一个基于基站轨迹数据的城市移动模式可视分析系统

李致昊[1)], 朱闽峰[1)], 黄兆嵩[1)], 丁铁成[2)], 罗月童[2)], 葛嘉恒[3)], 陈　为[1)*]

[1)] (浙江大学 CAD&CG 国家重点实验室　杭州　310058)
[2)] (合肥工业大学计算机与信息学院 VCC 研究室　合肥　230009)
[2)] (浙江高速信息工程技术有限公司　杭州　310007)
(lizhihao@zju.edu.cn, chenwei@cad.zju.edu.cn)

摘　要: 随着移动通信技术的发展, 手机基站轨迹数据在分析人类移动规律方面的优势日趋显著. 由于人群移动模式与其社会行为息息相关, 该模式能够直接反映各地理区块在不同时间段所具备的社会功能. 根据词嵌入模型, 首先将基站的时空信息映射为向量, 通过计算基站间的高层语义的相似规律来分析地理区域的功能性信息; 再将带有时空变化信息的手机用户移动轨迹映射至向量空间, 使基站地理坐标与轨迹相结合, 从而获取更加丰富的语义信息. 在交互方面, 设计了一个可视化分析系统 Trajectory2Vec 来探索城市区域功能和用户行为的关系, 案例分析证明了该系统可以有效地帮助用户分析移动人群与城市区域间关系的动态变化规律.

关键词: 轨迹可视化; 人群移动性; 词嵌入; 可视分析
中图法分类号: TP391.41　　**DOI:** 10.3724/SP.J.1089.2018.16921

Trajectory2Vec: A Visual Analytics Approach for Urban Mobility Patterns Based on Mobile Phone Data

Li Zhihao[1)], Zhu Minfeng[1)], Huang Zhaosong[1)], Ding Tiecheng[2)], Luo Yuetong[2)], Ge Jiaheng[3)], and Chen Wei[1)*]

[1)] (*State Key Lab of CAD&CG, Zhejiang University, Hangzhou*　310058)
[2)] (*CC Division, School of Computer and Information, Hefei University of Technology, Hefei*　230009)
[3)] (*Zhejiang High Speed Information Engineering Technology Co., Ltd, Hangzhou*　310007)

Abstract: Based on word embedding model, we map the spatio-temporal information of base stations to the vector space and calculate the similarity rule of the high-level semantics between base stations to analyze the social information of geographical areas. Moreover, we designed and implemented a visual analysis system to explore the relationship between urban interregional functions. Case studies show that the proposed system can effectively help users to analyze the dynamic changes of the relationship between urban area and local residents.

Key words: trajectory visualization; human mobility; word embedding; visual analysis

随着移动通信技术的迅猛发展, 智能手机的便捷性使其成为城市居民随身必备用品, 在各类人群中的覆盖面极为广泛[1]. 由于移动通信设备需要保持信号畅通, 因此手机总是搜索信号源稳定的基站保持连接, 连接的时间、位置信息也同时被记录下来. 通过提取涵盖时间、地点的基站连接

收稿日期: 2017-10-12; 修回日期: 2017-11-13. 基金项目: 国家重点基础研究发展计划(2015CB352503); 国家自然科学基金重点项目(61232012, U1609217); 国家自然科学基金(61422211, 61772456). 李致昊(1996—), 男, 本科生, 主要研究方向为数据可视化; 朱闽峰(1993—), 男, 博士, CCF 会员, 主要研究方向为城市数据可视化; 黄兆嵩(1993—), 男, 博士, CCF 会员, 主要研究方向为城市数据可视化; 丁铁成(1993—), 男, 硕士, 主要研究方向为可视化; 罗月童(1978—), 男, 博士, 教授, 硕士生导师, CCF 会员, 主要研究方向为科学可视化、可视分析及相关技术在核能领域的应用; 葛嘉恒(1986—), 男, 硕士, 主要研究方向为交通大数据应用; 陈　为(1976—), 男, 博士, 教授, 博士生导师, CCF 会员, 论文通讯作者, 主要研究方向为可视化、数据挖掘、及人工智能等相关技术.

记录，我们便能够获得手机用户的移动轨迹. 大规模的数据则能够更直观地反映城市居民的移动模式，该移动模式一方面能反映人群的活动内容，另一方面还能反映出某地理区块的社会功能信息[2]. 在活动内容方面，对于不同的人群类型，轨迹信息展现出的移动模式同样有着较大差异[3]. 以大学生为例，由于学业所需，其工作日的移动区域通常局限于大学校园内，而周末及节假日则可能前往繁华商圈或周边地市娱乐消遣. 另外，根据不同地理区块的功能分类(如商务写字楼区、购物商圈等)，轨迹分析结果还有助于推出用户兴趣分析系统、交通控制系统、城市区域热度分析系统等应用，具有广阔的研究前景[4].

Word2Vec 利用神经网络将单词训练为最优实数向量. 通过计算余弦距离，我们能够很容易地将比较语义相似度的过程转化为比较词向量在向量空间相似度的过程，这将有助于实现单词词性提取、单词聚类等高级应用[5]. 本文将基站的时间、地点编码为一个特征单词，通过多组特征的训练，获得每个单词的向量表达. 进一步，可计算得到基站特征间的向量相似度，以及基站语义的相关性. 此外，Doc2Vec 在词向量的基础上添加了段向量的概念，将上下文语义添加至单词预测的过程中，因此可为多个词向量赋予同一段落的向量值. 在针对用户轨迹的分析中，单条用户轨迹途径的多个基站记录共享同一用户编号，即对应于Doc2Vec模型的段落特征. 由此，可计算用户轨迹实数向量间的余弦距离，获取轨迹相似度，从语义角度更深入地分析城市大规模人群的运动模式.

本文将可视分析技术与词嵌入模型结合，提供有效的用户交互手段，可以让人们充分参与到分析该可视化结果的过程中来，并利用人的认知能力从数据中挖掘有效信息. 此外，每个基站除地理和时间信息外，还包含周边标志性建筑与公共设施等其他复杂属性，可以辅助分析过程. 通过交互技术，人们还可以选取具有代表性的地理区域进行重点分析，以提升可视化结果的有效性.

本文工作的主要贡献如下:

(1) 利用词嵌入模型对轨迹进行建模，将人群和基站训练为实数特征向量，通过向量距离的计算，挖掘区域与区域间、人群与区域间的移动模式的异同.

(2) 基于轨迹向量特征与其在地理空间中的对应关系，实现了可视化分析系统. 系统结合基础交互操作，将选中轨迹同时投影至二维向量空间及地理空间中，用于发掘随时空动态变化的人群移动模式.

1　相关工作

1.1　轨迹可视化

借助于先进的物体追踪技术，如社交网络、交通运输、GPS 信号等，大规模的轨迹时空数据在当今有多样的采集渠道. 所获得的轨迹信息有着非常广阔的应用前景，如交通管理、军事化应用等[6]; 有学者基于社交网络的社交信息计算用户轨迹相似性[7]，并完善 POI 算法的开发[8]. 此外，目前学界也有较多针对城市用户移动轨迹的可视研究及城市流量模拟，数据内容涵盖各个领域，如手机基站轨迹[9]、船舶轨迹、机动车轨迹、行人轨迹等. 这些轨迹信息通常包含多维度属性，其中的复杂属性不易通过可视化手段予以清晰呈现[10]. 对此，学界主要有 4 种典型的可视化策略，即基于空间因素、时间因素、时空因素和多属性因素[11].

