基于异步信息的匿名移动数据集的用户身份识别

张宏基 李文中 陆桑璐
(南京大学计算机软件新技术国家重点实验室 南京大学计算机科学与技术系 南京 210093)

摘 要 为了保护用户的隐私,大部分公开数据集都采用隐藏真实ID和引入噪声信息的方法来进行匿名处理。这些匿名处理即使在异步参考信息的攻击下也是脆弱的:即使只有部分位置信息暴露给攻击者并且暴露信息和公开数据集的收集过程不在同一时段内,攻击者依然能够识别出节点在公开数据集中的身份。首先,实验证明已有算法在异步信息情况下不适用;然后,提出针对异步信息的热点矩阵算法。采用3个真实移动数据集验证了识别算法的准确率。实验证明,热点矩阵法在异步信息条件下能够取得远高于已有方法的准确率。

关键词 匿名化,移动路径,异步信息,身份识别,地理空间热点
中图法分类号 TP301 **文献标识码** A

Identifying User's ID from Anonymous Mobility Trace Set via Asynchronous Side Information

ZHANG Hong-ji LI Wen-zhong LU Sang-lu
(State Key Laboratory for Novel Software Technology, Department of Computer & Technology, Nanjing University, Nanjing 210023, China)

Abstract With the development of social network applications, and to meet the demand of mobile system designing and scientific research, plenty of location trace information has been collected and published. Most public traces utilize the anonymous ID and adding noise to protect the privacy of users. However, these processes are vulnerable when facing asynchronous attack. Even if partial information was exposed to adversary and the collections of side information and public trace set are not in the same duration, the adversary can also identify user's ID with high accuracy. Our experiment shows that the existing method is not applicable when facing asynchronous side information. A novel method applicable in asynchronous condition was proposed which is called hot-matrix method. To verify this method, we employed experiments in three different mobility trace sets, whose subject is taxi, bus and human beings respectively. Experiments show that hot-matrix method performs much better than existing approach.

Keywords Anonymilization, Mobility trace, Identification, Geographical hot spots

1 引言及相关工作

目前,苹果应用商店有超过6400个位置相关的应用,Android应用商店有超过1000个位置相关应用,并且这个数字还在继续增长[1]。地理位置相关的社会网络服务也越来越被人们所关注[2-8]。为了协助移动系统和移动应用的设计和研究,许多人和交通工具的位置路径信息被收集并且发布。这些信息可以通过诸如Crawdad[9]这样的专业网站获取,或者也可以从独立的科研机构获取,如Reality Mining[10]、SUVnet-trace[11]。

为了保护用户的隐私,位置路径通常在发布之前进行匿名处理。匿名处理方法包括用随机ID替换真实ID,以及引入噪声等。然而,即使采取这些匿名处理,用户的隐私在攻击者面前依然是脆弱的。

用匿名处理来保护用户隐私这种做法的脆弱性,在文献[12]已经有所讨论。攻击者可以通过诸如相遇观察或者收集社会网络的"check in"功能中的位置信息等方法获取目标节点的部分移动路径信息。我们称被攻击者获取的目标用户的部分移动路径信息为参考信息。文献[12]中提出,基于匿名公开数据集和参考信息,通过一个基于贝叶斯定理的方法,可以比较准确地从匿名数据集中识别出目标节点。因此,通过参考信息,这个用户的完整路径信息可以被攻击者从匿名数据集中获取。

本文主要关注异步参考信息场景,即公开数据集和参考数据在不同时间段被收集。例如,参考数据的收集过程是在公开数据集的收集过程之后一个月才发生的。在现实情况下,这种异步的场景是更常见的,因为公开数据集的收集者在收集过程中会采取措施防止信息泄露,但是当收集活动结束后,一般就不会再采取保护措施。那么现在的关键问题是:基于匿名公开数据集和异步参考信息,能否有效地识别出参与节点的身份?

