计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202220417
61(2): 385−412, 2024

在线社交网络中用户地理位置预测综述

刘乐源[1] 代雨柔[1] 曹亚男[2] 周 帆[1]
[1](电子科技大学信息与软件工程学院 成都 610054)
[2](中国科学院信息工程研究所 北京 100093)
(leyuanliu@uestc.edu.cn)

Survey of User Geographic Location Prediction Based on Online Social Network

Liu Leyuan[1], Dai Yurou[1], Cao Yanan[2], and Zhou Fan[1]
[1](*School of Information and Software Engineering*, *University of Electronic Science and Technology of China* , *Chengdu* 610054)
[2](*Institute of Information Engineering*, *Chinese Academy of Sciences*, *Beijing* 100093)

Abstract With the increasing popularity of intelligent mobile terminals, people increasingly use social network platforms (such as Twitter, Sina Weibo, etc.) for information acquisition, comments, and exchanges. Although GPS devices can accurately obtain location information, many users do not directly share their location information for privacy and security considerations. Therefore, obtaining the geographic location of online users has become an important topic in both academia and industry and is the foundation of many downstream applications, such as location-based targeted advertising, event/location recommendations, early warning of natural disasters or diseases, and criminal tracking, etc. We survey in detail the methods, data types, evaluation metrics, and fundamental algorithms for predicting the geographic location of social network users. First, we discriminate different online user geolocation tasks and corresponding evaluation protocols. Subsequently, we assess the data structures and fusion methods used for individual geolocation tasks. Besides, we analyze the existing information extraction and feature selection approaches, as well as their advantages and disadvantages. Moreover, we provide a taxonomy to categorize existing user geolocation models and algorithms, followed by a thorough analysis of different methods from three aspects: geographic dictionary, traditional machine learning, deep learning and graph neural networks. Finally, we summarize the difficulties and challenges in user location prediction while outlining the possible research trend and opportunities to shed light on future work in this field.

Key words social network; geographic location prediction; data analysis and fusion; information extraction; temporal and spatial characteristics

摘 要 随着智能移动终端的日益普及,人们越来越多地利用社交网络平台(如Twitter、新浪微博等)获取信息、评论和交流.虽然全球卫星定位系统(GPS)设备能够精确获取位置信息,但是大量用户出于隐私和安全的考虑不会直接共享自己的位置信息.因此,如何获取在线用户的地理位置成为了一个前沿的研究领域以及学术界和工业界共同关注的重要课题,并且成为众多下游应用的基础,例如基于位置的定向广告投放、事件/地点的推荐、自然灾害或疾病预警和网络犯罪的追踪等.详细总结了预测社交网络用户

收稿日期: 2022-05-21; **修回日期**: 2023-04-03
基金项目: 国家自然科学基金项目(62176043, 62072077, 62272089); 四川省自然科学基金项目(2022NSFSC0505); 四川省科技计划项目(2022YFSY0006)
This work was supported by the National Natural Science Foundation of China (62176043, 62072077, 62272089), the Natural Science Foundation of Sichuan province (2022NSFSC0505), and the Sichuan Science and Technology Program (2022YFSY0006).
通信作者: 周帆(fan.zhou@uestc.edu.cn)

地理位置的方法、数据、评价体系和基础算法. 首先, 归纳了不同的定位任务以及相应的评价指标; 其次, 针对不同的任务梳理所用的数据类型和数据融合方式, 并且, 详尽分析了已有的信息抽取和特征选择方式及其优缺点; 再次, 对现有定位模型和算法进行分类, 从地名词典、传统机器学习和深度学习 3 个方面对用户定位方法进行阐述和分析; 最后, 总结了社交网络用户地理位置预测的难点和面临的挑战, 并展望该领域的发展趋势和未来研究所需要关注的方向.

关键词 社交网络; 地理位置预测; 数据分析与融合; 信息抽取; 时空特征

中图法分类号 TP391

过去的十多年, 在线社交网络(online social network, OSN)平台取得了空前的发展, 每天都吸引着成千上万的用户加入. 据统计, Twitter 用户人数已超 5 亿, 每月有 330 亿活跃用户, 每天产生超过 500 亿的对话; 新浪微博在 2020 年第 2 季度的微博日活跃用户人数也增长至 2.29 亿, 每天发布微博的数量超过 2 500 万条, 每秒可以生成 785 条微博. 大量用户生成了大量的个人数据, 如心情的分享、对某件事情的评论以及在某些景点的签到记录. 这些用户数据语义丰富, 获取简单、灵活、低成本, 数据的准确性和粒度有保证.

这些 OSN 平台可以分为 3 类: 1)通用平台, 如 Twitter、新浪微博、Instagram 和 Facebook 等. 在这些平台上, 用户可以通过文字、照片、视频以及打卡等方式记录自己的生活日常、参与话题讨论和关注热点事件, 也可以与他人建立社交关系并分享个人喜好, 如分享个人经验心得、关注新产品的发布、讨论社会热点事件、了解最新的交通拥堵情况等. 2)基于位置服务的平台, 如 Foursquare、大众点评、Yelp 和 Gowalla 等. 这些平台需要用户主动签到(check-in)以记录自己所在的位置, 且可以绑定用户的其他社交软件; 如 Twitter 和 Facebook 等以同步分享用户的地理位置信息. 此外, 用户可以对地点, 如电影院、餐厅等进行评价并展示给其他用户. 3)图片社交服务平台, 如 Flickr, Pinterest, Snapchat 等图片社交服务网站. 此类平台为用户提供图片的分享、管理等功能, 同时可以添加联系人, 使用户能够结交更多的朋友, 增强用户之间的交流. 用户可以在这 3 类平台中自由发布各种信息, 其中的文本消息文本统称为帖子.

在针对这些 OSN 平台的科学研究中, 准确地预测用户地理位置具有重要的意义: 首先, 预测用户地理位置是很多下游应用开发的基础, 下游应用基于此可对社交媒体上的动态事件进行更全面的时空分析[1], 提升管理部门的社会治理能力, 如新冠病毒传播预测[2-3]、政治话题检测[4]、紧急位置识别[5]、可疑人物监测[6]、灾难预警[7-9] 等. 其次, 准确的位置预测, 可以极大地提升运营商针对用户的个性化服务能力, 提升用户体验, 为企业带来巨大的经济效益, 如产品营销[1, 10]、精准广告投放、信息推荐、打车调度优化等.

已有关于 OSN 的用户位置预测研究综述, 如文献 [11−12] 发表时间较为久远, 最近几年 OSN 地理位置预测技术已经有了很大发展. 例如, 随着图神经网络的提出, 图神经网络的各种变形及其与其他网络模型组合得到的方法被应用于社交网络用户位置预测. 此外这些综述仅仅针对 Twitter 数据总结了位置预测方法, 且没有详细深入地讨论数据信息的特征以及基础模型的划分. 文献 [13] 是一篇关于用户地理位置预测的综述, 但该文献只总结了基于位置的 OSN 的工作, 并没有对其他非基于位置的 OSN 做出总结. 所以, 现在迫切需要一个新的 OSN 中用户地理位置预测研究综述. 与之前的关于 Twitter 位置预测和基于位置的社交网络(geolocation-based social network, GSN)的研究综述不同, 本文总结了几乎所有的 OSN 数据的研究情况, 同时加入了多源数据融合的分析、特征的选择以及采用时间顺序对该领域的基础算法模型进行了总结.

本文回顾了目前 OSN 数据中用户位置预测的最新文献, 并总结了该任务面临的 4 个现实挑战:

1) 居住在不同城市的用户可能会用方言或特定用语来讨论当地热点事件和地点等, 这时社交平台内容的自由性和随意性给地理位置预测带来了一些困难.

2) 用户经常以非常随意的方式编写社交动态内容, 如缩略词、汉语拼音缩写、拼写错误、特殊标记等, 这让文本内容变得嘈杂, 使得很多为正式的规范化文档内容所开发的文本处理技术在社交内容的应用上更容易出错.

3) 用户的注意力和关注点变化很快, 很多历史社交动态中存在大量与地理位置无关的信息, 这使

得对该任务而言，很多文本内容变得无用.

4）由于 OSN 用户是随意地手动或通过 GPS 附上自己的位置，导致 OSN 上的位置信息往往不完整和不准确. 以 Twitter 为例，文献 [14] 发现在 Twitter 平台的 U.S.数据集[15] 上，只有 21% 的用户在个人档案信息中提供了自己所居住的城市，5% 的用户给出了家庭住址的详细经纬度信息.

本文以 OSN 数据为重点，立足于 OSN 中地理位置预测任务时关注的 3 个问题：结果如何评价（标准）、数据有何特征（数据）及采用何种方法（模型），并回顾了用户位置预测问题的研究进展，特别是近 4 年的进展. 本文基于前述 3 个问题对 OSN 中的地理位置预测研究进行了综述. 此外，本文还分析了当前研究中存在的主要问题，并展望了未来的研究方向.

1 任务分类与评价指标

为了更清晰地理解 OSN 中地理位置预测的研究内容，本文首先介绍了位置预测的问题定义、分类以及评价指标.

1.1 任务描述

用户位置预测问题的目的是通过分析用户在 OSN 中生成的各种地理标签，即用户附加到各种媒体类型中的地理信息，挖掘其背后隐藏的各种地理信息、用户移动模式和个人偏好，从而确定用户的位置.

1.2 任务分类

本文着眼于 4 种类型的 OSN 相关位置的预测，即居住（home）位置、文本（text）位置、提及（mentioned）位置和未来（future）位置.

1.2.1 居住位置

居住位置即 OSN 用户的长期居住地址. 根据应用程序的具体要求，居住位置可以用不同的粒度表示. 通常，居住位置的粒度有 3 种：

1）行政区域，即用户所在的国家或城市.

2）地理网格，即地球被划分成面积大小相等的单元格，住址位置由用户所在的单元格表示.

3）地理坐标，即房屋的坐标，用它们的纬度和经度来表示. 坐标可以通过其行政区域或者单元转换获得.

