软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(7):2018−2045 [doi: 10.13328/j.cnki.jos.005576]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

时空数据语义理解:技术与应用*

姚 迪[1], 张 超[2], 黄建辉[1], 陈越新[3], 毕经平[1]

[1](中国科学院 计算技术研究所,北京 100190)

[2](University of Illinois at Urbana-Champaign, US)

[3](盲信号处理重点实验室,四川 成都 610041)

通讯作者: 毕经平, E-mail: bjp@ict.ac.cn

摘 要: 随着移动互联网的发展与手持智能终端的普及,海量带有用户时空属性的数据被生成.理解这些数据表达的语义信息对推测用户需求,分析用户偏好,进而提供精准时空推荐和预测服务具有重要作用.因此,近些年来,时空数据语义理解正成为时空数据挖掘领域的研究热点.从技术和应用两个层面,对近些年来国内外研究者在该领域的研究成果进行了系统的归类和总结.技术层面上,依据语义理解的不同任务,提出了时空数据语义理解的研究框架;并依次从地理位置语义理解、用户行为语义理解、热点事件语义理解 3 个主要任务,归纳了时空数据语义理解所包含的相关研究成果和关键技术.应用层面上,分别总结了时空数据语义理解在时空推荐和时空预测中的应用.最后,从数据质量、算法模型和计算模式 3 个方面,归纳了时空数据语义理解面临的主要挑战以及未来的研究方向.

关键词: 时空数据挖掘;语义理解;语义偏好;用户画像;机器学习

中图法分类号: TP182

中文引用格式: 姚迪,张超,黄建辉,陈越新,毕经平.时空数据语义理解:技术与应用.软件学报,2018,29(7):2018−2045. http://www.jos.org.cn/1000-9825/5576.htm

英文引用格式: Yao D, Zhang C, Huang JH, Chen YX, Bi JP. Semantic understanding of spatio-temporal data: Technology & application. Ruan Jian Xue Bao/Journal of Software, 2018,29(7):2018−2045 (in Chinese). http://www.jos.org.cn/1000-9825/5576.htm

Semantic Understanding of Spatio-Temporal Data: Technology & Application

YAO Di[1], ZHANG Chao[2], HUANG Jian-Hui[1], CHEN Yue-Xin[3], BI Jing-Ping[1]

[1](Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100190, China)

[2](University of Illinois at Urbana-Champaign, US)

[3](National Key Laboratory of Science and Technology on Blind Signal Processing, Chengdu 610041, China)

Abstract: With the development of mobile internet and widespread use of mobile phones, a large amount of data that contains user' time and space attributes has been generated and collected. Investigating the semantic information of the collective data plays an important role in understanding the needs, analyzing preference of the user, even recommending and predicting space and time. Recently, many researchers all over the world have turned their focus on understanding the spatio-temporal semantic data. This paper summarizes the related works regarding the spatio-temporal semantic data. Firstly, according to the tasks, the basic concepts and research frameworks are introduced; then, the works of location semantic understanding, user behavior semantic understanding and event semantic understanding are summarized. Additionally, the application scenarios of recommending and predicting space and time field are described. Finally, the future research directions of spatio-temporal data semantic understanding are discussed.

Key words: spatio-temporal data mining; semantic understanding; semantic profiling; user portrait; machine learning

* 基金项目: 国家自然科学基金(61472403, 61303243, 61702470)
Foundation item: National Natural Science Foundation of China (61472403, 61303243, 61702470)
收稿时间: 2017-06-13; 修改时间: 2018-03-16; 采用时间: 2018-04-04; jos 在线出版时间: 2018-04-16
CNKI 网络优先出版: 2018-04-16 11:00:00, http://kns.cnki.net/kcms/detail/11.2560.TP.20180416.1059.011.html

随着移动互联网、位置服务等技术的高速发展和移动设备的普及,在基于位置的社交网络(微博、Twitter等)、共享出行(Uber、滴滴)、共享单车(摩拜、ofo)的运营过程中,产生了海量的时空数据.据统计,Twitter 每天产生约 1 000 万条带有位置信息的消息;Foursquare 中已经积累了超过 100 亿条的签到记录;滴滴日订单量接近 2 000 万.这些数据表明,我们已经进入时空大数据时代[1].丰富的时空数据,从不同的粒度、层面和视角记录人的活动信息.然而,单纯的数据本身没有任何意义,只有被赋予含义的数据才能够被使用.而数据的含义就是语义,其可以看作是数据对应的现实事物所代表的概念和含义以及这些含义之间的关系,是数据在某个领域上的解释和逻辑表示[2].理解时空数据表达的语义对于深度挖掘数据商业价值、提升社会工作效率、提前预测并防范突发事件有着关键作用.例如:有了时空数据语义信息的支持,在广告投放中可以更精准地对用户建模,提高广告点击转化率;在车辆调度规划中可以感知不同功能区域的不同需求,从而进行合理调度;在社交网络事件监测中可以分析事件成因(how)、事件种类(what),提高监测中事件识别精度.总之,准确理解时空数据中的语义信息有着十分重要的意义.其结果可以广泛应用于诸多领域.近些年来,随着时空数据体量和类别的持续增长以及对时空数据应用的不断深化,时空数据语义理解受到越来越多研究者的关注,将会成为一个热点研究方向.

目前,已有许多研究人员从不同角度总结了时空数据挖掘研究进展情况.刘大有等人[3]对时空数据挖掘的研究进展进行了概括,将时空数据挖掘任务归纳为时空模式发现、时空聚类、时空异常检测、时空预测和分类、时空推理这 5 项任务,并对每类任务的研究进展作了介绍.吉根林等人[4]也对时空数据挖掘作了相似的综述.郑宇[5]对与城市计算相关的时空数据处理技术和相关应用作了总结,不仅定义了城市计算的核心问题和应用场景,而且结合时空数据挖掘的技术解决城市生活中存在的问题,为城市建设和规划提供建议.另外,时空数据中关于轨迹数据挖掘的综述比较多.郑宇[6]对轨迹数据挖掘的框架和应用进行了系统的总结,并对轨迹处理各个流程中涉及到的研究工作作了汇总.高强等人[7]对轨迹大数据处理的关键技术进行了综述,并按照轨迹数据处理的流程来论述:首先,概述了轨迹数据的产生累积过程,总结了轨迹数据的特点并给出轨迹数据处理的框架.随后,依次对轨迹预处理、轨迹索引与检索、轨迹数据挖掘、轨迹隐私保护等方面的研究工作进行了总结.最后,论述了轨迹大数据处理支撑技术和轨迹数据与量子计算关系,并对相关研究工作作了归纳.Ni 等人[8]总结了轨迹数据挖掘的框架,框架从预处理、数据管理、查询处理、挖掘分析、应用场景等方面对研究工作进行分类.文献[9]对轨迹数据预测研究工作总结作了综述.文献[10]介绍了基于距离的轨迹聚类方法.文献[11]对基于出租车数据的动态社群发现作了综述.文献[12]总结了基于外部数据理解轨迹语义的框架,并对带有语义信息轨迹的应用进行了深入阐述.

尽管与时空数据处理相关的综述类文章很多,但目前还未见研究人员从语义理解的角度梳理时空数据挖掘中的相关任务和研究工作.本文从技术和应用两个层面对时空数据语义理解进行了总结.在技术层面上,根据语义理解任务将相关研究划分为地理位置语义理解、用户行为语义理解、热点事件语义理解 3 类,在应用层面,根据不同应用场景将应用分为面向推荐的时空语义应用、面向预测的时空语义应用两大类.最后,本文从数据稀疏性、多模态融合、数据实时分析等方面归纳了时空数据语义理解中存在的问题及未来可能的研究方向.

本文第 1 节对时空数据语义理解涉及到的概念定义进行总结,整理并提出按任务分类的时空数据语义理解的研究框架.第 2 节对地理位置语义理解的研究成果及关键技术进行总结.第 3 节对用户行为语义理解的研究成果及关键技术进行总结.第 4 节对热点事件语义理解的研究成果及关键技术进行总结.第 5 节整理归纳时空数据语义理解的应用领域,分别总结并对比时空推荐类、时空预测类应用.本文最后总结全文,并对时空数据语义理解需要解决的问题和值得关注的热点研究方向进行讨论.

1 时空数据语义理解概述

时空数据刻画了用户的时间和空间属性,富含有价值的信息.时空数据语义理解是指在融合多源/多类信息的基础上,通过逻辑推理和知识发现等方法,理解时空数据产生过程中所反映出的用户行为、状态和偏好等语义信息.利用传统的序列挖掘方法,虽然可以直接挖掘到用户活动的频繁模式、关联模式等规律,从而预测用户

未来的时空属性.但是,由于缺失语义信息,这类挖掘结果通常可解释性差.然而,时空数据的大部分应用,需要根据挖掘结果,进行针对性的推荐和与预测.因此,准确理解时空数据的语义对时空数据应用至关重要.本节首先总结了时空数据元素名称,并给出了其在本文中的含义.然后,总结了各类时空数据的产生方式和特点.最后,从任务分类角度概述了时空数据语义理解的研究框架.

1.1 时空数据元素

时空数据是指同时具有时间维度属性和空间维度属性的数据.通常时空数据通过 GPS 定位设备、通信基站、银行卡交易、公共交通卡、RFID 传感器等设备采集,也可以通过提取 Web 文本或图像中的时间维度属性和空间维度属性获得.时空数据中包含位置、轨迹等多种类型的时空元素.为方便下文叙述,本节总结了下文涉及到的时空数据元素及其含义,见表 1.

Table 1 The meaning of spatio-temporal elements

表 1 时空元素含义

名称	英文名	含义	备注
POI	Point of Interest	对用户有用,或使用户感兴趣的一个具体的位置	
ROI	Region of Interest	对用户有用,或使用户感兴趣的一个具体的区域	
路径	Path	对用户有用,由一个位置到另一个位置的道路	
兴趣元素集	Elements of Interest	将 POI、ROI 和路径统称为兴趣元素集	这些元素的共同特点是其属性都包括两部分,即位置属性和语义属性,语义属性大多都是客观存在的
轨迹	Trajectory	用户在某一时间段内的时空数据记录	轨迹的语义信息主观性更强
停留点	Stop Point	用户轨迹中,用户在一段时间内保持静止的位置	
事件	Event	某个区域、时间段内发生的、有用户参与的热点模式	本文中事件大多指与地理位置相关的区域事件或局部事件
用户	User	时空数据所刻画的数据产生对象	本文用户是指所有被刻画的对象,不但包括人,还包括机动车、动物

目前,研究人员关注的时空数据主要包括人类活动产生的数据、动物活动产生的数据、交通工具产生的数据、自然现象产生的数据.表 2 总结了各类数据的产生方式、数据特点及典型数据集.

Table 2 The classification of public spatio-temporal dataset

表 2 时空数据公开数据集及分类

数据种类	产生方式	主要场景	数据特点	典型数据集
人类活动	被动收集:人类佩戴位置及其他传感器设备在人类活动中持续采集的数据	健康监控、运动监控、导航	采集频率固定;数据密集;误差受采样精度影响;	Geolife[13]
	主动发布:人类主动公开发布的、表现人类活动位置及状态的数据	位置社交网络、银行卡记录、RFID(公交卡)	数据密度低,特定用户数据稀疏;伴随丰富的其他数据(文本、图片、视频)	Twitter[15]、Flickr[16]
交通工具	安装在交通工具上的采集设备为了监控其运行状态采集得到的数据	车联网、船舶监控、飞机航线监控	数据采集频率较高且固定;数据格式规整且质量较高;存在固定航线、路径,数据规律性强	AIS[19]、T-Drive[14]
动物活动	生物研究人员为了研究动物活动规律,为动物安装传感器,这些传感器回传得到的数据	鸟类活动、草原动物活动	受功率限制,数据稀疏且质量不高;数据随机性强	候鸟迁徙[17]、斑马活动[18]
自然现象	气象学研究人员需要监控自然现象的活动,通过卫星遥感等方式可以获得某些自然现象(台风、洋流)的产生变化数据	台风预报、空气污染监控	采集成本高;数据应用实效性强	台风数据[20]、空气污染数据[21]

1.2 时空数据语义理解研究框架

通过总结时空数据的产生过程和数据特点可以发现,时空数据的产生伴随着丰富的语义,并且不同种类的空间数据表达的语义信息也各不相同.一方面,受采集技术、接收技术、传感器种类、功耗等方面的限制,许多有用的与语义相关的信息往往不能完全被收集.另一方面,在分析利用时空数据挖掘知识、做出决策时需要更

加精确的语义信息.现实的技术局限和应用的迫切需求形成了一对矛盾.时空数据语义理解就是为解决这一矛盾而产生的新兴研究方向.

从2001年Tsoukatos等人[22]首次提出时空数据挖掘开始,时空数据的语义理解作为时空数据挖掘的子领域吸引了越来越多研究者的关注.数据库顶级会议SIGMOD、VLDB、ICDE中都设有专门研究时空数据的专题.如表3所示,本文统计了三大会议近3年来关于时空数据研究的论文,可以看出,相关论文数量逐年上升.表中每一项以*A*/*B*形式表示,*A*表示与时空数据相关的论文数(包括长文、短文和演示),*B*表示会议收录论文总数.

Table 3 The numbers of research paper of spatio-temporal data

表3 时空数据研究论文数量统计

年份 会议	2015	2016	2017
VLDB	12/208	15/170	22/196
SIGMOD	8/158	9/137	10/105
ICDE	24/166	27/214	24/231

特别是近些年来,随着可穿戴设备、互联网的发展,越来越多的富含语义信息的时空数据被收集,这些信息为时空数据语义理解的进一步发展提供了新的契机.可以预见,未来时空数据的语义理解依然会是时空数据挖掘的一个热点研究方向.

时空数据的语义具有领域性特征和主观特征.领域性特征是指用户对于时空数据在不同领域上的解释,例如速度属性,对于用户来说可以刻画用户的驾驶习惯;对于交通领域来说,可以表示相关路段是否拥堵.另一类特征是主观特征,对于不同用户来说其用户行为偏好各不相同.例如,对于同一地点,一定有某些用户喜欢而另一些用户不喜欢.正是由于时空数据语义的两类特征,很难将现有工作按照待理解的语义属性进行分类.因此,本文将从研究对象的角度梳理时空数据语义理解的研究工作.在已有的时空数据语义理解研究中,根据研究对象的不同可将研究工作划分为3类,即:地理位置语义理解、用户行为语义理解、热点事件语义理解.研究框架如图1所示.

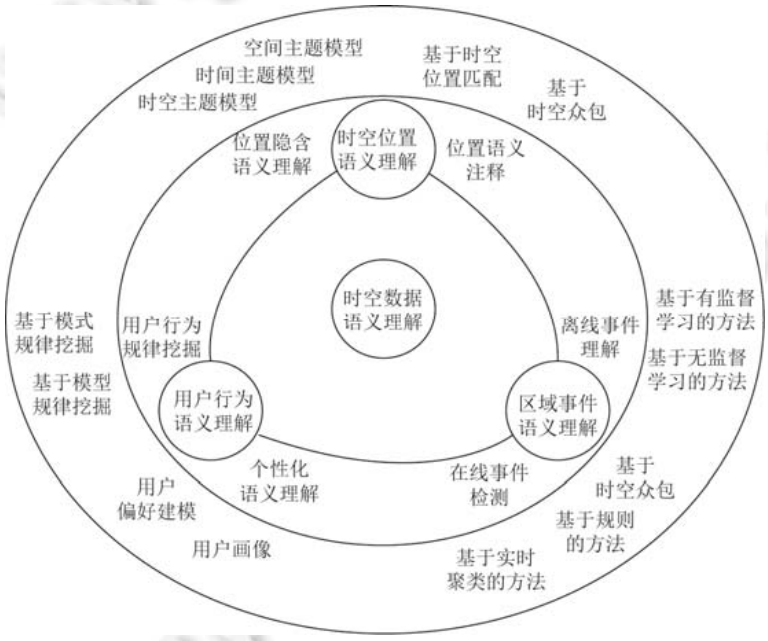

Fig.1 The research framework of semantic understanding of spatio-temporal data

图1 时空数据语义理解研究框架

地理位置语义理解.地理位置数据是时空数据中的重要类别.地理位置的语义理解是指通过时空众包、外部数据推断等方式,准确地理解地理位置语义属性蕴含的领域特征和主观特征.按照语义属性的类型,地理位置

语义理解可以分为地理位置语义注释和地理位置隐含语义理解.

