基于呼叫详情记录的社会角色推测可视分析

蔡梦杰[1],李学俊[1],王桂娟[1],周　锐[1],谭博友[1],赵韦鑫[1],吴亚东[2]
(1. 西南科技大学 计算机科学与技术学院,四川 绵阳 621010;
2. 四川轻化工大学 计算机科学与工程学院,四川 自贡 643002)

摘　要:城市居民的社会角色感知对城市规划策略制定与城市安全方案设计具有重要的辅助价值,对于后疫情时代疫情的防控具有重要价值。知晓患者用户的角色,可以对用户的接触人群进行更好地分析,做好疫情防控。该文提出了一种结合基站语义和用户时空状态序列的交互式用户社会角色可视分析框架。首先,基于序列数据建模方法,提出了考虑序列顺序的基站嵌入模型 Pos-Cell2Vec 对基站语义信息进行识别;然后,提出一个基于轨迹序列嵌入的用户聚类方法,获得用户聚类结果,进而采用高维可视化方法对基站以及用户的聚类结果进行可视化;最后,基于多视图协同可视分析技术,设计并实现了基于海量通话数据的用户社会角色推测可视分析系统。结合现实数据案例分析结果发现,分析者能够通过该系统结合用户状态序列、用户的通话特征、移动特征以及基站信息,对用户的社会角色进行推测,目前可以通过系统和模型推测出司机、学生以及推销人员等角色。

关键词:呼叫详情记录;社会角色;轨迹嵌入;群体行为模式;用户聚类;可视化分析
中图分类号: TP311.52;TN929.5　　**文献标识码**:A　　**文章编号**:1673-629X(2023)01-0165-08
doi:10.3969/j.issn.1673-629X.2023.01.025

Visual Analysis of Social Role Projections Based on Call Detail Records

CAI Meng-jie[1], LI Xue-jun[1], WANG Gui-juan[1], ZHOU Rui[1],
TAN Bo-you[1], ZHAO Wei-xin[1], WU Ya-dong[2]
(1. School of Computer Science & Technology, Southwest University of Science & Technology, Mianyang 621010, China;
2. School of Computer Science & Engineering, Sichuan University of Science & Engineering, Zigong 643002, China)

Abstract: The perception of the social roles of urban residents is an important aid to the development of urban planning strategies and the design of urban safety programmes, as well as to the prevention and control of epidemics in the post-epidemic era. By knowing the roles of patient users, a better analysis of the user's contact group can be achieved for epidemic prevention and control. In this paper, we propose a framework for interactive user social role visibility analysis that combines base station semantics and user spatio-temporal state sequences. Firstly, a base station embedding model Pos-Cell2Vec is proposed based on the sequence data modelling approach to identify the semantic information of base stations. Then, a user clustering method based on trajectory sequence embedding is proposed to obtain the user clustering results, and a high-dimensional visualization method is used to visualize the clustering results of base stations and users. Finally, based on the multi-view collaborative visual analysis technique, a visual analysis system for user social role inference based on massive call data is designed and implemented. The results of the analysis combined with real-life case studies show that the system allows analysts to infer the social roles of users by combining the user state sequences, their call characteristics, mobile characteristics and base station information.

Key words: call detail records; social role; trajectory embedding; crowdy behavior pattern; user clustering; visualization analysis

0　引　言

随着移动电话的普及,大规模的通话数据给人们提供了研究城市结构和动态的机会。王桂娟等人[1]归纳了通信数据的来源、特征以及数据处理方法,并总结了基于通信数据的城市可视分析任务方法和特点。城市中的基站,能够侦测和记录人类的移动以及通信行为。通过对城市中所有移动通话基站使用记录进行获取和分析,管理者能够获悉每个用户的行为模式以及

收稿日期:2022-01-12　　**修回日期**:2022-05-17
基金项目:国家自然科学基金资助项目(61802320,61872304)
作者简介:蔡梦杰(1994-),男,硕士研究生,研究方向为城市计算可视化;通讯作者:王桂娟(1981-),女,博士研究生,研究方向为自动可视化、可视化与可视分析。

用户标签,能够有效地对用户进行精准营销,而且大规模的用户通话记录能够以感知人类的行为模式为基础辅助城市管理者进行交通规划[2]。由于社会角色是指在社会生活中与位置相关联的一套个人行为模式,所以可以通过对用户行为模式的研究来探索用户的社会角色。通过使用智能算法对大规模通话记录进行分析,分析者能够有效地捕捉城市居民的潜在行为模式,并且行为模式能够反映用户的社会角色。在此基础上结合可视分析方法,分析者能够在与系统交互的过程中解释用户行为以及探索用户的社会角色。

人类活动分析对人的社会角色分析具有较大的参考意义,Zhu 等人[3]基于时序序列向量化技术设计并开发了基于 B/S 架构的可视分析系统,该系统能够对手机用户的位置进行预测,基于可视分析以及视觉隐喻的方法对城市的交通状态进行呈现以及评估,帮助城市管理者对大规模人群的移动性进行分析。李致昊等人[4]设计并开发了 Trajectory2Vec 系统,他们使用文本分析中的主题识别模型对城市每个区域的功能进行识别,在此基础上对城市中人群的跨区移动性进行宏观的探索分析。Cao 等人[5]提出了一种新型的基于张量的异常分析算法,该算法具有可视化和交互设计,可以动态地产生上下文的、可解释的数据摘要,并允许根据用户的输入对异常模式进行交互排序,由此来分析人群的行为模式。