文献[12]中提到的基于贝叶斯定理的方法是不适用于异

到稿日期:2013-01-28 返修日期:2013-06-04 本文受国家重点基础研究发展规划(973项目)(2009CB320705),国家自然科学基金委创新研究群体科学基金项目(61021062),国家自然科学基金项目(61003213,61073028)资助。

张宏基 男,硕士生,主要研究方向为无线网络、基于位置的服务,E-mail:zhanghongji@dislab.nju.edu.cn;李文中 男,博士,副教授,主要研究方向为移动计算、普适计算、社交网络;陆桑璐 女,博士,教授,主要研究方向为分布式计算、并行处理。

步场景的。要把此方法用于异步场景中必须对它进行改进。一个最直接的改进方法是引入时间偏移量 Δt，通过 Δt 的偏移，使异步的参考信息与公开数据集在时间轴上处于同一时期。不幸的是，在实际的情况下这种改进的效果不是很好。为了验证这种改进的实际效果，我们将改进的贝叶斯方法应用在出租车数据集[9]中，尝试用这种方法来处理只有异步参考信息的情况。识别准确率情况如图 1 所示。横坐标表示参考信息的长度，纵坐标表示识别准确率。可以看到，只有小部分节点能够被成功地识别，准确率在 1% 至 15% 之间徘徊。不管我们怎么调整 Δt 和参考信息的长度，准确率都没能超过 15%。基于这个实验，我们认为基于贝叶斯定理的方法不适用于异步场景，理由如下。在同步参考信息的情况下，参考信息本来就是公开数据集的一部分，可以利用极大似然估计法基于贝叶斯定理进行识别。但是，在异步参考信息的情况下，参考信息和公开数据集合的相关性相对较差，因为人类的行为会随着时间的变化而发生改变，不会准确地重复同一个移动轨迹。极大似然估计法在识别异步参考信息的情况下就自然会失效。

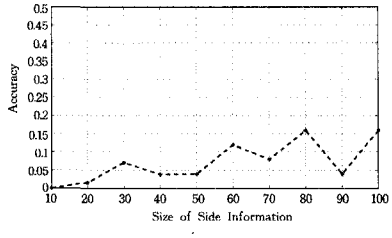


图 1　贝叶斯方法在异步场景中的准确率

据我们所知，还没有已有的工作专门处理异步参考信息的情况。本文将介绍解决异步参考信息情况下识别用户身份的方法，即基于地理热点分布特征的识别方法。首先，将用户节点的移动特征量化为一个二维的特征矩阵，称之为热点矩阵；然后，引入算法来计算特征矩阵的相似度；最后，用特征矩阵的相似度计算两个节点的相似度，即选择一个与参考信息的特征矩阵相似度最高的节点，作为最后的识别结果。我们在 3 个不同主体的真实数据集上进行了实验验证，结果证明在异步情况下我们的方法能够达到较高的识别率。

本文主要贡献如下：

- 提出了由异步参考信息引起的异步情况下匿名公开路径数据集中的隐私问题。
- 提出了能够比较好地解决异步情况下身份识别问题的热点矩阵方法。
- 在大量的真实数据集上，验证了热点矩阵方法的准确性。

本文第 2 节将本篇文章所要研究的问题理论化；第 3 节提出并且描述提取热点矩阵的方法；第 4 节给出计算热点矩阵相似度的方法，并且给出总体的识别算法；第 5 节验证本文所提算法的表现；最后总结全文。

2　问题形式化

本节给出本文相关的假设和问题定义。

2.1　相关定义和假设

在描述本篇文章的问题之前，引入下列概念，如表 1 所列。

表 1　模板图像的性质示例

概念	定义
路径	一个记录移动节点时间和相应位置的集合
节点	移动路径集合中的一条路径
取样时间	一个移动节点的位置信息被记录下来的时间
参考信息	被攻击者所掌握的部分移动位置信息
异步参考信息	参考信息的取样时间和公开数据集合的取样时间不在同一个时间区间内
公开数据集	公众可以自由获取的匿名路径信息集合
受害者	参考信息和 ID 被暴露给了攻击者的节点

在这篇文章中，我们关注如下这个问题：基于匿名移动路径集合和异步的参考信息，能否有效地识别出节点在公开数据集合中的真实 ID。

2.2　问题定义

我们引入下列符号标记来形式化地描述这个问题。

$T=\{(RID_i, T_{i,j}, L_{i,j})\}_{i=1,2,\cdots,N,j=1,2,\cdots,M}$：表示公开数据集合中的一条匿名路径。其中 RID_i 代表第 i 个用户的匿名 ID；$T_{i,j}$ 代表第 i 个用户的第 j 个位置信息的记录时间；$L_{i,j}$ 指 $T_{i,j}$ 所对应的位置，由地理坐标表示；N 是数据集中记录的用户总数；M 代表该条路径中记录的位置信息总数。