用户行为语义理解.时空数据可以有效地反映用户行为的特征.用户行为语义理解是指通过分析用户时空数据及相关外部数据,发现用户行为的规律,理解不同用户的行为特点,进而推断用户主观意图.按照用户行为的规律性或特殊性,用户行为语义理解可分为:用户行为规律挖掘和用户个性化行为理解.

热点事件语义理解.时空数据中的热点事件是指在某个区域某个时间段内发生的、有单个或群体用户参与的、对用户后续行为有影响的热点模式.热点事件的语义理解主要是挖掘事件产生和发展过程中描述事件状态的语义属性.按照对事件检测的实时性,热点事件的语义理解可以分为热点事件的离线理解与热点事件的在线检测.

2 地理位置的语义理解

地理位置是时空数据中的基本元素,具体类型包括兴趣点、兴趣区域、道路、轨迹等.根据语义信息的变化情况,地理位置的语义信息可划分为两类:地理位置静态语义属性和地理位置隐含语义属性.地理位置静态属性是指地理位置本身具有的随时间变化不明显的信息.以体育场这一地理位置为例,体育场的建筑面积、可容纳观众数量都属于静态语义属性.对于这类属性可以通过地理位置语义注释获得.地理位置隐含语义属性是指难以直观发现、需要通过推理得到的地理位置语义信息,该类位置语义信息会随外部因素的变化而变化.例如,体育场每年都会举办各种各样的活动,如演唱会、运动会等.在不同活动期间,体育场会被赋予演唱会举办地、运动会举办地等不同的语义.这类语义属性随着时间因素变化大且常常不能通过直观总结得到.因此与理解地理位置的静态语义不同,隐含语义需要通过隐含语义理解的特定技术来解决.地理位置语义理解的研究框架如图 2 所示.

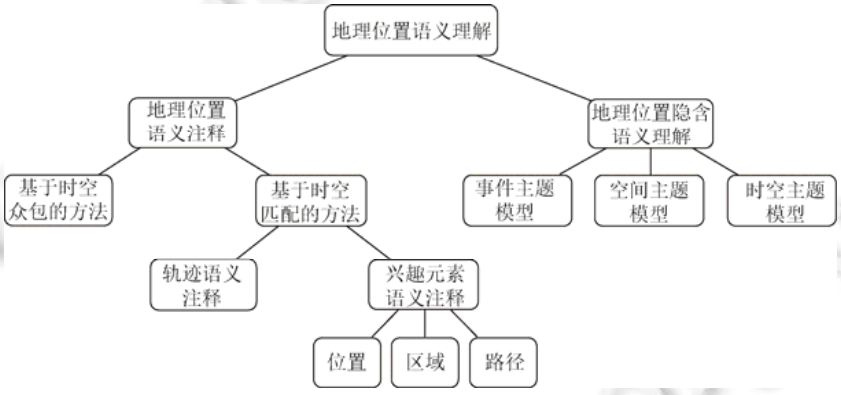

Fig.2 The research framework of semantic understanding of locations

图 2 地理位置语义理解研究框架

2.1 地理位置语义注释

对地理位置的语义注释大多通过人工语义标注实现.这种方式可以准确得到质量相对较高的时空数据的语义信息.然而,人工标注方式理解时空数据语义也有其局限性:(1) 效率低,时空数据中需要语义标注的数据体量庞大,完全依靠人工标注,处理速度慢且错误率高;(2) 成本高,人工语义标注虽然对技术水平要求不高,但劳动量较大,人力成本和质量管理成本较高.另外,当时空数据中地理位置的语义信息发生改变时,需要重新标注,进一步推高了人工标注的综合成本.针对上述局限性,目前有以下两种解决思路.

(1) 增加标注人员,提高语义标注工作的并行程度,从而提高标注效率;

(2) 基于已知语义信息的外部数据来推断未知的语义信息,通过自动化降低数据标注的成本.

与这两种思路对应,目前解决地理位置语义注释的研究工作可以分为两类:基于时空众包(spatio temporal crowdsourcing)的地理位置语义注释和基于时空匹配的地理位置语义注释.

2.1.1 基于时空众包的地理位置语义注释

时空众包是以时空数据管理平台为基础,将具有时空属性的众包任务按一定策略分配给非确定的众包工作者群体,要求工作者以主动或被动的方式来完成任务,并满足任务所指定约束条件的一种新型众包计算模式.文献[23,24]归纳了已有的时空众包数据管理技术,将时空众包的核心研究问题总结为任务分配、质量控制、隐私保护 3 类,并依次综述了每类问题的研究现状和可用技术,最后总结了时空众包的未来研究方向.基于时空众包的地理位置语义注释是指通过时空众包的方式,增大人工标注任务参与人员的规模,并通过任务分配、质量控制等方法保证语义标注的准确率.

基于时空众包对地理位置进行语义注释后形成的地图数据集称为众包地图(crowdsourced map).Open StreetMap[25](OSM)作为众包地图的典型代表被称作地图版的维基百科.OSM 通过招募众包参与者对空间地图进行编辑,从而获取了大量的标注建筑物、街道等地理位置的语义信息,并且随着注册编辑的不断增多,这些信息会变得越来越完善.截止 2014 年 3 月,OSM 的注册编辑已接近 150 万,含有语义信息的兴趣点数据条目超过 7 800 万条[26].

目前,基于时空众包地理位置语义注释的研究工作集中在众包任务分配算法和真实语义推断两个方面.任务分配算法主要解决的问题是通过合理地分配众包任务,提高众包工作者效率,并得到对地理位置语义的候选结果.真实语义推断解决的主要问题是从多个对同一地理位置的语义候选结果中推断出正确结果.为了达到对地理位置语义的准确注释,设计众包系统时,通常需要对这两类问题综合考虑.文献[27]研究了任务分配算法中如何为众包参与者提供合适的任务候选集的问题.其目标在于减轻参与者编辑工作量,提高编辑效率.文献[27]的作者提出了一种路网结构拓扑与轨迹数据融合的叠加推断模型,该模型基于对道路拓扑特征和轨迹特征的 logistic 回归模型以及路网连通性的朴素贝叶斯分类模型方面的考虑而形成.作者在实验中使用真实数据,结果表明,该模型能够有效提高编辑效率.文献[28]研究了 POI 标注质量提升方法,将标注任务拆分成两个子任务:首先将任务分配给多个合适的众包参与者,然后从多个众包标注结果中推断出正确的结果.作者针对这两个子任务分别提出了在线任务分配算法和语义推断模型.推断模型考虑了参与者自身的任务完成偏好、参与者与 POI 的空间距离以及 POI 的影响力等因素,构建了一个概率图模型,用于描述真实语义的生成过程.任务分配算法以提高推断的准确率为目标,对分配策略进行优化.两个模型交替优化,不断提高 POI 数据的标注质量.文献[29,30]研究了城市交通中的众包问题,发现对城市交通情况监控效果较差的主要原因是由于通过物理传感器采集到的交通情况数据比较稀疏.作者利用时空众包的方法解决了这一问题,通过对语义结果不确定、不一致的区域附近的多位众包参与者进行多次询问,对反馈结果进行汇集,推断得到准确结果.文献[31]研究了城市道路拥堵情况估计的问题.目前,只能获取到一小部分道路中的通行速度,需要对多数道路的通行速度情况进行估计.作者提出了一种基于众包的通行速度计算方案:首先为整个城市路网做种子道路选择,选取 K 条道路作为种子;然后,采用众包的方式得到这些道路的通行速度;最后,通过已知种子道路的速度,估计整个路网每条道路的通行速度.

2.1.2 基于时空匹配的地理位置语义注释

基于时空匹配的地理位置语义注释的主要思路是借助已知语义信息的数据(如知识库、文本或地图),将未知的地理位置匹配到已知语义信息的数据上,进而建立语义推断模型,得到地理位置的语义.根据语义注释的对象不同,该类任务分为轨迹语义注释和兴趣元素集语义注释.

(1) 轨迹语义注释.轨迹数据是时空数据中的一类重要元素,常见的轨迹数据直接通过位置传感器采集得到,没有任何语义信息,需要通过外部数据推断其语义.如图 3 所示,轨迹语义注释的建模和分析方法可总结为从序列轨迹(raw trajectory)到语义轨迹(semantic trajectory)的过程.其中,处理步骤包括轨迹清洗与预处理、轨迹行为挖掘、轨迹语义匹配等阶段.文献[12,32,33]分别对轨迹语义注释的计算方式和架构进行了总结,提出了轨迹语义注释的两个步骤的计算方式.

第 1 步.轨迹分段,根据轨迹中的停泊特点,将原始位置轨迹划分为轨迹段.

第 2 步.语义注释,结合文本、路网、区域功能等信息挖掘每个轨迹段的语义.

如图 3 所示,本文总结了上述计算方式和架构,结合每一步数据特点,提出了轨迹语义注释的框架.

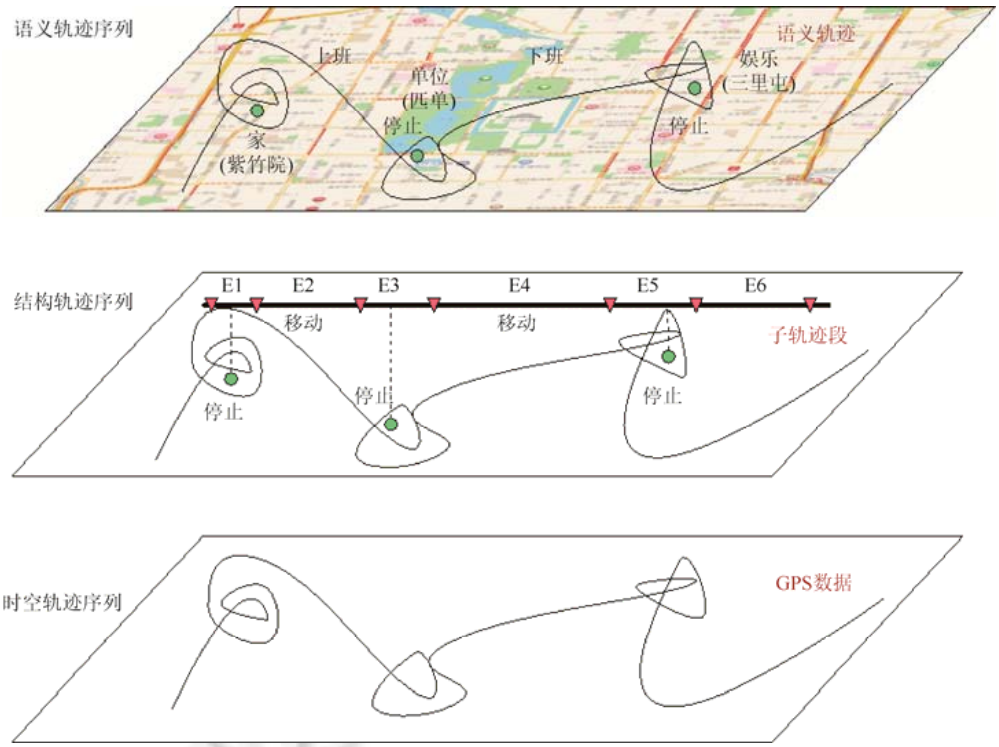

Fig.3 The framework of trajectory semantic annotation

图 3 轨迹语义注释框架

轨迹注释框架分为 3 层,第 1 层是时空轨迹序列,对应传感器收集到的原始数据;第 2 层是结构轨迹序列,根据轨迹中位置点的变化特性,将轨迹划分成片段(episode),每个片段内轨迹的结构特征相同.如图 3 所示,根据移动、停止特征可将轨迹划分成 E1~E6 这 6 段;第 3 层是语义轨迹序列,对结构轨迹中的每一个片段作语义注释,通过匹配地图或其他语义知识库得到语义轨迹序列,图 3 中标明某位上班族一天部分时间的语义轨迹.

轨迹语义注释的研究工作大部分都可以归纳到此框架下,主要关注的问题有两点:第一,如何由时空轨迹得到结构轨迹;第二,如何由结构轨迹得到语义轨迹.针对第 1 个问题,文献[35]首次定义了轨迹语义注释的问题,并针对利用停留点语义来注释轨迹语义的问题提出了解决方案.文献[34,36]从轨迹概念模型和用户个人模型的语义角度生成轨迹的注释.文献[37]也将轨迹中的停留点作为描述轨迹语义的重要因素,结合速度、移动方向等因素推断停留点的语义信息.利用停留点序列的语义,表达整个轨迹的语义.文献[38]研究了由时空轨迹序列到结构化 POI 序列的匹配方法,指出目前大部分基于最近邻查询(nearest neighbor query)的研究工作在 POI 点密集情景下不够准确的问题,作者提出了一种基于 GOIs(geometries of interest)的方法,这种方法可以更好地匹配轨迹点与 POI,生成 POI 序列.

针对第 2 个问题,文献[39]发现在结构轨迹长度较长时,分别对每个片段都作语义注释会造成整条轨迹语义理解的偏差.作者提出了一种基于条件随机场的轨迹语义划分方法,采用序列标注的方式,为整条轨迹生成语义注释.通过调整条件随机场的参数,可以将轨迹划分为不同粒度的子轨迹.然后,使用统一的语言模板表达每个子轨迹的语义,并综合所有子轨迹语义得到整条轨迹语义.文献[40]研究了由结构轨迹推断用户轨迹中频繁出现位置语义(家、工作地等)的问题.作者首先对单一用户的轨迹数据做物理位置抽取,得到结构轨迹序列;然后对结构序列中的每个物理位置分析其时间维度、空间维度、序列维度的属性,得到轨迹物理位置的属性序列;最后使用隐马尔可夫模型对序列隐状态进行标注.实验结果表明,该模型对家、工作单位、旅馆、超市等特征显

著的地点标注效果比较好.文献[41]使用社交网络中用户发布的消息对轨迹作注释,由于社交网络消息的频率远远低于轨迹,如何把轨迹与社交网络内容序列作语义匹配成为了最大的问题.文中提出衡量两者差异的空间匹配系数(SMC)和局部匹配系数(LMC),基于这两个系数找出匹配的社交网络序列,使用其中的文本信息对轨迹序列作注释.