在稀疏轨迹研究方面,Zheng 等人[6]对社会中涉及移动的数据进行了整理研究,针对以出租车数据为主的密集数据以及以通话数据为主的稀疏数据分别进行了总结以及对比,并提出了一个能够提高轨迹精度的数据融合框架。由于基站只会在用户与基站之间触发通信活动的时候记录用户的行为,GPS 会定时对实时位置进行记录生成密集的轨迹信息,相较于手机的 GPS 地理轨迹,基站轨迹的信息量较少,信息密度较低,不确定性较大[7]。用户基站轨迹属于稀疏轨迹,稀疏轨迹的不确定性给用户行为分析带来了一定的挑战。但如果用户数量足够大,记录时间跨度够长,稀疏轨迹信息密度过低的缺陷可以被弥补,合理地对基站序列建模能够捕获用户的宏观移动行为和模式,进而识别其用户角色。Al-Dohuki 等人[8]提出了一种基于词嵌入的出租车轨迹建模方法 SemanticTraj,SemanticTraj 可以直观高效地来管理和可视化出租车轨迹数据。该方法会对出租车的轨迹进行编码处理,处理为文档中句子的形式,然后会基于文本查询相关的算法对大规模的轨迹进行挖掘和分析。关海潮[9]基于文本编辑距离对用户之间的轨迹序列相似度进行计算,接着基于相似度的用户聚类分析对异常用户进行发现。但以上的轨迹序列建模或者嵌入的方法都没有考虑站点轨迹中的顺序问题,而轨迹的顺序对于其含义十分重要。该文基于通话记录数据对社会角色进行推测,对于稀疏轨迹采用了一种基于序列的方式来进行研究,并考虑了顺序问题。

在用户角色分析方面,胡亚慧等人[10]将用户角色定义为能够使用区域访问频次、区域亲密度、区域跳变性等方面进行识别的用户属性,例如用户的职业等。王峰等人[11]通过对用户所发微博提取必需的情景要素,对用户的移动规律建立用户角色与城市地域结构的互推断模型。在社交行为方面,Lee 等人[12]定义用户角色为能够使用社交行为进行识别的用户属性,例如用户的社交影响力、用户的社会地位等。用户角色决定了其移动和社交特征,反过来,可以基于移动特征和社交特征反向推导用户的社会角色。用户在城市中整体移动频率,移动范围,通话频率乃至通话时间的分布都与用户职业或者角色高度相关。结合以上的信息,分析者能够有效地推断用户的角色,从而更具针对性地对城市进行管理规划。

该文的目标是对社会角色进行分析,使用词嵌入以及位置嵌入共同构建基站语义的处理模型,然后结合可视分析对用户的社会角色进行分析。提出的基于用户轨迹和基站语义的社会角色推测可视分析方法的总体实现框图如图 1 所示。

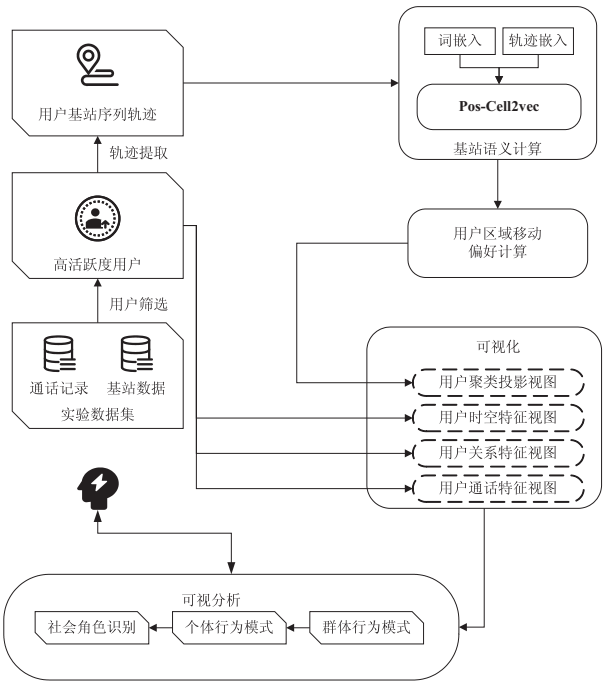

图 1 基于用户轨迹和基站语义的用户社会角色推测可视分析流程

1 数据处理和方法

实验数据由中国某市的匿名移动通信运营商提供,包含某市半年中 2 万左右手机用户的话单数据,此

话单数据集包含通话开始时间、通话持续时间、主叫用户 ID、被叫用户 ID、通话服务基站等,在该数据集中参与服务基站有 1 万左右,位置覆盖该市的所有地区。

1.1　轨迹处理及高活跃度用户筛选

张兰云等人[2]对话单数据进行压缩并对通话次数进行筛选,提取用户的轨迹,选择高活跃度用户,该文借鉴这种思想进行轨迹提取和高活跃度用户的选择。该文关注的是用户的行为模式而非每一次通话记录,这样就需要对用户的冗余通话记录进行处理,使得用户基站序列中的相邻元素各不相同。所以,对用户基站轨迹进行如下定义:假设用户 u 的第 i 次通话是经由基站 c_i 进行服务的,那么该用户在某一段时间 T 内的轨迹可以表示为:$\tau = \{c_1, c_2, \cdots, c_n\}$ 。