$S=\{(TID, \tau_i, \xi_i)\}_{i=1,\cdots,k}$：表示被攻击者观测到的某个受害者的参考信息。其中 TID 指这个用户的真实 ID；τ_i 指参考信息中的第 i 个位置记录的时间；ξ_i 表示相应的位置记录；K 是参考信息中的记录数目，代表着参考信息的大小。

假设 $T_{\min}=\mathrm{MIN}\{T_{i,j}\}_{i=1,2,\cdots,M}$ 并且 $T_{\max}=\mathrm{MAX}\{T_{i,j}\}_{i=1,2,\cdots,N;j=1,2,\cdots,M}$ 是某个路径相应的最小和最大时间。对于异步参考信息，满足对 $\forall i, \tau_i \notin [T_{\min}, T_{\max}]$，即前文所说，参考信息的获取和原路径信息的获取不属于同一时段。

用上述符号，我们的问题可以被定义成如下形式。已知 $T=\{(RID_i, T_{i,j}, L_{i,j})\}_{i=1,2,\cdots,N;j=1,2,\cdots,M}$ 和 $S=\{(TID, \tau_i, \xi_i)\}_{i=1,\cdots,k}$，其中 S 和 T 从不重合的时间区域选择，即 $\{\tau_i\}_{i=1,\cdots,k} \cap \{T_{i,j}\}_{i=1,2,\cdots,N;j=1,2,\cdots,M}=\Phi$，找到一个匿名 ID $RID \in (RID_i)_{i=1,\cdots,N}$ 使得 RID 为 TID 在公开数据集合中对应的匿名 ID。

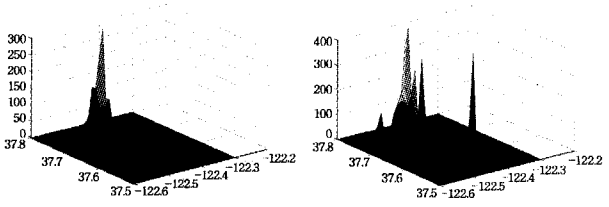
这个问题并不容易解决，在下面章节中，我们将提出一个基于地理位置热点的识别模型，这个模型包括两个部分。第一步，根据路径节点在空间分布上的特性，提取二维空间热点矩阵；第二步，计算二维热点矩阵相似度，并且以此估算路径节点相似度，从公开数据集中选择与参考信息相似度最高的节点作为识别结果。

3　热点矩阵

在异步信息条件下识别目标节点，首先需要一个提取移动模式的方法。已经被广泛认可的事实是，人类活动有着明显的周期性[13]，以天为周期显著地复着某些模式。比如每天早上会起床、去上班、吃午饭、回家吃晚饭等。在位置相关的路径信息中，这些特征反应为地理空间上的一些有自己特色的分布。

为了更好地说明这个问题，我们进行了下面的观察。我们从旧金山的出租车数据集[9]中随机地选择了两个匿名路径，并且计算在 30 天内所访问的空间位置的频率分布，其统计结果如图 2 所示。图中，x 和 y 坐标分别代表经纬度，z 代表节点在这个位置的出现频率。根据我们的统计，观察到如

下几点特征：(1)节点的移动在空间上并不是随机分布。对于每个用户它的移动都被限制在一个相对较小的主要活动区域之内；而对于不同的用户，这个主要活动区域不同。例如，图2中036号节点(见图2(a))主要分布在区域经纬度坐标[37.7,37.8]*[−122.5,−122.4]之间；而129号节点(见图2(b))主要分布在[37.6,37.8]*[−122.5,−122.3]之间。(2)每个节点都有有限个高频访问节点(即图中的极点)，而且不同的节点极点分布不同。例如，图中节点036(见图2(a))有一个极点，而节点129(见图2(b))有多个极点，并且这些极点的位置也各不相同。这些都说明节点的移动有非常高的位置相关性。地理位置特性可以反映出节点的自身特性。



(a) 出租车路径的位置分布统计图　(b) 出租车路径的位置分布统计图

图2

为了在异步参考信息的情况下识别目标节点，我们首先需要找到一个提取移动模式的方法。基于上述分析，我们提出了热点矩阵作为描述节点移动特性的特征矩阵。所谓热点矩阵是一个二维矩阵，基于节点的位置信息而产生。首先，假设节点移动氛围分布在$[LT_{\min},LT_{\max}]*[LG_{\min},LG_{\max}]$之内，其中$LT_{\min}$、$LT_{\max}$、$LG_{\min}$、$LG_{\max}$分别表示维度和经度的最小和最大值。我们将整个区域网格化为边长为δ的正方形，则网格总数为

$$L=M*N=\left\lfloor\frac{LT_{\max}-LT_{\min}}{\delta}\right\rfloor*\left\lfloor\frac{LG_{\max}-LG_{\min}}{\delta}\right\rfloor \tag{1}$$