另外,还有研究人员对轨迹语义注释提出了一些新的问题,文献[42]提出了一个语义轨迹的近似关键词检索问题,在有一个已经构建完成的语义数据库的前提下,语义轨迹的近似关键词检索主要解决检索一个关键词集合、语义轨迹数据库返回与这些关键词最相关的 k 条轨迹采用检索方法的问题.这种检索模式与传统时空语义检索的最大区别是这种检索没有空间约束,需要考虑数据库中所有轨迹.该文首次提出了空间-文本效用函数(spatio-textual utility function)以用于度量关键词与轨迹的相关性.随着可用的语义数据集数量的增加,如何选取合适的语义数据集用作轨迹注释成为一个问题.文献[43]从外部数据集选择的角度研究了如何选取最合适的数据集用于生成轨迹的语义注释,该文作者认为,轨迹访问过的位置对轨迹的语义作用比较大,提出了一种访问位置与语义数据集的关联算法,为轨迹的语义数据集做排序.文献[44]研究了如何在事件信息缺失或者在轨迹绝对路径缺失的条件下完成轨迹语义分割,假设轨迹生成过程满足马尔可夫性,文章提出了一个基于轨迹条件熵的模型,通过计算轨迹的条件熵,在熵最大处做轨迹分割.实验结果表明,该方法对数据缺失的情况有很好的效果,比单纯使用地理信息分割的准确率提高 20%.

(2) 兴趣元素集语义注释.我们定义兴趣元素集包括兴趣点(POI)、兴趣区域(ROI)和路径(path).基于位置匹配的兴趣点语义注释的研究工作可分为:POI 语义注释、ROI 语义注释、Path 语义注释 3 类.对每类涉及的研究工作总结如下.

POI 语义注释.POI 语义注释工作根据数据集是否带有文本信息可以分为两类,第 1 类是带有文本信息的数据集,这类数据主要有 tweet、微博、签到数据等,对于此类数据集,POI 语义注释的主要任务是考虑时空相关性、用户偏好等因素,为地理位置匹配或生成语义标签.主要工作有:文献[45]使用带位置的 tweet 消息(geo-tweets)为 POI 作语义注释,作者提出了一种基于概率图模型的方法,用于判断一条 tweet 消息是不是与 POI 相关.模型综合考虑了文本的信息、空间特征和用户行为特征,可以准确地判断 tweet 与 POI 的关联关系.实验结果表明,通过这种方法生成的 POI 注释比仅仅通过位置排序或聚类的效果要好.文献[46,47]利用 geo-tweet 信息对城市范围内的位置的语义关键词作注释,这个任务需要处理含有大量噪声的 tweet 文本,提取能表达语义的关键词,然后将位置信息与关键词匹配,作者将这个任务转换为估计单词在空间上分布的概率,并提出基于频率(tf-idf)的方法、基于高斯混合模型的方法、基于核密度估计的方法,通过实验发现,使用核密度估计能够很好地估计单词分布,解决位置注释的问题.文献[48]提出一种使用用户签到数据,结合用户偏好,自动生成位置数据的语义标签的方法.方法首先对位置与标签的关系建立概率模型,然后使用层次聚类生成与位置最相关的标签.实验对比了自动生成的标签与人工标签,表明这种方法可以得到有意义的语义标签.文献[49]将用户的签到数据与位置数据作匹配,用于推断位置的名称.文献[50]使用用户签到数据,从中提取兴趣点的特征(包括明显模式特征和隐含模式特征),最后使用 SVM 为兴趣点分配事先规定好的语义注释标签.

第 2 类是纯 GPS 数据集,这类数据集主要有人类移动、动物迁徙等.对于这类数据,POI 语义注释的主要任务是将数据集中的特征点(如:停泊点、弯曲点等)与语义地图中的地点作匹配.主要工作包括:文献[51]利用手机收集到的用户数据,判断 POI 的语义信息,作者认为,语义注释过程中有 3 个重点:位置点距离度量、区域属性、语义类别.文章中采用 Haversine 距离作为两个位置的距离度量,采用 DBSCAN 方法对兴趣点区域聚类,采用核密度估计与核判别分析的方法对兴趣点的语义信息分类.由于用户停留地点大多在室内,室内 GPS 信号比较弱,匹配用户停留点与已知语义 POI 的准确率往往不高.文献[52]针对上述问题,提出通过时空密度估计和行计数推断(line count inference)的方法,实现停留点与 POI 的精确匹配,实验结果表明,该方法匹配准确率高达 96.5%.

ROI 语义注释.与 POI 语义注释不同,ROI 有明显的区域边界.因此,对于 ROI 语义注释的工作需要研究如何合理地划分 ROI 边界的问题.针对不同的应用场景,ROI 语义注释的方式往往也不相同.文献[53]提出了一种简

单、有效的方法检测 ROI 的范围.之前大多数研究工作都是采用基于密度聚类的方法圈定兴趣区域的范围.对于某些计算速度要求高、准确度要求相对较低的应用,这种方式不太适用.作者将 ROI 的范围划定分为两步:首先,选取一个 POI,以点位置为中心划定圆型区域;然后,判断区域内 POI 的个数,如果大于阈值,则确定这些 POI 是一个集合,进行下一步,否则,重新选 POI.集合确定之后找出 POI 集合的一个凸包,将其作为 ROI 的边界.文献[54]使用人类移动数据和已知语义的 POI 数据推断城市的功能区(商业区、工作区、居住区等).作者利用自然语言处理中的主题模型(topic model)将区域当作一个文档,区域功能作为主题,将 POI 当成源数据,例如文档作者、单位、关键词等,将用户的移动模式当作文档中单词.这样,可以表达成区域功能的分布,得到不同区域的不同功能.实验中利用北京轨迹数据集验证了方法的有效性.文献[55]使用 Foursquare 社交网络的数据,提出了一种基于谱聚类的算法,利用地点的分类数据对用户行为和区域的语义情况建模.文献[56]使用核密度估计的方法对区域人口密度大小进行推断,文章对比了使用 KDE 和不使用 KDE 两种方法,发现使用 KDE 可以更好地过滤噪声,对区域人口密度推断的效果更好.文献[57]利用带有位置信息的社交网络(LBSN)的数据,分析城市范围内有哪些商业中心,并为商业中心所在商圈确定范围.作者指出,地理位置空间上的分布情况可以通过核密度估计的方法确定,但是仅仅通过地理位置的密度轨迹不能反映用户在两个商业中心之间移动的偏好,对用户这个移动模式建模可以更准确地估计城市中的商业中心及其范围.文献[58]融合多模态数据,生成区域的时空表达,并将其应用于城市的动态模型.

Path 语义注释.Path 语义注释主要以城市中的道路为研究对象,主要研究数据集中移动轨迹与道路匹配、道路状态匹配等问题.主要研究工作有:文献[59]研究了路网匹配中从带有噪音、稀疏 GPS 序列中找到路径的方法,发现传统的基于隐马尔可夫模型的方法计算复杂度高的瓶颈在于计算转移概率过程中需要对最短路径作检索.作者提出了一种基于截断的方法,在不影响准确率的情况下提高检索计算效率.文献[60]发现,在用户使用导航的过程中,对详细理解路径转弯点语义有迫切的需求,作者提出了一种基于用户偏好的面向用户导航应用路径语义注释的系统,利用用户历史导航记录、历史路径偏好、历史经过地标点推荐用户熟悉的路径.文献[61]研究了目前对路径消耗(通行时间、油耗)描述的模型,指出目前的路径描述模型是一个带权图的模型,将一条路径划分成许多小段,为每一个小段分配权重.作者提出了一个新的描述模型,称作混合图(hybrid graph),使用这个模型可以对路径作更准确、更有效率的消耗分布估计.文献[62]通过轨迹聚类的方法发现路径,作者提出了一种新的轨迹相似度计算方法,该计算方法可以感知轨迹的线索(clue).在真实数据和仿真数据上的实验结果表明,基于这种相似度的轨迹聚类方法比已有方法效果要好.文献[63]研究了利用轨迹发现道路的问题,文中定义了 *K*-基础路径(*k*-primary corridors)问题,给定一个轨迹集合,将集合中的轨迹分成 *K* 组,每一组的中心路径代表一条基础路径.传统的解决方案在计算轨迹相似度时消耗的计算资源太多,作者提出了一种基于最短路径的算法以降低资源消耗.实验结果表明,在不影响实验结果的基础上,基于最短路径的算法有效地降低了计算开销.

2.2 地理位置隐含语义理解

地理位置隐含语义理解是指通过融合先验知识或外部数据集挖掘地理位置中的隐含语义信息.由于决定地理位置隐含语义的因素非常多,并且在不同的数据集中各种因素的影响力也各不相同,一直以来,地理位置的隐含语义理解都是一个非常困难的任务.但是,随着基于位置的社交网络的流行,带有文本、图片信息和位置信息的数据可以被大量收集,融合多源数据为理解地理位置的隐含语义提供了新的机会.本文总结目前地理位置隐含语义理解的研究工作,根据隐含语义的变化情况,把研究工作分为基于时间主题模型的隐含语义理解、基于空间主题模型的隐含语义理解、基于时空主题模型的隐含语义理解.本节将分别介绍这 3 类方法的相关研究工作.地理位置隐含语义理解分类如图 4 所示.

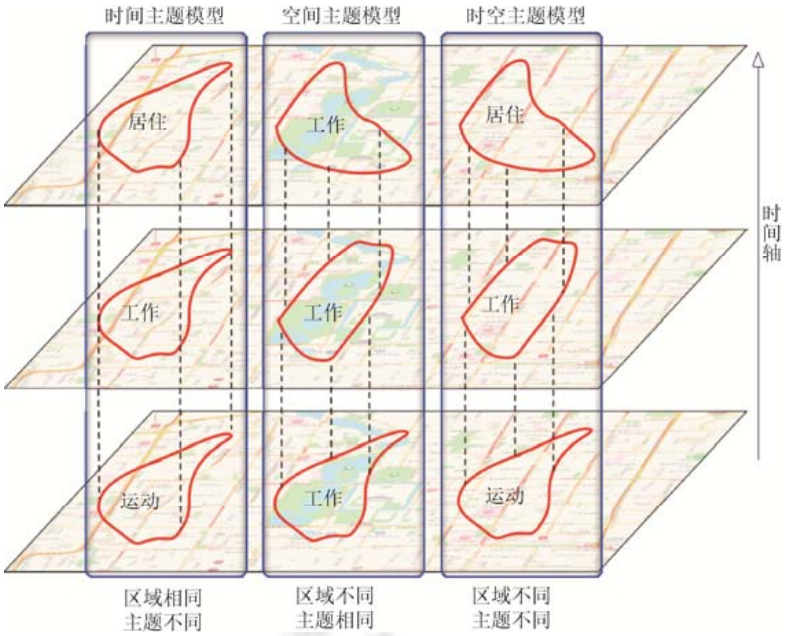

Fig.4 The classification of latent semantic understanding of locations

图 4 地理位置隐含语义理解分类

2.2.1 基于时间主题模型的隐含语义理解

基于时间主题模型的地理位置的隐含语义理解是指,考虑时间因素对固定区域或位置隐含语义的影响,对时间维度上区域内由于 POI 变化、轨迹变化、用户活动等因素造成的地理位置的隐含语义变化建模,用于发现同一区域随时间变化产生的不同主题.经典的基于隐含语义分析(latent dirchlet allocation,简称 LDA)[65]的主题模型都假设数据集中的文档都是顺序无关的,这个假设在面向地理位置的隐含语义理解中就很适用.文献[64]提出了一个动态主题模型(DTM),该模型考虑文档的先后顺序,设计了比 LDA 更丰富的后验分布,使之更加适用于时间序列的主题挖掘.利用 DTM 挖掘目标区域隐含主题,首先分区域提取不同区域中的文本数据,之后利用 DTM 生成随时间变化的区域主题.文献[66]考虑到区域的主题变化不一定满足马尔可夫性,提出了时间主题模型 ToT(topic over time),利用这个模型可以发现连续时间下区域的主题变化情况.

2.2.2 基于空间主题模型的隐含语义理解

基于空间主题模型的地理位置隐含语义理解是指,考虑到同一隐含主题出现的区域不同,发现隐含语义相同或相似的不同空间区域.文献[54]利用类似文档主题模型的方法对空间数据建模,从而发现城市范围内不同的功能区域(商业区、工作区等).文献[67]提出了 LATM(location aware topic model),这个模型在 LDA 的基础上加入了位置信息,保证生成的文档的聚类都是在空间接近的区域内.文献[68]使用带有位置信息的 tweet 数据,提出了发现不同主题的连续空间区域(coherent geographical region)方法,以及比较不同区域之间的隐含主题方法.作者提出了基于 PLSI(probabilistic latent semantic indexing)的主题模型(latent geographical topic analysis,简称 LGTA),将 tweet 中的位置与文本结合,认为区域中的 POI 和单词都是由区域生成的,这样就可以生成对区域的聚类,得到主题类似空间区域.这样得到的区域主题是静态的,只随训练数据集的改变而改变.在 Twitter 和 Gowalla 数据集上的结果表明,上述方法可以很好地发现隐含语义相似的区域,作者还将这种方法应用到 POI 推荐和用户预测任务中,取得了很好的效果.文献[69]与文献[68]相似,也假设目标区域生成了其内部的元素,作者在 LGTA 的基础上,对 Twitter 数据中的多样性进行建模,考虑到文本主题多样性、地理区域多样性、用户偏好多样性等因素,并且将具有马尔可夫性(用户的下一个位置仅与上一个位置相关)的用户位置以参数的形式加入主题模型.实验结果表明,这种方法可以发现基于位置的有趣的主题.文献[70]也是基于 LGTA 的扩展,作者假设文

本的位置信息服从二维的高斯分布,区域服从一个多项式分布,将位置与区域的信息加入图模型中,用于推断隐含语义相似的区域.文献[71]使用 Twitter 和 Yelp 数据,考虑到用户的移动偏好对区域的隐含语义信息有一定的影响,例如喜欢运动的用户常常出现在与隐含语义带有运动的区域或地点,作者将用户移动偏好信息加入到图模型中,用于准确地生成语义相似区域,并且将模型用于用户的位置推荐,准确率比已有方法提高 50%.文献[72]提出了一个集成的生成模型,用于对地点、主题、用户偏好的联合分布进行建模.不同于之前都将数据建模成一个扁平化模型的方法,作者提出的模型 nCRF(nested Chinese restaurant franchise)可以自动地发现位置和区域之间的层次关系.实验结果表明,利用这个模型可以将位置估计的不确定性降低 40%.

2.2.3 基于时空主题模型的隐含语义理解

基于时空主题模型的地理位置隐含语义理解是指,在区域划分方法和主题选取规则不变的情况下,考虑到地理位置的隐含语义信息可能存在时间上的变化,也可能存在空间上的变化,即同一区域的隐含语义会随着时间的变化而变化,相似语义的区域范围也会随着时间的变化而变化.时空主题模型就是对上述变化情况建模,动态地生成地理位置隐含语义信息.目前,对于这个问题的研究工作不多,主要有:文献[73]分别对上述两种变化情况进行建模,基于概率图模型建立了 DSTTM(downstream spatio-temporal topic model)和 USTTM(upstream spatio-temporal topic model),这两个模型都可以用于发现地理位置的主题和区域划分,通过综合两种模型的结果解决动态位置隐含语义生成问题.文献[74]提出了基于LBSN挖掘地理位置时空隐含语义的问题,并将这个问题分解为 3 个子问题,即主题抽取、位置主题生命周期生成、主题快照生成.文中提出了一个概率模型来解决上述问题.实验验证基于 3 个 LBSN 数据集,结果表明,模型可以有效地发现时空主题模式.文献[75,76]从用户的角度研究社交网络用户数据的生成过程,作者提出用户位置生成的过程受区域、主题、时间这 3 个因素的影响,利用这种模型可以生成地理位置的隐含主题.

