot\cos\beta+\min(|L_k^i|,|L_k^j|)\cdot\sin\beta, & \pi/2\leqslant\alpha\leqslant\pi,\beta=\alpha-\pi/2\end{cases}\\&(1)\end{aligned}$$

其中,L_k^i 和 L_k^j 分别为轨迹 T_i,T_j 的对应轨迹段,α 为轨迹段 L_k^i 和 L_k^j 的方向夹角.

那么,轨迹的方向距离表示为:

$$d_{\mathrm{Dir}}(T_i,T_j)=\frac{1}{p}\cdot\sum_{k=1}^{n}\mathrm{DirDist}(L_k^i,L_k^j)\qquad(2)$$

其中,T_i,T_j 是同步轨迹,轨迹 $p\%$-相交,$p>0$,L_k^i 和 L_k^j 分别为 T_i,T_j 的轨迹段,n 为轨迹段个数.

4.3 轨迹转角距离

定义 7 转角.对于给定的轨迹,其相邻轨迹段的转向角即为轨迹段的转角.

图 3 给出了转角的实例.其中,θ_1、θ_2和 θ_3是转角,α为轨迹段间的夹角.由于轨迹段顺时针或逆时针偏转对于轨迹相似性有较大影响(方向相同的轨迹段顺时针和逆时针偏转 90 度会产生方向截然相反的两条轨迹),本文对转角给出如下公式:

$$\theta=\begin{cases}\pi-\alpha, & \text{轨迹段顺时针偏转}\\ \alpha-\pi, & \text{轨迹段逆时针偏转}\end{cases} \tag{3}$$

因此,由图 3 得,$\theta_3=\pi-\alpha$;$\theta_1=\alpha-\pi$.

轨迹段转角距离反应了轨迹内部的方向变化以及波动程度,用 AngDist 表示.

$$\mathrm{AngDist}(L_k^i,L_k^j)=(|\theta_k^i-\theta_k^j|)/(|\theta_k^i|+|\theta_k^j|) \tag{4}$$

其中,L_k^i和 L_k^j为 T_i,T_j的对应轨迹段,θ_k^i和 θ_k^j为轨迹段 L_k^i和 L_k^j与下个邻近轨迹段间的转角.

那么,轨迹的转角距离表示为:

$$d_{\mathrm{Ang}}(T_i,T_j)=\frac{1}{p}\cdot\sum_{k=1}^{n-1}\mathrm{AngDist}(L_k^i,L_k^j) \tag{5}$$

其中,T_i,T_j是同步轨迹,轨迹 p%-相交,$p>0$,L_k^i和 L_k^j分别为 T_i,T_j的轨迹段,n 为轨迹段个数.

4.4 轨迹速度距离

轨迹段速度距离表示移动对象在对应轨迹段上移动速度的差异性,用 SpeDist 表示.

$$\mathrm{SpeDist}(L_k^i,L_k^j)=\left|\frac{\sqrt{(x_{\mathrm{end}}^i-x_{\mathrm{begin}}^i)^2+(y_{\mathrm{end}}^i-y_{\mathrm{begin}}^i)^2}}{\Delta t_i}-\frac{\sqrt{(x_{\mathrm{end}}^j-x_{\mathrm{begin}}^j)^2+(y_{\mathrm{end}}^j-y_{\mathrm{begin}}^j)^2}}{\Delta t_j}\right| \tag{6}$$

其中,L_k^i和 L_k^j分别为 T_i,T_j的对应轨迹段,x_{begin}^i,x_{end}^i和 y_{begin}^i,y_{end}^i分别为轨迹段 L_k^i的开始和结束位置的 x、y 坐标,Δt_i为轨迹段 L_k^i的持续时间.

那么,轨迹的速度距离表示为:

$$d_{\mathrm{Spe}}(T_i,T_j)=\frac{1}{p}\cdot\sum_{k=1}^{n}\mathrm{SpeDist}(L_k^i,L_k^j) \tag{7}$$

其中,T_i,T_j是同步轨迹,轨迹 p%-相交,$p>0$,L_k^i和 L_k^j分别为 T_i,T_j的轨迹段,n 为轨迹段个数.