用户稀疏的行为序列会导致建模较大的不确定性。所以要对用户进行筛选,这里对用户每周访问基站的频次进行了计算。统计结果为:用户一周基站访问频次最小值为0,最大值为37,平均值为9,中位数为10。将具有低活跃度的用户进行排除。该文将一周接打电话频次为 6 及以上的用户作为高活跃度用户,并使用这些用户的基站轨迹作为训练样本对提出的考虑基站序列顺序的基站嵌入模型进行训练。这样不仅可以兼顾用户规模,还可以保证模型的准确性。

1.2　考虑序列顺序的基站嵌入模型 Pos-Cell2Vec

该文提出的 Pos-Cell2Vec 模型是基于词嵌入和位置嵌入模型。结合两种模型,对用户的基站轨迹信息进行识别分析。Word2Vec 采用的语义模型是 N-Gram 模型+词袋模型,也就是说假设一个单词只与周围若干个单词有关且不考虑单词间的顺序关系。而该文提出的 Pos-Cell2Vec 模型能够考虑人群稀疏轨迹的顺序,从而提高模式识别的准确率以及模式的多样性。

(1)词嵌入模型。

首先是词嵌入部分,词嵌入通常定义一个映射 $f_{we}: \mathbb{N} \to \mathbb{R}^D$ 从离散的词索引到 D 维的实值向量,并且 $\mathbb{N} = \{0,1,2,\cdots\}$ 。借鉴词嵌入方法的思想,该文将高活跃用户轨迹中的基站视为词语,轨迹视为句子。而轨迹嵌入就是定义一个映射 $f_{we}: \mathbb{N} \to \mathbb{R}^D$ 从离散的基站索引到 D 维的实值向量,并且 $\mathbb{N} = \{0,1,2,\cdots\}$ 。这里将构建轨迹嵌入模型将手机用户的基站轨迹转换为向量。

通过以下几个步骤保证 Pos-Cell2Vec 模型的正确训练。

第一步,从现有的用户轨迹中产生足够的正样本,将 Sliding Window 设置为 n ,也就是设置滑动窗口,这表示模型在对某个基站进行预测的时候,该基站的前面和后面的 n 个基站都是与它相关的上下文。当 Sliding Window 滑动的时候,模型就会利用此时出现在 Sliding Window 中的所有基站创建一个正样本。每滑动一次都会产生一个正样本。简而言之就是使用 CBOW 模型的思想,将基站 c_{n-1} 和基站 c_{n+1} 作为上下文输入,最终得到中心基站 C_n 。

第二步,对目标函数进行定义和优化,该文使用最大似然估计的方法对目标函数进行优化,目的是为了样本的条件概率之积最大化。综上,定义 Pos-Cell2Vec 模型的目标函数如下:

$$\frac{1}{M}\sum_{i=1}^{M}\sum_{-n\leqslant j\leqslant n, j\neq 0}\log k(c_{i+j} \mid c_i) \tag{1}$$

式中,M 表示样本大小,n 表示基站序列的长度。

由于关于中心基站的预测属于多分类问题,使用 Softmax 函数作为条件概率函数,即 $\log k(\cdot) = \mathrm{softmax}(\cdot)$ 。该模型会对每一个基站语义信息进行处理,每一个基站的语义信息都可以使用一个向量进行表示。那么两个基站之间的相似性可以使用欧氏距离进行衡量。于是,中心基站的条件概率的定义为:

$$k(a_i \mid a_j) = \frac{\exp(v'^{\mathrm{T}}_{a_i} v_{a_j})}{\sum_{a=1}^{A}\exp(v'^{\mathrm{T}}_{a_i} v_{a_j})} \tag{2}$$

该模型可以为每一个基站分配一个定长的向量,其中向量中的每一个维度表示相应基站某个语义的对数概率,简而言之表示的是该基站与这个语义之间的关联程度。这里将序列窗口大小 n 设置为 2,语义向量长度 N 设置为 100,通过使用高活跃度用户的基站序列,Pos-Cell2Vec 模型可以将所有的基站处理为 100 维的语义向量。

(2)位置嵌入模型。

然后是位置嵌入部分,以上的词嵌入方法并没有考虑句子中词语的顺序,基于同样想法的基站轨迹嵌入方法也没有考虑基站在轨迹中的顺序。顺序这个信息对于处理时间序列数据来说非常重要,它可以表示局部的结构甚至可以表示全局的结构。不仅是时间顺序位置顺序同样重要,它们两个并不严格相同。如果没有学习到顺序信息,将会影响到学习的效果。根据常识可知,基站在轨迹中的访问顺序具有重要的意义,不同的顺序可能具有截然不同的意义,例如从学校到住宅的语义是放学,而从住宅到学校的语义则是上学。该文希望对基站在轨迹中的位置信息进行建模从而提高基站语义识别的准确性,因此,引入位置嵌入模型是为了能够让词包含位置的信息,从而让 Pos-Cell2Vec 能够更好地处理基站轨迹信息。

位置嵌入(PE)类似于词嵌入(WE),位置嵌入则是定义了另外一个映射关系 $f_{pe}: \mathbb{N} \to \mathbb{R}^D$,从离散位置索引映射到向量,能够更加准确地捕获基站之间的关系。该文使用 Gehring 提出的位置嵌入算法[13]对基