3 用户行为的语义理解

大部分时空数据是由用户的活动产生或收集用户活动得到.其中,含有大量的用户行为特征信息.通过上述地理位置语义理解,可以得到大量的带有语义信息的 POI、ROI、语义轨迹数据.基于这些数据,用户行为的语义理解的目标在于通过分析带有用户或用户群身份信息的时空数据,总结用户活动规律和移动模式,理解用户个性化的语义.因此,用户行为的语义理解包含两个核心任务:用户行为模式挖掘和用户个性化语义理解,其研究框架如图 5 所示.

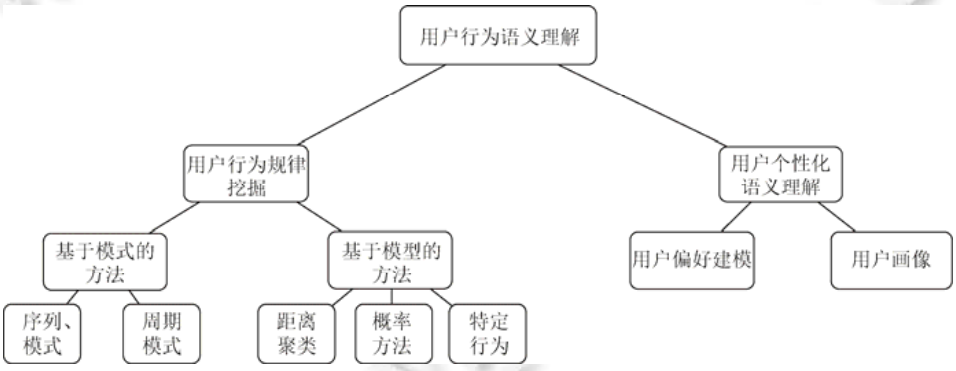

Fig.5 The research framework of semantic understanding of user behavior

图 5 用户行为语义理解研究框架

3.1 用户行为规律挖掘

用于行为规律挖掘的主要目标是,挖掘用户或用户群行为活动中的隐含规律.目前针对用户行为模式挖掘的研究工作有两种思路:(1) 基于模式的用户行为规律挖掘;(2) 基于模型的用户行为规律挖掘.

3.1.1 基于模式的用户行为规律挖掘

基于模式的用户行为规律挖掘方法,通常基于某一类预先定义好的模式去发现用户时空数据中隐含的行为规律.现有预先定义的用户行为模式有序列模式和周期模式,对应两种模式的挖掘方法特点的对比情况可见表 4.

Table 4 The comparison of user mobility patterns

表 4 用户移动模式比较

移动模式	英文名	轨迹用户数量	轨迹长度
序列模式挖掘	SPM (sequential pattern mining)	多用户多轨迹	较短
周期模式挖掘	PPM (periodic pattern mining)	单用户	较长

用户序列模式挖掘.用户行为的先后顺序常常有隐含的规律,序列模式挖掘就是从含有语义的时空数据中挖掘出有规律的序列.自从序列模式挖掘被文献[77]提出以后,一直都是一个热点的研究问题.由于时空数据在空间位置上的连续性使得传统序列模式挖掘的方法不适用于时空数据.文献[78–80]是时空数据中的序列模式挖掘的先驱性工作,作者利用空间划分策略,将时空数据按提前定义好的空间粒度划分小网格,之后统计用户在网格内的移动情况用于挖掘的序列模式.尽管这样的方法简单、高效,但网格划分会带来尖锐边界问题(sharp boundary problem),即对处于网格边界附近的数据处理效果不好,这个问题会随着网格粒度细化变得越来越严重.文献[81]首先利用轨迹语义注释中停留点检测技术,并将停留点与背景地图作匹配,得到语义轨迹,然后抽取频繁出现的位置序列作为序列模式.然而,由于空间连续性的问题,这样的方式不适用于对大范围内的序列模式进行检测.文献[82]将用户的序列模式定义为语义的标签(例如学校-公园),通过频繁模式挖掘的方法发现频繁语义序列.作者认为这样的语义序列可以反映用户的行为偏好,基于这种方法,可以更好地预测用户下一个出现的位置.文献[83]考虑到之前的研究工作对用户行为模式的发现大多只考虑时空因素、语义因素的研究工作,作者提出了一种基于 PrefixSpan 的序列模式发现算法:STS-TPs.由于 PrefixSpan 需要符号化(symbolization)的序列,作者提出了 SS 和 ASS 两种算法,将包含空间、时间、语义属性的序列符号化.文献[84]研究了基于用户语义轨迹的序列模式挖掘问题后,发现目前已有的序列挖掘算法不适用于语义轨迹序列中的模式挖掘.其主要原因在于已有的序列挖掘算法(如 PrefixSpan 等)都假设序列中点的每一项(item)是相互独立的,然而在用户语义轨迹中,用户上一个位置与下一个位置并不是相互独立的.作者定义了语义轨迹的细粒度序列模式挖掘问题,并指出解决这个问题需要考虑空间紧密性(spatial compactness)、语义一致性(semantic consistency)及时间连续性(temporal continuity)3 个因素.作者提出了 SPLITTER 方法以解决上述问题.方法分为两步:第 1 步,从语义轨迹数据集中抽取粗糙的空间模式,每一个模式对应一个语义轨迹集合,作者提出了加权片段转移(weighted snippet shift)方法用于检测语义轨迹集合;第 2 步:对每一个粗糙模式采用自顶向下的方式,利用分治策略渐进地将其拆分成细粒度的模式.最后,通过仿真数据实验和真实数据实验验证了 SPLITTER 的有效性.文献[85]定义了序列模式中的空间共同演化模式,并提出了一种检测方法:Assembler,用于高效、准确地挖掘上述模式.

用户周期模式挖掘.用户常常有一些周期性的活动行为,例如在家与工作单位之间往返等.挖掘用户周期性的活动规律可以广泛应用于用户位置预测、用户推荐等领域.文献[86]研究了从移动物体的轨迹数据中挖掘周期性的规律,文章将周期模式挖掘拆分成两个问题,即如何从复杂的物体移动数据中检测周期数据段;如何理解周期内的移动行为.作者首先将用户轨迹数据与地图数据作匹配,发现轨迹中的参考点(reference spot),得到参考点与参考点之间的数据段,并以此作为用户移动周期的基本组成单元.然后,对周期行为规律建立概率模型,针对一个特定的数据段,通过层次聚类的方法将数据段聚集在一起,形成用户移动周期.通过仿真实验和真实数据实验表明,上述方法可以有效地发现用户移动行为模式中的周期规律.文献[87]考虑到目前时空数据存在冲突、缺失、噪音大的缺点,如何从不完整的观测数据中发现用户移动行为的周期模式变得非常重要.作者基于这个问题提出了一种考虑时空数据缺陷的时空数据周期度量方法和一种检测算法,用于发现周期规律.通过仿真数据和真实数据的实验,验证了方法的有效性.文献[88]对城市路网的交通情况建模,分析路网交通情况的周期性变化,作者首先通过单条道路上传感器采集到的速度数据,发现周期模式,并用概率分布表达速度的周期变

化,然后通过基于密度聚类的方法,对路网上所有传感器的周期行为属性及空间距离属性聚类,合并属性相同或相似的类,得到含有少量节点的路网交通情况.其结果可以用于道路交通预测等应用.文献[89]考虑到基于位置社交网络用户的签到行为具有周期性,并且用户的移动行为常常受到其附近朋友移动行为的影响,作者提出了一个概率模型,该模型利用基于周期衰减核(periodic decaying kernel)的双随机点过程(doubly stochastic point process),对用户行为记录的时间因素建模,利用时变多项式分布对用户位置建模,并利用 EM 算法学习模型的参数.实验结果表明,运用上述方法对 Foursquare 数据集中用户的周期性行为建模比原有方法效果要好.

基于模式的方法可以快速、有效地挖掘出含有预先定义好的行为模式,其优势在于算法简单,计算开销相对较小.但是,这种方法存在两点局限性:首先,行为模式的预先定义需要依赖先验知识;其次,用户的行为非常复杂,常常受各种因素(天气、突发状况、心情等)的影响,基于模式的方法只能理解预定义模式下的行为规律,无法对用户行为的不确定性建模,即对于未知模式的发现能力较弱.

3.1.2 基于模型的用户行为规律挖掘

基于模型的用户行为规律挖掘方法,通常基于对时空数据分布的假设(如:用户位置服从二维正态分布,行为发生时间服从泊松分布),对数据的生成过程建模,利用统计模型(如隐马尔可夫模型、概率图模型等)发现用户的行为规律.根据使用方法不同可分为基于距离聚类的方法、基于概率模型的方法和基于特定行为的方法.

基于距离聚类方法.轨迹聚类作为传统轨迹数据挖掘的一项重要任务,可以应用于用户行为规律挖掘.聚类得到的轨迹簇可以看作是簇内的用户行为相同或相似.基于距离的轨迹聚类是轨迹聚类中最常用的一种.对反映用户行为规律的轨迹聚类需要衡量轨迹之间的相似性,常见的轨迹相似性度量有 EDR[90]、DTW[91]、LCSS[92]、HD[93]、CPD 和 SPD,其计算方式和特点见表 5.

Table 5 Distance measurement of trajectory

表 5 轨迹距离度量

名称	英文全称	计算方式	特点
EDR	Edit distance on real sequence	计算将一条轨迹编辑为另一条的开销	对噪声鲁棒,但计算量较大
DTW	Dynamic time warping	复制记录,排列轨迹使轨迹对齐后计算开销	可处理变长轨迹,但重排列增加了计算噪声
LCSS	Longest common subsequence	计算两条轨迹的最长子序列记录长度	对噪声鲁棒,但计算量较大,适用于密集轨迹
HD	Hausdorff distance	将轨迹看作点集,计算最小最大距离	有几何意义,但得到的距离具有方向性、不对称
CPD	Closet-Pair distance	计算两条轨迹的最小距离记录	计算简单,但没有考虑整个轨迹的分布情况
SPD	Sum of pairs distance	计算两条轨迹记录对应点的距离之和	计算简单,但不能处理记录长度不同的轨迹

在具体应用中,使用自定义轨迹相似性度量的研究工作有:文献[94]对轨迹聚类算法和轨迹相似度测量进行总结,首先对数据聚类算法进行了分类与归纳,同时系统化地介绍了轨迹数据算法,从基于空间聚类、基于时间聚类、基于路网匹配聚类和基于语义轨迹聚类等方面进行介绍.文献[95]提出了在线挖掘轨迹数据流中移动模式的框架,整个框架分为 4 步:第 1 步,基于滑动窗口,划分实时轨迹数据中的自轨迹段;第 2 步,基于轨迹距离度量,对自轨迹段内的轨迹聚类,每类轨迹维护一个小组(micro-group);第 3 步,作者提出一种增量更新算法,算法随时间不断更新组内的轨迹段和用户;第 4 步,基于小组内用户和轨迹的变化情况,总结用户行为的变化规律.文献[96]中,Lee 等人提出先划分后聚合的框架,按照最小描述长度原则划为子轨迹,利用基于密度的聚类方法处理.

基于概率模型方法.现有的基于概率模型的方法都是对用户的移动行为建模,得到用户的移动模型,基于这个模型预测用户未来的位置或行为.Younghoon 在文献[97]中研究了面向用于行为规律发现的用户轨迹的主题发现问题,作者提出基于隐含主题的 geo-tweetd 聚类模型,使用图模型结合 HMM 对数据进行建模,用于发现主题相似的时空区域和轨迹模式.Chao[98]分析了 geotweet 中的行为偏好相似用户的行为模式通常也比较接近,考虑到这个因素,作者提出了 GMove 方法,将用户的移动行为建模拆分成两个任务,即用户的移动位置预测与用户分组.作者使用 HMM 模型同时对两个任务训练,为每一组的用户建立并训练一个 HMM 预测用户位置,预测结果反馈到用户分组,不断迭代到模型收敛.实验结果表明,基于这种方式预测用户移动位置的准确率比原有方

法要高.Hongzhi[99]基于签到数据研究用户移动的行为规律,用户的签到数据非常稀疏并且签到伴随的文本数据质量往往不高,因此需要结合用户所在位置的时间、空间、口语化、上下文信息对用户的签到数据建模,但是目前缺乏将这些因素结合起来的方法.作者提出了基于联合概率分布的生成模型,用于描述签到数据的生成过程,并将模型应用在 POI 推荐中.通过对在用户“居住地”和“旅行”两个场景的实验表明,上述模型推荐结果比原有模型要好.文献[100]利用手机记录的用户地点数据和两个基于位置的社交网络数据,分析了影响用户移动的因素,指出用户的社交关系影响用户 10%~30%的移动行为,周期性的行为可以解释 50%~70%的行为,并且这两类行为往往不重合.因此,作者建立了一个结合周期性移动和用户社交关系的用户移动行为预测模型,实验结果表明,上述模型对用户预测有比较好的效果.为了解决用户行为数据稀疏的问题,文献[101,102]使用深度表征学习的方法,融合多种外部数据,学习用户行为表达,并将其应用于构建时空用户行为模型.

基于特定行为方法.在现实应用中,人们往往只关心某种特定行为(如:交通事故、出租车绕路等).基于特定行为的方法针对某一类行为的特征,挖掘并发现行为规律.文献[103]研究了如何利用大量异构数据推断城市交通事故的问题.作者通过人在城市中的移动数据来推断出现交通事故的风险,并分析了东京 7 个月内的事故数据及 160 万人的移动数据,用于训练模型.该模型基于 Stack Denoise Autoencoder,学习层次化的用户移动特征表达.当实时数据输入时,此模型可以输入交通事故风险的预测结果.文献[104]利用移动 GPS 数据和交通工具网络数据,研究了城市人员交通出行模式预测的问题.文中提出了 DeepTransport 模型,其核心是利用深度卷积神经网络来理解用户移动和交通出行的模式.输入一段任意时间段内的用户移动数据,模型会输出用户未来的移动方向及交通方式.文献[105]通过分析车载 GPS 数据,识别跟车行为,并分析驾驶模式.作者提出了一种先划分后聚类的方法.首先,对驾驶行为进行特征提取,文中提取了纵向加速度、横向加速度、偏航率、偏航角度、车道偏移、车辆速度等 8 个因素.然后,基于这些属性对跟车行为定义规则,从数据集中自动检测出符合跟车行为的数据段.最后,对提取出的数据段聚类,根据聚类结果识别驾驶模式.Siyuan[106]研究了如何检测城市中出租车欺诈行为的问题,发现由于定位不准、路况复杂、中途停车等因素,现有的轨迹异常行为检测算法不适用于检测出租车欺诈.考虑到有欺诈行为的出租车往往会人为调整计价器,使得其记录距离比真实距离要远,从而多收取费用.但这也会造成计算得到的出租车平均速度变快,而 GPS 定位产生的距离和速度是准确的.基于上述特点,作者提出了一个基于速度的出租车欺诈检测系统,用于检测欺诈行为.

3.2 用户个性化语义理解

由于用户差异性的存在,时空数据中相同的位置对不同用户也会产生截然不同的语义.准确理解用户个性化语义信息对于针对用户偏好的推荐和预测起着重要的作用.用户个性化语义理解按任务可以分为用户偏好建模和用户画像.