假设一个位置$L_{i,j}$的对应经纬度坐标为$L_{i,j}=(Lx_{i,j},Ly_{i,j})$，其中$i,j$对应$L$在网格中的位置序号，则

$$i=\left\lfloor\frac{Lx-LT_{\min}}{\delta}\right\rfloor,j=\left\lfloor\frac{Ly-LG_{\min}}{\delta}\right\rfloor \tag{2}$$

如此可以得到如下热点矩阵X：

$$X=[X_{i,j}]_{M*N} \tag{3}$$

其中，i、j、M、N分别按照式(1)、式(2)计算。这样就提供了一种将移动路径信息提取为二维矩阵信息的方法。下面需要考虑的问题是，基于这个二维矩阵，如何计算节点相似度，进而如何识别目标节点。

4 相似度计算及目标识别

至此计算两个路径节点的相似度问题，转换为计算两个二维矩阵的相似度。本节将首先介绍计算热点矩阵相似度的方法；然后基于热点矩阵相似度计算，给出识别目标节点的算法。

4.1 热点矩阵相似度计算

在此我们讨论如何比较热点矩阵相似度的问题。考虑到热点矩阵实际上是一个二维矩阵，等价于一个灰度图像，比较灰度图像的相似度问题已经有一些比较常用的方法。在这里，我们就选择两个比较灰度图像相似度最为常用的方法进行试验。

4.1.1 向量夹角余弦值

将二维矩阵简化为一维向量，用向量夹角余弦值来计算两个二维矩阵的相似度。对于热点矩阵$X=[X_{i,j}]_{M*N}$，按照从上到下从左到右的顺序一次生成一维向量的元素。具体计算见式(4)，将其转化为一维向量$Y=[Y_t]_T$，其中$T=M*N$。

$$Y_{i+j*m}=X_{i,j} \tag{4}$$

向量$y1=[y1_t]_T$和$y2=[y2_t]_T$的相似度按照式(5)，取其夹角余弦值计算。

$$\Psi=y1*y2/(|y1|*|y2|) \tag{5}$$

式中，$|y1|$和$|y2|$分别指代$y1$和$y2$的模。用夹角余弦值计算向量相似度可以将相似度的值约束在[0,1]的区间内，便于进行比较。

4.1.2 频率分布向量

提取二维矩阵的频率分布，构建频率分布向量α，用α的相似度作为二维矩阵的相似度。对于热点矩阵$X=[X_{i,j}]_{M*N}$，其频率分布向量α用来描述$x_{i,j}$取值的分布情况，由于灰度图像中频率分布往往是一个重要的识别特征，因此这里我们也用它来进行实验。具体计算方法见算法1，其中N表示频率分布向量对频率的细分程度。α_N的相似度采用式(5)计算。

算法1 计算热点矩阵的频率分布向量α

输入：热点矩阵 $X=[X_{i,j}]_{M*N}$

输出：频率分布向量 α_N；

$\alpha_N=$zeros(N)；

$e=\lfloor \max\{X_{i,j}\}/N \rfloor$；

for $0\leqslant i\leqslant M$

  for $0\leqslant j\leqslant N$

    $\alpha_N(\lfloor X_{i,j}/e \rfloor)=\alpha_N(\lfloor X_{i,j}/e \rfloor)+1$；

  end for

end for

4.2 目标节点识别算法

在热点矩阵的提取算法和热点矩阵相似度比较算法的基础上，我们给出目标节点的识别算法。为了方便表述，将提取热点矩阵过程抽象为fh，其输入为路径$Trace$，经过第3节中所述方法的处理，得到热点矩阵$HotMatrix$，如式(6)所示。

$$HotMatrix=fh(Trace) \tag{6}$$

同样，将比较两个热点矩阵相似度的过程抽象为函数fs，其输入为两个热点矩阵$M1$和$M2$，经过4.1节中所述方法的处理，得到$M1$和$M2$的相似度Ψ，如式(7)所示。

$$\Psi=fs(M1,M2) \tag{7}$$

具体的识别方法如下。首先，利用fh计算参考信息的热点矩阵和公开数据集中每个路径的热点矩阵；其次，比较公开数据集中每个路径的热点矩阵与参考信息的热点矩阵的相似度，选择相似度最高的节点作为识别结果。其形式化描述如算法2所示。