站在轨迹中的位置进行编码。对于基站位置的编码有两种方式,一种是使用 one-hot 独热向量输入到 embedding 层获得连续的向量,另外一种是使用公式进行计算。该文使用第二种方式进行位置编码计算,奇偶位置情况需要分开讨论,在生成位置的向量中,如果是偶数就使用公式 3 中的第一个公式,如果是奇数就使用公式 3 中的第二个公式。

$$\begin{cases} \mathrm{PE}(\mathrm{pos},2i)=\sin\left(\dfrac{\mathrm{pos}}{10\ 000^{2i/d_{\mathrm{model}}}}\right) \\ \mathrm{PE}(\mathrm{pos},2i+1)=\cos\left(\dfrac{\mathrm{pos}}{10\ 000^{2i/d_{\mathrm{model}}}}\right) \end{cases} \tag{3}$$

其中,d 是位置嵌入结果的向量长度,pos 是基站在序列中所在的位置,i 是 embedding 维度,例如词向量的长度为 512,则 i 属于 512。参数 i 指用于分别奇偶情况。由于基站映射关系 f_w 和位置嵌入关系 f_p 都仅取整数值作为单词索引或位置索引,因此单独训练单个单词或位置的嵌入矢量。每个单词向量的独立训练是合理的,因为单词索引基于给定任意词汇的顺序,并且不会捕获与其相邻单词之间的任何特定顺序关系。

(3)Pos-Cell2Vec 模型。

最终,结合词嵌入模型和位置嵌入模型,Pos-Cell2Vec 模型对用户轨迹的编码如公式 4,用户最终的轨迹编码是词嵌入和位置嵌入共同编码获得。位于句子中第 pos 个基站 c_j 的最终嵌入结果可以表示为:

$$f(c_j,\mathrm{pos})=f_{\mathrm{ce}}(c_j)+f_{\mathrm{pe}}(\mathrm{pos}) \tag{4}$$

其中,$f(c_j,\mathrm{pos})\in\mathbb{R}^D$,最终的基站轨迹编码结果能够同时包含基站嵌入信息和位置嵌入信息。在使用词向量构建序列向量的时候无法考虑基站的顺序,Vaswani 等人[14]经过实验验证使用正弦函数能够使每个位置都提供一个唯一的向量,这样可以让模型通过对应位置来学习。只需要将词向量与位置向量叠加便能得到嵌入位置信息的词嵌入结果。

1.3 基于用户轨迹嵌入的聚类方法

上一节在使用 Pos-Cell2Vec 算法对用户的基站移动轨迹进行建模之后,本节根据用户的历史移动轨迹,结合上一章的用户轨迹嵌入方法提出了一个聚类方法。首先,该方法会创建用户移动轨迹喜好向量,然后,使用 t-SNE 方法对该向量进行降维分析。该算法的流程如算法 1 所示。第一,它会考虑用通话记录中出现的所有基站;第二,每个基站的权重参数使用用户对每个基站使用的频率表示;第三,对每一个基站语义向量的对数概率进行加权求和处理,最终获得轨迹向量 τ;第四,使用 Softmax 对轨迹向量进行标准化处理;最后,使用 T 分布随机近邻嵌入(t-SNE)[15]根据用户的区域移动偏好向量对用户基站序列进行处理。该函数最终的输出就是需要得到的用户轨迹嵌入向量。t-SNE 算法的降维结果的特征是区域移动偏好类似的用户会在可视化结果中处于相邻的位置,算法如下:

算法 1:基于 Pos-Cell2Vec 模型用户有效基站语义偏好向量的聚类算法。

1. 记函数 Pos_EB 为基站位置嵌入函数;函数 Cell_EB 为基站语义向量的查找函数;List 为一个空数组,变量 $\bar{v}$ 为全零向量,$|\bar{v}|=N$;

2. For u in 手机用户集合 U:

3. 记 T_{u_i} 为手机用户 u_i 的基站轨迹;

4. For c in pos j of T_{u_i}:

5. v_c =Cell_EB(c)+Pos_EB(c,j), | v_c | = N;

6. $\bar{v}=\bar{v}+v_c$;

7. $V_{\mathrm{sum}}=\mathrm{softmax}(\bar{v})$;

8. List[i] = V_{sum};

9. PL=t-SNE (List);

10. return PL /* PL 是所有用户坐标的数组,后续的可视化交互视图将使用该算法输出的数组 */

2 可视分析系统设计

对于该文的分析任务,设计了基于通话数据的用户角色推测及群体行为模式可视分析系统,如图 2 所示。图中,A 部分是用户轨迹嵌入结果使用 t-SNE 算法的聚类投影结果;B 部分是用户的时序相关的特征以及移动相关的特征统计结果;C 部分是用户行为甘特图,用以呈现用户在长时间区间内的基站访问情况;D 部分是基站列表,呈现了部分重要基站在不同小时的使用频次特征;系统的基础地理信息图层,即 E 部分是基站分布及用户移动特征地图。接下来会对每个部分的设计以及功能任务进行描述。