4.5 轨迹位置距离

轨迹位置距离反应轨迹的时空相似性,表示为:

$$d_{\mathrm{Loc}}(T_i,T_j)=\frac{1}{p}\cdot\sqrt{\sum_{t_s\in\mathrm{ot}(T_i,T_j)}\frac{(x_s^i-x_s^j)^2+(y_s^i-y_s^j)^2}{|\mathrm{ot}(T_i,T_j)|^2}} \tag{8}$$

其中,T_i和 T_j是同步轨迹,轨迹 p%-相交,$p>0$,$\mathrm{ot}(T_i,T_j)$为轨迹 T_i,T_j的重叠时间间隔.

4.6 轨迹结构相似性

本文基于轨迹方向、速度、转角和位置特征属性,提出如下轨迹结构距离度量模型:

$$\begin{aligned}&\mathrm{Dist}(T_i,T_j)\\&=W_D\cdot\mathrm{Form}(d_{\mathrm{Dir}}(T_i,T_j))+W_S\cdot\mathrm{Form}(d_{\mathrm{Spe}}(T_i,T_j))\\&+W_A\cdot\mathrm{Form}(d_{\mathrm{Ang}}(T_i,T_j))+W_L\cdot\mathrm{Form}(d_{\mathrm{Loc}}(T_i,T_j))\end{aligned} \tag{9}$$

其中,W_D,W_S,W_A,W_L分别为轨迹结构特征的权重,取值在 0~1 之间,并满足 $W_D+W_S+W_A+W_L=1$,权重的取值可根据实际情况指定,本文使用 $W_D=W_S=W_A=W_L=0.25$.Form 为归一化函数,由于轨迹结构中每个特征的值域不同,结构距离必须进行归一化处理.在此基础上,轨迹结构相似性表示如下:

$$\mathrm{Sim}(T_i,T_j)=1-\mathrm{Dist}(T_i,T_j) \tag{10}$$

如果 T_i,T_j轨迹 0%-相交,则需要判断轨迹集合中是否存在轨迹 T_k,满足 T_i和 T_k是相交轨迹,并且 T_k和 T_j是相交轨迹.如果存在这样的轨迹 T_k,则轨迹 T_i,T_j间的相似性表示为:

$$\mathrm{Sim}(T_i,T_j)=\sum_{\max T_k}\mathrm{Sim}(T_i,T_k)+\mathrm{Sim}(T_k,T_j) \tag{11}$$

否则,不能度量 T_i,T_j间的相似性,$\mathrm{Sim}(T_i,T_j)=0$.

本文使用无向带权图[20]来描述轨迹间距离,通过求轨迹最短距离来计算轨迹相似性.

定义 8 轨迹距离图(Trajectory distance Graph,TG).轨迹距离图是一个无向带权图,满足:(1)节点代表轨迹;(2)节点 T_i,T_j直接相连当且仅当 T_i,T_j轨迹 p%-相交,$p>0$;(3)T_i,T_j的边权重代表对应轨迹间的距离.

给定轨迹集合,首先将其转化为同步轨迹集合,并计算其中任意 2 条轨迹的距离,构造轨迹距离图.轨迹距离图中任意两点间的最短距离即为对应轨迹的距离.本文使用 Floyd 算法[22]来计算轨迹距离图中任意两

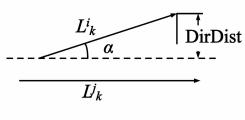


图1 轨迹段方向距离(α<=90)

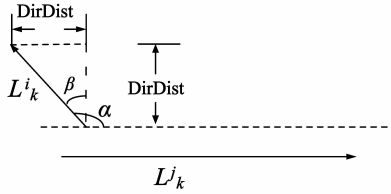


图2 轨迹段方向距离(α>90)

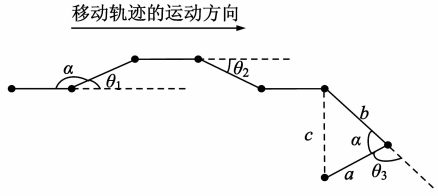


图3 轨迹段的转角

点间的最短距离,然后根据式(10)、(11)计算轨迹间的相似性.

5　稀疏化最小生成树聚类

5.1　稀疏化

尽管任意对象间都有某种程度的邻近性,但对大部分数据集,对象只与少量对象高度相似.据此,可以先对轨迹距离图进行稀疏化处理,只保留移动对象与其最近邻间的连接.不仅有效地处理噪声和离群点问题,更大大地压缩了数据量,提高了处理问题的规模.