了解 OSN 用户的居住位置可以实现许多基于此的服务，如本地内容推荐、基于位置的广告、公共健康监测和民意调查等. 然而，在几乎所有的 OSN 中，用户个人资料都是选填内容，其居住位置大多数情况下都是缺失或包含噪声的，即不准确或虚假的，因此很多研究者致力于预测用户的居住位置，且预测的粒度是城市级别.

1.2.2 文本位置

文本位置，即发送帖子的位置. 有了文本位置，即可得到更完整的用户移动模式，也能更好地估计事件的发生位置. 不同于从用户资料和地理标签中收集的住址位置，文本位置通常是基于文本内容的地理标签. 由于文本位置具有独特的视图，研究者们广泛采用兴趣点（points of interest, POI）或坐标作为文本位置的表示，而不是行政区域或网格.

根据文献 [16] 的分析，用户发送帖子的主要目的是为了分享或寻找信息. 例如一个人可能会在推文或微博上说自己正在餐厅用餐. 如果餐厅的名字作为标签与 OSN 平台有明确的关联，这些信息将有助于推广这家餐厅；用户可能会在 Instagram 或新浪微博上发帖说自己迷路了，在这种情况下，他/她的朋友可以在 OSN 平台的评论区或通过私信给以准确的指示. 不幸的是，用户很少在文本中明确提及了自己的地理位置，例如，只有不到 1% 的 Twitter 帖子带有明确的地理标签[17].

1.2.3 提及位置

用户在发帖时，可能会在内容中标记或提到一些地点的名字. 例如用户会发帖子评论餐厅、购物中心或电影院，把 OSN 平台当作一个生活记录平台等. 当游行或灾难发生时，用户可能会发出大量包含地理标签的帖子来通知其他人. 用户还可以通过在帖子中提及的地点名称来透露相关地理信息. 对位置名的预处理是为用户和事件[18] 积累信息和执行后续分析的关键步骤[19]. 在目前的研究中，位置预处理包括 2 个步骤：1）提及位置识别，对潜在的位置信息块进行标注提取；2）歧义消除[20]，将已识别的地理位置映射到位置数据库中的正确条目，从而确定它们所指的位置. 这 2 个问题的难点在于地点实体提及的可变性和模糊性是实体识别和关联. 可变性是指一个实体可能以不同的表面形式被提及，模糊性是指一次提及可能涉及多个实体. 不幸的是，这 2 个难题因为帖子简短和噪声变得更具挑战性. 由于 OSN 平台语言的固有噪声和歧义的特点，提及地点的实况数据在很大程度上依赖于人的注释. 对于提及位置的 2 个步骤，位置的粒度既涉及行政区域，又涉及 POI.

1.2.4 未来位置

未来位置即用户将来可能会去的位置. 未来位置的预测即利用用户过去的轨迹来预测用户将来可能会去的 1 个或者多个地点，其本质是深入挖掘数

据背后隐藏的时空特征、序列属性和用户个人偏好，然后利用基于机器学习的技术或大数据分析来整合多维信息，从而推断用户未来的访问地点.

在对未来位置进行预测时，由于轨迹是在一定的时间段内生成，用户可能只进行了少量的签到，存在数据稀疏问题，增加了预测难度；另外，用户的历史移动模式会影响后续足迹，且个人签到是复杂和个性化的行为，这导致数据有很强的复杂性. 因此，在进行此类预测时，数据存在的这些特点是学者们首要面临的挑战，例如复杂的序列过渡规律、高度的顺序性、人类移动的多层次周期性以及收集的轨迹数据的异质性和稀疏性等.

1.3 评价指标

本节回顾了文献中用于评价最终预测结果的常用指标. 根据预测结果的表示方式，常用指标分为基于距离的度量、基于标记的度量和通用度量.

1.3.1 基于距离的度量

家庭位置预测或文本位置预测的目标是对每个用户或文本的位置进行预测. 令h表示用户或帖子，H为预测结果的集合，$l(h)$为h的真实位置，系统可以对每个h预测其位置$l'(h)$. 无论采用何种粒度，所有的真实位置$l(h)$和预测位置$l'(h)$都可以转换为坐标形式. 直观地，可以使用真实位置与预测位置之间的欧氏距离来度量预测误差距离(error distance, ED)，记为$ED(h)$. 定义如式(1)所示:

$$ED(h) = distance(l(h), l'(h)). \tag{1}$$

由于评估是在一组用户或帖子上进行的，所以可以取所有误差距离的平均值或中值，以获得语料库级别的误差度量. 这就产生了平均误差距离(mean error distance, *MeanED*)和中值误差距离(median error distance, *MedianED*). 其定义分别如式(2)和式(3)所示:

$$MeanED = \frac{1}{|H|}\sum_{h\in H} distance(l(h), l'(h)), \tag{2}$$

$$MedianED = \underset{h\in H}{median}\{distance(l(h), l'(h))\}, \tag{3}$$

当预测结果不准确时，*MedianED* 通常不如 *MeanED* 敏感. 因此，大量研究常用 *MeanED* 作为评估度量. 另外，一些研究[21]使用均方误差(mean squared error, *MSE*)代替 *MeanED*. 定义如式(4)所示：

$$MSE = \frac{1}{|H|}\sum_{h\in H} distance^2(l(h), l'(h)), \tag{4}$$

MSE 和 *MeanED* 的唯一区别是前者取误差距离的平方.

除 *MeanED* 和 *MedianED* 外，还有另一种被广泛采用的语料库级度量，称为基于距离的准确度，记为 *Acc@d*. 与精确匹配不同，*Acc@d* 认为在一定距离误差阈值 d 之内的预测结果都是“正确值”. 语料库上的 *Acc@d* 度量被定义为可容忍正确预测的比例，如式(5)所示：

$$Acc@d = \frac{|\{h\in H : ED(h) \leqslant d\}|}{|H|}, \tag{5}$$

通常采用的距离阈值 d 是 161 km (100 英里)[21].

1.3.2 基于标记的度量

与有精确地理距离度量的方法相比，基于标记的度量有更广阔的应用场景. 对于本文 1.2 节中所述的 4 个地理位置预测问题，最简单的基于标记的度量是准确率. 假设$l(h)$和$l'(h)$分别为真实位置和预测位置. 在这种情况下，当预测与真实情况一致时，预测才被认为是正确的. 因此将准确率定义为 H 中正确预测的比例，记作 *Acc*，如式(6)所示：

$$Acc = \frac{|\{h\in H : l'(h) = l(h)\}|}{|H|}. \tag{6}$$

在某些情况下，预测结果并非是单个位置，可能是基于某种标准的排名列表$L'(h)$. 按照准确率度量的标准，直观地，位于列表首位的位置即为预测结果(准确率最高). 但是这种方法忽略了列表中的其他位置预测，而实际应用中，其他预测对于下游应用程序可能颇有价值. 所以，有研究者设计了基于排名的准确率度量，称之为 *Top@k*. 该方法认为：如果真实位置位于前 k 个结果$L_k'(h)$内，则认为预测结果是正确的. 认定一个排序列表为正确列表的计算方法如式(7)所示：

$$Top@k = \frac{|\{h\in H : l(h)\in L_k'(h)\}|}{|H|}. \tag{7}$$

1.3.3 通用度量

在一些情况下，如提及位置消歧，系统可能无法为给定的位置找到合适的条目. 在这种情况下，采用精确率(*Precision*)、召回率(*Recall*)和 *F*1 作为指标.h表示给定用户、tweet 或可识别的提及位置，如果系统无法做出任何预测，则$l'(h)$=null.

评估语料库H上的精确率定义为所有预测的位置中正确预测所占的比例，如式(8)所示：

$$Precision = \frac{|\{h\in H : l(h)\in L_k'(h)\}|}{|\{h\in H : l'(h)\neq \text{null}\}|}. \tag{8}$$

召回率定义为正确预测位置占所有真实位置的比例，如式(9)所示：

$$Recall = \frac{|\{h\in H : l'(h) = l(h)\}|}{|H|}. \tag{9}$$

*F*1 定义为精确率和召回率的调和均值，如式(10)所示：

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall}. \quad (10)$$

最后，需要注意的是精确率、召回率和$F1$在位置预测领域都是适用的，并且被广泛采用. 另外，在某些信息提供不足的情况下，预测系统可能无法做出预测[22-24].

2 数据类型

OSN不断地以极高的速度积累着大量的异构数据，这些数据也将作为后续位置预测的输入. 按照OSN上数据的类型，本文将其分为4类：1)由用户发布的简短而嘈杂的消息文本，例如推文、微博和点评等；2)图像数据；3)用户在基于地理位置的社交网络(location based social network, LBSN)上的签到数据；4)评分数据.

2.1 文本数据

文本信息是OSN中最为普遍的一类信息类型. 随着社交平台的日益流行，OSN中发布的帖子数量庞大，OSN平台发布的内容可以描述用户想传递的任何信息：发布原创的帖子、转发他人发布的帖子，同时用户的帖子也会被推送给关注他的用户. 在撰写消息文本内容时，可能存在使用特殊字符的现象，即用“#”开头的单词或不间隔的短语给自己发布的内容添加标签；用户还可能在消息文本中使用“@”来提到另一个用户的名字，被提及的用户将会获得通知，用户之间能够以这种方式开始对话.

除了形式多样以外，OSN平台上发布的内容往往有许多非标准缩写、排印错误、使用的表情符号和热门话题被称为标签. 这种非常规的、非结构化的文本被称为噪音，由于标准的自然语言处理(natural language processing, NLP)工具[25]不能很好地处理这些问题，给社交网络中的内容分析带来了不小的挑战.

2.2 图像数据

目前，越来越多OSN平台的用户将地点与图像关联起来. 图像能包含更多的视觉信息，可以更好地反映用户对位置的偏好从而预测事件发生的地点. 现有位置与图像相联合进行位置预测的研究大多是将相关区域划分为网格，并预测图像所在的精确网格[26-27]，或者用一个多边形区域，例如用圆形或三角形区域来表示用户发帖数较多的地点[28]，将图像和地点相关联. 例如，文献[27]利用视觉内容将Flickr图片映射到地球上的网格中；文献[29]利用Instagram上基于地理位置的图片，对病患是否遵循了COVID-19防疫政策进行跟踪. 这些将图像与位置相关联的研究表明，图像可以用来描述位置. 但是到目前为止，基于图像本身视觉特征的用户位置预测工作还较少.