2.1 用户轨迹嵌入投影

用户轨迹聚类投影不好进行呈现,该文设计了一个可交互的探索视图,它是一个带有坐标信息的圆盘,上面的散点表示所有手机用户(如图 3 所示),每个散点的 x 和 y 坐标都是由基于用户轨迹嵌入的聚类方法处理得到。可以通过刷选的选择方法进行交互,用户可以通过鼠标控制一个小圆盘(小圆盘的半径可以调整),然后对感兴趣的点进行刷选,刷选后的散点将使用红色着色,作为候选兴趣用户。选择完成之后,分析者选择的兴趣用户将顺时针排列在视图中取代用户轨迹嵌入结果,如图 3 右所示。如果用户之间有通话联系,将使用虚线将他们之间进行连接。此外,多维特征统计图将呈现该用户群体的相关特征值。在分析者点击某个用户之后,地图视图将使用热力图显示兴趣用户的移动范围,散点图标注所有到访过的基站,基站列表和用户活动甘特图将根据选中用户的相关信息进行更新。使用者点击用户聚类视图中的中心能够返回到

聚类可视化结果。

2.2　用户多维特征统计结果可视化

该文希望呈现用户时序和空间相关的多维统计特征,帮助分析者对用户角色进行分析,如图 2B 所示。在系统用户选择感兴趣的手机用户之后,矩阵图将呈现其时序特征。左矩阵图中的散点用以表征一个用户在日间和夜晚的通话频率对比,右矩阵图中散点用以表征一个用户的工作日以及休息日之间的对比情况。对于空间相关的特征,使用单维散点图对兴趣用户的每日平均移动次数、每日平均移动举例等特征进行呈现。使用者点击基站列表中的某个基站能够在地图中高亮其所在的位置,并在甘特图中高亮与其相关的移动事件。

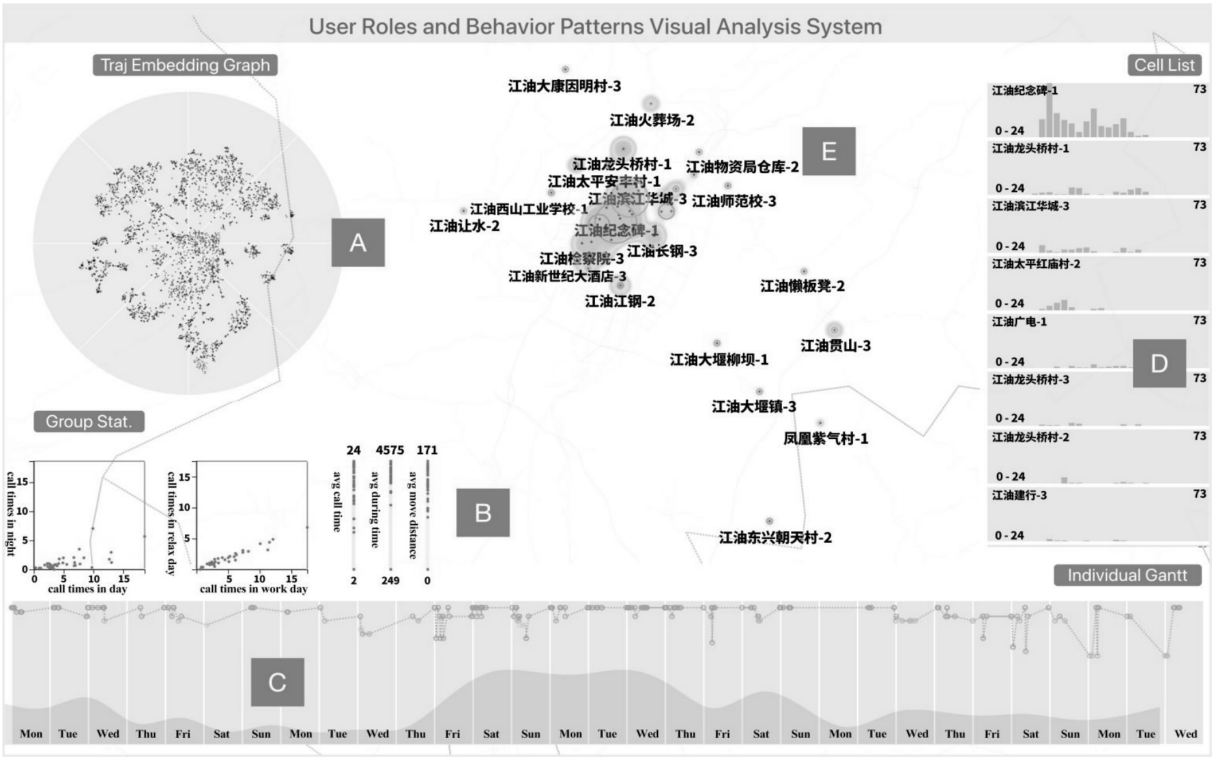

图 2　基于通话数据的用户角色推测及群体行为模式可视分析系统

2.3　用户时空特征可视化

手机用户在不同基站上的时序使用偏好能够反映出基站对于用户的功能,该系统采用基站列表(图 2D 部分)对用户在高相关基站上的时序访问特征使用频率图进行可视化。

地图将对用户悬浮的基站使用 Icon 进行展示,帮助其对基站相关信息进行连结。通过观察该列表,分析者能够快速发现用户高频访问的基站以及基站相对应的功能。

2.4　用户活动甘特图

该系统使用甘特图(图 2C 部分)对用户长时间的行为记录进行可视化。在用户行为活动甘特图中,纵轴表示不同的基站或者基站所属的聚类,横轴代表时间的递进。甘特图中的连接线用于辅助呈现通话事件之间的时序关系。在这个空间中,使用圆环描述用户的每一次移动事件,圆环的颜色基于基站或者基站所属簇 ID 决定。甘特图的下层呈现了以用户频次为高度的面积图,用以描述用户在整个时段内的移动频次变化。