3.2.1 用户偏好建模

时空数据中用户差异性最直观的表现就是,不同的用户反映出的偏好不同.如何利用特定用户历史时空数据,建模用户偏好,对于用户行为语义理解的应用有重要的意义.利用 twitter 数据,文献[107]首次从空间、时间、活动这 3 个方面发现单个用户的移动行为偏好.作者基于时空活动特征提出了一个刻画单个用户 Who、Where、When、What 的概率图模型,用于在给定 tweet 内容和发布时间的前提下推断用户的位置.实验数据集涵盖世界范围和美国范围 geo-tweet 数据集.结果表明,该模型可以准确地反映单个用户的偏好,在用户位置预测任务上的表现比已有方法要好.文献[108]研究了如何利用多维度知识对用户签到偏好进行建模.考虑特定时间因素、位置流行度因素、用户类型因素、地理偏好因素影响,作者提出了一种基于最大熵判别的优化模型.该模型充分考虑到时间变化和聚集行为对用户签到偏好的影响.实验结果表明,该模型在用户签到、地点推荐和用户位置预测任务上效果比现已有方法要好.文献[109]利用用户在公交和地铁产生的轨迹数据,融合站点附近的 tweets 数据,估计用户对交通工具中投放广告的偏好.基于概率模型 LDA,作者提出 gLDA 建模数据的生成过程,并提出一种高效的 top-k 检索算法用于公共交通系统中的广告推荐.同样,基于线下新加坡公共交通数据和线上 twitter 数据,文献[110]提出 CO^2 来打通线上线下两类数据,并用于建模用户偏好.实验结果表明,CO^2 在区域主题发现和用户偏好推断两个任务中的准确率显著高于其他方法.另外,作者通过问卷调查的方式验证了 CO^2 的有效性.

文献[111]结合用户移动轨迹数据和 geo-tweet 数据,理解用户移动记录的语义.作者首先从轨迹数据中提取特征点;然后对特征点附近 geo-tweets 的单词概率分布建模,找出热点单词;最后综合用户全部轨迹,使用马尔可夫随机场生成用户移动记录的个性化注释.考虑到不同用户在导航应用中对路径提示的不同需求,文献[60]提出了基于路径摘要的个性化导航框架,该框架的目标在于生成更直观、更贴近用户习惯的路线规划信息,并将其转换为语音提示.作者从用户历史导航数据中抽取出用户经常访问的地标(landmark)和路径(path),生成面向单个用户的路径个性化摘要,从而生成个性化导航.文献[112]针对用户偏好建模的任务,提出了面向用户偏好的用户行为模型,该模型从带有地理信息的文本数据中挖掘用户的行为偏好,并应用于用户个性化地图生成.考虑到用户活动中产生的数据具有序列性,但现有的对基于用户偏好的空间对象(spatial item)推荐中没有用到活动位置的序列性信息,文献[113]提出了基于序列的空间目标个性化推荐框架(SPORE).此框架利用一个描述“主题区域”的隐变量,对用户的个性化偏好和序列模式建模,可以缩小推荐空间区域进而解决数据稀疏问题.在算法实现层面,作者设计了一种非对称地理位置敏感哈希算法,用于加速在线 top-K 空间对象推荐.实验结果表明,对比原有方法,该框架在空间对象推荐的准确度和算法效率两个层面都取得了显著的提升.

3.2.2 用户画像

有许多研究工作希望通过理解时空数据蕴含的语义信息,生成用户画像.这一技术可应用于用户身份识别和用户分组.文献[114]研究了如何利用签到数据建立用户画像(profile).文中指出,虽然已有许多工作利用社交网络关系、tweet 数据推断用户信息,但是这些工作都没有利用位置信息.作者提出了一个利用用户位置推断用户属性的框架(location to profile),该框架从用户签到数据中抽取单个用户的空间、时间和位置知识,并利用张量分解(tensor factorization)的方法生成用户对签到位置的偏好,进而推断用户属性.实验结果表明,该方法的推断准确率明显高于原有方法.文献[115]研究了利用用户分组信息,判断不同数据源的轨迹是否属于同一个用户.作者提出了一个基于 MapReduce 的用户自动识别 AUI(automatic user identification)框架,针对数据源可能存在不同的采样率和噪音模式,设计了一个信号层面的用户相似度度量(siginal based simulity measure),用于衡量从不同数据源采集到轨迹的相似度.文中指出,AUI 框架不仅能为单个用户多数据源的数据集成提供一个有效的方式,还从更深层面揭示了异构数据源中相同用户的确存在的特定移动方式.文献[116]利用用户移动数据中访问的特定位置信息来识别用户,从用户移动的速度、方向、距离这 3 个因素入手,推断用户身份.文献[117]研究了如何利用社交网络中带有位置信息的数据识别社交网络中的用户,并通过识别 POI 对不同用户的特定含义(家、单位)来识别用户身份.

4 热点事件的语义理解

时空数据中的热点事件是指在特定区域和时间段发生的、有用户参与的、对用户后续行为有影响的事件.热点事件的语义理解主要是挖掘事件产生和发展过程中描述事件状态的全部或部分属性.事件的全部属性可以概括为事件发生的时间(when)、地点(where)、人员(who)、内容(what)、成因(how),即事件的 W4H1 属性.本文根据时效性将热点事件的语义理解分为两个主要任务,即热点事件的离线分析和热点事件的在线监测. 不同的事件检测任务对事件的状态属性有不同的侧重.本节将分别对这两类任务的研究工作和技术进展作总结.热点事件语义理解研究框架如图 6 所示.

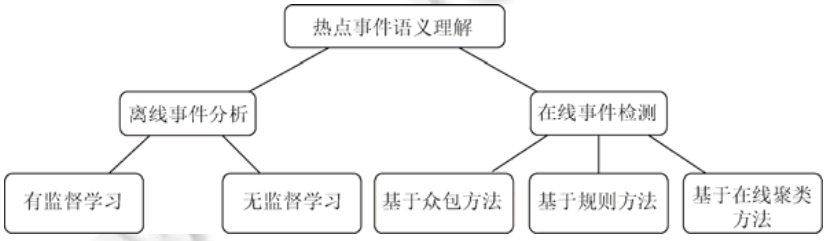

Fig.6 The research framework of semantic understanding of event

图 6 热点事件语义理解研究框架

4.1 热点事件的离线分析

时空数据的热点事件离线分析主要是基于事件历史数据,提取事件产生及发展过程涉及到的相关数据段,分析挖掘其中模式.根据用于训练模型数据集中的事件是否有标签,离线理解的方法可以分为有监督和无监督两类.本节分别总结了这两类方法的相关研究工作.

4.1.1 基于有监督学习的热点事件离线分析

利用有监督学习的方式检测离线事件是最为常见的问题.这类问题中的事件带有标签,可以通过历史数据集中检测与已发生事件类似的事件,抽取事件属性,其结果可用于事件预测.由于历史数据有标签,因此这类问题不关注事件的内容(what)和成因(how),更侧重挖掘事件的时间(when)、地点(where)、人员(who).此类问题相关研究工作有:文献[118]指出,传统的事件序列分析主要采用马尔可夫的方法或标记时间点过程(marked temporal point processes)的方法对事件的生成过程建模,但这些方法都将事件点的时间、位置、内容独立建模,并且对事件的生成过程有诸多分布假设,而这些假设也许不能反映现实情况,作者将事件的产生建模为历史数据的一个非线性函数,提出了一种新的事件序列建模方法——循环标记时间点过程(recurrent marked temporal point process).该方法通过循环神经网络预测下一个事件出现的时间和地点.针对模型的学习,作者提出了一种随机梯度的方法,并在纽约出租车数据集、金融交易数据集、电子医药记录数据集、Stack overflow 数据集上验证了算法的有效性.结果表明,采用这种方式预测事件的准确率优于原有方法.基于社交网络中的事件数据库,文献[119]研究了事件时空语义类型抽取和分类问题.作者从事件参与者、语义类型、发生区域、与其他事件的关系这 4 个方面出发,分析 Flickr 中的事件,提出基于集成学习的方法,用于判断每个事件的语义类型.实验结果表明,基于上述模型的方式比基于“词袋”模型的效果要好.文献[120]首次提出基于 geotweets 数据从局部区域中抽取事件的框架,作者提出了一种基于网格划分的方法,并通过实验证明这个框架在局部区域事件检测上的准确率约 70%.文献[121]研究了如何从大量 Web 页面中抽取区域事件的问题.文中指出,目前的信息抽取框架大多都需要领域知识指导,这类方法不适用于对大规模 Web 数据的抽取.作者提出了基于间隔指导(distant supervision)事件属性(名称、时间、地点)的独立评分机制,用于页面分组,然后利用一种结构化群组优化方法,将群组中的页面对应到事件记录.实验结果表明,上述方法比原有方法在准确率和召回率上均有大幅提升.

4.1.2 基于无监督学习的热点事件离线分析

在更多的实际应用场景下,有标签的训练数据集常常很难得到,且用户更关心特征明显但未发生过的事件.基于无监督学习的方法正是为了解决此类问题而被提出.对比有监督学习的工作,此类工作更加注重事件的成因(how)和内容(what).考虑到当一个空间事件发生时,围绕发生位置的用户讨论的话题相比其他地点更为频繁,文献[122]提出了一种基于图主题搜索统计的方法 Graph-TSS(graph topic scan statistic),为空间划分区域,这种方法是主题模型对数似然率测试的泛化,作者首先证明了基于 Graph-TSS 的空间划分是一个 NP-hard 问题,然后提出了一种近似算法来解决.回溯阿根廷动荡事件、智利地震、美国传染病爆发等事件可以证明,该方法可以有效地检测此类事件.考虑到时空数据与信号类似,存在各种各样的噪音,文献[123]将时空数据看作信号,利用信号处理中噪音滤波器的方法解决噪音对区域事件发现准确率的影响.区域事件以关键词聚类簇表示,交替地进行关键词聚类和噪音滤波,并利用关键词的变化反映事件的变化.文献[87]对事件的周期性建模,提出算法挖掘周期性事件,并对如何在数据不完整的情况下发现事件的周期性进行了讨论.文献[124]利用用户签到的数据分析了体育赛事前后用户的消费信息存在的模式.基于 2010 年~2014 年间 21 个北美城市的 32 个篮球队数据的分析结果,作者指出,比赛签到时间较晚的顾客更倾向于来自花费较高的宾馆或酒吧.文献[125]从事件与位置的关系角度分析新闻事件产生及发展过程,通过考虑事件发生的位置和提及此事件的用户所在位置两个因素聚类,发现层次化的事件簇和参与事件用户所在的区域,并用于检测区域范围内影响力最大的事件.文献[126]通过分析用户的移动模式检测城市中的事件并估计事件影响力,作者指出,现有的事件检测及影响评估方法有两个局限:首先,大都需要依赖于社会调研等方式对事件进行标注;其次,目前方法大多适用于检测国家范围内的事件,对于城市粒度的事件检测没有作专门优化.作者提出基于用户移动模式的事件检测及影响评估框架,该方法分为 3 步:第 1 步,从轨迹中抽取用户移动的动态流;第 2 步,利用基于移动关系变化异常的事件检测方法,该方

法不仅能发现事件的时间地点,还能检测事件波及的范围;第 3 步,提取区域变化信息,通过其变化情况评估该事件的影响力.

4.2 热点事件的在线检测

由于热点事件影响力具有实效性,发现越早价值越大,因此,事件在线检测一直以来都是研究人员关注的热点问题.由于此问题针对在线数据,因此检测这类事件不必关心事件的发生时间(when).面向时空数据的热点事件在线检测的方法可以分为 3 类:基于众包的方法、基于规则的方法和基于实时聚类的方法.本节分别总结了这 3 类方法相关的研究工作.

4.2.1 基于众包的热点事件在线检测

与基于众包的语义注释相似,基于众包的热点事件在线检测是指通过众包平台实时发布众包任务,汇总众包工作者的结果,用于检测热点事件.文献[127]首次提出了基于众包的在线区域事件检测系统 iSee,系统用户使用手机标注发现的事件.为了解决用户定位过程中出现的 GPS 点漂移的情况,作者在标注过程中考虑到用户的朝向,并在汇总时采用区域网格化的方法降低误差.文献[128]研究了基于众包的在线区域事件检测中质量保证的问题,在任务分配中将众包系统与在线事件检测系统相结合,每一个区域由多名 worker 标注事件,并提出一种真值发现模型,用于判断哪一名 worker 的成果是真实事件.

4.2.2 基于规则的热点事件在线检测

最直观的热点事件在线检测方法是基于规则的方法,用户通过制定事件规则筛选出感兴趣的事件.文献[113]将区域内发布 geo-tweets 数量的变化与区域事件相关联,提出了一个 geo-tweets 数量预测模型,当预测值与真实值差距较大时,认为此区域发生了事件.文献[129]研究了海洋监视领域的事件识别,提出了船舶追踪和复杂事件识别模型,对实时数据做滑动窗口分割,基于定义的规则发现事件.

4.2.3 基于实时聚类的热点事件在线检测

基于实时聚类的热点事件在线检测方法是当前时空数据事件在线检测的研究热点.该类方法大多基于 geo-tweet 数据,通过对数据中文本主题的聚类、位置关键词的抽取等技术,可以从 Who、Where 和 What 这 3 个维度检测热点事件.相关工作有:文献[130,131]提出了一个由区域候选事件生成和区域事件排序组成的区域事件在线检测框架,并以聚类中心的代表性 tweets 的形式输出事件.考虑到现有的工作大多基于 tweet 中的内容信息,对位置信息考虑不多,文献[132]提出了关键词在空间上分布的在线计算方法,然后根据关键词的空间聚类检测区域事件.文献[133]研究了从社交媒体流数据中检测区域事件的问题,利用聚类的方法生成消息发布的时空热点簇.文献[134]研究了从大量的在线文本流中发现空间区域事件的问题.文中指出,现有方法大多基于主题模型检测文本流中的主题变化来检测事件,而这种方法在针对巨大实时数据量和区域事件检测上效果不佳.作者提出了一种基于语义浏览(semantic scan)的方法来解决上述问题.这种方法不需要人工介入或带标签的训练.由于这种方法关键在于分辨新文档簇与旧文档簇中的异常模式,而非仅仅检测某一个文档中的异常,使得这种方式对真实世界的噪音比较鲁棒.实验结果表明,与 online LDA、Topic over Time 等传统方法相比,上述方法可以显著提升检测的准确率并缩短检测所需时间.考虑到社交网络中紧急事件的在线提前检测中已有方法在含有大量噪音的短文本数据中表现较差,且没有考虑事件发生的位置等因素,文献[135]提出了一种基于在线主题聚类的方法以解决上述问题.作者利用 tweet 中提到的位置和发表 tweet 用户所在的位置来推测紧急事件发生的位置.实验结果表明,上述方法可以从不同粒度检测突发的紧急事件.文献[136]研究了区域事件的在线检测方法.文中指出,利用 tweet 数据中文本的主题在小地理范围内的变化可以检测区域的事件,但是由于带有位置标签的 tweet 信息只占所有信息的 0.7%,仅仅通过带位置信息的 tweet 不足以支撑对区域事件的推断.作者提出了一种自动化的文档位置推断方法,实时地推断 tweet 发表的位置.文献[137]同样研究了区域事件的在线检测问题,文中将检测方法分为事件候选集生成和在线事件检测两步.作者利用贝叶斯混合模型生成候选事件,利用表达学习的方法生成事件表达并检测事件.文献[138]指出,通过提取多条 tweet 消息中的关键词可以更容易地发现其中蕴含的有意义的信息.作者提出了基于 STREAMCUBE 的 tweet 事件在线检测框架,该框架使用层次化时空哈希标签聚类的方法,实现了对不同粒度空间区域、不同粒度时间段的事件发现,

并提出基于信号传递的事件分辨算法,用于生成对用户友好事件的语义哈希标签.最后,作者提出了一种事件排序算法,用于输出给定区域范围和事件段内的实时事件.