定义 9　稀疏化.在实际聚类过程开始前,将许多低相似度(高相异度)的值置为 0(无穷大),这个过程就是稀疏化的过程.

稀疏化的方法有 2 种:(1)断开相似度低于指定阈值的边;(2)保留连接到点的 m 个最近邻的边,创建 m-最近邻图.本文采用第二种稀疏化方法,具体过程见算法 1.

算法 1　Create m-nearest neighbor graph
输入:轨迹距离图 TG(包括同步轨迹集合 D'),最近邻图参数 m
输出:稀疏化轨迹距离图 TG'
BEGIN
1.　TG' = Φ
2.　将 TG 中的顶点加入 TG'
3.　for all $T_i \in D'$　do
4.　　对 T_i与其他轨迹的距离进行升序排序
5.　　将与 T_i距离前 m 的顶点所对应的边加入到 TG'
6.　End for
7.　Return TG'
END

5.2　算法描述

为了满足用户多尺度的隐私需求,降低最小生成树聚类的时间,本文将轨迹 k-匿名集合的选择问题,转换成稀疏化图的个性化 k-节点划分问题,定义如下:

定义 10　个性化 k-节点划分.给定轨迹距离图 TG,TG 的个性化 k-节点划分是一系列不相交的连通子图 $V_1, V_2, V_3, \cdots, V_n$,并满足:(1) $V_1 \cup V_2 \cup V_3 \cup \cdots \cup V_n$ = TG;(2)对于任意 $i(1 \leqslant i \leqslant n)$,有 $k_i \leqslant |V_i| \leqslant 2k_i - 1$,其中,$k_i$是 V_i的轨迹中最大的轨迹隐私级别;(3)对于任意 $i, j(1 \leqslant i \leqslant n, 1 \leqslant j \leqslant n, i \neq j)$,有 $V_i \cap V_j = \Phi$.在个性化 k-节点划分中,$V_1, V_2, V_3, \cdots, V_n$构成匿名集合.

为了最小化匿名信息损失,个性化 k-节点划分要求划分内部的边权重之和最小.本文提出稀疏化最小生成树聚类(the Sparse Minimum Spanning Tree Clustering-SMSTC)算法来近似的划分轨迹距离图,下面给出具体描述.

算法 2　The sparse minimum spanning tree clustering-SMSTC
输入:轨迹距离图 TG,轨迹隐私级别集合 $L_{TG} = \{l_1, l_2, \cdots, l_{max}\}$,最近邻图参数 m
输出:轨迹等价类集合 Q
BEGIN
1.　构造 TG 的 m-最近邻图
2.　$Q = \Phi$;
3.　VS = Φ;//VS 表示没有达到隐私级别的轨迹集合
4.　RS = Φ;//RS 表示删除的轨迹集合
5.　For all $T_i \in$ TG do
6.　　State[i] = False;
7.　End for
8.　$L_{max} = \max(L_{TG})$;//获取当前轨迹集合中最大的轨迹隐私级别
9.　While |TG| $\geqslant L_{max}$ do
10.　　$C = \Phi$;//C 表示某个轨迹等价类
11.　　Mark = 0;
12.　　$TR_l = \{tr_i \in TG \mid l(tr_i) = L_{max}\}$;
13.　　If ($\exists\, e = (V_a, V_b)$,满足:e 的权重最小,$V_a \in TR_l$, State[a] = False, State[b] = False;) do
14.　　{$C = C \cup V_a$; $C = C \cup V_b$; State[a] = True, State[b] = True;}
15.　　Else
16.　　Mark = 1;
17.　　End if
18.　　While |C| < L_{max} && mark ≠ 1 do
19.　　If ($\exists\, e = (V_i, V_j)$,满足:e 的权重最小,$V_i \in C$, State[i] = True, State[j] = False;) do
20.　　$C = C \cup V_j$; State[j] = True;
21.　　Else
22.　　VS = VS $\cup$ {V_a}; Break;
23.　　End if
24.　　End While
25.　　If |C| $\geqslant L_{max}$ do
26.　　$Q = Q \cup C$; TG = TG − C;
27.　　Else
28.　　TG = TG − {V_a}; VS = VS $\cup$ {V_a};
29.　　End if
30.　　$L_{max} = \max(L_{TG})$;
31.　End while
32.　While VS ≠ Φ do
33.　　$L_{max} = \max(L_{VS})$;//获取 VS 轨迹集合中最大的轨迹隐私级别
34.　　$TR_l = \{tr_i \in VS \mid l(tr_i) = L_{max}\}$;
35.　　While $TR_l \neq \Phi$ do
36.　　T_r为 TR_l中任意轨迹
37.　　$TR_l = TR_l - \{T_r\}$
38.　　VS = VS − {T_r}
39.　　If ($\exists\, T$,满足:$e = (T_r, T)$,且 e 的权重最小,$T \in C$, $C \in Q$, C 中轨迹与 T_r相交,$2 \cdot L_{cmax} > |C| \geqslant L_{max} - 1$) do
//L_{cmax}表示集合 C 中最大的轨迹隐私级别
40.　　$C = C \cup \{T_r\}$
41.　　Else
42.　　RS = RS $\cup$ {T_r}
43.　　End if
44.　　End while
45.　End while
46.　Return Q
END