2.3 签到数据

在LBS中，用户到达每个场所可以进行签到，表明到达过这个地方. 通常，一个签到记录可以用一个三元组(u,p,t)定义，表示用户u在时间t时的签到地点为p. 在LBS中的地点被定义为唯一标识的特定地点，如电影院、酒吧或咖啡店. 将用户的这些签到点按时间顺序连接起来，就能得到该用户的一条专属轨迹，现在很多研究，如文献[30−31]，就是利用用户的历史轨迹来预测其将来要去的场所.

2.4 评分数据

一些OSN，如Yelp、大众点评等，允许用户在他们访问的地点添加评分，一般从1~5分不等. 数字评分包含丰富的信息，可以在位置预测任务中提高预测的准确度. 从情感分析的角度来看，评分越高，表示积极情绪越高，说明用户再次访问该地方的可能性越大；相反，较低的评分代表消极情绪，这意味着用户未来访问这里的可能性较小. 在文献[32]中，讨论了情感信息对捕获用户签到行为的潜在影响. 文献[33]基于Yelp数据集衍生的用户POI评级分矩阵，使用语义标签对每个用户的评级偏好进行建模，并将该因素纳入矩阵分解框架用于下一次访问地点预测. 文献[34]通过计算余弦相似度来估计给定用户在候选地点的签到概率. 通常情况下，这种用户场所偏好仅使用用户配置文件中的积极评分(评分大于3.0)进行建模. 实验结果表明，使用高评分的方法比所有只访问过地点的方法能更好地提高用户位置预测的准确性. 到目前为止，很少有研究考虑用户位置预测的数字评分. 如何有效地整合这类数据来提高预测性能仍有待探索.

3 信息来源和特征选择

3.1 信息来源

用户在OSN中发布各种公开的信息，除了帖子本身的内容类型不同以外，还有可能在帖子上附加各种信息或是在用户的个人介绍中提及有价值的信息等. 如何使用这些不同来源的信息进行位置预测或辅助进行地理位置预测也是越来越多研究者关注的问题.

本文从地理标签、社交网络结构、用户配置文

件和语义信息这 4 个方面, 对数据信息特征的类型进行划分. 在这 4 个特征类型中, 地理标签基于直接地理坐标; 语义信息和用户配置文件需要字符串处理来提取有用的数据信息. 这些特征都能影响位置预测的精度, 因此需要不同的数据分析技术, 并且每种技术都面临不同的挑战.

3.1.1 地理标签

现在大多数智能手机都配备了 GPS 功能, 能够准确地获知用户的经纬度坐标信息, 部分用户也习惯于在其帖子中直接附加地理位置信息. 突发事件的目击者, 往往会在其 OSN 上第一时间发布相关的内容, 包括地理信息. 这些信息在时间上要早于官方媒体, 且有助于估计突发事件的精确位置. 例如用户在新浪微博或微信朋友圈发布交通事故或火灾的信息并带有地理标签. 因此, 在 OSN 一条帖子中最直接和精确的位置获取方式, 即是从它的 GPS 地理标签和位置附件中获取.

然而实际的情况却并非如此简单, 帖子位置信息面临着数据稀疏性问题. 文献 [35−36] 报道地理标记过的帖子只占 OSN 平台中所有帖子的一小部分, 其中收集帖子的方式也会对稀疏性有影响. 为了减少稀疏性带来的影响, 文献 [37] 按照时间线将每个用户发布的帖子划分为多个集群, 并将每个集群分类为一个城市级的位置类来解决此问题, 但此类方法显然无法满足细粒度预测位置的需求. 另外, 也有学者通过预测地理位置属性来解决稀疏性, 如文献 [38] 描述和分析了 Facebook 的大规模公共页面的位置属性之后, 提出了一个基于广度优先搜索(breadth-first search, BFS)的框架, 利用页面图的连通性来预测缺失的地理位置信息.

除了稀疏性影响, 另一个问题是带有地理标签的帖子的可用性也会受到区域内用户分布差异的阻碍[39-40]. 由于城市地区的人口密度远高于农村地区, 如果不进行相应的处理, 使用地理标签帖子的方法往往倾向于将人口密集的地区作为某些事情发生的地点. 为了最小化这种偏差, 文献 [40] 根据随机选择的 OSN 用户分布重新计算了算法中的权重来反映感兴趣区域的人口分布. 文献 [41] 将用户可能的位置定义为聚类坐标区域, 从而处理人口分布中异质性的影响. 聚类的结果显示, 人口密集的地区被小的细颗粒区域所代表, 而人口稀少的地区被更大的区域定义所覆盖.

3.1.2 社交关系网络

除了个体发布文本内容这种行为以外, 用户也会在 OSN 平台上与他人建立多种联系, 例如关注、转发帖子和回复社交动态等, 这些社交关系构成了用户的社交关系网络.

用户社交关系网络是 OSN 的重要属性之一, 研究表明有社交关系的朋友比陌生人更有可能在同一区域[42], 在研究用户个体发布的信息的同时, 发现用户之间的社交网络关系能提供额外的信息, 此种关系亦被广泛用于用户位置预测[43]. 例如, 文献 [44] 研究了 Facebook 平台上用户在线朋友关系与用户地理位置的关系, 结果发现, 任何一对用户存在线下朋友关系的可能性会随着地理距离的增加而单调下降. 文献 [45−46] 也表明, 如果 2 个用户居住在同一个城市, 那么他们很可能经常通信, 反之亦然.

除了直接的好友关系, 社交平台上用户之间的间接好友关系在用户住址预测中也很有价值. 文献 [47] 发现, 如果用户 A 和用户 B 经常与多个相同的第三方用户联系, 那么用户 A 和用户 B 极有可能也存在好友关系. 更具体来说, 文献 [48] 证实, 如果 2 位用户间的共同好友都超过了他们各自全部好友数目的一半, 那么这 2 位用户在现实生活中地理位置距离在 10 km 以内的可能性为 83%, 但当共同好友占比小于 10% 时, 其概率就下降到 2.4%.

社交网络上的朋友关系无论是单向关注还是双向关注, 与现实生活中是否具有朋友关系没有必然联系. 例如 OSN 上的明星并不关注他/她们的绝大多数粉丝, 即粉丝单向关注明星. 但是, 研究者也发现生活中的朋友会经常在网上提到彼此[23,45-46,48], 并且 2 个相距遥远的陌生人也可能成为朋友. 在探究用户网络上的关联关系对位置预测问题的影响时, 本文对社交网络用户之间的关注和提及行为不进行区分, 并将该网络称为社交关系网络. 一个典型的 OSN 如图 1 所示, 网络中没有边的用户被称为孤立节点, 其中没有定位标识的用户表明不知道其地理信息. 图 2 中只有单向箭头的表示关注了箭头指向的用户, 例如用户 10 关注了用户 1, 双向箭头表示这些用户之间是相互关注的关系, 例如用户 8 和用户 9.

3.1.3 个人资料

在注册 OSN 平台账户时, 用户个人资料属性中一般会有其所在位置的选填项, 文献 [49] 的调查称 66% 的个人资料包含有效的地理信息. 这些个人资料中的地理信息能够帮助广告商准确定位其目标用户, 还可以帮助应急救援人员寻找被救助者, 或帮助追踪网络犯罪嫌疑人. 用户个人资料中的位置属性允许用户在自由文本字段中指定他们的位置. 与消

息文本不同的是，一些 OSN 属性有字符数的限制，如 Twitter 最多包含 30 个字符，大众点评最多包含 50 个字符. 由于用户被要求在这个字段中输入一个位置名称，且只提供了有限的空间，所以大多数情况下都不会使用复杂的短语.

OSN 平台个人资料中的地理信息属性的一个特征是地理位置字段并未对用户输入提出约束，可以是自由文本. 部分用户在这个文本里输入非地理位置信息，增加了获取其位置信息的难度. 图 2 以 Twitter 为例展示了用户在 OSN 平台的个人简介以及资料编辑时可以选择性添加地理标签的情形. 另一个特征是地理位置大多为粗粒度位置[35,50-51]，如省级或者市 1 级. 这样会导致个人资料中的地理标签缺乏细粒度信息，如街道名称，在某些事件，如建筑火灾或交通事故中需要对用户进行定位时，可能无法得到满意的预测结果.

除了直接的地理信息外，用户的个人主页或相关网页也可以列在个人资料中，这些数据可以为地理位置预测提供辅助支撑. 例如用户列出的网站由居住在其本国的提供商托管，则可能提供国家级别的信息. 但是实际上，由于信息技术全球化带来的影响，用户也有可能在一个地点托管他们的网站，而住在另一地点，这就导致此类信息在地理位置预测中的辅助效果不佳.

此外，如将 OSN 应用程序接口（application programming interface, API），捕获信息中的时间戳、时区和个人资料等一起作为数据源进行位置预测，可以使预测至少达到国家级别的粒度[52]，从而缩小预测范围，为下一步更精确地进行预测提供支持. 文献 [53] 在进行国家层次的预测后，能进一步用国家层次的预测指导城市级的预测.

随着多功能社交网站的普及，用户跨越多个社交网络拥有账号的现象越来越普遍[54]，例如大众点评允许使用新浪微博、QQ 或微信等进行连接登录，Foursquare 允许用户将他们的 Twitter, Instagram 等账户与 Foursquare 贴子连接，这些贴子通常带有地理标签，因此即使用户没有在 Twitter[55] 上透露自己的位置信息，也可以从 Foursquare 的贴子中推断出用户位置. 使用这种方式获得的位置信息可以被认为是个人资料中的一种.

3.1.4 语义信息

OSN 平台中发布的帖子的语义信息被定义为能够涉及地理位置的信息和反映用户情绪的信息. 由于帖子相比个人资料会包含更多的信息，因此帖子文本的语义信息构成了位置预测领域中大多数研究的主要对象. 使用自然语言处理的方法理解语义，提取其中的用户地理信息，就可以作为地理信息预测的数据来源. 但是实际情况却并非如此简单. 由于用户习惯和 OSN 平台的限制，给语义识别带来了很多挑战.