5 时空数据语义理解应用领域

通过对时空数据中地理位置、用户行为、热点事件语义的理解,可以生成语义地图知识库、用户画像知识库及事件状态知识库,并服务于多种应用.按场景可将其分为推荐类应用及预测类应用.本节分别总结了两类应用的相关研究进展情况.

5.1 基于时空数据语义理解的推荐类应用

理解时空数据中的语义信息对不同的应用场景、不同用户偏好下的用户推荐有非常大的指导意义.与时空数据语义理解的研究框架相对应,基于时空数据语义理解的推荐应用可以根据推荐对象分为:位置推荐应用、用户推荐应用、事件推荐应用这 3 类.本节分别总结了 3 类应用的主要研究成果.

5.1.1 位置推荐类应用

位置推荐类应用主要根据对用户偏好、用户当前位置、时间偏好等因素,为用户推荐 POI 或路径.文献[139]综述了 LBSN 中的 POI 推荐研究.文献[140]研究了如何个性化地为用户推荐位置,文中指出,现有方法都是基于用户访问历史、地理和时间影响,没有考虑到位置内容(context)的语义.作者提出了一种基于 Skip-gram 的模型,以用户的位置访问记录为输入,为每个位置生成一个隐含语义表达,并提出基于 pair-wise 的方法,汇集用户偏好,用于推荐 top-*N* 的位置.文献[141]研究了用户兴趣转移对用户 POI 推荐的影响,文中指出,已有方法都是基于用户地理偏好为用户推荐 POI,没有考虑用户兴趣转移的因素,即用户在不同的地理区域对 POI 的兴趣也不同.作者对用户签到记录的生成过程进行建模,提出 ST-LDA.利用 ST-LDA 可以分析出,人群在不同地理区域的不同偏好.基于地理偏好信息与用户偏好信息,联合为用户推荐 POI.实验结果表明,上述方法比原有方法更准确、有效.考虑到现有的位置推荐中存在的如用户 POI 决策建模、数据稀疏、冷启动等问题,文献[142]提出了一种基于辅助数据的方法,利用用户好友数据为用户推荐 POI.作者将用户分为 3 类,即社交类朋友、位置类朋友、邻居类朋友,并设计了一个框架为用户推荐 POI.考虑到现有的基于用户偏好、社会或地理影响的 POI 推荐方法无法解决用户在不同时间段内对 POI 的偏好不同的问题,文献[143]提出了 WWO 框架,传统的推荐框架可以被视作 WWO 框架在无限时间段长度下的扩展.文献[144]研究了在历史数据量少的情况下的位置推荐问题,作者通过分析用户的移动行为模式,发现用户的移动总是围绕若干个中心,且不同类型的用户围绕的中心也不相同.基于上述发现,作者提出了一个多中心高斯模型(multi-center Gaussian model),对用户的活动规律建模,基于这个模型可以为用户推荐 POI.文献[145]将用户的签到偏好和用户的 POI 推荐联合建模,并考虑到非城市(out-of-town)用户的特殊性为用户推荐 POI.考虑到城市中的 POI 可以按照语义信息分类(参观、博物馆、公园等),文献[146]定义了一个新的 POI 推荐问题,即基于 top-*K* 位置分类的用户 POI 推荐,作者证明了为城市中的 POI 分类问题是一个 NP 完全问题,并提出了一种近似方法以求解.

5.1.2 好友推荐类应用

好友推荐是指根据用户的语义偏好为用户推荐其感兴趣的用户.考虑到用户当前状态的社会关系、个人偏好、位置语义等信息,文献[147]提出了一种基于 random walk 的用户推荐算法(random walk based context-aware friend recommendation,简称 RWCFR),作者将基于位置的社交网络建模成一个无向无权重的图,算法基于带重启的 random walk 在图中搜索推荐结果.实验结果表明,上述方法的结果优于现有的基于流行度、基于专家、基于好友等方法.文献[148]提出基于语义信息的好友推荐系统 Friendbook,作者考虑到现有的基于社交网络图结构模型的推荐方法可能不能准确地反映用户的偏好,提出了基于传感器数据的、以用户为中心的用户建模方法.该方法利用 LDA(latent Dirichlet allocation)从用户的日常行为中提取用户生活方式,并在此基础上提出了用户相似性度量方法,将其应用于用户好友推荐.

5.1.3 事件推荐类应用

事件推荐是指基于用户的兴趣偏好,为用户推荐在其附近、符合其兴趣的事件.文献[149]最早研究如何基

于移动设备的位置数据为用户推荐事件,作者采样了波士顿 100 万用户的行为轨迹与事件映射.通过分析发现,最有效的事件推荐方法是基于区域的推荐算法,即推荐与用户所在位置相近区域发生的事件.文献[150]提出了一种基于共同贝叶斯泊松分解的方法,解决用户事件推荐中的冷启动问题,作者考虑到现有的基于事件的社交网络的事件发布数量大,且事件存在生命周期的特点,基于贝叶斯泊松分解,分别从用户对事件、社会关系、文本内容角度进行建模,然后采用共同矩阵分解的方式将各个独立部分组合起来,用于为用户推荐事件.

5.2 基于时空数据语义理解的预测类应用

根据待预测的目标,基于时空数据语义理解的预测应用可以分为:位置预测、行为预测、事件预测这 3 类.本节分别总结了 3 类应用的主要研究成果.

5.2.1 位置预测类应用

位置预测是时空数据语义理解的一个重要应用场景.根据时空数据的语义信息预测用户位置、自然现象演变等.文献[151]分析了航空管理领域决策支持系统对飞机位置预测的需求,研究了在复杂的天气条件下,预测飞机轨迹的问题.作者将空间划分成 3D 网格网络,以网格为中心构造立方体,整个空间就可以被看作立方体的集合.这样,每一个立方体都可以由立方体中心、源网格观测点、观测天气变量来定义.在此基础上,作者将飞机轨迹映射到空间立方体中,然后利用隐马尔可夫模型,在考虑飞行不确定性的条件下,预测飞机轨迹.文献[152]研究了空间变化下的预测问题,文中指出,现有的基于仿真的方法存在适用性差的问题,作者提出了基于空间变化趋势的方法,这种方法比基于仿真的方法更灵活,更适用于不确定性大的空间变化预测场景.文献[153,154]研究了基于车辆的部分轨迹预测行程目的地的问题,作者利用轨迹聚类的方式发现用户的行为模式,利用二维高斯分布描述交通流,并且生成每类模式的目的地分布.当新轨迹输入模型后,首先判断轨迹属于哪一种模式,然后根据模式内部的轨迹特征预测轨迹目的地.文献[155]研究了在稀疏数据场景下的目的地预测问题,指出目前解决数据稀疏性的方法大多依靠外部信息,然而在绝大部分场景下这些信息不可用.作者提出了不依赖外部数据的目的地预测方法.预先将区域划分网格,从历史数据中学习网格到网格的转移概率,然后预测轨迹目的地在哪个网格内.文献[156]研究了人类活动的规律及位置预测问题,文中指出目前有研究工作表明,人类活动具有一定的规律性,且会被其他人的活动影响(一致性),但目前没有方法将活动的规律性和一致性统一建模.作者针对上述问题,提出了一种对移动规律性和一致性统一建模的方法,并在若干个城市范围内的居民移动数据集上验证了方法的有效性.文献[157]介绍了 ECML/PKDD 竞赛中出租车目的地预测问题的解决方案,作者首先将目的地位置进行聚类,得到热度较高的目的地,作为预测输出结果的候选集.然后对轨迹的初始部分编码,并将编码后的数据输入神经网络,得到预测结果.作者对比了 CNN、RNN、bio-RNN 等方法,证明使用神经网络模型可以在该场景下准确地预测轨迹目的地.文献[82]提出一种基于语义轨迹的目的地预测方法,首先通过提取用户语义轨迹聚类得到用户的地点访问模式.当待预测用户到达时,查询历史数据中的相似用户,得到对下一个地点预测的结果.文献[158]利用 Flickr 中带有位置信息的标签,描述用户在城市范围内的偏好位置.作者提出了一种基于高斯核卷积的方法以衡量两个用户的 geotag 分布的相似性,并基于相似用户预测用户未来可能到达的位置.

5.2.2 行为预测类应用

此类应用基于时空数据及其语义信息建模用户的行为规律,从而预测用户未来行为.文献[159]利用用户群组检测的方法预测用户的行为.作者提出了一种基于用户时空数据和语义信息的群组检测方法 UCGT(user community geo-topic),用于仿真用户群组的生成过程,并将 UCGT 应用在用户签到预测和用户社交链接预测中,取得了比原方法更好的效果.文献[104]利用人在城市移动的数据和交通网络数据,预测单个用户在未来可能乘坐的交通工具或出行方式.

5.2.3 事件预测类应用

根据时空数据及其语义信息,结合已产生的事件,预测未来事件发生的时间、地点、规模一直以来都是研究的热点问题.文献[160]指出,时空事件的产生受许多因素的影响,包括经济、政治、文化等,传统的基于单一数据集的事件预测方法很难覆盖事件影响因素的所有方面,限制了模型的性能.而基于多源数据的事件预测方法存在数据缺失和数据稀疏的问题.作者提出了一种基于特征学习的方法,通过融合异构数据源在不同地理层面上

的数据,由低层次的特征得到高层次的特征.基于这些特征预测时空事件的产生时间、位置、内容等属性.文献[161]研究了社交媒体中时空事件预测问题,文中指出,现有研究工作大多针对时序事件作预测,而没有考虑事件的空间属性及时间与空间的隐含关系.作者考虑了上述因素,提出一种对事件的结构内容属性和时空突发属性联合建模的方式,为时空事件的生成过程建立图模型.实验结果表明,基于上述模型的方法在事件预测任务上的表现比已有模型要好.文献[162]研究了时空事件预测的问题,指出由于事件内容特征的动态变化(事件关键词变化)及地理信息异构(空间关联、样本不平衡、位置流行度不同)等原因,目前时空事件预测存在困难,作者提出了一个基于多任务学习的框架,将从领域知识库中抽取的静态特征和从查询变化中抽取的动态特征,集成到一个多任务特征学习框架中,并采取多种可选的策略平衡上述两种特征的同质性和多样性,用于预测时空事件.

6 总结与展望

时空数据语义理解是智能手持设备大范围普及、位置服务技术飞速发展背景下催生的一个新兴研究领域.与传统时空数据挖掘任务不同,时空数据语义理解不仅仅关注时空数据中存在的规律和模式,而且致力于从语义层面来理解规律、模式背后隐含的地理位置特点、用户行为偏好、事件属性及影响等.虽然有许多研究人员在该领域做了大量的工作并取得了突出的成绩,但是该领域在数据质量、算法模型、计算模式等方面仍面临许多问题和挑战,尤其是以下的研究方向值得研究人员深入探究.

(1) 时空数据的数据稀疏问题.数据质量上,由于受数据采集设备功耗限制和用户主动发布数据的随机性,数据稀疏是时空数据中普遍存在的问题.数据的稀疏性导致了时空信息流中大量的信息空洞,进而造成信息缺失时段内用户属性的不确定,这种不确定会造成理解用户行为语义时可用样本的不连贯,从而影响语义理解的准确性,这个问题成为时空数据语义理解的一大挑战.现有的解决方法是通过引入外部数据源进行缺失信息的推断和补全.但利用外部数据还需解决数据源选取、用户或位置匹配、数据源噪声过滤等一系列问题.因此,如何解决时空数据稀疏性仍是一个重要的研究方向.

(2) 时空数据的噪声处理问题.数据质量上,时空数据大部分通过传感器收集,传感器采集和传输过程中的噪声造成了时空数据存在信号噪声.此外,由于用户输入错误或输入不全而产生的无意义的数值造成了时空数据中存在语义噪声.这些噪声会造成时空属性的错误描述,极大地影响了时空语义理解.现有的处理手段通常利用信号处理方式来解决噪声问题,如卡尔曼滤波、粒子滤波等.但在数据稀疏的情况下,现有的处理方法效果不理想.因此,如何处理时空数据中的噪声需要进一步研究.

(3) 时空数据的多源多模态数据融合问题.算法模型上,时空数据的数据源非常丰富,由于单一数据源数据稀疏或含有噪声,在理解时空数据的语义信息时,常常需要融合多源多模态数据.多源数据相互印证、相互补充,可以更好地推断语义.现有研究工作大量处理的是基于空间数据与自然语言或图片的融合问题,如 geo-Twitter 和 Flickr.对于语音数据或视频数据则缺乏有效的融合手段.最近兴起的表达学习有完成这一任务的潜力,但这方面的研究工作刚刚起步,如何针对不同数据源、数据模态设计有效的表达学习方法仍未得到解决,还需深入研究.

(4) 大规模流式时空数据实时处理问题.计算模式上,时空数据语义理解中的某些任务(事件发现、异常检测)对结果实效性要求很高,需要处理流式时空数据,得到实时的语义理解的结果.但是随着实时数据量的增大,时空数据语义理解的实时性保证成为一个非常困难的问题.现有的实时分析算法大多基于规则过滤或实时聚类,分别存在分析能力弱和计算复杂度高的问题,无法满足大规模流式时空数据实时理解的需求.因此,如何在有限的计算资源约束下完成大规模流式时空数据中语义信息的实时理解仍然需要算法和理念上的重大突破.

(5) 时空数据交互式语义理解问题.计算模式上,由于从时空数据中抽取语义信息需要大范围地尝试模型和调整参数,在许多实际应用中,事先确定待使用模型的参数是一件非常困难的事情.因此,需要与用户进行交互,反馈语义理解效果.由于时空数据在时间、空间语义上存在上下文约束,现有的在线学习方法不适用于处理时空数据.针对时空数据的特点,在线学习模型参数还未有可行的解决方案.因此,如何展现时空数据及其语义信息以及如何将交互结果反馈到模型进而优化模型参数,仍是时空数据语义理解中存在的一大问题.

本文首次从地理位置的语义理解、用户行为的语义理解、热点事件的语义理解这 3 个方面归纳总结了近些年来关于位置语义注释、位置隐含语义、用户行为规律、用户个性化语义、热点事件检测等方面的研究成果.最后总结了时空数据语义理解在时空推荐和时空预测中的应用.希望本文所做的工作可以为致力于时空数据语义理解与挖掘的相关研究人员提供参考.

References:

[1] Eldawy A, Mokbel MF. The era of big spatial data. In: Proc. of the 32nd IEEE Int'l Conf. on Data Engineering (ICDE). IEEE, 2016. 1424–1427.

[2] https://baike.baidu.com/item/%E8%AF%AD%E4%B9%89/9716033?fr=aladdin

[3] Liu DY, Chen HL, Qi H, Yang B. Advances in spatiotemporal data mining. Journal of Computer Research and Development, 2013,50(2):225–239 (in Chinese with English abstract).

[4] Ji GL, Zhao B. A survey of spatiotemporal data mining for big data. Journal of Nanjing Normal University (Natural Science Edition), 2014,37(1):1–7 (in Chinese with English abstract).

[5] Zheng Y. Introduction to urban computing. Geomatics and Information Science of Wuhan University, 2015,40(1):1–13 (in Chinese with English abstract).

[6] Zheng Y. Trajectory data mining: An overview. ACM Trans. on Intelligent Systems and Technology (TIST), 2015,6(3):29.

[7] Gao Q, Zhang FL, Wang RJ, Zhou F. Trajectory big data: A review of key technologies in data processing. Ruan Jian Xue Bao/ Journal of Software, 2017,28(4):959–992 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5143.htm [doi: 10.13328/j.cnki.jos.005143]

[8] Feng Z, Zhu Y. A survey on trajectory data mining: Techniques and applications. IEEE Access, 2016,4:2056–2067.