2.5　用户活动空间特征可视化

基于 DBSCAN 算法提出地图放缩级别敏感的基站算法 Zoom Level Sensitive (ZoLeSe) – DBSCAN,该算法至少需要两个参数(ε ,MinPts)用来描述邻域的样本分布紧密程度。在设置了初始参数之后,参数 ε 能够跟随地图当前的放缩级别进行缩放。如图 4 所示,该方法能够自动聚合在当前放缩因子下地理位置相近的基站,接着计算这些基站的中心位置作为聚合基站的新位置,最后提取被聚类基站名字的公共子串作为聚合基站的新名字。为了呈现基站的地理特征以及手机用户的移动行为特征,该文基于 Mapbox. GL 框架对地图进行开发。在兴趣用户被确定后,地理信息视图(图 2E 部分)会提取该用户群体频繁访问的基站,并将其展示在地图中,基站之间的语义相似度也被使用散点图进行呈现。一方面,采用动态基站缩放聚类的方式对基站相关信息进行刻画以帮助分析者对手机用户的移动行为模式进行多层次分析。另一方面,系统亦使用热力图对兴趣用户相关的地理空间分布进行展示,从而帮助分析者快速获取兴趣用户的行为活动特

征和移动趋势。

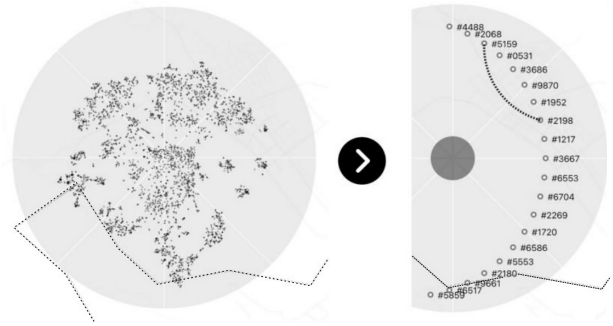

图 3 基于用户轨迹序列的用户嵌入投影结果图

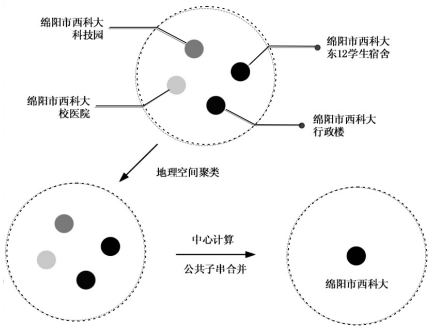

图 4 基于 ZoLeSe-DBSCAN 的基站聚类方法

3 案例分析

为了展示该系统在对角色识别以及行为模式发现上的能力,本章选择的是一个月的数据结合三个案例对根据移动偏好、根据时空特征以及营销角色的有效性进行验证。

3.1 结合用户时空特征的角色发现

该案例对文中方法结合用户的时空特征对用户的角色发现的能力进行评估,从而进一步验证文中方法的可行性。首先,在用户聚类视图中选择具有聚类倾向的一组用户。从图 5 中的多维特征统计图中可以发现,该组用户的通话频率相对较高,每日平均移动距离也较大。从中随机选择用户 U1132 对其进行详细的剖析,如图 5 所示。