6 轨迹可用性度量

为了准确度量匿名轨迹的可用性,本文用泛化区域面积之和与空间总面积的比值,作为信息损失,来度量匿名轨迹的可用性.表达式如下:

$$\text{InfoLoss} = \sum_{i=1}^{l}\sum_{j=1}^{m}\frac{\text{Area}(x_j, y_j, t_j)}{\text{MaxArea}} + \sum_{i=1}^{h}|T_i| + \sum_{i=1}^{n}|L_i| \tag{12}$$

产生信息损失的原因包括:(1)位置信息泛化;(2)轨迹被删除;(3)位置被删除.在式(12)中,l 代表划分的个数;m 代表该划分中轨迹的位置数量.Area(x_j, y_j, t_j)表示位置(x_j, y_j)在时刻 t_j泛化的区域面积,MaxArea 表示空间的总面积.h 代表删除的轨迹数量;n 代表删除的位置数量;其中,删除轨迹的损失为该轨迹的位置的数量.

7 实验及结果分析

本文使用 2 个数据集来分析算法的可用性:(1)合成数据集.Brinkhoff 轨迹生成器生成的 1000 条合成轨迹,共包含德国奥尔登堡市的 46425 个位置,记为 OLDENBURG;(2)真实数据集.收集自希腊雅典的卡车运送混凝土的轨迹数据集[23],包含 50 辆卡车在 33 天的共 276 条轨迹信息.经过预处理,得到 2708 条轨迹信息,其中每条轨迹平均包含 41.4 个位置,记为 TRUCKS.为了对 SMSTC 算法及其可用性进行分析,本文实现了文献[10]的算法,记为 GKNP 作为对比.

7.1 可用性比较

图 4 和 5 描述了 SMSTC 算法在 OLDENBURG 和 TRUCKS 数据集中 k 取 10 时,随 M(最近邻图参数)变化的信息损失.k 一定时,随着 M 增加,SMSTC 的信息损失先减少后增加.因为在 M 较小时(M < 200),M-最近邻图存储的轨迹距离信息较少,聚类过程中,会造成较大的信息损失;在 200 < M < 600 时,随着 M 增加,信息损失减少的幅度变小,因为轨迹的 M-最近邻图基本上可以表示轨迹间的距离信息;在 M > 600 时,随着 M 增加,信息损失有小幅增长,因为在轨迹距离图中,不是所有轨迹两两之间都能求得距离,轨迹超过一定数量(600),形成的 M-最近邻图不能真实地反应轨迹间的距离.

图 6 和 7 描述了 OLDENBURG 和 TRUCKS 数据集中 M = 100 时,SMSTC 和 GKNP 随最大隐私级别 k 变化的信息损失.随着 k 增加,SMSTC 和 GKNP 的信息损失都在增加,因为随着 k 增加,聚类过程中要考虑的轨迹数量增加,造成信息损失增加;在增加趋势方面,SMSTC 比 GKNP 更加曲折,因为 SMSTC 有机会将未被聚类的轨迹加入到其他类中,降低信息损失;在信息损失大小方面,对于同样的 k,SMSTC 比 GKNP 有更小的信息损失,因为与 GKNP 相比,SMSTC 是个性化轨迹隐私保护算法,它按照轨迹隐私级别,个性化地为每条轨迹定制合适的等价类,有效地降低信息损失,这也佐证了本文提出的轨迹结构相似性能很好地度量轨迹之间的相似性.TRUCKS 数据集的信息损失要大于 OLDENBURG.因为相对于 OLDENBURG 数据集,TRUCKS 数据集更加稀疏,聚类过程中损失信息更多.