[9] Sabarish BA, Karthi R, Gireeshkumar T. A survey of location prediction using trajectory mining. In: Artificial Intelligence and Evolutionary Algorithms in Engineering Systems. Springer India, 2015. 119–127.

[10] Besse P, Guillouet B, Loubes JM, *et al*. Review and perspective for distance based trajectory clustering. arXiv Preprint arXiv:1508.04904, 2015.

[11] Castro PS, Zhang D, Chen C, *et al*. From taxi GPS traces to social and community dynamics: A survey. ACM Computing Surveys (CSUR), 2013,46(2):17.

[12] Parent C, Spaccapietra S, Renso C, *et al*. Semantic trajectories modeling and analysis. ACM Computing Surveys (CSUR), 2013,45(4):42.

[13] Zheng Y, Xie X, Ma WY. GeoLife: A collaborative social networking service among user, location and trajectory. IEEE Data Engineering Bulletin, 2010,33(2):32–39.

[14] Yuan J, Zheng Y, Zhang C, *et al*. T-Drive: Driving directions based on taxi trajectories. In: Proc. of the 18th SIGSPATIAL Int'l Conf. on Advances in Geographic Information Systems. ACM, 2010. 99–108.

[15] Kwak H, Lee C, Park H, *et al*. What is Twitter, a social network or a news media. In: Proc. of the 19th Int'l Conf. on World Wide Web. ACM, 2010. 591–600.

[16] http://www.mobile.yahoo.com/flickr

[17] https://nationalzoo.si.edu/scbi/migratorybirds/research/data/

[18] https://catalog.data.gov/dataset/zebra-and-quagga-mussel-distribution-in-north-america-direct-download

[19] http://www.marinetraffic.com/

[20] https://catalog.data.gov/dataset/joint-typhoon-warning-center-storm-wallets

[21] http://datacenter.mep.gov.cn/

[22] Tsoukatos II, Gunopulos D. Efficient mining of spatiotemporal patterns. In: Proc. of the Int'l Symp. on Spatial and Temporal Databases. Berlin, Heidelberg: Springer-Verlag, 2001. 425–442.

[23] Tong YX, Yuan Y, Cheng YR, Chen L, Wang GR. A survey of spatiotemporal crowdsourced data management techniques. Ruan Jian Xue Bao/Journal of Software, 2017,28(1):35–58 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5140.htm [doi: 10.13328/j.cnki.jos.005140]

[24] Zhao Y, Han Q. Spatial crowdsourcing: Current state and future directions. IEEE Communications Magazine, 2016,54(7):102–107.

[25] Haklay M, Weber P. Openstreetmap: User-Generated street maps. IEEE Pervasive Computing, 2008,7(4):12–18.

[26] https://www.mapbox.com/osm-data-report/

[27] Ding Y, Zheng J, Tan H, *et al*. Inferring road type in crowdsourced map services. In: Proc. of the Int'l Conf. on Database Systems for Advanced Applications. Springer Int'l Publishing, 2014. 392–406.

[28] Hu H, Zheng Y, Bao Z, *et al*. Crowdsourced poi labelling: Location-Aware result inference and task assignment. In: Proc. of the ICDE. 2016.

[29] Schnitzler F, Artikis A, Weidlich M, *et al*. Heterogeneous stream processing and crowdsourcing for traffic monitoring: Highlights. In: Proc. of the Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg: Springer-Verlag, 2014. 520–523.

[30] Artikis A, Weidlich M, Schnitzler F, *et al*. Heterogeneous stream processing and crowdsourcing for urban traffic management. In: Proc. of the EDBT. 2014. 712–723.

[31] Hu H, Li G, Bao Z, *et al*. Crowdsourcing-Based real-time urban traffic speed estimation: From trends to speeds. In: Proc. of the 32nd IEEE Int'l Conf. on Data Engineering (ICDE). IEEE, 2016. 883–894.

[32] Yan Z, Chakraborty D, Parent C, *et al*. SeMiTri: A framework for semantic annotation of heterogeneous trajectories. In: Proc. of the 14th Int'l Conf on Extending Database Technology. ACM, 2011. 259–270.

[33] Yan Z, Chakraborty D, Parent C, *et al*. Semantic trajectories: Mobility data computation and annotation. ACM Trans. on Intelligent Systems and Technology (TIST), 2013,4(3):49.

[34] Spaccapietra S, Parent C, Damiani ML, *et al*. A conceptual view on trajectories. Data & Knowledge Engineering, 2008,65(1): 126–146.

[35] Alvares LO, Bogorny V, Kuijpers B, *et al*. A model for enriching trajectories with semantic geographical information. In: Proc. of the 15th Annual ACM Int'l Symp. on Advances in Geographic Information Systems. ACM, 2007. 22.

[36] Guc B, May M, Saygin Y, *et al*. Semantic annotation of gps trajectories. In: Proc. of the 11th AGILE Int'l Conf. on Geographic Information Science. 2008,38(6):1–9.

[37] Moreno B, Times VC, Renso C, *et al*. Looking inside the stops of trajectories of moving objects. In: Proc. of the Geoinfo. 2010. 9–20.

[38] Mousavi SM, Harwood A, Karunasekera S, *et al*. Geometry of interest (GOI): Spatio-Temporal destination extraction and partitioning in gps trajectory data. arXiv Preprint arXiv:1603.04110, 2016.

[39] Su H, Zheng K, Zeng K, *et al*. Making sense of trajectory data: A partition-and-summarization approach. In: Proc. of the 31st IEEE Int'l Conf. on Data Engineering. IEEE, 2015. 963–974.

[40] Lv M, Chen L, Xu Z, *et al*. The discovery of personally semantic places based on trajectory data mining. Neurocomputing, 2016, 173:1142–1153.

[41] Nabo RGB, Fileto R, Nanni M, *et al*. Annotating trajectories by fusing them with social media users posts. In: Proc. of the GeoInfo. 2014. 25–36.

[42] Zheng B, Yuan N J, Zheng K, *et al*. Approximate keyword search in semantic trajectory database. In: Proc. of the 31st IEEE Int'l Conf. on Data Engineering. IEEE, 2015. 975–986.

[43] Leme LAPP, Renso C, Nunes BP, *et al*. Searching for data sources for the semantic enrichment of trajectories. In: Proc. of the Int'l Conf. on Web Information Systems Engineering. Springer Int'l Publishing, 2016. 238–246.

[44] Kafsi M, Grossglauser M, Thiran P. Traveling salesman in reverse: Conditional Markov entropy for trajectory segmentation. In: Proc. of the 2015 IEEE Int'l Conf. on Data Mining (ICDM). IEEE, 2015. 201–210.

[45] Zhao K, Cong G, Sun A. Annotating points of interest with geo-tagged Tweets. In: Proc. of the 25th ACM Int'l on Conf. on Information and Knowledge Management. ACM, 2016. 417–426.

[46] Wu F, Li Z, Lee WC, *et al*. Semantic annotation of mobility data using social media. In: Proc. of the 24th Int'l Conf. on World Wide Web. ACM, 2015. 1253–1263.

[47] Wu F, Wang H, Li Z, *et al*. SemMobi: A semantic annotation system for mobility data. In: Proc. of the 24th Int'l Conf. on World Wide Web. ACM, 2015. 255–258.

[48] Hegde V, Parreira JX, Hauswirth M. Semantic tagging of places based on user interest profiles from online social networks. In: Proc. of the European Conf. on Information Retrieval. Berlin, Heidelberg: Springer-Verlag, 2013. 218–229.

[49] Lian D, Xie X. Learning location naming from user check-in histories. In: Proc. of the 19th ACM SIGSPATIAL Int'l Conf on Advances in Geographic Information Systems. ACM, 2011. 112–121.

[50] Ye M, Shou D, Lee WC, *et al*. On the semantic annotation of places in location-based social networks. In: Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2011. 520–528.

[51] Sarda S, Eickhoff C, Hofmann T. Semantic place descriptors for classification and map discovery. arXiv Preprint arXiv:1601.05952, 2016.

[52] Bhattacharya T, Kulik L, Bailey J. Automatically recognizing places of interest from unreliable GPS data using spatio-temporal density estimation and line intersections. Pervasive and Mobile Computing, 2015,19:86–107.

[53] Vu DD, Shin WY. Low-Cmplexity detection of POI boundaries using geo-tagged tweets: A geographic proximity based approach. In: Proc.of the 8th ACM SIGSPATIAL Int'l Workshop on Location-Based Social Networks. ACM, 2015.

[54] Yuan J, Zheng Y, Xie X. Discovering regions of different functions in a city using human mobility and POIs. In: Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2012. 186–194.

[55] Noulas A, Scellato S, Mascolo C, *et al*. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. The Social Mobile Web, 2011,11:2.

[56] Lichman M, Smyth P. Modeling human location data with mixtures of kernel densities. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2014. 35–44.

[57] Lloyd A, Cheshire J. Deriving retail centre locations and catchments from geo-tagged Twitter data. Computers, Environment and Urban Systems, 2017,61:108–118.

[58] Zhang C, Zhang K, Yuan Q, *et al*. Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In: Proc. of the 26th Int'l Conf. on World Wide Web. Int'l World Wide Web Conferences Steering Committee, 2017. 361–370.

[59] Imamichi T, Osogami T, Raymond R. Truncating shortest path search for efficient map-matching. In: Proc. of the 25th Int'l Joint Conf. on Artificial Intelligence (IJCAI 2016). 2016. 589–595

[60] Li Y, Su H, Demiryurek U, *et al*. PerNav: A route summarization framework for personalized navigation. In: Proc. of the 2016 ACM SIGMOD Int'l Conf. on Management of Data. 2016. 2125–2128.

[61] Dai J, Yang B, Guo C, *et al*. Path cost distribution estimation using trajectory data. Proc. of the VLDB Endowment, 2016,10(3): 85–96.

[62] Hung CC, Peng WC, Lee WC. Clustering and aggregating clues of trajectories for mining trajectory patterns and routes. The VLDB Journal, 2015,24(2):169–192.

[63] Evans MR, Oliver D, Shekhar S, *et al*. Summarizing trajectories into *k*-primary corridors: A summary of results. In: Proc. of the 20th Int'l Conf. on Advances in Geographic Information Systems. ACM, 2012. 454–457.

[64] Blei DM, Lafferty JD. Dynamic topic models. In: Proc. of the 23rd Int'l Conf. on Machine Learning. ACM, 2006. 113–120.

[65] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003,3:993–1022.

[66] Wang X, McCallum A. Topics over time: A non-Markov continuous-time model of topical trends. In: Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2006. 424–433.

[67] Wang C, Wang J, Xie X, *et al*. Mining geographic knowledge using location aware topic model. In: Proc. of the 4th ACM Workshop on Geographical Information Retrieval. ACM, 2007. 65–70.

[68] Yin Z, Cao L, Han J, *et al*. Geographical topic discovery and comparison. In: Proc. of the 20th Int'l Conf. on World Wide Web. ACM, 2011. 247–256.

[69] Hong L, Ahmed A, Gurumurthy S, *et al*. Discovering geographical topics in the twitter stream. In: Proc. of the 21st Int'l Conf. on World Wide Web. ACM, 2012. 769–778.

[70] Eisenstein J, O'Connor B, Smith NA, *et al*. A latent variable model for geographic lexical variation. In: Proc. of the 2010 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010. 1277–1287.

[71] Hu B, Ester M. Spatial topic modeling in online social media for location recommendation. In: Proc. of the 7th ACM Conf. on Recommender Systems. ACM, 2013. 25–32.

[72] Ahmed A, Hong L, Smola AJ. Hierarchical geographical modeling of user locations from social media posts. In: Proc. of the 22nd Int'l Conf. on World Wide Web. ACM, 2013. 25–36.

[73] Liu Y, Ester M, Hu B, *et al*. Spatio-Temporal topic models for check-in data. In: Proc. of the 2015 IEEE Int'l Conf. on Data Mining (ICDM). IEEE, 2015. 889–894.

[74] Mei Q, Liu C, Su H, *et al*. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In: Proc. of the 15th Int'l Conf. on World Wide Web. ACM, 2006. 533–542.

[75] Yuan Q, Cong G, Ma Z, *et al*. Who, where, when and what: Discover spatio-temporal topics for twitter users. In: Proc. of the 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2013. 605–613.

[76] Hu B, Jamali M, Ester M. Spatio-Temporal topic modeling in mobile social media for location recommendation. In: Proc. of the 13th IEEE Int'l Conf. on Data Mining. IEEE, 2013. 1073–1078.

[77] Agrawal R, Srikant R. Mining sequential patterns. In: Proc. of the 11th Int'l Conf. on Data Engineering. IEEE, 1995. 3–14.

[78] Tsoukatos II, Gunopulos D. Efficient mining of spatiotemporal patterns. In: Proc. of the Int'l Symp. on Spatial and Temporal Databases. Berlin, Heidelberg: Springer-Verlag, 2001. 425–442.

[79] Wang J, Hsu W, Lee ML. Flowminer: Finding flow patterns in spatio-temporal databases. In: Proc. of the 16th IEEE Int'l Conf. on Tools with Artificial Intelligence. IEEE, 2004. 14–21.

[80] Cao H, Mamoulis N, Cheung DW. Mining frequent spatio-temporal sequential patterns. In: Proc. of the 5th IEEE Int'l Conf. on Data Mining (ICDM 2005). IEEE, 2005. 8.

[81] Alvares LO, Bogorny V, Kuijpers B, *et al*. Towards semantic trajectory knowledge discovery. Data Mining and Knowledge Discovery, 2007.

[82] Ying JJC, Lee WC, Weng TC, *et al*. Semantic trajectory mining for location prediction. In: Proc. of the 19th ACM SIGSPATIAL Int'l Conf. on Advances in Geographic Information Systems. ACM, 2011. 34–43.

[83] Chen CC, Kuo CH, Peng WC. Mining spatial-temporal semantic trajectory patterns from raw trajectories. In: Proc. of the 2015 IEEE Int'l Conf. on Data Mining Workshop (ICDMW). IEEE, 2015. 1019–1024.

[84] Zhang C, Han J, Shou L, *et al*. Splitter: Mining fine-grained sequential patterns in semantic trajectories. Proc. of the VLDB Endowment, 2014,7(9):769–780.

[85] Zhang C, Zheng Y, Ma X, *et al*. Assembler: Efficient discovery of spatial co-evolving patterns in massive geo-sensory data. In: Proc. of the 21st ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2015. 1415–1424.

[86] Li Z, Ding B, Han J, *et al*. Mining periodic behaviors for moving objects. In: Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2010. 1099–1108.

[87] Li Z, Wang J, Han J. ePeriodicity: Mining event periodicity from incomplete observations. IEEE Trans. on Knowledge and Data Engineering, 2015,27(5):1219–1232.

[88] Jindal T, Giridhar P, Tang LA, *et al*. Spatiotemporal periodical pattern mining in traffic data. In: Proc. of the 2nd ACM SIGKDD Int'l Workshop on Urban Computing. ACM, 2013. 11.

[89] Zarezade A, Jafarzadeh S, Rabiee HR. Spatio-Temporal modeling of check-ins in location-based social networks. arXiv Preprint arXiv:1611.07710, 2016.

[90] Chen L, Özsu MT, Oria V. Robust and fast similarity search for moving object trajectories. In: Proc. of the 2005 ACM SIGMOD Int'l Conf. on Management of Data. ACM, 2005. 491–502.

[91] Berndt DJ, Clifford J. Using dynamic time warping to find patterns in time series. In: Proc. of the KDD Workshop. 1994,10(16): 359–370.