作为空间因素的可视化表达，Lundblad 等[12]将航线投影为折线型航道，与当地天气共同映射在地图上，为航船公司提供船只信息监测和不良天气预警服务. 然而，由于静态地图具有无法展示轨迹时间序列的缺陷，Wang 等[13]利用时间线的方式呈现二维轨迹的属性差异，直观有效地展示时空信息，避免了对信息聚类所造成的内容缺失，并结合可视化手段展现了轨迹运动变化方面的特性. 通过对时空因素进行分类，Landesberger 等[14]针对地理位置随时间变化的规律，结合时间、空间双方面信息进行可视化，设计了动态分类数据视图，为用户提供了面向任务的时间阶段选择方法，来支撑有关类别变化的可视探索. 同样在交互手段方面，FromDaDy 通过交互式的查询，可处理及分析大规模航空轨迹信息[15].

为了改进传统模式中流聚类的可视化方法，Landesberger 等[16]设计的 MobilityGraph，是一种用于减少大规模移动轨迹所导致的数据杂乱性的优雅方法，可在时空图上展现了跨度较大时间维度中人群的移动模式. Vrotsou 等[17]通过利用轨迹属性段的方式，简化了轨迹结构的复杂度. 为了有效地揭示乘客在交通网络的再分布，Zeng 等[18]设计交换圆环图来探索基于时空的移动模式. TrajRank[19]则针对沿某条轨迹的动态旅行时间变化

展开研究.

除时空因素外，为将单个轨迹点的属性纳入考虑范围, Tominski 等[20]通过堆叠轨迹带的方法有效地在可视化结果中添加了这类信息. 对于多属性的可视呈现, Scheepens 等[21]将子属性聚类为深度域，借助分布式地图的交互方式，用户能够高效地选取所需深度域并获取结果. 上述研究的相通点是侧重描述轨迹所蕴含的时空多样特征，而非挖掘轨迹发起者移动模式的隐含信息.

1.2 轨迹数据挖掘

已有大量工作尝试挖掘轨迹数据中所蕴含的丰富语义信息[22]. Chu 等[23]通过出租车的移动数据反映了城市人群的移动模式和趋势，借以归纳模式中的隐含语义. 针对具体出租车轨迹, Al-Dohuki 等[24]设计的方法十分直观且语义丰富，其将轨迹转化为文档模型，实现对于轨迹的文本搜索方法，对于挖掘出租车轨迹形成的动机进行了有益探索.

在分析移动模式时，Ma 等[25]通过手机数据提取地理和社交网络信息，借助欧拉方法对人群的运动进行研究. 在近年研究中，针对大规模人群移动趋势的同现性[26-27]，人们开始考虑轨迹的分布与用户兴趣点的联系，并据此分析某城市区域的功能[28]. 另有学者基于新加坡真实交通轨迹数据与兴趣点展开研究，并取得了长足进展[29-30]. 进一步，我们能够分析城市中热门区域与用户兴趣的潜在联系，如 Yuan 等[31]通过出入某区域的人口流量来研究对应地理位置的功能信息. 另外，还有通过研究轨迹分布规律分析得到区域热度信息，该信息能够用作城市内广告牌布局的参考性指标，同时也有助于对目标区域其他商业因素展开合理规划[32].

近年来研究发现，基于神经网络的 Word2Vec 模型对于捕捉单词序列的语义关系极其有效. Feng 等[33]提出的 POI2Vec 模型正利用这一点，将每个兴趣点映射为向量，兴趣点间的相关度则用向量的内积表示. 类似的, Liu 等[34]使用的 Skip-gram 模型根据位置信息的上下文(如先后抵达的位置集合) 来获悉潜在的前 N 个私人兴趣点. 除了将处理后的轨迹直接显示在地图上，并探索某用户在某一具体时刻的具体地理位置，以反映用户具体的移动方式外[18]，Yu 等[35]利用 Word2Vec 模型计算 1 592 562 条交通工具轨迹的相似性，并与卷积神经网络结合，来对道路交通流量进行预测. 由于手机基站轨迹的在时空维度上具有强烈的上下文相关性，使用词嵌入模型来分析用户轨迹的短暂时空特征的方法非常有效.

本文工作与上述工作有所差异. 我们结合词嵌入模型，基于人群运动轨迹的上下文相关性，将基站轨迹视为文档来考察用户移动模式特征，并提取轨迹中所蕴含的隐含语义，而非直接通过将人群轨迹流聚类进行展示. 与此同时，我们定义并计算轨迹的属性信息 (如经纬度最大值、最小值、平均值、移动速度、覆盖面积等)，分析向量空间中相似轨迹分布的规律性，从而推测大规模人群的移动模式.

2 方 法

2.1 概述

流行的针对词语的机器学习算法将词视为一个定长高维向量予以表达，最常见的特征基于词袋模型实现，而由于词袋模型丢失了单词在上下文中的顺序，因此无法得到单词语义信息. 本文基于 Doc2Vec 的词嵌入模型提出轨迹段向量的概念，能够在非监督模式下接收变长文本的数据输入，即基站轨迹记录. 因此，该轨迹段向量模型可以处理段落、文章等内容，后文将称其为 Trajectory2Vec.

根据生活中的实际情况，即城市中的密集人群常常在相同时间经过同一基站，每个基站与时间的捆绑表示均可被视为文章中出现频率较高的单词，词嵌入模型可以用于有效分析城市移动模式所蕴含的隐含语义. 换句话说，将每个基站视为单词，将每条轨迹(与用户相对应)视为段落. 在不同轨迹的相似时间段内常被记录的基站，往往在向量空间中的距离也十分相近；同理，具有相似运动模式的用户，其运动轨迹也具有相似的向量表达.

2.2 数据

2.2.1 数据编码

城市中的位置点常包含多类别的信息，如地理位置、POI 信息等，我们采集了中国浙江省温州市附近手机用户途径基站的时间序列数据，以及每个基站的地理坐标信息. 每条轨迹记录将包含多个属性：用户编 u_{id}、基站编号 s_{id}、经过时间 t. 按照用户依次经过的基站序列，定义用户轨迹

$$T=\{(u_{\mathrm{id}1},s_{\mathrm{id}1},t_1),(u_{\mathrm{id}2},s_{\mathrm{id}2},t_2),\cdots\} \tag{1}$$

其中，时间依次递加，即 $t_i<t_{i+1}$. 由此，可获得

轨迹的时间变化信息.

2.2.2　单词的生成

原数据中的时间带有时分秒等信息, 为了保证每个单词具有较高的出现频率, 将时间信息聚合至每个小时整点, 形如 $w=(s_{\mathrm{id}},t)$.

2.2.3　段落的生成

针对每个用户连续的一条轨迹, 将其编码为由一系列单词组成的段落, 即每个单词为某个特定时间点轨迹所经过的位置, 即

$$p=\{u_{\mathrm{id}},w_1,w_2,w_3,\cdots\} \tag{2}$$

将轨迹 p 与用户 id 相结合, 使用 Doc2Vec 模型进行训练, 得到每条轨迹的实数向量.