首先，OSN 平台上没有明确的写作规则，因此帖子中会包含很多的非标准文本，并且大多数的帖子含有错别字、长度过短等现象，这些都会干扰对用户位置的预测. 另外，即使在帖子文本中确定了位置，也不能直接推断其为用户住所位置或是发帖的位置[56]. 如一条帖子包含了城市的名字“上海”，但这并不一定意味着用户来自上海，或者这条推帖子就是在上海发出的.

另外，另一种与文本中地名相关的干扰是地理歧义，这是由若干具有相同名称的不同地点造成的，如“巴黎”（Paris）有 140 种不同的可能性[57]. 为了确定用户的帖子中所指的位置，文献 [58] 使用了发帖距离、个人资料中的地理距离和文本中提到的位置之间的层次包容. 文献 [57] 利用某个主题，聚类所有的推文，根据该聚类中地名之间的可能解释的关系（如地理包容、发帖距离和地理距离）来解决歧义. 在其他研究中，文献 [35] 将用户个人资料中的粗粒度位置信息与帖子中的细粒度位置名称结合起来，以解决细粒度定义中的歧义.

除了对语义中的地理信息进行挖掘以外，对 OSN 平台上的帖子语义进行情绪分析，也可以对位置预测产生积极的帮助. 情绪信息通常嵌入在用户所发布的帖子中. 如果用户对某地使用积极的语言表达，

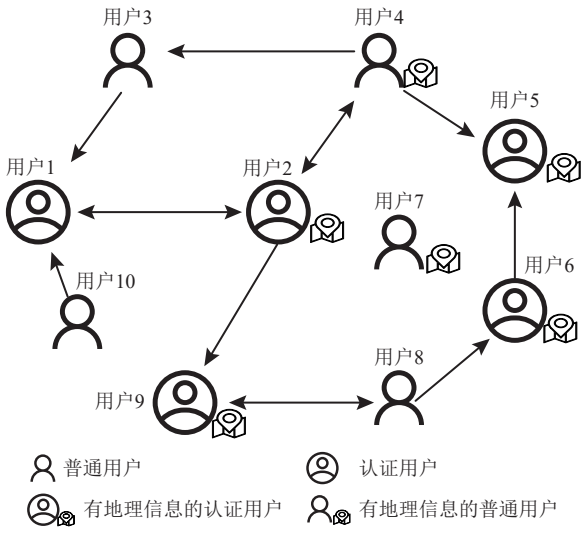


Fig. 1 Illustration of OSN

图 1 OSN 示意图

相应地，该地理位置重要程度就越高；否则，重要性则越低[32]. 例如在文献 [59] 中假设用户的帖子也可以作为预测其访问地点的依据，通过分析情绪的变化趋势，能够捕捉到用户的意图和心情，并进一步确定用户下一次最有可能访问的地点类型. 文献 [60] 只使用用户个人资料中的积极评价来衡量用户的地点偏好，因为积极评级比中性或负面评级更具有预测能力. 由于当前的研究集中在从帖子中的向量化词汇和句法特征中获取情绪特征，而没有进一步的上下文，因此文献 [61] 设计了矢量化位置信息与词嵌入相结合以产生混合表示，从而改进了情感分类任务，即使情感分析能够对地理位置预测起到积极的作用.

总体而言，集成语义信息可以有效地缓解数据稀疏性问题并了解用户偏好，但是无法单独通过情感分析去预测用户的地理位置，大多数情况下还是作为其他方法的辅助手段存在，以便提高预测效果. 现有的研究大多只考虑文本中涉及的地理位置，情感信息的影响仍有待探讨.

3.2 特征选择

对在线社交平台数据的信息特征分析显示，并不总是能从数据中获取高质量和可靠的信息. 这意味着，应该将信息缺失和不同信息源之间的差异性考虑在内. 数据的信息特征可以根据需求和期望以不同的方式被选择和使用.

根据位置预测算法中对特征的使用情况，本文将 OSN 数据的信息特征分为“主要(primary)”和“次要(secondary)”2 种类型. 主要信息特征是首先在帖子中被检查，并在位置预测算法中使用的特征. 实际应用中，应用程序可以选择使用单个主要特征或多个主要特征的组合. 次要特征可以对缺乏预期效果的主要属性的帖子起到补充作用. 例如一个应用程序可以主要使用帖子的 GPS 地理标签，对于没有地理标签的帖子，可以利用用户配置文件作为次要功能来推断用户的位置[40]. 本文对现有文献按照使用单一主要特征、次要特征来处理缺失的特征；使用多个特征的组合作为主要来源进行分类回顾.

1) 使用单一主要特征. 基于单一主要特征的标准确定帖子中的信息特征作为来源，并仅使用该特征进行位置估计. 例如文献 [5] 仅使用帖子中的位置名称来确定事件位置. 文献 [26] 仅利用 Twitter 帖子中的 GPS 坐标对其进行时空聚类. 类似地，文献 [62]

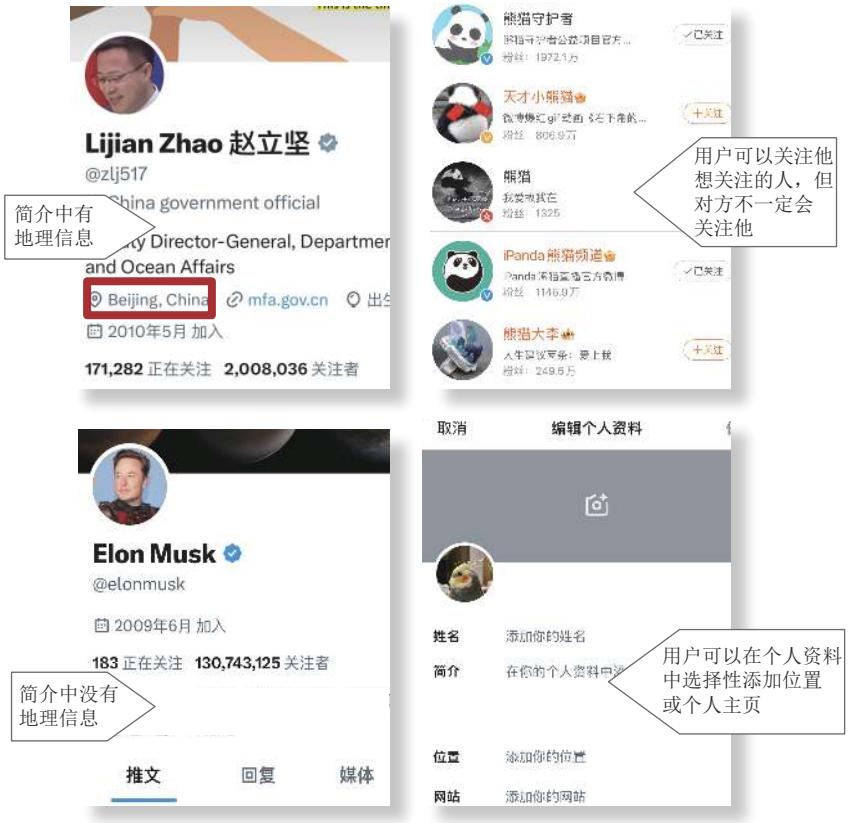


Fig. 2 Illustration of user profiles containing geo-information

图 2 用户资料中包含地理信息示意图

利用事件相关推文的GPS坐标来研究突发关键词的空间分布，以检测事件并预测其位置. 一般来说，选作主要特征的数据，都能较容易地提取出地理位置信息，其选用的提取或预测技术往往也较少，整个预测应用设计架构相对简单. 但采用单一的主要特征，对数据集要求较高，整个模型会受到社交网络数据特性的影响，例如数据稀疏、缩写简写等问题.

2）使用次要特征. 由于位置预测方法可以从多个特征中获益，如果选择的主要特征在一个帖子中不可用（缺少数据或者存在不可用数据），其他次要特征可以用来提取帖子的位置信息. 使用次要特征的一种常见做法是利用用户个人资料中的位置属性来代替无地理标签帖子的GPS数据[63]. 这意味着，无地理标签帖子的坐标可以从用户个人资料中特定的位置信息推断得出. 虽然次要特征可以在主要特征不可用时起到很好的补充作用，但作为次要特征的数据常常无法提供有效的信息，一定程度上降低了其可用性. 例如，用户在个人资料中常常不填写自己的准确地址.

3）使用多种特征. 第3种选择是利用多个特征作为主要的信息来源. 这种方法不考虑特征之间的相互替代，相反地将它们组合在一个模型中. 例如文献[64−65]基于GPS地理标记、消息文本和用户个人资料，首先为可能发生事件的位置分配基本概率，然后根据dempster-shafer理论中的组合规则将这些概率分配到单个解决方案中. 在另一项使用多个特征的研究中，文献[39]同时利用了微博的GPS地理标签、微博文本中的参考位置和用户资料，对新浪微博检测到的事件进行了本地化，从而提取用户坐标. 使用多种特征既能很好地缓解单一主特征面临的数据稀疏等问题，又能避免使用次要特征效率不高的缺陷，但这种情况下，往往需要更为复杂的模型或框架来实现预测功能，带来更大的计算压力.

4 基础算法和模型

用户位置预测领域已经有多种方法和技术被提出，近年来也持续不断地对已知方法改进和提出新的方法，例如，机器学习、统计学、概率学和深度学习等，取得了一定成果. 在自然语言处理领域，有一类重要的任务：命名实体识别（named entity recognition, NER）[66]，该任务用于从给定的文本中识别出实体的边界、类型（通常为人为预先确定）等内容，地理位置的识别也是其中一类重要的目标. 实际上很多研究者也将NER技术用于帖子中地理位置的提取. 与NER对其技术进行分类的标准类似，本文将地理位置预测的算法/模型分为3类：使用字典的方法、使用传统机器学习的方法和使用深度学习的方法. 本文在介绍当前对应的NER，尤其是其中与地理位置预测相关内容的基础上，还介绍了地理消歧等任务的研究现状. 本节所总结的算法类别如图3所示.