我们以旧金山出租车数据集[9]来验证这两种相似度计算方法的效果。根据我们的实验结果，方法1的识别率为50%以上，方法2的识别率均不足10%，表现差异巨大。我们认为方法2的表现不佳，有其内在的原因。热点矩阵是一个稀疏矩阵，大部分非零节点聚集在少数位置，并且频率分布相对集中，频率分布向量α的相似度较高，因此由α来计算相似度，难以达到好的效果。所以，在具体实验验证的过程中，我们采用方法1作为我们的最终选择。

算法2 识别算法

输入：异步参考信息Y，公开数据集T_s 输出：目标RID对应的匿名ID

```
RID=0,SIM=0;
M_side=fh(Y);
for each trace X_i∈T_s do
  M_X=fh(X_i);
  If fs(M_side,M_X)>SIM then
    RID=I,SIM= fs(M_side,M_X);
    end if
end for
```

5 实验

本节将介绍我们的实验验证工作，以及在实验中对算法细节的进一步探究。

5.1 实验数据集

在本节的实验中，我们基于3个公开匿名数据集，对我们识别方法的准确率进行了验证。

• Cabspotting[9]：旧金山的出租车数据集，发布在Crawdad网站上，包含近500个节点、30天左右的记录。

• SUVnet-trace[11]：上海交通大学无线传感器网络实验室收集公交车的位置路径信息，包含约100个节点、3个月的记录。

• Reality Mining[10,14]：MIT的Reality Mining组收集以人为主体的位置路径信息，包含约80个节点。

关于这3个数据集的信息总结于表2。

表2 公开匿名数据集信息汇总

主体	节点数目	时间区间	位置
出租车	500	17/05/08—10/06/08	旧金山
公交车	100	28/02/07—16/05/07	上海
人	80	26/07/04—05/05/05	MIT校园

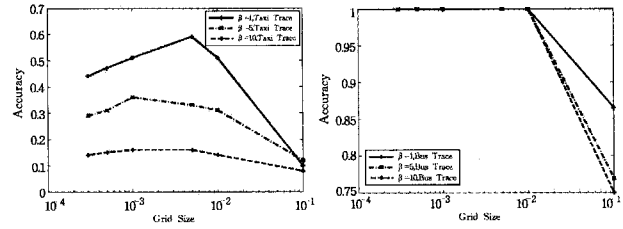
5.2 实验方法

将我们的识别算法应用到这3个数据集中进行验证。首先，从数据集中随机选择一个节点作为目标。从这个节点中选择一定比例的尾部节点作为参考信息，这部分信息从路径中删除，并且将数据集中其他节点晚于这个时间的信息也从其路径中删除，如此即构造成异步参考信息。然后，采用我们的识别方法，用这部分参考信息在公开数据集中进行识别。如果识别结果为原选定路径，则识别成功；否则识别失败。如此，重复此过程1000次，统计识别准确率。

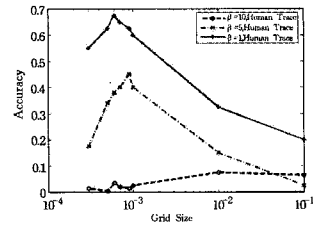
需要注意的是，我们算法的识别准确率与一些参数的选择有关，首先是生成热点矩阵时网格化所采用的网格边长δ；其次是公开数据集合中路径长度与参考信息路径长度的比率β。在下面的内容中，我们将具体地分析这两个参数的选择对结果的影响。

5.3 实验结果及参数分析

首先，我们探讨网格边长δ的变化对准确率的影响。为此，我们选择了$\beta=[1,5,10]$ 3个值，然后分析网格边长δ的变化对准确率的影响。其中，$\beta=1$代表参考信息足够充分，能够形成完整的热点矩阵；$\beta=10$代表参考信息不够充分，参考信息的热点矩阵不完整的情况；$\beta=5$作为折中的参考。实验结果如图3所示。图中，横坐标代表网格边长δ，纵坐标代表对应的识别准确率。从实验结果中可以看出以下几个特点。首先，随着δ的减小，识别的准确率先增加后降低，并且都会在某个值处达到顶点。其次，不同的主体到达顶点的位置不同，但一般出现在$[10^{-3},10^{-2}]$之间，大于或者小于这个区间识别准确率均有可能下降。