2.2.4　轨迹特征提取

为利用平行坐标轴视图显示每条轨迹数据的特征, 选取了如下几个属性值作为每条坐标轴的主题:

(1) 经纬度最大、最小值. 从轨迹所经过的所有基站中分别选取经纬度最大、最小值 ($lng_{\max}$, $lng_{\min}$, $lat_{\max}$, $lat_{\min}$) 并作为 4 个竖直坐标轴分别显示. 这一系列值将有助于我们理解每条轨迹所途径的范围.

(2) 经纬度均值 (lng_{avg}, lat_{avg}). 经纬度均值定义为某条轨迹所经过所有基站的经纬度平均值, 该数据将用于估计轨迹活动中心所处位置.

(3) 覆盖面积. 通过最大、最小经纬度来计算得到每条轨迹所覆盖的最大面积.

(4) 移动速度(v_{avg}). 与每个单词的时间因素结合, 能够得到一条轨迹的平均移动速度. 假设共有 n 个单词节点, 则

$$v_{\mathrm{avg}}=\frac{\sum_{i=1}^{n-1}\sqrt{(lng_{i+1}-lng_i)^2+(lat_{i+1}-lat_i)^2}}{t_n-t_1} \tag{3}$$

2.3　Word2Vec 与 Doc2Vec 模型

2.3.1　模型概念

Le 等[36]提到, 可将每个输入单词映射为向量, 并作为矩阵 $\boldsymbol{W}$ 的一列, 该矩阵的列标即代表词汇表中每个单词的下标. 词嵌入模型的目的是计算最大概率以获得单词的向量表示, 训练过程使用一系列单词 $w_1,w_2,w_3,\cdots,w_r$ 作为公式

$$\frac{1}{T}\sum_{T-k}^{t=k}\log p(\omega_t\mid\omega_{t-k},\cdots,\omega_{t+k}) \tag{4}$$

的输入.

对于每个输出单词, 能够得未经归一化的对数概率 y_i, 其计算公式为

$$y=b+Uh(w_{t-k},\cdots,w_{t+k};\boldsymbol{W}) \tag{5}$$

其中, U 和 b 是 softmax 分类器的输入参数, h 由从 $\boldsymbol{W}$ 中提取出的词向量通过均值或连接操作构造. 通过卷积神经网络的训练, 语义相似的单词在向量空间中的距离相近, 而语义差别较大的单词在向量空间中则相距较远. 借此特性, 能够使用向量对单词的语义做加减操作. 如 Mikolov 等[37]给出的范例所述, “国王”–“男人”+“女人”=“皇后”.

单个单词的向量只能应用于单词与单词的操作, 而无法获得段落前后文的语义, 段向量则弥补了这一不足. 与词向量作为矩阵 $\boldsymbol{W}$ 中的一列类似, 段向量也被映射为矩阵 $\boldsymbol{D}$ 中的一列, 与词向量一同训练. 对段落和单词进行连接或均值化操作后, 能够得到含有上下文语义的文章内容, 从而可以对接下来的单词进行预测. 在这个过程中, 这些被视为单词的段落, 因其效果非常像用于存储文章上下文语义的存储单元, 因此又被视为段向量分布式存储模型 (distributed memory model of paragraph vectors, PV-DM) [36].

总而言之, 本文算法本身有 2 个主要阶段:

(1)通过训练得到词向量 $\boldsymbol{W}$, 段向量 $\boldsymbol{D}$ 和 softmax 权值 U 与 b.

(2)根据固定的 $\boldsymbol{W}$, U 和 b, 对 $\boldsymbol{D}$ 使用梯度下降法添加新列, 从而产生新的未经输入的段向量 $\boldsymbol{D}$.

由于该算法通过无实义标签的数据对单词进行训练, 从而获得语义结果, 因此在训练不具备足够标签的单词数据集时, Doc2Vec 模型能够体现出明显优势.

2.3.2　差异比较

基于出现在相同的上下文语义 (或邻近单词)中的单词所含语义相似的假设, Word2Vec 能通过神经网络来表示分布式的词嵌入模型. 实验证明, 该算法在大型语料库中进行单词聚类、相似单词寻找过程中可行有效; 然而, 它只能应用于对单个单词的操作, 无法将上下文语义的实际内容纳入考虑.

Doc2Vec 的出现是为了对含有多个单词的语素(如句子、段落甚至整篇文档)进行语义提取. 通过为每个句子赋予 id, 此模型可应用于更高维度语素的相似度计算.

2.3.3　本文应用

通过将词嵌入模型应用到基站数据集中, 为每个基站训练生成一个实数向量, 同时为每条用户轨迹训练生成一个实数向量, 这是一种十分紧凑的表达方式, 并且对局部区域有着高敏感度的

反馈，不同区域的向量表达将具备明显的差异．本文将使用 cosine 距离来计算基站与基站、基站与轨迹或轨迹与轨迹间的相似度，如

$$Similarity(\boldsymbol{v}_m,\boldsymbol{v}_n)=\frac{\boldsymbol{v}_m\cdot\boldsymbol{v}_n}{\|\boldsymbol{v}_m\|_2\cdot\|\boldsymbol{v}_n\|_2} \tag{6}$$

由于采取对于轨迹编码的方式与时空因素相关，因此通过上述计算得到的基站和轨迹相似度不但包含了用户的运动模式，该相似度同时也蕴含了时空关系上的相似性．对于在时间和空间的变化规律相似的轨迹，可以认为其具有相似的社会行为与动机．借此，便能探索发现移动模式中所隐藏的语义信息．

3 城市移动模式可视分析系统设计

3.1 分析任务

本文采用词嵌入模型将手机轨迹和基站同时嵌入到向量空间中，从而计算轨迹和基站之间的相似性．为了探索轨迹和基站相似性随时间的变化，本文可视分析系统应当支持以下分析任务．

(1) T_1．分析基站和轨迹在向量空间的分布．为了分析基站与基站和轨迹与轨迹的关系，需要分析基站和轨迹向量在向量空间中的相似性．

(2) T_2．探索基站的功能随时间的变化．基站在不同时刻可能呈现出不同的功能，这和不同时刻经过基站的人群相关．因此，需要分析不同时刻和基站相似的人群的分布．

(3) T_3．探索人群移动行为随时间的变化．人的行为随着时间有着周期性的变化，如对于上班族来说，白天会在上班地点附近出没，晚上就会在家附近逗留．本文可视化系统需要支持用户在不同时刻的轨迹位置分布．

3.2 可视化系统

本文的可视分析系统如图 1 所示，主要包含 6 大视图：基站投影图、轨迹投影图、地图、控制面板、流量图和平行坐标轴。 其中，基站投影图和轨迹投影图提供了基站和轨迹在向量空间中的分布，通过对于投影图的刷选，其他 3 个视图将会联动更新其地理空间属性及其他多种属性．图 2 所示为该系统的局部展示．