7.2 执行时间比较

图 8 和 9 描述了在 OLDENBURG 和 TRUCKS 数据集中,k = 10 时,SMSTC 随 M 变化的执行时间.k 一定时,随着 M 增加,SMSTC 的执行时间逐渐增加,不过增加的趋势减小.因为 M 较小(M < 200)时,M-最近邻图存储的距离信息较少,聚类过程中,算法计算的次数较少,因此 M < 200 时,随着 M 增加,执行时间迅速增加;随着 M 继续增加,执行时间依然增加,不过增加速度减小.

图 10(a)和(b)描述了在 OLDENBURG 和 TRUCKS 中,M = 100 时,SMSTC 和 GKNP 随 k 变化的执行时间.分析图 10(a)发现,随着 k 增加,SMSTC 和 GKNP 的执行时间都呈减小的趋势,因为随着 k 增加,聚类数量减

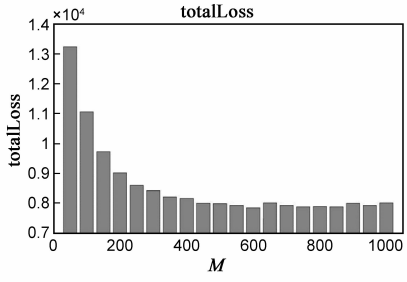


图4 OLDENBURG中随M变化的信息损失

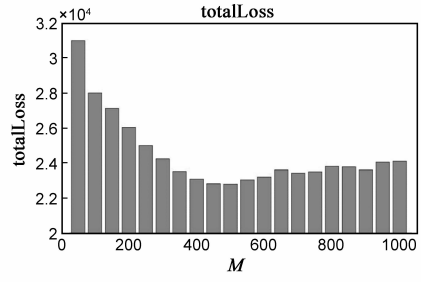


图5 TRUCKS中随M变化的信息损失

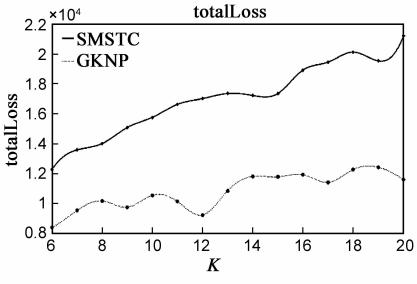


图6 OLDENBURG中随k变化的信息损失

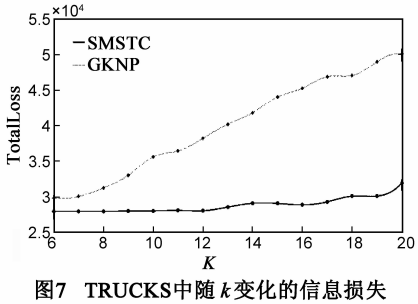


图7 TRUCKS中随k变化的信息损失

少,迭代次数减少,运行时间逐渐减少;在减少趋势方面,GKNP 比 SMSTC 曲折,尤其是 $k=8$ 时,执行时间有一个大的减少,然后上升,这是因为 GKNP 在预处理时产生的开始时间和结束时间近似的等价类,某些由于数量不足,会直接删除,因此减少了运行时间;在执行时间大小方面,对于同样的 k,SMSTC 比 GKNP 具有更小的执行时间,因为 SMSTC 是稀疏化的聚类算法,算法只需进行较小数量的运算,在保持聚类结果近似不变的条件下,显著减小执行时间,在轨迹数量多时效果更加明显. 图 10(b)与图 10(a)相似.