4.1 地名词典

地名词典是收录特定范围内地理实体的专有名词的词典. 一般提供自然、经济、人文和历史这些方面的信息，并对其进行说明. 地理数据库类似地名词典，也能很好地应用于这项研究，使用的一些数据库工具包括GeoNames, GeoNet, OpenStreetMap, WikiMapia等，这些工具允许以任何数据格式下载许多国家的空间定义[67]，如城市、城镇、街道、建筑、山脉、河流等. 基于地名词典的方法，其基本思路是提取帖子中的地点与地名词典中的地点进行匹配. 因此，此类方法的重点是地点（实体）提取.

许多研究者采用NER的方法来识别各个地点实体，该实体可以是基于段的表示，也可以是基于词的表示[55]. 前者在识别帖子中的地点方面表现出更高的效率，这种技术广泛使用的工具是Standford NER.但文献[68]发现，在OSN平台的帖子中使用Stanford NER并不能准确地检测出包括地名在内的实体，特别是实体被缩写时，会有更高的概率识别错误. 因此，文献[69]重新训练了Stanford NER, OpenNLP, Twitter-NLP等工具对2 878条与灾难相关的推文进行了10倍交叉验证，发现经过再次训练的Stanford NER的表现优于其他之前的方法. 文献[70]采用了一种混合方法，位置实体被提取并解析成地名词典，以准确地对推文中提到的地名进行地理编码.

基于字典的NER方法虽然在小数据集上可达到较高的准确率，但面对大量的数据集以及未知地点（不在字典内的地点），这种方式变得不再适用. 因此，文献[71−72]使用GeoNames，用地名词典来替代传统的基于字典的NER的方法，并认为地名词典更适合于有噪声的消息文本. 文献[55]也采用基于地名词典的方法，利用OSN帖子中的各种元数据信息来推断地理位置.

由于文本表达的信息的多样性，会产生“歧义”现象，即在帖子中发现的地理信息文本，通过联系该文本的上下文也可能表示其他非地理意义. 例如“England”既可以指代英国，也可以表示英国人[65]. 为了解决这种地理/非地理的歧义，文献[73]探索了在

有歧义的术语之前是否存在一个空间指标，如“在”“位于”等，使用 PipePOS 标记器和 USGS 位置数据库推断用户的当前位置，从而解析不明确的位置名称. 文献 [65] 开发了一种启发式方法，该方法在其他消息文本中寻找与本事件相关的位置名称，以消除这种歧义.

虽然地名词典很容易实现，且无需训练数据，但此类方法面临着诸多质疑：1)其处理速度慢[21]，无法简单通过缩小数据集的方法来提升其速度. 文献 [43] 发现，通过增加好友/粉丝关系的深度来改变 OSN 数据集的大小，并不会影响使用地名词典方法计算和检测用户位置所花费的时间.2)尽管基于地名词典的方法将语言依赖性最小化，但对语法的完全忽视可能会限制地名解析的性能. 例如使用诸如“A 在北京，B 在距 A 约 100 km 的城市”这样的短语，要得到准确可用的预测地点，需要分析完整的句子结构，不能通过简单地在数据库中查找地名来确定地点.3)由于这类句子在 OSN 平台上很少被观察到[74]，是否值得对这些短语进行深入分析也很有争议.

4.2 传统机器学习方法

随着机器学习研究的兴起，数据挖掘和传统机器学习方法同样被应用于检测 OSN 平台的用户位置预测中，并且已经被证明是一类很有效的用户位置预测方法[75]，如 *k*-邻近分类、模糊匹配[43]、朴素贝叶斯模型[76-77]、条件随机场(CRF)、矩阵分解以及马尔可夫模型等.

4.2.1 概率方法

概率方法的目的是根据观察，以概率来确定预测地点最可能的位置. 在本节中，将解释概率方法的 4 种主要实现：第 1 种方法采用最大似然估计(maximum likelihood estimation, MLE)；第 2 种方法采用期望最大化(expectation maximiation, EM)[78]；第 3 种方法基于贝叶斯滤波器给出 2 种实现，即卡尔曼滤波器(Kalman filters)和粒子滤波器(particle filters)；第 4 种是使用较为广泛的马尔可夫模型.

1) 最大似然估计

这种概率技术是基于学习帖子中的单词在一系列地点 X 上的概率分布. 例如一个国家的城市，使用与地点相关帖子中的单词计算某地点属于城市 $x \in X$ 的后验概率. 然后，选择后验概率最大的城市作为预测地点位置.

在文献 [78] 中，最大似然估计是作为地点定位的基准方法. 文献 [65] 中也提出了一种类似的方法作为监督的基准方法. 文献 [79] 提出了一个半监督框架，使用最大似然估计来估计用户的位置. 文献 [80] 用最大似然估计算法识别城市的本地居民. 极大似然估计要求定义一组表示可能的地点位置和这些位置上的注释已进行训练的选择. 因此，准确性可能会随训练数据的质量而改变. 虽然基于最大似然估计的方法具有简单易实现的特点，但其求解参数的方式决定了在面对参数较多的情况下难以求解.

2) EM 算法

EM 算法是一种迭代算法，用于含有隐变量的概率参数模型的最大似然估计或极大后验概率的参数估计. 该算法在面临参数多、一次性求解计算开销大的情况下具有优势. 但 EM 算法对需要初始化的参数初始值敏感，即参数的选择直接影响收敛效率以及是否能够的到全局最优解.

部分研究者基于热点话题和位置的不同组合作为潜在变量，提出了不同的概率模型，并使用基于 EM 等变分推理方法来预测位置. 文献 [81] 在整个美国的 Twitter 数据集上用用 MeanED 和 MedianED 报告了 494 km 的预测结果. 文献 [82] 通过考虑用户的话题倾向性，对文献 [81] 的方法进行了扩展，并通过假设每个用户有多个位置开发了一个概率模型. 文献 [43] 使用了额外的地理编码过程，对检测到的命名实体提取其位置引用. 该方法首先对事件在开始处近似局部化；其次采用基于 EM 的预测技术，利用模型中实体的位置关系密切度来更新和增强这些初始预测；最后，对位置层次结构中的每个层次，如国家、州、城市和街道级别分别执行预测. 这些解决方案在其预测模型中将位置视为一个连续变量. 因此，当涉及到更大的地理区域，如国家或大洲时，它们就会表现出扩展性问题.

3) 滤波器

除了上述方法外，滤波器也常用于位置预测任务中. 其中贝叶斯滤波虽然可以有效滤除噪声，得到比较精准的状态估计，但是在计算期望的时候，需要做无穷积分，大多数情况下没有解析解(精确解). 使用较多的方法是卡尔曼滤波器和粒子滤波器.

①卡尔曼滤波器. 卡尔曼滤波器[83-84]是以最小均方误差为估计的最佳准则来寻求一套递推估计的模型，其基本思想是：采用信号与噪声的状态空间模型，基于先前的估计和最近观察的位置更新来预测对象的位置，它表示坐标分布的均值和协方差值的不确定性. 卡尔曼滤波器包含 2 种更新规则：用于预测的时间更新和用于校正的测量更新. 卡尔曼滤波器适合于实时处理和计算机运算. 文献 [85] 使用卡尔曼