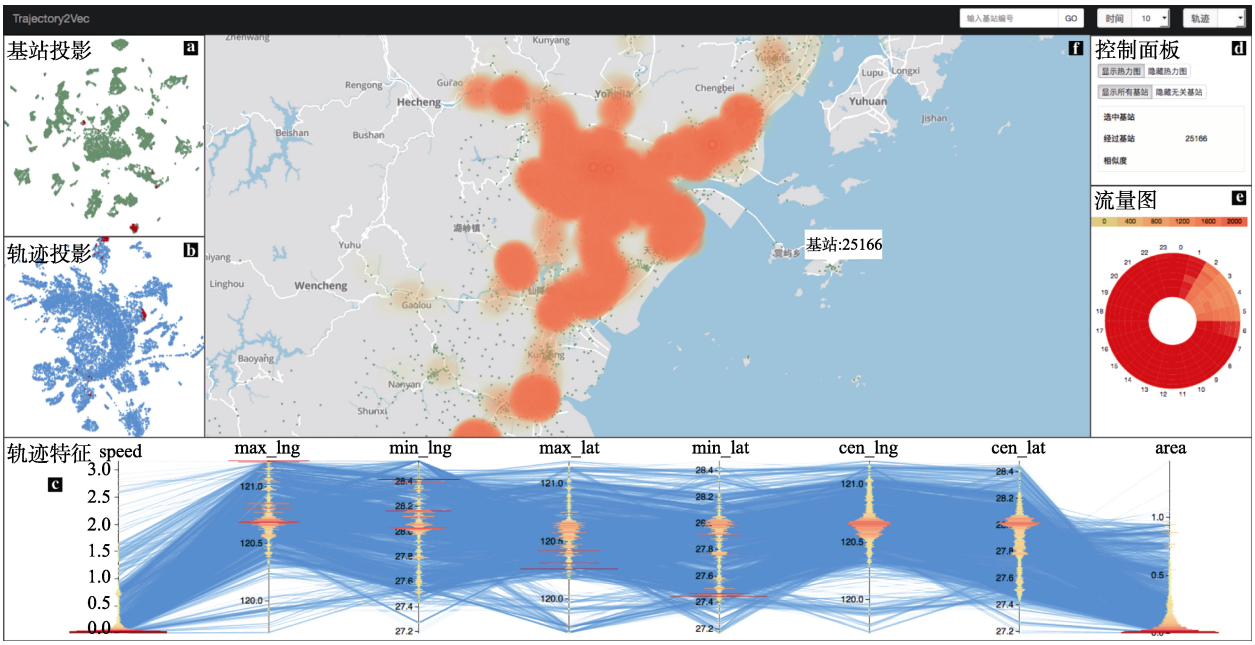


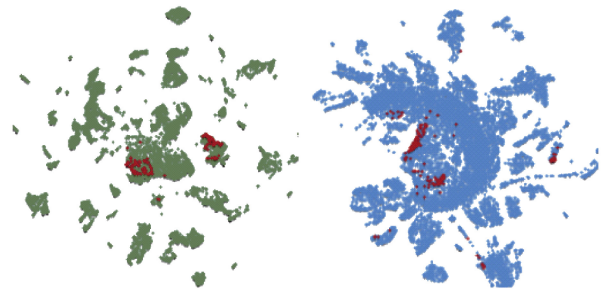
图 1 系统主界面

3.2.1 基站投影图和轨迹投影图

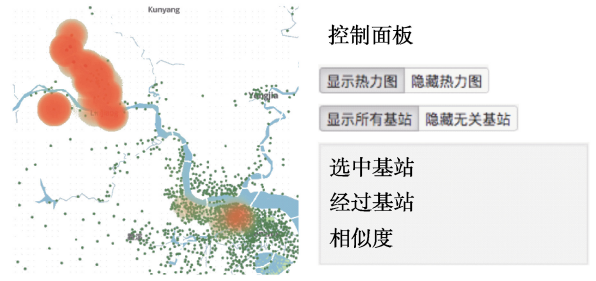
投影图展示了基站和轨迹在高维向量空间中的总体分布(T_1)．如图 2a 和图 2b 所示，本文使用降维算法 LargeVis[7]，将高维流形上的结构根据向量之间的相似性嵌入到二维平面上．由于基站之间的相似性也被保留到二维平面中，系统支持使用框选操作选择一系列相似的基站．

3.2.2 地图视图

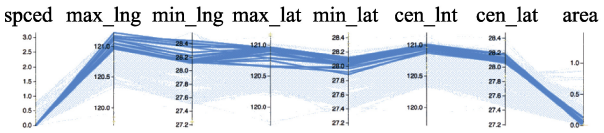
地图视图展示了用户选择的基站或者轨迹的地理空间属性分布，如图 2c 所示．它显示了真实世界的地理地图，展示了原始数据中所有基站在的位置，这些基站均支持点击显示与其最相似的基站．为了减少大量轨迹引起的视觉遮挡，本文采用热力图展示人群轨迹的分布情况 (T_2&T_3)，从而



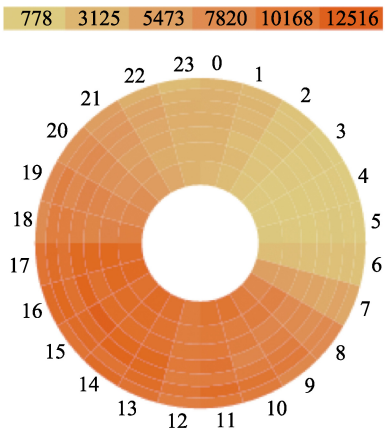
a. 基站投影　　b. 轨迹投影



c. 轨迹热力图　　d. 控制面板



e. 平行坐标轴



f. 流量图

图 2　可视化系统局部展示

较为直观地表现轨迹分布的区域密集程度，红色代表高流量，黄色代表低流量．当一个基站被选中时，基站的基本信息和对应的上下文轨迹会被呈现，相似的基站会通过流量图展示其不同时间段的流量信息．此外，系统还支持用户在地图上直接拉动选框，选择一定范围内的基站，得到与其相似的基站轨迹．

3.2.3　控制面板

如图 2d 所示，控制面板提供了操作的相关信息和可视化参数调整的相关功能．在控制面板中，显示了当前选中基站的 ID 和相似基站之间的相似度．可以选取感兴趣的时间段，探索基站与轨迹相似性随时间的变化(T_2&T_3).

3.2.4　流量图视图

流量图显示了所选一个或者多个基站在一周中每个小时的出入流量总和(T_2)，如图 2f 所示．由内到外的 7 层圆环代表 7 天的流量，圆环被平均分割为 24 等分，每一块的颜色代表 1 个小时的流量，颜色编码和地图上的热力图保持一致．

3.2.5　平行坐标轴

如图 2e，平行坐标轴展示了轨迹的属性分布(T_3)，包括速度、经纬度最大最小值、经纬度均值以及轨迹途径区域所覆盖最大面积．当用户在轨迹投影图中选取一类轨迹时，平行坐标轴会高亮显示这些轨迹在平行坐标轴中的线条；而选取某基站时，最相似的轨迹特征也会被高亮显示在平行坐标轴中．用户可拖动每个坐标轴改变其相对位置，有助于清晰地展示在某些特定属性间的线条变化．

3.3　可视化探索

本节将介绍该可视化系统具体工作流程．针对移动模式分析的切入点，本文提供了 3 种主要的交互手段供用户选择．

(1) 在基站投影视图中，通过划选向量距离相似的基站，过滤得到与其相似的轨迹，并显示在地图视图中；

(2) 在轨迹投影视图中，直接挑选某一向量聚类的轨迹，并在地图视图中显示；

(3) 在地图视图中划选目标基站，得到其相似轨迹热力图，并在投影视图中显示．

3.3.1　基站投影操作

首先，观察基站和轨迹投影视图，对基站和轨迹的整体移动状态有一个大致的掌握．在投影视图中常常会出现多个易于观察的小规模聚类，通过交互选择操作，可查看这些二维投影点在地理空间中所对应的具体基站或轨迹，图 3 中向量空间

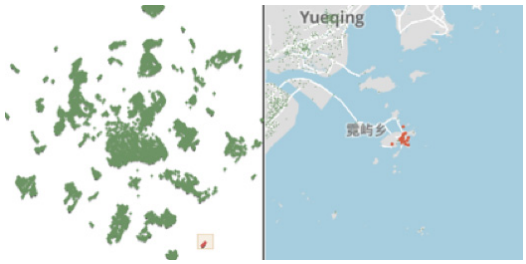


图 3　基站嵌入视图与地图视图的对应

的聚类对应了地理空间中集中在某小岛上的基站. 这些基站在向量空间和地理空间的距离都非常接近.