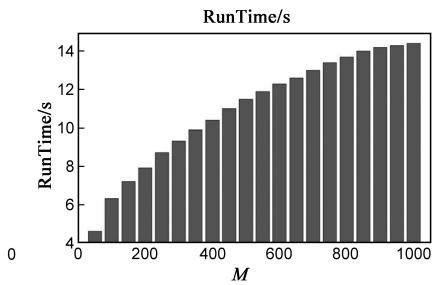


图8　OLDENBURG中随M值变化的执行时间

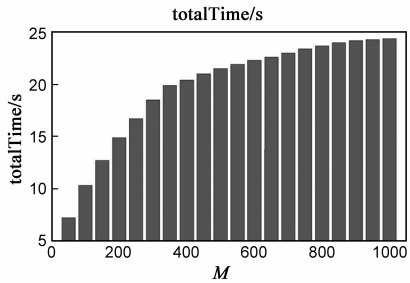
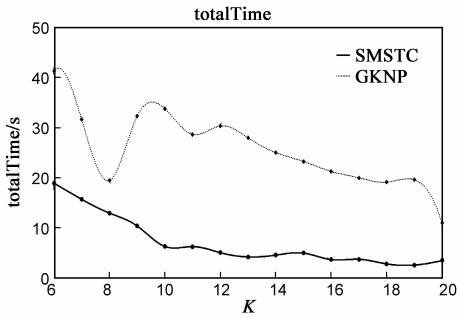
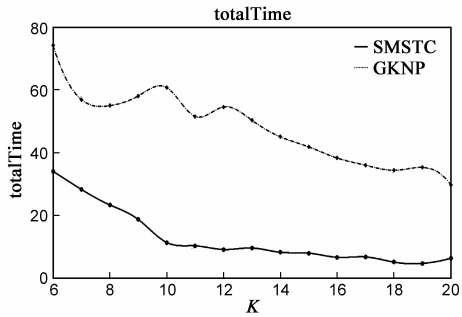


图9　TRUCKS中随k和M变化的信息损失比较



(a) OLDENBURG中随k值变化的执行时间



(b) TRUCKS中随k值变化的执行时间

图10

8　结论

本文研究了轨迹数据发布的个性化的隐私保护问题. 针对现有的轨迹匿名算法忽略了移动对象个性化的隐私需求和轨迹的内外在特征信息,产生的匿名轨迹集合可用性相对较低等问题,本文提出了个性化轨迹 k-匿名的概念,并提出了轨迹结构相似性度量模型,综合考虑轨迹的方向、速度、转角和位置等内外在特征信息;然后提出了基于稀疏化最小生成树聚类的个性化隐私保护算法,通过稀疏化的方法降低最小生成树聚类的执行时间,在有效地保护轨迹数据的同时,显著地提高了轨迹数据的可用性. 最后,在合成和真实轨迹数据集上的实验结果表明,本文提出的算法花费了更少的时间代价,具有更高的数据可用性.

参考文献

[1] 李建中,刘显敏. 大数据的一个重要方面:数据可用性[J]. 计算机研究与发展,2013,50(6):1147 – 1162.
Li Jianzhong, Liu Xianmin. An impotant aspect of big data: Data usability[J]. Journal of Computer Research and Development 2013,50(6):1147 – 1162. (in Chinese)

[2] 刘大有,陈慧灵,齐红,等. 时空数据挖掘研究进展[J]. 计算机研究与发展,2013,50(2):225 – 239.
Liu Dayou, Chen Huiling, Qi Hong, et al. Advances in spatiotemporal data mining[J]. Journal of Computer Research and Development 2013,50(2):225 – 239. (in Chinese)

[3] 韩建民,岑婷婷,虞慧群. 数据表 k-匿名化的微聚集算法研究[J]. 电子学报,2008,36(11):2021 – 2029.
Han Jianmin, Cen Tingting, Yu Huiqun. Research in microaggregation algorithm for k-anonymization[J]. Acta Electronica Sinica,2008,36(11):2021 – 2029. (in Chinese)

[4] 杨高明,杨静,张健沛. 半监督聚类的匿名数据发布[J]. 电子学报,2011,32(11):1489 – 1494.
Yang Gaoming, Yang Jing, Zhang Jianpei. Semi-supervised clustering-based anonymous data publishing[J]. Acta Electronica Sinica,2011,32(11):1489 – 1494. (in Chinese)

[5] 杨静,王波. 一种基于最小选择度优先的多敏感属性个性化 l-多样性算法[J]. 计算机研究与发展,2012,49(9):2603 – 2610.
Yang Jing, Wang Bo. Personalized l-diversity algorithm for multiple sensitive attributes based on minimum selected degree first[J]. Journal of Computer Research and Development, 2012,49(9):2603 – 2610. (in Chinese)

[6] 王波,杨静. 一种基于逆聚类的个性化隐私匿名方法[J]. 电子学报,2012,40(5):883 – 890.
Wang Bo, Yang Jing. A personalized privacy anonymous method based on inverse clustering[J]. Acta Electronica Sinaca,2012,40(5):883 – 890. (in Chinese)