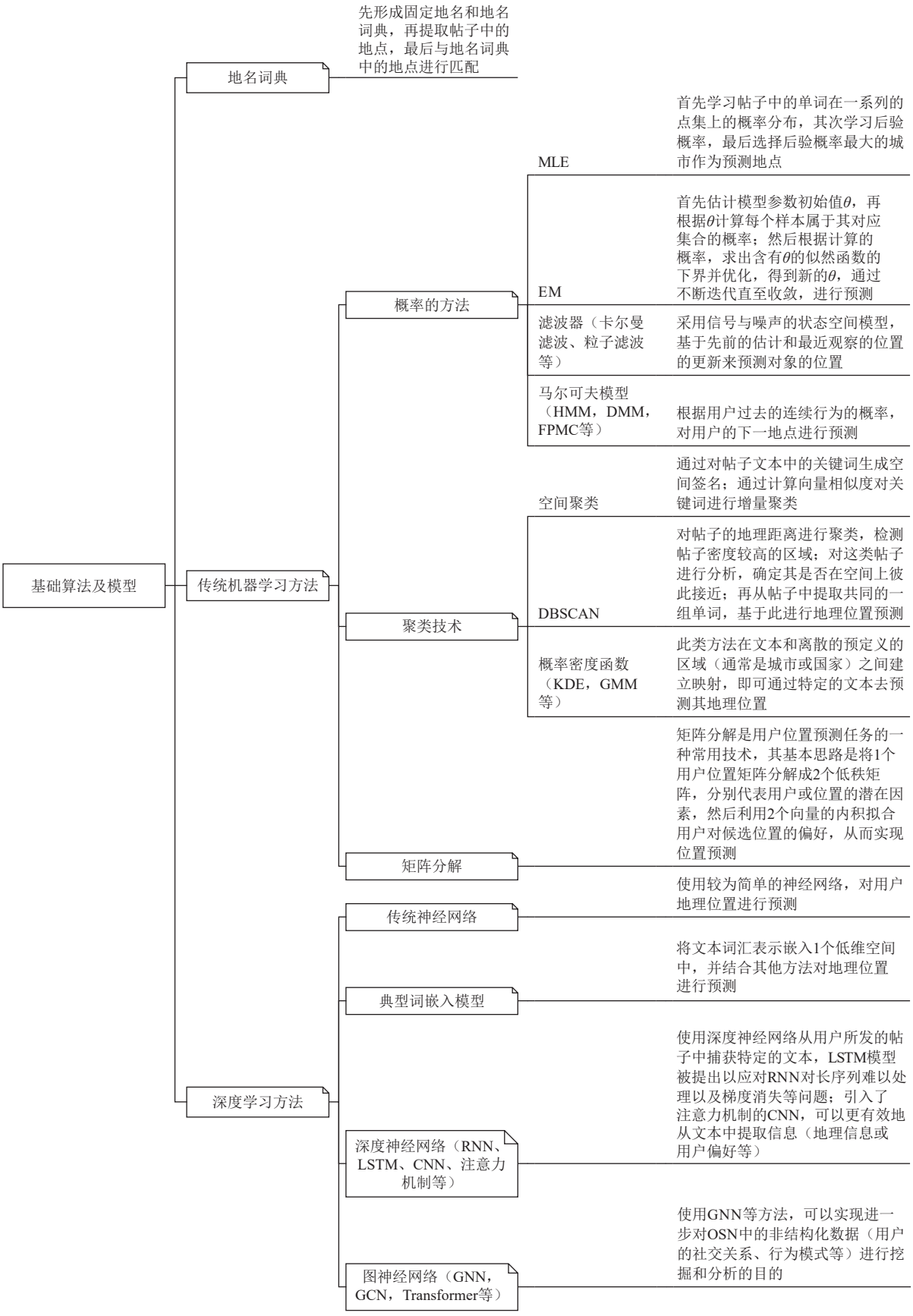


Fig. 3 Illustration of basic algorithms and models

图 3 基础算法及模型示意图

滤波器在 Twitter 中检测地震震中. 在估计检测值时, 假设没有出现时间转换的误差, 因此只使用了测量更新规则. 文献 [85] 指出: 卡尔曼滤波器在线性高斯环境中工作得更好, 由于 Twitter 帖子或微博动态不一定是线性的, 并且从遥远的位置发布的消息文本可能会对方差产生负面影响, 因此在预测地点时使用卡尔曼滤波器的可能性能不佳. 在这种情况下, 粒子滤波器被认为比卡尔曼滤波器更合适.

②粒子滤波器. 通过寻找一组在状态空间中传播的随机样本来近似表示概率密度函数, 用样本均值代替积分运算, 进而获得系统状态的最小方差估计的过程. 这些样本被形象地称为“粒子”, 故而叫粒子滤波. 为了预测 OSN 中的位置, 文献 [86] 描述了一种使用顺序重要性采样方法的粒子滤波器实现. 在实现中, 随机分布在感兴趣区域的粒子是在初始生成阶段产生的. 提供与预测地点相关的消息文本中的经纬度值, 算法迭代地重新采样粒子, 通过更新它们的位置、速度和加速度来预测粒子的下一个状态, 并根据它们与最近观察的距离为粒子分配权重. 最终确定事件的估计位置为粒子坐标的平均值. 文献 [61] 认为, 为了提高位置估计方法的准确率, 发布事件相关消息文本的用户应该是独立同分布的, 即关于事件的信息不应该在 Twitter 中传播太多. 该文献解决了 OSN 用户地理分布差异带来的偏差问题, 并提出了一种粒子滤波器的改进方法. 该方法基于随机抽样的消息文本集, 考虑传感器分布, 为粒子分配权重. 文献 [65] 中还采用了卡尔曼滤波和粒子滤波的实现作为基准的位置预测方法.

虽然实验证明了粒子滤波器在事件地点预测方面的估计精度优于卡尔曼滤波器, 由于使用的粒子数量具有不确定性, 所以粒子滤波器的执行时间并不一定适合对实时性有要求的任务.

4) 马尔可夫模型

在 OSN 最初的位置预测中, 许多研究者选择研究基于马尔可夫模型的各种地点预测情况, 如住所位置预测、Text 位置预测、下一个位置的预测等. 文献 [87] 提出了一种基于马尔可夫链的预测模型. 文献 [55] 使用隐马尔可夫模型(hidden Markov model, HMM)来建立用户在给定时间段内访问的不同地点之间的关系, 并预测用户访问的下一个位置. 在文本的位置预测中, 文献 [88] 提出了一个马尔可夫随机场概率模型, 根据用户的消息文本内容和他们的社交网络来推断用户的位置.

由于马尔可夫链(additive Markov chain)在序列数据建模方面具有天然的优势, 有研究者将基于马尔可夫链的方法应用于用户地理位置轨迹, 如 GSN 中用户历史轨迹具有明显的序列特性的预测中. 此类算法的基本思想是: 根据用户过去的连续行为预测用户的下一个行为. 文献 [89−91] 基于过去的行为建立一个估计的转移矩阵来表示一个行为的概率, 并使用马尔可夫链框架来建模 POI 之间的顺序影响. 文献 [91] 利用了只考虑用户访问序列中最后一个位置的 1 阶马尔可夫链来预测对序列的影响.

由于 n 阶马尔可夫链的复杂度随 n 的增加呈指数级增长, 而经验实验表明 1 阶马尔可夫链的有效性, 因此大多数文献都假设下一次访问的概率仅依赖于当前访问的概率[92]. 然而, 在现实中, 用户签到行为可能不仅依赖于最新的地点, 还依赖于用户最早访问的地点. 部分研究者也将用户过去访问地点的影响考虑到地理位置预测中. 文献 [89] 通过添加马尔可夫链来预测用户下一次访问的地点. 由于最近签入时间戳的位置通常比包含旧时间戳的位置具有更大的影响力, 它们根据具有最近签入时间戳的位置, 为以前的位置手动设置衰减率参数. 文献 [90] 结合马尔可夫链和人类活动的规律性进行了更复杂的研究. 该文献将移动规律和马尔可夫模型无缝地结合到一个用于位置预测的隐马尔可夫模型框架中, 在该框架中, 登记场所作为隐藏状态, 其他信息作为观察. 文献 [93] 提出 MDC-MCM 的模型, 该模型将签到周期、空间距离和其他影响位置预测的因素进行线性加权, 并与 n 阶马尔可夫链预测结果相融合, 使用 Brightkite 数据集的验证实验表明该模型有更好的预测结果.

有研究者使用基于马尔可夫模型的方法进行地理位置消歧, 提高预测精度. 文献 [94] 使用最大熵马尔可夫模型(maximum mentropy Markov models, MEMM)来查找位置和一组规则来消除地理位置的歧义. 据报道, 该模型在多个新闻报道数据集上的总体精度为 93.8%.

作为传统马尔可夫链的扩展, FPMC(factorizing personalized Markov chain)[95] 结合通用马尔可夫链模型成功地应用于用户位置预测. 然而 FPMC 假设所有的序列行为都是线性组合的, 因素间存在很强的独立性, 并且 FPMC 不能用于建模多个组件之间的关系(例如给定 $a\rightarrow b$, $b\rightarrow c$, 由于 a 和 c 都与 b 有密切的关系, 因此可以直观地认为存在 $a\rightarrow c$ 的转移, 但由于 FPMC 具有很强的独立性假设, 无法捕捉到这种转移).

文献 [96] 提出了 ComPredict 的方法，该方法集成了动态马尔可夫模型（dynamic Markov model, DMM）和从社交网络用户评论中挖掘的移动意图，用于实现用户位置预测. 该方法首先使用基于滑动时间窗口的 PrefixSpan 算法，对打卡轨迹中的频繁模式进行挖掘，同时通过多阶马尔可夫模型在轨迹前缀树中确定模型的正确的阶数；其次为了提升方法的效率，基于用户语义相似性和签到频率将用户聚类为不同的组，用邻接簇的信息避免直接降阶，根据用户当前轨迹序列和其历史轨迹模式，可以得到语义位置预测轨迹集合（semantic location prediction trajectory set, SLPset）；最后，将用户访问过的地点的 SLPset 映射到位置类别层次语义轨迹前缀树，算法在 SLPset 的用户评论信息和群组位置中进行挖掘，进而将其细分为不同类别.

除了上述 4 类方法以外，还有综合采用多种概率模型用于位置预测的研究. 文献 [81] 采用多种概率模型，将用户使用的地点词合并在一起，利用其朋友的地点词解决未提及足够地点词的用户的位置推断问题.

4.2.2　聚类技术

聚类方法的目的是将具有相似空间属性的对象分组成簇[97]. 在本节中，将基于聚类对 OSN 中实现位置预测聚的技术分为 3 类：1）空间聚类；2）具有噪声的基于密度的聚类算法 DBSCAN（density-based spatial clustering of applications with noise）；3）概率密度函数.

1）空间聚类

空间聚类，是对消息文本中的关键词生成空间签名，通过计算向量相似度[62] 对关键词进行增量聚类的方法. 在对 OSN 位置进行预测的方法中，多种聚类算法都有使用.

文献 [62] 中将感兴趣区域建模为具有预定带宽的网格，消息文本在基于时间的滑动窗口中处理，地理标记的消息文本根据它们的坐标分配给网格单元. 根据历史使用值识别当前时间窗口中的突发关键字，生成该关键字的空间签名. 该文献指出，与同一事件相关的关键字应该具有相似的空间签名[62]. 基于此，文献 [97] 采用了一种类似于 BIRCH 的单程聚类算法，可以增量地维护和更新在先前时间窗口中检测到的聚类，而无需提前获取整个数据集. 为了将事件集群与事件无关的事件集群区分开来，还应用了集群评分方案，并在单元大小等于体育场大小的网格系统在足球比赛的定位上验证其算法.

在另一项采用类似方法的研究中，文献 [26] 根据城市的 GPS 坐标将推文分配给网格单元，并使用内容中的网格、事件位置术语以及发布时间为每条推文定义特征向量. 然后根据空间、时间和主题一致性对这些特征向量进行聚类.

另外，文献 [98] 采用与 POI 相关的语义连贯的软聚类来预测用户未来的签到位置. 文献 [99] 开发了一个基于社交关系强度的解决方案，使用局部聚类系数（共同朋友的数量）加权关系. 文献 [100] 的方法首先聚类地理标签，帖子数量最多的组被认为是“住所集群”，并以集群中各点的几何中值作为住所坐标. 采用几何中值的方法避免了简单采用中位数作为居住地点而忽略了用户多个活跃区域的问题. 文献 [101] 采用 *k*-means 聚类算法对相关帖子进行信息特征的分组，考虑用户文本内容等多种因素，但更重要的是考虑其时空细节. 文献 [102] 提出了一种基于模糊 *C*-均值聚类算法和自组织映射方法的地理位置预测方法 PredicTour，该方法可以依据一定时间间隔下的用户打卡历史记录及其个人档案信息形成的用户行为模式，对用户未来可能访问的国家进行预测.