3.3.2 轨迹投影操作

图 4 中嵌入视图的聚类则对应于分布在沿海交通线上的用户轨迹. 在嵌入视图中, 这些轨迹呈一条直线分布, 而在地理空间视图中, 表现为沿海岸线分布. 即使我们无法理解经投影至二维平面的向量的坐标有何含义, 但是向量空间中距离较近点的地理特征却能够体现出某些相似的性质.

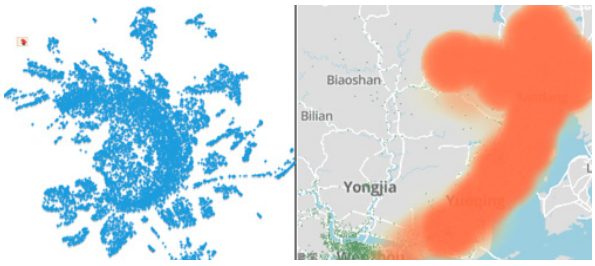


图 4 轨迹嵌入视图与地图的对应

3.3.3 地图划选操作

通过在地图视图上的选择操作, 能得到嵌入视图中的对应信息. 选取温州市乡村附近公路上的基站点, 得到了附近分布在公路上的轨迹热力图, 如图 5 所示, 相似轨迹总体沿南北向公路分布, 南端一直延伸到市区范围内.

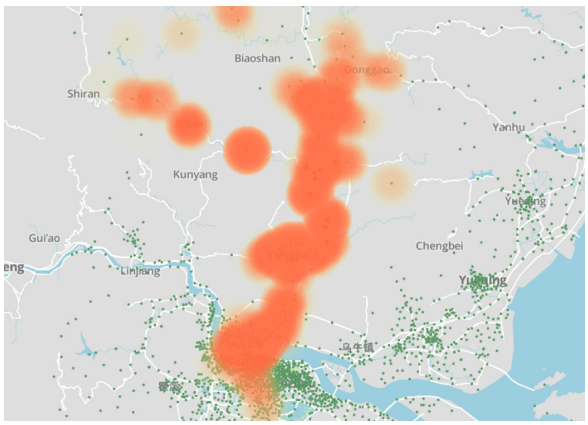


图 5 通过地图视图选择的沿某公路分布的热力图

4 案例分析

以下所有实验所使用数据存储在一台 24 核集群的 MySQL 数据库中, 通过 Gensim 提供的 Doc2Vec 算法生成本文中所使用的 Trajectory2Vec 模型, 并通过 LargeVis 投影将其映射在二维向量视图上。

4.1 郊区人群移动变化

本节将关注人群在一天不同时段内的移动规律, 尤其以城市附近一带居民的进出城的移动模式为重点进行分析. 为了判断某区域的人口密集程度, 可以将基站的密集度作为区域人口是否密集的指标. 进一步观察地图发现, 在基站密集的温州市西北方, 有一处人口较为密集的地带, 称其为郊区 *A*, 将东南方的温州市市区成为市区 *B*, 如图 6 所示.

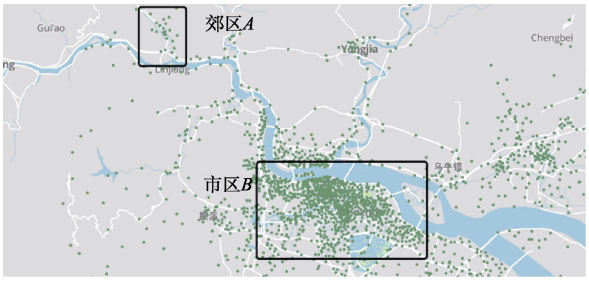


图 6 郊区 *A* 与市区 *B* 的相对位置关系

首先, 如图 7a 所示, 在基站投影图上选中一个聚类的相邻点, 系统将计算与这些点相似的轨迹, 并投影在轨迹投影图上. 轨迹投影结果显示后, 不难发现相似轨迹在向量空间距离同样较近, 并且呈现出了明显的分布特征. 观察图 7b 可发现, 除极少部分点分布在其他区域外, 大部分点分布在一个独立的直线状聚类中.

如图 7c 所示, 发现地图视图中, 所选基站在地理空间中也有着紧凑的分布, 即分布在郊区 *A* 的附近. 地理上距离较近的点, 在移动模式上也具有相似性, 我们据此展开后续探索.

a. 基站向量投影 b. 轨迹向量投影 c. 地理位置

图 7 所选数据在不同地图的呈现

通过观察图 8 发现, 选中基站的流量每日的分布情况基本相似, 但是在不同时间段内的分布差异较大. 具体表现为, 分布在日间至傍晚 (19:00~22:00) 明显多于深夜 (1:00~6:00), 该现象符合人们正常的作息起居习惯, 即在夜间通常处于安静的睡眠状态, 频繁移动极少发生.