(a) δ对准确率的影响(出租车)　(b) δ对准确率的影响(公交车)



(c) δ对准确率的影响(人类)

图3

然后，我们进一步探讨参考信息长度比率β对识别准确率的影响。设δ取值分别为相应数据集的最优点，即出租车数据集取$\delta=5*10^{-3}$、公交车数据集取$\delta=5*10^{-3}$、人类数据集取$\delta=8*10^{-4}$。结果如图4所示，横坐标表示长度比率β，纵坐标表示识别准确率。可以看出，随着β的增加，准确率逐渐降低。出租车和人类数据集在$\beta\leqslant4$时能够达到50%以上的识别准确率；公交车即使在$\beta=10$时依然能够达到近80%的准确率。这是因为出租车和人类的移动路径比较随机，需要更长的参考信息来提取准确的特征矩阵；而公交车的行驶路线相对固定，能够从较少的参考信息中提取特征矩阵。

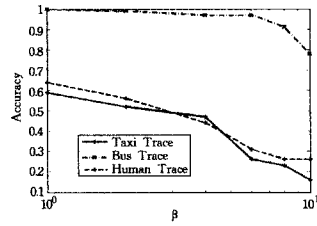


图4 β对准确率的影响

综上所述，网格边长δ和参考信息长度比率β能够对识别的准确率产生一定的影响。根据实验，我们得到如下的参数选择原则：1)网格边长δ在$[10^{-3},10^{-2}]$中取值，不宜过大也不宜过小；2)参考信息长度比率β在条件允许的情况下尽可能小。合理调整识别模型的参数，可以使得热点矩阵方法达到更好的效果，在出租车、公交车和人类的数据集上分别可以达到60%、70%和100%的识别准确率。

5.4 与贝叶斯方法比较

如引言中所述，已有的贝叶斯方法经过修改也可以用于异步参考信息的情况下。通过引入一个时间偏移量Δt可以将异步的参考信息调整到同步时间段内，本节对这种方法的效果进行了细致的实验验证。如图5所示，横坐标代表3个不同的数据集，纵坐标代表识别准确率，深蓝色的柱状图代表修改后的贝叶斯方法，浅蓝色的柱状图代表热点矩阵方法。可以看出，在出租车和人类数据集中，贝叶斯方法最多只能取得15%左右的识别准确率，而热点矩阵方法可以分别取得60%和70%的识别准确率，是贝叶斯方法的3倍。在出租车数据集中，贝叶斯方法最多能够取得78%的准确率，但是热

点矩阵能够取得近100%的准确率。可以得出，热点矩阵方法全面优于已有的贝叶斯方法。

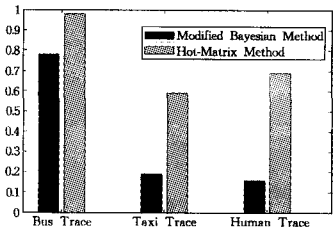


图5 β对准确率的影响

结束语 本文主要讨论了异步参考信息对匿名移动路径信息中节点身份的隐私攻击。我们认为攻击者即使只掌握了异步参考信息，也能够比较准确地识别出目标用户在匿名公开数据集中的ID。我们提出了基于移动位置分布特性的热点矩阵识别算法，并且在3个真实的数据集中对其准确率进行了验证。实验结果表明，热点矩阵识别算法在出租车、人类和公交车数据集上分别能够达到60%、70%和100%的识别准确率，远高于已有的方法。

参考文献

[1] Lba counts in apple store and android marketplace [OL]. http://www.skyhookwireless.com/locationapps/,2010

[2] Quercia D,Lathia N,Calabrese F,et al. Recommending social events from mobile phone location data[C]// Proceedings of the 2010 IEEE International Conference on Data Mining(ICDM'10). Washington, DC, USA, 2010. IEEE Computer Society, 2010:971-976

[3] Sohn T,Li K A,Lee G,et al. Place-its:A study of location-based reminders on mobile phones[C]// Fifth International Conference on Ubiquitous Computing(UbiComp'05). Berlin, Heidelberg, 2005. Springer Verlag,2005:232-250