2）DBSCAN

DBSCAN 可以将空间项分组成任意形状的聚类以找到高密度区域[30,103].DBSCAN 根据它们彼此之间的距离对条目进行聚类[104]. 文献 [105] 利用 DBSCAN 对 Twitter 帖子的地理距离进行聚类，并检测帖子密度较高的区域的聚类. 给出了 Twitter 帖子流中检索到的帖子列表，通过分析，确认包含相同单词的帖子是否在空间上彼此接近. 由此产生的帖子集群包含特定的一组单词，并在一个地区具有高密度. 在识别出空间集群作为潜在的地点区域之后，该文献选择了与某一地点事件相关的区域和词汇，这些区域和词汇在一定时间内始终保持活跃状态.

文献 [106] 提出了基于 DBSCAN 的扩展，称为基于密度的时空聚类，该算法可以识别消息文本以时间划分的簇和以空间划分的簇，如果 2 条消息文本的地理距离（从 GPS 坐标中获得）和发布时间的差值在指定的阈值范围内，则认为它们在规定的邻域内，从而从大量的地理参考文档中提取热点区域. 文献 [107] 关注带有地理标签的 Twitter 帖子，进一步扩展了基于密度的时空聚类算法，以支持增量和实时执行. 当接收到新的消息文本时，并不是从头开始执行整个聚类算法，而是通过部分重新聚类和合并的过程来更新受影响的集群. 其算法可以提取热点区域和观测紧急事件的话题.

3）概率密度函数

概率密度函数是基于过去的观测结果，用来估计给定点的连续随机变量的概率密度[108-109]. 此类方法在文本和离散的预定义的区域之间建立映射，即可通过特定的文本去预测其地理位置，但此类方法一般只能将预测精确到训练前指定的范围较大的区域，例如国家或城市. 高斯混合模型（Gaussian mixture model, GMM）和核密度估计（kernel density estimation, KDE）是 2 种成熟而广泛应用的概率密度估计技术.

GMM 是用高斯概率密度函数（正态分布曲线）精确地量化事物，它是一个将事物分解为若干的基于高斯概率密度函数（正态分布曲线）形成的模型.

文献 [21] 提出了统一的判别影响模型，该模型基于社交关系的概率算法，使用高斯分布来捕捉节点的影响. 文献 [110] 根据用户的兴趣对不同的地理位置进行表征并通过在地理坐标上应用双变量高斯分布来模拟 Twitter 数据，文中提出当文本信息与位置信息结合时可以提供有价值的模式. 文献 [111] 将用户的影响建模为以用户所在位置为中心的高斯分布，将分布的方差解释为用户的影响范围，并将家庭位置预测扩展到多个位置分析. 文献 [84] 在预测用户的家庭位置问题上使用 GMM 来适应空间单词的使用，该模型为了缓解以词为中心的家庭预测方法的数据稀疏性，使用高斯或 GMM 实现单词使用分布的平滑. 同样，文献 [112] 也采用了 GMM 进行 Twitter 帖子的位置预测. 文献 [24] 也利用高斯模型对空间 N-Gram 进行建模. 最近，文献 [113] 提出了一种结合文本特征和网络特征的 GMM 定位预测混合方法. 文献 [1] 先使用基于 GCN 和注意力机制的方法对推文中的命名实体进行平滑嵌入，之后采用基于 GMM 的模型对用户发布 Twitter 帖子的位置进行预测，与其他工作不同，其输出一个二元高斯分布的混合，并非以前工作的坐标.

由于高斯分量的数量可能在不同的数据集上有很大的差异，确定高斯分量的数量和调整高斯分量的混合权重可能会带来额外的挑战. 因此，在估计概率密度时，KDE 通常被认为是 GMM 的有力替代.

KDE 是一种统计分析工具，用于生成非参数概率密度函数估计. KDE 是局部核密度估计的线性组合，其中内核的平滑度由以纬度/经度为单位的带宽参数集 θ 控制. 核密度估计[109,114] 是 OSN 在地点预测过程中广泛使用的密度估计技术之一.

KDE 对观测数据的分布做了不严格的假设，这使得它更适合于任意的数据分布[109]. 文献 [115−116] 广泛讨论了 KDE 相对于 GMM 的优点. 文献 [115] 开发了一种混合 KDE 方法，根据活动历史预测个体的位置. 该文献计算了个体水平、区域水平和人口水平 3 种密度分布的加权组合，并使用验证数据集调整方法中的参数，为所有用户应用所选值. 文献 [116] 利用 KDE 建模用户访问地点的地理分布，以便在基于位置的社交网络中提出建议. 文献 [109] 解决了基于网格的文本定位模型中的数据稀疏性问题，并应用 KDE 平滑区域内的文档和单词计数，以提高估计的准确性. 文献 [28] 提出了一种利用 KDE 对 Flickr 照片和维基百科文章进行地理标记的特征选择技术，该技术使用核 KDE 将每个术语建模为地球表面上的 2 维概率分布，并在文献 [117] 中描述的鲁棒扩散 KDE 方法上进行了扩展. 在文献 [114] 将 KDE 用在地图上生成可视化图形. 文献 [28] 在基于内容的微博定位问题中，提出了一个本地化 KDE，它使用信息理论度量，并且不需要对这些设置进行任何参数调优，此方法会根据每个词的局部强度分别自动确定核带宽和权值. 文献 [118] 提出了一种基于 KDE 的混合框架用于完成社交网络用户账户连接任务，并通过对该框架的扩展，实现了对用户地理位置的预测. 文献 [119] 提出了一种基于加权距离 KDE 和张量分解的 2 阶段从粗到精的 POI 推荐算法，第 1 阶段同时考虑序列上下文的长期偏好和人群的偏好来估计粗粒度的用户类别兴趣，第 2 阶段设计一个考虑空间距离的加权 KDE 来确定细粒度的用户位置兴趣.

4.2.3 矩阵分解

矩阵分解是用户位置预测任务的一种常用技术，其基本思路是将 1 个用户位置矩阵分解成 2 个低秩矩阵，分别代表用户或位置的潜在因素，然后利用 2 个向量的内积拟合用户对候选位置的偏好，从而实现位置预测.

已有研究者提出了基于矩阵因子分解，利用用户的明确反馈（例如评级）来建模用户的偏好，从而有效地预测用户未来访问地点的方法. 文献 [120] 从 2 维核密度估计的角度捕捉空间聚类现象，并将空间聚类现象无缝地融合到加权矩阵分解框架中，这样用户对 POI 的偏好就被建模为在向量空间中的内积. 在此基础上，文献 [121] 进一步提出了一个可扩展、灵活的联合地理建模和基于隐式反馈的矩阵分解框架. 文献 [60,122] 充分利用用户文本评论，提出了一种新的基于矩阵分解的方法来联合建模用户偏好和场地特征. 在文献 [123] 将地理影响量化为用户和场所之间的签到概率，然后运用矩阵分解模型得到用

户和场所的向量表示. 文献 [124] 在早期地理贝叶斯非负矩阵分解(geographical Bayesian non-negative matrix factorization, Geo-BNMF)构想[125] 的基础上, 进一步开发了 Geo-PFM 框架, 以捕捉地理对用户签到行为的影响, 并有效地模拟用户移动模式, 文中证明, 当用户签到数据有偏差时其方法依然有效.

矩阵分解模型联合建模多个实体方面也被证明是有效的. 文献 [126] 联合对用户内容偏好矩阵、用户空间偏好矩阵和 POI 特征矩阵进行矩阵分解, 提取用户和 POI 影响中的层次结构, 以实现细粒度用户位置预测.

近年来, 也有研究人员对矩阵分解模型扩展了其适用场景. 其中文献 [127−128] 中的 3 阶张量分解模型和文献 [129] 中的跨区域矩阵分解模型具有代表性.

基于矩阵分解的模型可以获得最好的预测精度. 然而, 基于矩阵分解的模型通常难以处理训练数据中的新用户或新位置, 换言之, 它们在冷启动(cold-start)场景[130] 中不够灵活.

4.3 深度学习方法

近年来, 人工智能技术和计算机硬件水平不断发展: 一方面, 随着自然语言处理方法的不断改进(如词嵌入), 给 OSN 位置预测问题带来了新的工具; 另一方面, 随着硬件能力的提升, 使深度神经网络模型的训练时间大大缩短, 其实用性大大增强. 因此, 越来越多的研究者尝试使用深度神经网络模型来解决 OSN 的位置预测问题. 本节将分别介绍传统神经网络结构、典型词嵌入以及深度神经网络(deep neural network, DNN)模型和基于图神经网络等模型在 OSN 用户位置预测上的应用.

通常用于从文本中提取地名的文本挖掘技术是 NER 的任务之一. 此类方法基于词性标注对文本进行分析, 将某一组词语标注为人、组织或位置等实体类别. 深度学习的出现同时也使 NER 的研究重点转向 DNN.

4.3.1 传统神经网络

传统神经网络的特点是网络层数量少, 往往只有输入输出层以及中间的几个全连接层, 导致模型的拟合度低. 文献 [131] 使用了一个简单的神经网络结构, 其包含完全连接的层和平均池化过程, 将一个帖子或一个用户分类到一个城市. 文献 [132] 使用带有一个隐含层的多层感知器(multilayer perceptron, MLP)来分类用户的家庭位置. 文献 [133] 利用神经网络模型和混合密度网络将 2 维地理标签转换为连续向量空间输入. 文献 [134] 将用户友谊信息编码成神经网络模型. 与其他研究不同的是, 该文献将联网的用户和他们所在的城市分开, 分别分配用户嵌入和城市嵌入. 采用注意力机制添加用户和城市嵌入信息, 以获得对家庭位置预测的有用信息. 文献 [135] 提出基于神经网络的命名实体识别方法, 该方法在大量未标记的训练数据上学习内部表示, 同时该方法中每个单词具有固定大小的窗口, 但没有考虑长距离单词之间的有效信息.