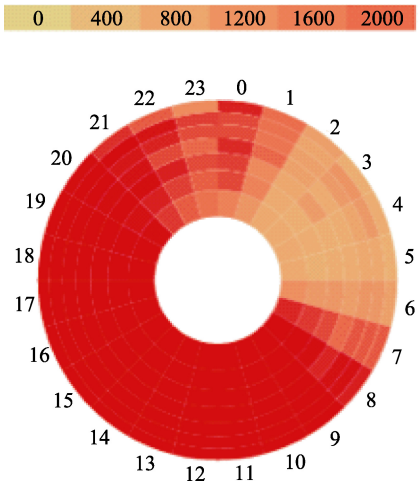


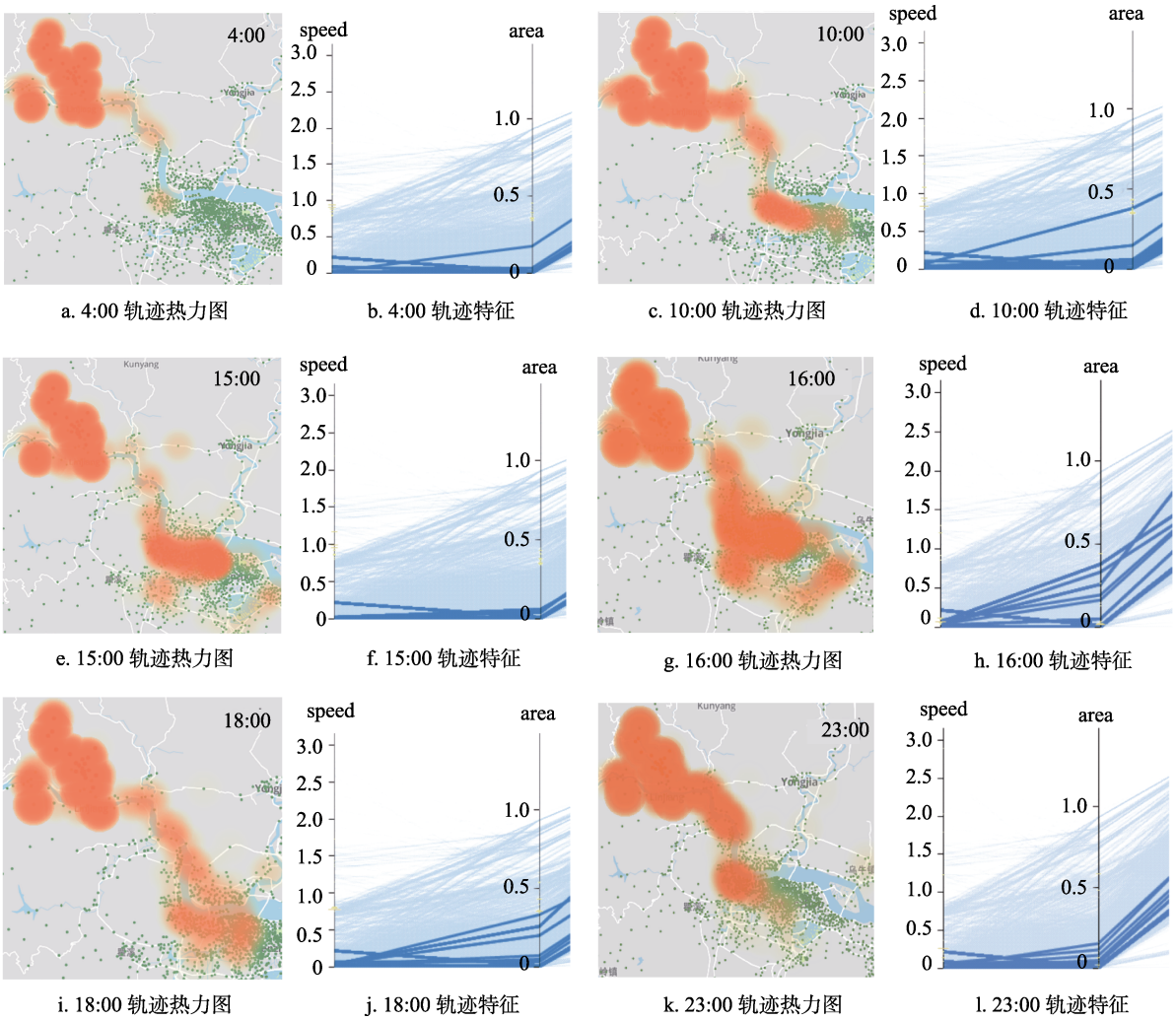
图 8　流量图

在接下来的分析过程中，通过选取不同的时间段，与先前所选取基站在向量空间中相似的轨迹热力图非常直观地表现出了 2 个人口密集区域间人群的移动模式.

图 9a 和图 9b 的时间为 4:00，根据人的正常作息时间，大多数用户应正处于睡眠状态，热力图也反映了这一点. 由图 9 可以看出，此时用户的轨迹大多分布在郊区 *A* 内，有极少量轨迹分布在城区 *A* 和 B 间的公路上；另外，平行坐标轴反映了此时用户的移动速度和面积范围均较小.

图 9c 和图 9d 的时间为上午 10:00，由地图视图可见郊区 *A* 与市区 *B* 间公路的热力图颜色明显加深，说明上午时段有部分用户开始往返于郊区与城区之间. 同时，用户的移动范围也有明显的增长，这也印证了该移动模式的含义为城区间的往返.

观察图 9e 和图 9f 所体现的移动模式，市区 *B* 的轨迹热力图颜色明显加深，而郊区 *A* 与市区 *B* 间的热力图颜色则有所变浅，这反映了多数用户在上午时段来到市区后，主要活动范围局限在市区范围内，该时段远距离的移动有所减少. 平行坐标轴中的移动面积也较上一时段有所减少，进一步印证了该分析结果.



a. 4:00 轨迹热力图　b. 4:00 轨迹特征　c. 10:00 轨迹热力图　d. 10:00 轨迹特征

e. 15:00 轨迹热力图　f. 15:00 轨迹特征　g. 16:00 轨迹热力图　h. 16:00 轨迹特征

i. 18:00 轨迹热力图　j. 18:00 轨迹特征　k. 23:00 轨迹热力图　l. 23:00 轨迹特征

图 9　不同时段内往返于 2 个城区间的移动状况

将 16 时的轨迹热力图与 15:00 进行对比，出现有趣现象：此时段郊区 A 与市区 B 的主要往返道路上未出现明显连续轨迹，但是市区 B 内的热力图的半径开始向周边区域扩散．结合平行坐标轴也能明显看出，此时轨迹的覆盖范围大幅增加．这样的特征暗示了下午 16:00 来自郊区的人群在城市中的运动轨迹较不规律，活动丰富．由于城市商业能否取得较高利润的关键因素在很大程度上取决于顾客的数量，因此该移动模式为城市中的商业布局提供了具有经济价值的参考意见．

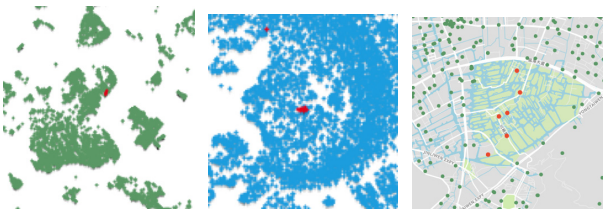
图 9i 和图 9j 为傍晚 18:00 居住在郊区 A 的用户开始陆续返回，沿 2 个城区间公路的轨迹热力图颜色再次加深，另外市区 B 的热力图颜色同样开始变浅，反映了典型的回程模式．图 9k 和图 9l 为夜间 23:00，在市区 B 内的热力图颜色明显变浅，活跃轨迹大幅减少，只有靠近郊区 A 的部分区域仍然分布有较多轨迹．很明显，此时大多数人群已返回郊区 A．至此，往返两地市人群一天内的主要活动接近尾声．

本案例通过选取一天中不同的 4 个时间段来分析人们的移动模式，展现了词嵌入模型在提取轨迹语义方面的优势．随着时间的变化，与最初所选取基站最相似轨迹表现出了明显的移动规律，而非局限于郊区 A 附近，间接表达了用户的活动内容语义(如上班、驻留、回家等内容)．即使某时刻某轨迹与基站在地理空间联系不强，我们仍然能够分析该轨迹与关注区域的语义相似度．