[92] Vlachos M, Kollios G, Gunopulos D. Discovering similar multidimensional trajectories. In: Proc. of the 18th Int'l Conf. on Data Engineering. IEEE, 2002. 673–684.

[93] Chen J, Wang R, Liu L, *et al*. Clustering of trajectories based on Hausdorff distance. In: Proc. of the 2011 Int'l Conf. on Electronics, Communications and Control (ICECC). IEEE, 2011. 1940–1944.

[94] Yuan G, Sun P, Zhao J, *et al*. A review of moving object trajectory clustering algorithms. Artificial Intelligence Review, 2016, 1–22.

[95] da Silva TLC, Zeitouni K, de Macêdo JAF, *et al*. A framework for online mobility pattern discovery from trajectory data streams. In: Proc. of the 17th IEEE Int'l Conf. on Mobile Data Management (MDM). IEEE, 2016,1:365–368.

[96] Lee JG, Han J, Whang KY. Trajectory clustering: A partition-and-group framework. In: Proc. of the 2007 ACM SIGMOD Int'l Conf. on Management of Data. ACM, 2007. 593–604.

[97] Kim Y, Han J, Yuan C. TOPTRAC: Topical trajectory pattern mining. In: Proc. of the 21st ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2015. 587–596.

[98] Zhang C, Zhang K, Yuan Q, *et al*. GMove: Group-Level mobility modeling using geo-tagged social media. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2016.

[99] Yin H, Zhou X, Shao Y, *et al*. Joint modeling of user check-in behaviors for point-of-interest recommendation. In: Proc. of the 24th ACM Int'l Conf. on Information and Knowledge Management. ACM, 2015. 1631–1640.

[100] Cho E, Myers SA, Leskovec J. Friendship and mobility: User movement in location-based social networks. In: Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2011. 1082–1090.

[101] Zhang C, *et al*. Spatiotemporal activity modeling under data scarcity: A graph-regularized cross-modal embedding approach. In: Proc. of the AAAI. 2018.

[102] Zhang C, Zhang K, Yuan Q, *et al*. ReAct: Online multimodal embedding for recency-aware spatiotemporal activity modeling. In: Proc. of the 40th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2017. 245–254.

[103] Chen Q, Song X, Yamada H, *et al*. Learning deep representation from big and heterogeneous data for traffic accident inference. In: Proc. of the 30th AAAI Conf. on Artificial Intelligence. 2016.

[104] Song X, Kanasugi H, Shibasaki R. Deeptransport: Prediction and simulation of human mobility and transportation mode at a citywide level. In: Proc. of the IJCAI. 2016.

[105] Higgs B, Abbas M. Segmentation and clustering of car-following behavior: Recognition of driving patterns. IEEE Trans. on Intelligent Transportation Systems, 2015,16(1):81–90.

[106] Liu S, Ni LM, Krishnan R. Fraud detection from taxis' driving behaviors. IEEE Trans. on Vehicular Technology, 2014,63(1): 464–472.

[107] Yuan Q, Cong G, Zhao K, *et al*. Who, where, when, and what: A nonparametric bayesian approach to context-aware recommendation and search for twitter users. ACM Trans. on Information Systems (TOIS), 2015,33(1):2.

[108] Wang J, Li M, Han J, *et al*. Modeling check-in preferences with multidimensional knowledge: A minimax entropy approach. In: Proc. of the 9th ACM Int'l Conf. on Web Search and Data Mining. ACM, 2016. 297–306.

[109] Zhang D, Guo L, Nie L, Shao J, Wu S, Shen HT. Targeted advertising in public transportation systems with quantitative evaluation. ACM TOIS, 2017,35(3):20:1–20:29. [doi:10.1145/3003725]

[110] Guo L, Zhang D, Wu H, Cui B, Tan KL. From raw footprints to personal interests: Bridging the semantic gap via trip intention aggregation. In: Proc. of the 33rd IEEE Int'l Conf. on Data Engineering (ICDE). IEEE, 2007. 123–126. [doi:10.1109/ICDE.2017.55]

[111] Wu F, Li Z. Where did you go: Personalized annotation of mobility records. In: Proc. of the 25th ACM Int'l Conf. on Information and Knowledge Management. ACM, 2016. 589–598.

[112] Zhao K, Liu Y, Yuan Q, *et al*. Towards personalized maps: Mining user preferences from geo-textual data. Proc. of the VLDB Endowment, 2016,9(13):1545–1548.

[113] Wang W, Yin H, Sadiq S, *et al*. SPORE: A sequential personalized spatial item recommender system. In: Proc. of the 32nd IEEE Int'l Conf. on Data Engineering (ICDE). IEEE, 2016. 954–965.

[114] Zhong Y, Yuan N J, Zhong W, *et al*. You are where you go: Inferring demographic attributes from location check-ins. In: Proc. of the 8th ACM Int'l Conf. on Web Search and Data Mining. ACM, 2015. 295–304.

[115] Cao W, Wu Z, Wang D, *et al*. Automatic user identification method across heterogeneous mobility data sources. In: Proc. of the 32nd IEEE Int'l Conf. on Data Engineering (ICDE). IEEE, 2016. 978–989.

[116] Rossi L, Walker J, Musolesi M. Spatio-Temporal techniques for user identification by means of GPS mobility data. EPJ Data Science, 2015,4(1):1.

[117] Rossi L, Williams M J, Stich C, *et al*. Privacy and the city: User identification and location semantics in location-based social networks. arXiv Preprint arXiv:1503.06499, 2015.

[118] Du N, Dai H, Trivedi R, *et al*. Recurrent marked temporal point processes: Embedding event history to vector. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2016. 1555–1564.

[119] Van Canneyt S, Schockaert S, Dhoedt B. Categorizing events using spatio-temporal and user features from Flickr. Information Sciences, 2016,328:76–96.

[120] Krumm J, Horvitz E. Eyewitness: Identifying local events via space-time signals in twitter feeds. In: Proc. of the 23rd SIGSPATIAL Int'l Conf. on Advances in Geographic Information Systems. ACM, 2015. 20.

[121] Foley J, Bendersky M, Josifovski V. Learning to extract local events from the Web. In: Proc. of the 38th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2015. 423–432.

[122] Liu Y, Zhou B, Chen F, *et al*. Graph topic scan statistic for spatial event detection. In: Proc. of the 25th ACM Int'l on Conf. on Information and Knowledge Management. ACM, 2016. 489–498.

[123] Liang Y, Caverlee J, Cao C. A noise-filtering approach for spatio-temporal event detection in social media. In: Proc. of the European Conf. on Information Retrieval. Springer Int'l Publishing, 2015. 233–244.

[124] Hristova D, Liben-Nowell D, Noulas A, *et al*. If you've got the money, I've got the time: Spatio-Temporal footprints of spending at sports events on foursquare. In: Proc. of the 10th Int'l AAAI Conf. on Web and Social Media. 2016.

[125] Quezada M, Peña-Araya V, Poblete B. Location-Aware model for news events in social media. In: Proc. of the 38th Intl ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2015. 935–938.

[126] Zhang W, Qi G, Pan G, *et al*. City-Scale social event detection and evaluation with taxi traces. ACM Trans. on Intelligent Systems and Technology (TIST), 2015,6(3):40.

[127] Ouyang RW, Srivastava A, Prabahar P, *et al*. If you see something, swipe towards it: Crowdsourced event localization using smartphones. In: Proc. of the 2013 ACM Int'l Joint Conf. on Pervasive and Ubiquitous Computing. ACM, 2013. 23–32.

[128] Ouyang RW, Srivastava M, Toniolo A, *et al*. Truth discovery in crowdsourced detection of spatial events. IEEE Trans. on Knowledge and Data Engineering, 2016,28(4):1047–1060.

[129] Patroumpas K, Artikis A, Katzouris N, *et al*. Event recognition for maritime surveillance. In: Proc. of the EDBT. 2015. 629–640.

[130] Zhang C, Zhou G, Yuan Q, *et al*. GeoBurst: Real-Time local event detection in geo-tagged tweet streams. In: Proc. of the 39th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2016. 513–522.

[131] Zhang C, Lei D, Yuan Q, *et al*. GeoBurst+: Effective and real-time local event detection in geo-tagged tweet streams. ACM Trans. on Intelligent Systems and Technology (TIST), 2018,9(3):34.

[132] Abdelhaq H, Sengstock C, Gertz M. Eventweet: Online localized event detection from twitter. Proc. of the VLDB Endowment, 2013,6(12):1326–1329.

[133] Walther M, Kaisser M. Geo-Spatial event detection in the twitter stream. In: Proc. of the European Conf. on Information Retrieval. Berlin, Heidelberg: Springer-Verlag, 2013. 356–367.

[134] Maurya A, Murray K, Liu Y, *et al*. Semantic scan: Detecting subtle, spatially localized events in text streams. arXiv Preprint arXiv:1602.04393, 2016.

[135] Unankard S, Li X, Sharaf MA. Emerging event detection in social networks with location sensitivity. World Wide Web, 2015,18(5): 1393–1417.

[136] Watanabe K, Ochi M, Okabe M, *et al*. Jasmine: A real-time local-event detection system based on geolocation information propagated to microblogs. In: Proc. of the 20th ACM Int'l Conf. on Information and Knowledge Management. ACM, 2011. 2541–2544.

[137] Zhang C, Liu L, Lei D, *et al*. Triovecevent: Embedding-Based online local event detection in geo-tagged tweet streams. In: Proc. of the 23rd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2017. 595–604.

[138] Feng W, Zhang C, Zhang W, *et al*. STREAMCUBE: Hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream. In: Proc. of the 31st IEEE Int'l Conf. on Data Engineering. IEEE, 2015. 1561–1572.

[139] Yu Y, Chen X. A survey of point-of-interest recommendation in location-based social networks. In: Proc. of the Workshops at the 29th AAAI Conf. on Artificial Intelligence. 2015.

[140] Liu X, Liu Y, Li X. Exploring the context of locations for personalized location recommendations. In: Proc. of the 25th Int'l Joint Conf. on Artificial Intelligence (IJCAI 2016). 2016. 1188–1194.

[141] Yin H, Zhou X, Cui B, *et al*. Adapting to user interest drift for poi recommendation. IEEE Trans. on Knowledge and Data Engineering, 2016,28(10):2566–2581.

[142] Li H, Ge Y, Zhu H. Point-of-Interest recommendations: Learning potential check-ins from friends. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2016. 975–984.

[143] Liu Y, Liu C, Liu B, *et al*. Unified point-of-interest recommendation with temporal interval assessment. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2016. 1015–1024.

[144] Cheng C, Yang H, King I, *et al*. A unified point-of-interest recommendation framework in location-based social networks. ACM Trans. on Intelligent Systems and Technology (TIST), 2016,8(1):10.

[145] Yin H, Zhou X, Shao Y, *et al*. Joint modeling of user check-in behaviors for point-of-interest recommendation. In: Proc. of the 24th ACM Int'l Conf. on Information and Knowledge Management. ACM, 2015. 1631–1640.

[146] Chen X, Zeng Y, Cong G, *et al*. On information coverage for location category based point-of-interest recommendation. In: Proc. of the AAAI. 2015. 37–43.

[147] Bagci H, Karagoz P. Context-Aware friend recommendation for location based social networks using random walk. In: Proc. of the 25th Int'l Conf. Companion on World Wide Web. Int'l World Wide Web Conferences Steering Committee, 2016. 531–536.

[148] Wang Z, Liao J, Cao Q, *et al*. Friendbook: A semantic-based friend recommendation system for social networks. IEEE Trans. on Mobile Computing, 2015,14(3):538–551.

[149] Quercia D, Lathia N, Calabrese F, *et al*. Recommending social events from mobile phone location data. In: Proc. of the 2010 IEEE Int'l Conf. on Data Mining. IEEE, 2010. 971–976.

[150] Zhang W, Wang J. A collective Bayesian Poisson factorization model for cold-start local event recommendation. In: Proc. of the 21st ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2015. 1455–1464.

[151] Ayhan S, Samet H. Aircraft trajectory prediction made easy with predictive analytics. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2016. 421–434.

[152] Hai NT, Nguyen HH, Thai-Nghe N. A mobility prediction model for location-based social networks. In: Proc. of the Asian Conf. on Intelligent Information and Database Systems. Berlin, Heidelberg: Springer-Verlag, 2016. 106–115.

[153] Besse PC, Guillouet B, Loubes JM, *et al*. Destination prediction by trajectory distribution based model. arXiv Preprint arXiv:1605.03027, 2016.

[154] Chen M, Liu Y, Yu X. Predicting next locations with object clustering and trajectory clustering. In: Proc. of the Pacific-Asia Conf. on Knowledge Discovery and Data Mining. Springer Int'l Publishing, 2015. 344–356.

[155] Xue AY, Qi J, Xie X, *et al*. Solving the data sparsity problem in destination prediction. The VLDB Journal, 2015,24(2):219–243.

[156] Wang Y, Yuan NJ, Lian D, *et al*. Regularity and conformity: Location prediction using heterogeneous mobility data. In: Proc. of the 21st ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2015. 1275–1284.

[157] de Brébisson A, Simon É, Auvolat A, *et al*. Artificial neural networks applied to taxi destination prediction. arXiv Preprint arXiv:1508.00021, 2015.

[158] Clements M, Serdyukov P, De Vries AP, *et al*. Using flickr geotags to predict user travel behavior. In: Proc. of the 33rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2010. 851–852.

[159] Yin H, Hu Z, Zhou X, *et al*. Discovering interpretable geo-social communities for user behavior prediction. In: Proc. of the 32nd IEEE Int'l Conf. on Data Engineering (ICDE). IEEE, 2016. 942–953.

[160] Zhao L, Ye J, Chen F, *et al*. Hierarchical incomplete multi-source feature learning for spatiotemporal event forecasting. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2016. 805–814.

[161] Zhao L, Chen F, Lu CT, *et al*. Spatiotemporal event forecasting in social media. In: Proc. of the SDM. 2015,15:963–971.

[162] Zhao L, Sun Q, Ye J, *et al*. Multi-Task learning for spatio-temporal event forecasting. In: Proc. of the 21st ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2015. 1503–1512.

附中文参考文献:

[3] 刘大有,陈慧灵,齐红,杨博.时空数据挖掘研究进展.计算机研究与发展,2013,50(2):225–239.

[4] 吉根林,赵斌.面向大数据的时空数据挖掘综述.南京师范大学学报:自然科学版,2014,37(1):1–7.

[5] 郑宇.城市计算概述.武汉大学学报(信息科学版),2015,40(1):1–13.

[7] 高强,张凤荔,王瑞锦,周帆.轨迹大数据:数据处理关键技术研究综述.软件学报,2017,28(4):959–992. http://www.jos.org.cn/1000-9825/5143.htm [doi: 10.13328/j.cnki.jos.005143]

[23] 童咏昕,袁野,成雨蓉,陈雷,王国仁.时空众包数据管理技术研究综述.软件学报,2017,28(1):35–58. http://www.jos.org.cn/1000-9825/5140.htm [doi: 10.13328/j.cnki.jos.005140]

姚迪(1990－),男,河南许昌人,博士,CCF学生会员,主要研究领域为数据挖掘,机器学习.

张超(1988－),男,博士,助理研究员,主要研究领域为 data mining,machine learning.

黄建辉(1977－),男,博士,高级工程师,主要研究领域为移动机会网络,大数据应用.

陈越新(1983－),男,博士,工程师,主要研究领域为计算机网络,大数据应用技术.

毕经平(1974－),女,博士,研究员,博士生导师,CCF 高级会员,主要研究领域为计算机网络,数据处理.