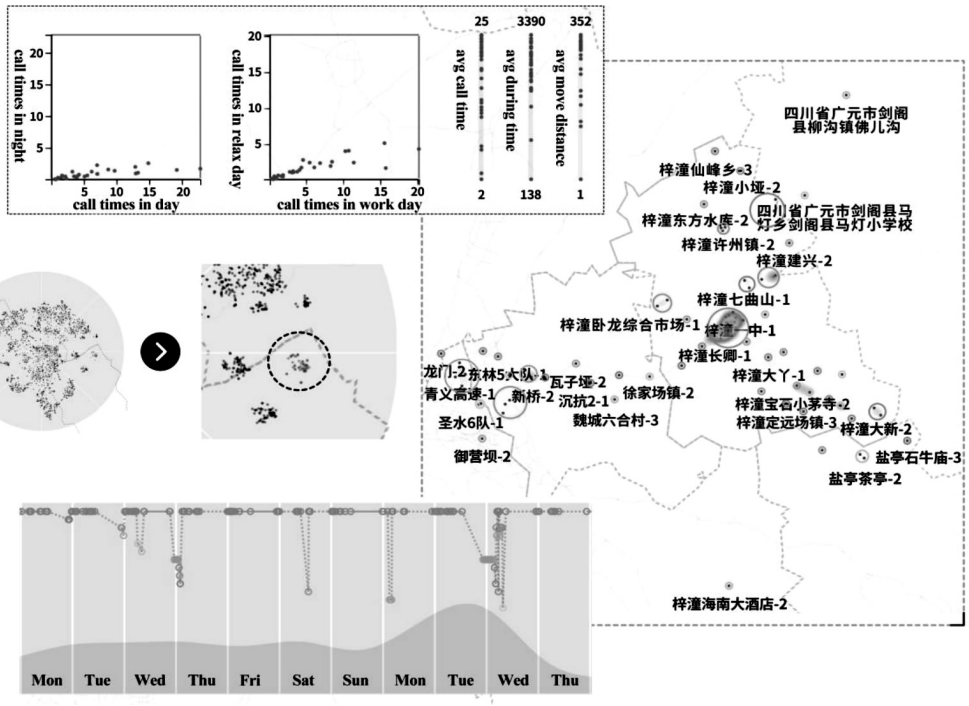

图 5 结合用户时空特征的角色发现

从地图中可以发现,该用户的移动区域非常广,其中访问最频繁的区域是梓潼的城区,主要活动区域都是从此为中心向外辐射。接着,观察该用户在用户活动甘特图上的表现可以发现,该用户在工作日和非工作日会频繁地离开城区,空间转移的频率非常高,并且空间的转移在地图上呈现出显著的连续性特征。可以推断,该用户群体很可能是运输相关的司机。通过网络上公开的信息对该手机用户号码进行查询,发现该用户为梓潼市的出租车司机,从而印证了推断。这个案例有效地证明了文中方法对群体用户以及单个用户的行为模式的识别以及解释能力。

3.2 结合用户移动偏好的角色发现

该案例对文中方法结合用户移动偏好对用户的角色发现的能力进行评估,从而验证文中方法的可行性。选择大学生这个具有显性行为模式的群体进行分析。因为校园一般都有时间管制措施,学生必须在这个时间内返校或者出校、大多数学生统一上课下课,在下午 5:00-7:00 之间出入校园的大多是学生群体。所以选

择这个时间段常在大学区域的用户进行筛选,对用户的行为模式进行观察,结合多维统计视图可以发现,该组手机用户在白天通话频率高于夜晚,工作日通话频率高于休息日,表现出较为自由的行为特征。进一步对该组用户的角色进行分析,从该用户群中选择用户 U3719 对其详细的情况进行分析。图 6 呈现了该用户时空相关的特征。结合用户行为甘特图可以发现,该用户在工作日活动的位置较为固定,其出现在大学附近的概率远高于其他位置,平均通话次数相较休息日更高。该用户白天主要出现在教学楼附近,在夜晚长时间停留在学校且基站大多靠近宿舍附近,可以推测该用户具有较大的概率是学生。同时,从地图中能够发现,该组用户工作日主要停留在大学附近,而在休息日常访问火车站、商业区等,这是由于该类用户的消费频率和出行频率较大且周末的时间较为自由。

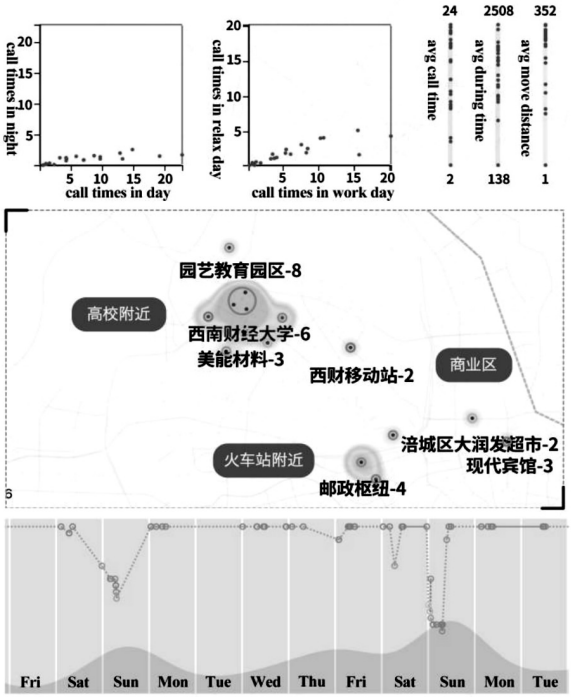

图 6　结合用户移动偏好的角色发现-学生

从该行为特征来看,该用户十分符合大学生群体的行为模式。该案例有效验证了文中方法对城市区域人群的行为模式的识别以及解释能力。

3.3　营销及骚扰用户识别

该案例通过先验知识对文中设计的可视分析系统发现异常用户的能力进行验证评估。异常用户的定义就是那些频繁出现异常行为的用户,比如经常进行大量短通话,在特定的时间进行大量的短通话,移动行为没有规律,每次都会在特定的地点进行连续大量通话。

在用户聚类视图中选择一组比较偏离中心的用户群,该组用户相关的可视化结果呈现在图 7 中。从多维统计特征视图中可以发现,该组用户的通话频率十分高,且平均通话时间较短。从中选择手机用户 U8193 进行观察,从图中地图部分结合基站列表可以发现,该用户移动区域较为固定,主要集中在 M 市的三台县内。观察用户行为甘特图可以发现,其运动模式十分单一,通话频率十分高。根据以上的线索可以推测,该组用户的活动模式与营销及骚扰电话的拨打者十分匹配。该案例对文中系统对于营销及骚扰用户的识别的能力进行了有力的验证。