4.2 本地居民用户的移动模式

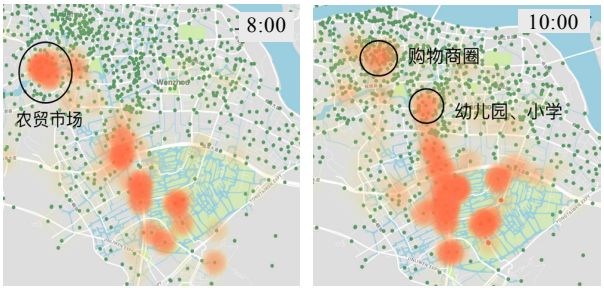
我们希望通过系统来探寻本地居民的在市区中移动模式，但是若选择主市区内的某块居民居住区域，其轨迹容易受到过路用户的影响，代表性会有所损失．经过观察发现，在温州市的三垟湿地内分布着一些居民区，由于天然湿地的限制，此地受到城市化的影响较小，这些居民区与外界也相对较为独立．因此如图 10c 所示，在地图上划选一些点，来进行深入分析．由图 10a 和图 10b 可见，在向量空间中，所选点的距离也十分相近．

通过对不同时段用户移动模式的分析发现了如下现象：在上午 8:00 以前，大多用户正在休息，移动轨迹多分布在居民区内．8:00 开始，移动轨迹首先明显趋向于居民区的西北部，该位置主要分布着大型农贸市场以及写字楼，如图 11a 所示．结合地图分析，由于三垟湿地内主要居民区为城中村，因此该区域区民有可能具有清晨买菜购物的习惯．同时，也有部分轨迹反映，有用户前往位于相同区域的写字楼等办公场所，开始一天的工作．

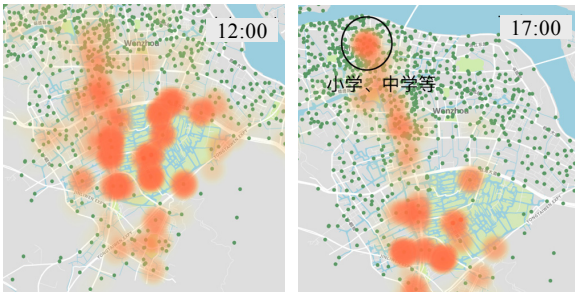


a．基站向量投影　b．轨迹向量投影　c．地理位置分布

图 10　某居民区用户在向量空间中的分布位置



a. 8:00 轨迹热力图　b. 10:00 轨迹热力图



c. 12:00 轨迹热力图　d. 17:00 轨迹热力图

图 11　三垟湿地居民的移动模式

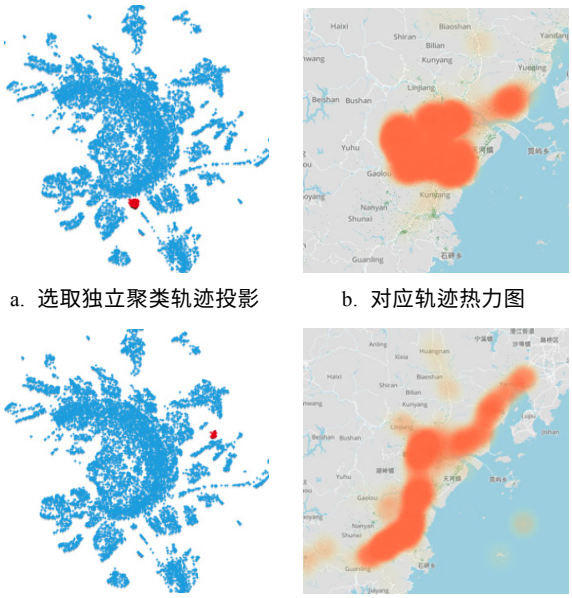
由图 11b 可见，上午 10:00 轨迹开始发生明显变化，有用户开始前往市中心的核心商圈，该地分布有沃尔玛等多个购物场所．同时，还有一些轨迹往返于幼儿园、小学等教育机构与三垟湿地间，说明该时段存在家长接送学生或学生自行前往学校的移动．中午 12:00，由三垟湿地向外部的轨迹有所减少，主要轨迹分布在三垟湿地内，这表现了中午时段人们的外出行为减少，开始享用午餐．

与上午 10:00 相对应的，下午 17:00 又出现了往返于学校、写字楼与三垟湿地间的轨迹，这再次印证了该时段为上下班高峰期，大量人群踏上返程．

通过本例，我们容易通过 Trajectory2Vec 模型来分析得到某地区居民日常不同时段内的活动主题，因此，该模型能够为安排城市中基础设施布局、商业设施选址等任务提供参考性建议．

4.3　轨迹向量聚类的差异分析

选取时间为中午 12:00，在轨迹向量投影图中选取较为独立的聚类，如图 12a 所示．该聚类的相似轨迹被投影至地图视图中，如图 12b．由地图可见，相似轨迹主要聚拢在温州市区周围，总体呈团状分布，中心地带热力图颜色与周边区域相比明显较深，且无明显的大范围转移．与之形成鲜明对比的是图 12c 的向量聚类；在图 12d 中，相似轨迹无明显抱团现象，而是沿海边交通干线呈条带状分布．在轨迹分布处，确有温州绕城高速、甬台温高速复线等高速公路存在．



a．选取独立聚类轨迹投影　　b．对应轨迹热力图

c．另一聚类的轨迹投影　　d．对应轨迹热力图

图 12　不同聚类结果

由此可见，向量空间中的不同聚类明显地反映出了轨迹差异明显的移动模式，轨迹的转移范围、覆盖面积等属性均有较大差异，这些聚类在反映移动模式的差异性上具有重要意义．

5　结　语

本文结合具体案例，挖掘并分析不同时段内基站轨迹的隐含语义，证明了基于词嵌入模型的可视化系统之有效性．

在数据处理方面，预计算所需进行的主要步骤如下：Step1．将大规模轨迹信息训练为 Doc2Vec 向量模型．Step2．使用 LargeVis 将高维向量投影至二维空间．这些操作均需要耗费大量时间及计算资源，因此本系统目前暂不支持实时的数据导入及显示．

在未来的研究中，将对各步预计算所耗时间进行统计，并深入研究如何进一步提升模型训练、向量投影的计算效率，以期能够向实时的用户移动模式分析系统逐步靠拢，满足相关用户的探索需求．此外，希望针对向量空间更复杂的计算进行延伸，如通过轨迹向量的“加、减”等计算得到轨迹向量的和与差，并利用该运算结果计算与其最相似的“运算轨迹”．我们还将利用向量视图及地图视图，考虑范围覆盖、时间变化等因素，挖掘行为异常的轨迹．通过抽取这些异常轨迹的属性，分析其存在异常的原因，并探究这些轨迹是否能为我们带来新的发现．

参考文献(References):

[1] Toole J L, Ulm M, Bauer D, *et al*. Inferring land use from mobile phone activity[C] //Proceedings of the ACM SIGKDD International Workshop on Urban Computing. New York: ACM Press, 2012: 1-8

[2] Arvind T. Probabilistic models for mobile phone trajectory estimation[D]. Cambridge: Massachusetts Institute of Technology. Department of Electrical Engineering and Computer Science, 2011

[3] Jiang X, Zheng C, Tian Y, *et al*. Large-scale taxi O/D visual analytics for understanding metropolitan human movement patterns[J]. Journal of Visualization, 2015, 18(2): 185-200

[4] Zhao S, King I, Lyu M R. A survey of point-of-interest recommendation in location-based social networks[C] //Proceedings of the 29th AAAI Conference on Artificial Intelligence. Palo Alto: Association for the Advancement of Artificial Intelligence, 2016: 53-60