[4] Gaonkar S,Li J,Choudhury R R,et al. Micro-blog:sharing and querying content through mobile phones and social participation [C]// Proceedings of the 6th international conference on Mobile systems,applications, and services(MobiSys'08). New York, NY,USA,2008. ACM,2008:174-186

[5] Motani M,Srinivasan V,Nuggehalli P S. Peoplenet:engineering a wireless virtual social network[C]// Proceedings of the 11th annual international conference on Mobile computing and networking(MobiCom'05). New York, NY, USA, 2005. ACM, 2005:243-257

[6] NarayananA, Shmatikov V. De-anonymizing social networks[C]// Proceedings of the 2009 30th IEEE Symposium on Security and Privacy(SP'09). Washington,DC,USA,2009. IEEE Computer Society,2009:173-187

[7] Sala A, Zhao X, Wilson C, et al. Sharing graphs using differentially private graph models[C]// Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference (IMC'11). New York,NY,USA,2011. ACM,2011:81-98

[8] The foursquare website[OL]. https://foursquare.com/,2012

[9] Piorkowski M,Sarafijanovic-Djukic N,Grossglauser M. CRAWDAD data set epfl/mobility[OL]. http://crawdad.cs.dartmouth.edu/epfl/mobility,2009-02-24

[10] Eagle N, Pentland A. Reality mining: sensing complex social systems[J]. Personal and Ubiquitous Computing,2006,10(4):255-268

[11] University S J. Suvnet-trace data[OL]. http://wirelesslab.sjtu.edu.cn

[12] Ma C Y,Yau D K,Yip N K,et al. Privacy vulnerability of published anonymous mobility traces[C]// Proceedings of the sixteenth annual international conference on Mobile computing and networking(MobiCom'10). New York,NY,USA,2010. ACM, 2010:185-196

[13] Gonzalez M C,Hidalgo C A,Barabasi A-L. Understanding individual human mobility patterns[J]. Nature,2008,453:779-782

[14] Eagle N,Pentland A S. CRAWDAD data set mit/reality[OL]. http://crawdad.cs.dartmouth.edu/mit/reality,2005-07-01

(上接第73页)

结束语 复杂网络的聚类分析是一个很重要的问题，在很多领域都有广泛的应用。本文提出了一种基于数据场的复杂网络聚类算法，它将物理学的数据场概念引入复杂网络中，通过一种基于互信息的复杂网络节点重要计算方法计算出节点的重要性，然后根据节点的势来划分网络的簇结构。本文提出的聚类算法有效地对复杂网络的簇结构进行划分，并且可以根据实际的需求决定划分的粒度。基于数据场的复杂网络聚类算法虽然在计算精确度和计算复杂度上具有一定的优势，但是过分依赖事先给定的先验条件。下一步应该研究能根据网络的实际情况智能地决定划分条件、不需要人为地给出参数并且有待进一步推广到大型复杂网络中的聚类算法。

参考文献

[1] Brandes U,Delling D,et al. On modularity clustering[J]. IEEE Transaction on Knowledge and Data Engineering,2008,20(2):172-188

[2] 杨博，刘大有，等. 复杂网络聚类方法[J]. 软件学报，2009，20(1):54-66

[3] 孙吉贵，刘杰，赵连宇. 聚类算法研究[J]. 软件学报，2008，19(1):48-61

[4] 淦文燕，李德毅，王建民. 一种基于数据场的层次聚类方法[J]. 电子学报，2006，34(2):258-262

[5] 戴晓军，淦文燕，李德毅. 基于数据场的图像数据挖掘研究[J]. 计算机工程与应用，2004(26)

[6] 陈勇，胡爱群，胡啸. 通信网中节点重要性的评价方法[J]. 通信学报，2004，25(8)

[7] Page L,Brin S. The PageRank Citation Ranking:Bringing Order to the Web[C]// Stanford Digital Libraries Working Paper. 1998

[8] Wasserman S, Faust K. Social network analysis: methods and applications [M]. Cambridge: Cambridge University Press, 1994:218

[9] 傅祖芸. 信息论-基础理论与应用[M]. 北京：电子工业出版社，2001

[10] 胡钢锋，李德毅，陈桂生，等. 一种网络化数据挖掘方法研究[J]. 微电子学与计算机，2006，23(9):126-128

[11] Zachary W W. An information flow model for conflict and fission in small groups[J]. Journal of Anthropological Research,1977,33(4):452-473