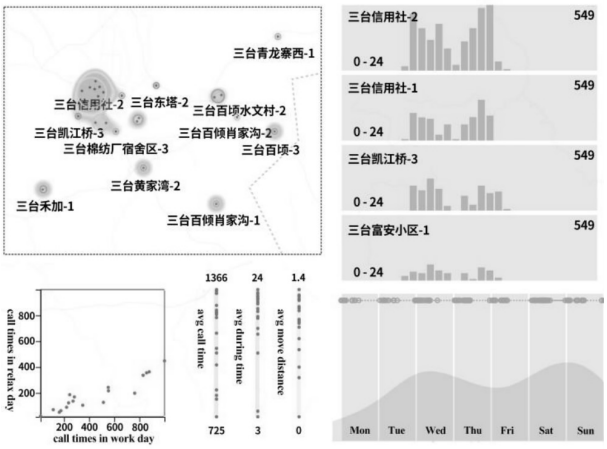

图 7　结合用户移动偏好的角色发现-营销/骚扰人员

4　模型对比

Pos-Cell2Vec 模型相较于 Cell2Vec 模型,能够考虑序列的顺序,理论上能够更加准确地捕获用户的移动行为模式。为了验证算法在用户移动行为模式捕获上是否有有效的提升,选择了 100 个学生用户,使用用户之间的相似度作为算法准确度的度量对 Pos-Cell2Vec 与 Cell2Vec 模型进行对比。该度量由以下公式计算得到:

$$S = \sum_{i,j \in N}^{N} \frac{|v_i \odot v_j|}{N} \tag{5}$$

其中,N 为用户集合,v_i 为用户 i 的轨迹嵌入向量,符号 ⊙ 表示两个向量间的欧氏距离,该度量值越小表示这些用户的移动行为模式越相似。已知这些用户拥有同样的社会角色,这意味着用户应该有更高的相似度,也就是通过该模型得到的度量值越小,算法越能正确捕获用户的移动行为模式。

这里选择了三个不同的窗口大小,基于相同的训练数据集,对 Pos-Cell2Vec 和 Cell2Vec 模型进行了对比,对比结果如表 1 所示。

表 1　模型对比

窗口大小	Pos-Cell2Vec	Cell2Vec
2	27.245	31.112
3	25.681	28.335
4	25.509	28.334

从结果中可以发现,Pos - Cell2Vec 模型相较 Cell2Vec 模型具有更加优秀的用户行为移动模式捕获能力。

5 结束语

结合真实数据的案例分析表明,文中方法能够有效地通过用户的轨迹、基站语义以及通话特征来探索用户的行为模式,进而推测用户的社会角色。对于未来的工作,将继续改进该方法,通话数据规模一般比较大,对于聚类降维图的处理相对耗时,当数据的可用时间跨度增大或者城市的规模更大时,实时交互将面临较大挑战。未来将继续优化 Pos-Cell2Vec 模型,希望可以达到交互式对用户数据进行快速处理的目标,以便推测出更多的社会角色。

参考文献:

[1] 王桂娟,周 锐,蔡梦杰,等. 基于移动通信数据的城市可视分析研究[J]. 大数据,2021,7(2):32-60.

[2] 张兰云,蒋宏宇,赵韦鑫,等. 基于用户轨迹及基站语义的城市活动模式可视分析[J]. 计算机应用研究,2021,38(6):1884-1888.

[3] ZHU M,CHEN W,XIA J,et al. Location2vec:a situation-aware representation for visual exploration of urban locations[J]. IEEE Transactions on Intelligent Transportation Systems,2019,20(10):3981-3990.

[4] 李致昊,朱闽峰,黄兆嵩,等. 一个基于基站轨迹数据的城市移动模式可视分析系统[J]. 计算机辅助设计与图形学学报,2018,30(1):68-78.

[5] CAO N,LIN C,ZHU Q,et al. Voila:visual anomaly detection and monitoring with streaming spatiotemporal data[J]. IEEE Transactions on Visualization and Computer Graphics,2018,24(1):23-33.

[6] ZHENG Xinhu,CHEN Wei,WANG Pu,et al. Big data for social transportation[J]. IEEE Transactions on Intelligent Transportation Systems,2016,17(3):620-630.

[7] CHOI J. Comparison of CDR and GPS data for estimating the individual activity space[D]. Tartu:University of Tartu,2020.

[8] AL-DOHUKI S,WU Yingyu,KAMW F,et al. SemanticTraj:a new approach to interacting with massive taxi trajectories[J]. IEEE Transactions on Visualization and Computer Graphics,2017,23(1):11-20.

[9] 关海潮. 一种基于行为序列的网络用户异常行为分析方法[D]. 南京:南京邮电大学,2018.

[10] 胡亚慧,李石君,余 伟,等. 基于时空感知的用户角色推理[J]. 电子与信息学报,2016,38(3):517-522.

[11] 王 峰,钟宝荣,余 伟,等. 基于用户角色与城市地域结构的互推断模型[J]. 计算机科学与探索,2016,10(12):1662-1672.

[12] LEE A J T,YANG F C,TSAI H C,et al. Discovering content-based behavioral roles in social networks[J]. Decision Support Systems,2014,59:250-261.

[13] GEHRING J,AULI M,GRANGIER D,et al. Convolutional sequence to sequence learning[J]. arXiv:1705.03122,2017.

[14] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[J]. arXiv:1706.03762,2017.

[15] MAATEN L V D,HINTON G E. Visualizing data using t-SNE[J]. Journal of Machine Learning Research,2008,9:2579-2605.