[5] Bakshev S, Spinsanti L, Macêdo J A F, *et al*. Trajectory semantic visualization[C] //Proceedings of the 13th International Conference on Enterprise Information Systems. Lisbon: SciTePress, 2011: 326-332

[6] Tang L A, Zheng Y, Yuan J, *et al*. On discovery of traveling companions from streaming trajectories[C] //Proceedings of the 28th IEEE International Conference on Data Engineering. Los Alamitos: IEEE Computer Society Press, 2012: 186-197

[7] Zhang Ying, Li Zhi, Zhang Sheng. Users trajectory similarity algorithmic research on location-based social network[J]. Journal of Sichuan University: Engineering Science Edition, 2013, 45(S2): 140-144(in Chinese)
(张　莹, 李　智, 张　省. 基于位置的社交网络用户轨迹相似性算法[J]. 四川大学学报: 工程科学版, 2013, 45(S2): 140-144)

[8] Li Po, Hua Yixin, Li Xiang, *et al*. Research on POI personalized recommendation algorithm based on users' trajectory[J]. Geomatics & Spatial Information Technology, 2016, 39(11): 55-58(in Chinese)
(李　坡, 华一新, 李　响, 等. 基于用户轨迹的 POI 个性化推荐算法研究[J]. 测绘与空间地理信息, 2016, 39(11): 55-58)

[9] Wu Jiansheng, Huang Li, Liu Yu, *et al*. Traffic flow simulation based on call detail records[J]. Acta Geographica Sinica, 2012,

67(12): 1657-1665(in Chinese)
(吴健生, 黄 力, 刘 瑜, 等. 基于手机基站数据的城市交通流量模拟[J]. 地理学报, 2012, 67(12): 1657-1665)

[10] Tang J, Liu J, Zhang M, *et al*. Visualizing large-scale and high-dimensional data[C] //Proceedings of the 25th International Conference on World Wide Web. New York: ACM Press, 2016: 287-297

[11] Chen W, Guo F, Wang F Y. A survey of traffic data visualization[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(6): 2970-2984

[12] Lundblad P, Eurenius O, Heldring T. Interactive visualization of weather and ship data[C] //Proceedings of the 13th International Conference Information Visualization. Los Alamitos: IEEE Computer Society Press, 2009: 379-386

[13] Wang Z, Yuan X. Urban trajectory timeline visualization[C] // Proceedings of the International Conference on Big Data and Smart Computing. Los Alamitos: IEEE Computer Society Press, 2014: 13-18

[14] Landesberger T, Bremm S, Andrienko N, *et al*. Visual analytics methods for categoric spatio-temporal data[C] //Proceedings of IEEE Conference on Visual Analytics Science and Technology. Los Alamitos: IEEE Computer Society Press, 2012: 183-192

[15] Hurter C, Tissoires B, Conversy S. FromDaDy: spreading aircraft trajectories across views to support iterative queries[J]. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6): 1017-1024

[16] Landesberger T, Brodkorb F, Roskosch P, *et al*. MobilityGraphs: visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering[J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1): 11-20

[17] Vrotsou K, Janetzko H, Navarra C, *et al*. SimpliFly: a methodology for simplification and thematic enhancement of trajectories[J]. IEEE Computer Society, 2015, 21(1): 107-121

[18] Zeng W, Fu C W, Arisona S M, *et al*. Visualizing interchange patterns in massive movement data[J]. Computer Graphics Forum, 2013, 32(3): 271-280

[19] Lu M, Wang Z, Yuan X. TrajRank: Exploring travel behaviour on a route by trajectory ranking[C] //Proceedings of IEEE Pacific Visualization Symposium. Los Alamitos: IEEE Computer Society Press, 2015: 311-318

[20] Tominski C, Schumann H, Andrienko G, *et al*. Stacking-based visualization of trajectory attribute data[J]. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(12): 2565-2574

[21] Scheepens R, Willems N, Wetering H, *et al*. Interactive visualization of multivariate trajectory data with density maps[C] // Proceeding of IEEE Pacific Visualization Symposium. Los Alamitos: IEEE Computer Society Press, 2011: 147-154

[22] Yu L, Wu W, Li X, *et al*. iVizTRANS: Interactive visual learning for home and work place detection from massive public transportation data[C] //Proceedings of IEEE Conference on Visual Analytics Science and Technology. Los Alamitos: IEEE Computer Society Press, 2015: 49-56

[23] Chu D, Sheets D A, Zhao Y, *et al*. Visualizing hidden themes of taxi movement with semantic transformation[C] //Proceedings of IEEE Pacific Visualization Symposium. Los Alamitos: IEEE Computer Society Press, 2014: 137-144

[24] Al-Dohuki S, Wu Y, 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

[25] Ma Y, Lin T, Cao Z, *et al*. Mobility Viewer: An eulerian approach for studying urban crowd flow[J]. IEEE Transactions on Intelligent Transportation Systems, 2016, 17(9): 2627-2636

[26] Wu W, Xu J, Zeng H, *et al*. TelCoVis: Visual exploration of co-occurrence in urban human mobility based on telco data[J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1): 935-944

[27] Schwamborn M, Aschenbruck N. Towards an extensive map-oriented trace basis for human mobility modeling[C] //Proceedings of 35th IEEE International Performance Computing and Communications Conference. Los Alamitos: IEEE Computer Society Press, 2016: 1-10

[28] Wang M, Yang S, Sun Y, *et al*. Predicting human mobility from region functions[C] //Proceedings of IEEE International Conference on Internet of Things and IEEE Green Computing and Communications and IEEE Cyber, Physical and Social Computing and IEEE Smart Data. Los Alamitos: IEEE Computer Society Press, 2016: 540-547

[29] Zeng W, Fu C W, Arisona S M, *et al*. Visualizing the relationship between human mobility and points of interest[J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 18(8): 1-14

[30] Jiang S, Ferreira J, Gonzalez M C. Activity-based human mobility patterns inferred from mobile phone data: a case study of Singapore[J]. IEEE Transactions on Big Data: A Case Study of Singapore, 2017, 3(2): 208-219

[31] Yuan J, Zheng Y, Xie X. Discovering regions of different functions in a city using human mobility and POIs[C] //Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2012: 186-194

[32] Liu D, Weng D, Li Y, *et al*. SmartAdP: visual analytics of large-scale taxi trajectories for selecting billboard locations[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1): 1-10

[33] Feng S, Cong G, An B, *et al*. POI2Vec: geographical latent representation for predicting future visitors[C] //Proceedings of AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2017: 102-108

[34] Liu X, Liu Y, Li X. Exploring the context of locations for personalized location recommendations[C] //Proceedings of the 25th International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016: 1188-1194

[35] Yu D, Liu Y, Yu X. A data grouping CNN algorithm for short-term traffic flow forecasting[C] //Proceedings of Web Technologies and Applications: Asia-Pacific Web Conference. Cham: Springer International Publishing, 2016: 92-103

[36] Le Q, Mikolov T. Distributed representations of sentences and documents[C] //Proceedings of the 31st International Conference on International Conference on Machine Learning. Stroudsburg: International Machine Learning Society, 2014: 1188-1196

[37] Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations[C] //Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta: Association for Computational Linguistics, 2013: 746-751