[7] 韩建民,于娟,虞慧群,等. 面向敏感值的个性化隐私保护[J]. 电子学报,2010,38(7):1723 – 1728.
Han Jianmin, Yu Juan, Yu Huiqun et al. Individuation privacy preservation oriented to sensitive values[J]. Acta Electronica Sinaca,2010,38(7):1723 – 1728. (in Chinese)

[8] Abul O, Bonchi F, Nanni M. Never walk alone: uncertainty for anonymity in moving objects databases[A]. Proceedings of the

24th IEEE International Conference on Data Engineering[C]. Cancun, Mexico, 2008. 376 – 385.
[9] Abul O, Bonchi F, Nanni M. Anonymization of moving objects databases by clustering and perturbation[J]. Information. Systems, 2010, 35(8): 884 – 910.
[10] Huo Z, Huang Y, Meng X. History trajectory privacy-preserving through graph partition[A]. Proceedings of the First International Workshop on Mobile Location-based Service[C]. Beijing, China, ACM, 2011. 71 – 78.
[11] Chen L, Özsu MT, 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]. Baltimore, Maryland, USA: ACM Press; 2005. 491 – 502.
[12] Tiakas E, Papadopoulos AN, Nanopoulos A, Manolopoulos Y, Stojanovic D, Djordjevic-Kajan S. Searching for similar trajectories in spatial networks[J]. Journal of Systems and Software, 2009; 82(5): 772 – 788.
[13] Gao S, Ma J, Sun C, et al. Balancing trajectory privacy and data utility using a personalized anonymization model[J]. Journal of Network and Computer Applications, 2014, 38: 125 – 134.
[14] 袁冠,夏士雄,张磊,等.基于结构相似度的轨迹聚类算法[J].通信学报,2011,32(9):103 – 110.
Yuan G, Xia S X, Zhang L, et al. Trajectory clustering algorithm based on structural similarity[J]. Journal on Communications, 2011, 32(9): 103 – 110. (in Chinese)
[15] You TH, Peng WC, Lee WC. Protecting moving trajectories with dummies[A]. 8th International Conference on Mobile Data Management (MDM'07)[C]. Mannheim, Germany: IEEE; 2007. 278 – 282.
[16] Gao S, Ma J, Shi W, et al. LTPPM: a location and trajectory privacy protection mechanism in participatory sensing[J]. Wireless Communications andMobile Computing, 2012.
[17] Terrovitis M, Mamoulis N. Privacy preservation in the publication of trajectories[A]. Ninth International Conference on Mobile Data Management(MDM'08)[C]. Beijing, China: IEEE; 2008. 65 – 72.
[18] Chen R, Fung B, Mohammed N, et al. Privacy-preserving trajectory data publishing by local suppression[J]. Information Sciences, 2013, 231: 83 – 97.
[19] M E Nergiz, M Atzori, Y Saygin, B Guc. Towards trajectory anonymization: a generalization-based approach[J]. Transactions on Data Privacy, 2009, 2(1): 47 – 75.
[20] Josep Domingo-Ferrer, Rolando Trujillo-Rasua. Microaggregation-and permutation-based anonymization of movement data [J]. Information Sciences, 2012. 208: 55 – 80.
[21] Mahdavifar S, Abadi M, Kahani M, et al. AClustering-based Approach for Personalized Privacy Preserving Publication of Moving Object Trajectory Data[M]. Springer Berlin Heidelberg: Network and System Security, 2012: 149 – 165.
[22] R. W. Floyd, Algorithm 97: shortest path[J]. Communications of the ACM 1962, 5(6): 345 – 350.
[23] Frentzos E, Gratsias K, Pelekis N, et al. Nearest Neighbor Search on Moving Object Trajectories[M]. Springer Berlin Heidelberg: Advances in Spatial and Temporal Databases, 2005: 328 – 345.

作者简介

王 超 男,1988 年生于河北省沧州市.哈尔滨工程大学计算机科学与技术学院博士研究生.主要研究方向为数据库与知识工程、数据挖掘、隐私保护.
E-mail: wangchao0605@hrbeu.edu.cn

杨 静 女,1962 年生于黑龙江哈尔滨.哈尔滨工程大学计算机科学与技术学院教授、博士生导师.主要研究方向为数据库与知识工程、数据挖掘、隐私保护、软件理论等.
E-mail: yangjing@hrbeu.edu.cn