还有研究者将传统神经网络模型与其他技术相结合对用户位置进行预测. 文献 [136] 采用神经网络模型预测文本的位置, 结合 GMM, 利用由消息文本内容提供的卷积混合密度网络来估计 GMM 的参数, 并将估计的密度的模态值作为文本位置的预测坐标, 文中还指出不同的损失函数确实会影响模型的性能.

4.3.2 典型词嵌入模型

词嵌入(word embedding)就是将词汇表示嵌入到一个低维空间中, 便于计算机理解和处理. 其优势之一是它在识别非常规词汇方面的灵活性(这在 OSN 上很常见), 因为单词之间的相似性检查可以用来识别关键词列表中列出的实体[56]. 目前的研究表明, 词嵌入技术, 特别是 Word2Vec[136] 和 Doc2Vec[137] 等词嵌入技术被证明能够有效地捕捉到项目, 如句子中的词、轨迹序列中的轨迹点等之间相互作用的语言规律. 受最近自然语言处理和文本挖掘中词汇嵌入的成功案例的启发, 许多研究尝试应用词嵌入模型来推断位置嵌入, 即 OSN 中位置的向量表示[132,138].

文献 [132] 采用对给定用户的推文内容的 l_2 标准化词袋(bag of words)表示作为输入, 输出是由 k-d 树或 k-means 生成的预定义离散化区域. 文献 [139] 采用基于自然语言处理的混合词嵌入模型进行语义理解, 并结合余弦相似度和 Jaccard 相似度度量进行特征向量提取和降维. 文献 [140] 介绍了用于分析任何语言文本的 Unicode 卷积神经网络, 并在字符级卷积神经网络中加入了一种新的字符嵌入和语言估计器. 文献 [141] 采用 Word2Vec 技术中流行的 Skip-gram 模型, 通过将采样序列中的每个地点作为一个单词来学习低维向量空间中的位置嵌入. 文献 [53] 采用字符感知的词嵌入层处理噪声文本并捕获词汇表外的词, 并利用变分器编码器, 学习了不同特征字段之间的相关性. 文献 [138] 测试了单词嵌入、字符嵌入等特征对分类结果的影响, 提出了一种带有条件随机场(conditional random fields, CRF)层的双向长短期记忆神经网络来识别社交媒体信息中的地理实体, 特别是很少知道的局部位置, 并探索了使用正字法、语

义和句法特征来获得最佳性能的方法. 文献 [20] 提出了一个新的隐含表示模型 POI2Vec, 该方法能够结合地理影响, 并联合建模用户偏好和 POI 顺序转移影响的方法, 以预测特定 POI 的潜在访问者.

典型词嵌入模型虽然被广泛研究, 但也存在一定的缺陷. 现有的词嵌入模型不能很好地对本地时间上下文进行建模; 另外, 由于典型词嵌入模型中的参数数量通常非常大, 框架通常是不可分割的, 即使采用随机梯度下降(stochastic gradient descent, SGD)优化, 依然存在计算成本较高的问题.

4.3.3 深度神经网络

传统的机器学习模型的输出和输入之间只能学习到一个线性关系, 因此只能用于 2 元分类, 且无法学习比较复杂的非线性模型, 因此在实际应用中可用性不强. 而深度神经网络模型则在感知机的基础上做了扩展, 加入了隐藏层, 多层的好处是可以用较少的参数表示复杂的函数, 同时输出层的神经元也可以有多个输出. 近几年, 在 OSN 用户位置预测领域, 深度学习取得了有效的发展. 文献 [142] 使用深层回归神经网络(deep neural network for regression, DNN-R), 负责从多个位置记录中选择一个用户家庭位置轨迹; 深层神经网络分类(deep neural network for classification, DNN-C), 用于选择批准或拒绝选择记录, 从而控制预测家庭位置的数据子集.

1) 基于 RNN 及其变体的方法

循环神经网络(recurrent neural network, RNN)是一种适合序列分类的方法, 能够将一个节点的输出传递给它的后继节点, 这可以理解为一个单词对后续单词的影响. 与传统方法相比, RNN 在自然语言处理中表现出了更好的性能, 一些研究成功地将基于 RNN 的模型应用于用户位置预测[134,138]. 文献 [134] 扩展了文献 [131] 中的方法, 按时间顺序对用户的消息进行排序, 并应用序列 RNN 模型对内容进行编码.

RNN 中的过去输入与当前输入一起处理, 以产生一种临时记忆形式, 这种形式在处理顺序信息时有效. 然而 RNN 由于短时记忆难以处理长序列、反向传播过程中的梯度消失等缺点, 研究人员开发了更多的 RNN 变体, 其中就包括 LSTM.自然地, 研究者也在进行将 LSTM 应用于用户地理位置预测的任务中.

LSTM 是一种特殊的 RNN, 它在长序列上表现得更好. LSTM 采用一种门机制来控制神经网络中的信息流. 因此, 文献 [138] 提出了一种带有条件随机场层的双向长短时记忆神经网络模型(bidirectional - long short-term memory, BiLSTM)来识别社交媒体信息中的地理实体, 特别是很少被人知道的局部位置. 文献 [53] 利用 BiLSTM 模型学习了不同特征字段之间的相关性, 并输出了 2 种分类表示, 分别用于国家和城市的预测. 该模型首先计算国家层面的预测, 并进一步用于指导城市层面的预测. 文献 [37] 使用卷积 LSTM 模型将用户生成的内容及其相关位置视为序列, 采用双向 LSTM 和卷积运算进行位置推断. 文献 [143] 采用了带有条件随机场输出层的双向长短时记忆(BiLSTM-CRF)训练 POI 识别器, 用于微博中的细粒度位置识别和链接. 文献 [144] 首先使用了 HiSpatialCluster 算法从打卡地点中识别出 CA(clustering areas), 再使用 LINE(large-scale information network embedding)取得 CA 的向量表示, 最后使用 BiLSTM-CNN(convolutional neural network)进行地理位置预测. 文献 [145] 提出了基于 LSTM 的人类地理位置预测模型, 该模型在社交网络中, 依据社交关系预测其位置, 并将结果用于移动城市传感器网络的路由协议设计中. 文献 [146] 提出了一种基于 BiLSTM-CNN 的架构, 可以自动检测单词和字符级别的特征. 该技术在结构化良好的句子和知名地点上表现良好, 但 OSN 中帖子的文本通常以非正式或随机的格式写入, 产生地理或非地理歧义, 从而导致 NER 的方法在从社交媒体中提取地理实体上表现不佳.

2) 基于 CNN 的方法

由于 RNN 适用于顺序或长文本数据. 而 OSN 的消息文本中有短句, 这有利于使用 CNN 而不是 RNN. 使用 CNN 的原因是卷积层可以自动学习输入数据的更好表示, 然后全连接层可以利用这些输入表示来识别位置.

文献 [147] 训练了一个基于 CNN 的系统, 将消息文本表示为正常的句子, 并突出显示包含位置信息的单词并将之提取. 文献 [148] 使用 CNN 将文本和用户资料信息元数据集成到单个模型中, 实验显示其优于堆叠朴素贝叶斯分类器的性能. 字符级卷积神经网络目前已经成功应用于只包含英语或有限数量的欧洲语言的文本语料中, 但不能直接应用于可能出现多种语言的 OSN 平台文字分析中.

统一字符编码标准的卷积神经网络(Unicode-CNN)用于分析多语言社交媒体的文本. 该模型直接从输入文本中的 Unicode 字符生成特性, 不需要标记、词干提取或其他预处理. 文献 [140] 探索了 Unicode-CNN 在 Twitter 帖子内容的地理定位任务中的有效性, 提出的 UnicodeCNN 可以将任何 Unicode 字符串作为输入, 理论上能用于所有语言.

3）注意力机制

基于CNN方法的主要思想是通过使用许多不同长度的过滤器来捕获一些特定的的文本. 但由于卷积核滤波器的长度限制, CNN只善于处理局部信息建模. 为了捕获更全面的信息, 一些研究者们开始采用注意力机制[149]来实现OSN的用户位置预测. 注意力机制同时拥有RNN和CNN的优点, 可以对所有文本进行并行计算.

文献[103]在文献[150]的基础上, 将注意力机制改为变分注意力, 同时使用轨迹CNN来学习历史移动性, 与经常使用的RNN相比, 效率得到了更多的提升. 文献[53]从文本嵌入向量中提取位置特征之后, 使用单词级的注意层来生成这些文本字段的表示向量. 文献[151]提出了一个用于文本表示的多头自注意力模型, 为了进一步提高对非正式语言的识别效果, 其将子词(subword)作为模型的一个特征. 然后, 该模型将城市和乡村数据联合用于训练, 以整合来自不同标签的信息. 文献[152]提出了一种基于注意力机制的联合深度学习算法, 该算法可在用户及与其有交互的节点组成的异构网络中捕捉语义元路径, 以进行细粒度的位置信息预测, 特别地, 文中还采用了Pairwise Learning的方法对预测过程进行了优化. 文献[153]设计了一种时空自注意力网络, 基于自注意力机制, 结合用户轨迹的时空信息实现对用户的动态轨迹进行预测. 该模块由3个部分构成: 地点注意力模块, 通过自注意力机制捕捉位置序列迁移; 空间注意力模块, 捕捉用户对地理位置的偏好; 时间注意力模块, 捕获用户活动时间偏好. 最终通过真实世界的社交网络打卡数据集, 验证了时空信息确实能够显著提升模型的性能. 文献[154]提出了一种基于注意力机制的双向门控循环单元模型, 该模型可以用于预测POI种类而非预测精确的POI, 文中提出模糊的POI种类比精确的POI更能反映用户的兴趣, 实验证明该模型能够减轻数据系数问题并保护用户的位置隐私. 另外, 由于引入了注意力机制有选择地关注历史打卡记录, 也使得模型的可解释性得到了改进.

4.3.4 图神经网络

真实生活中, 非结构化的数据结构普遍存在, 例如, 社交网络拓扑结构、化学分子结构和知识图谱等. 语言文本的内部实际上是复杂的树状结构, 它也是一种图结构, 而RNN和CNN很难有效处理此类型的数据. 最近2年, 随着图神经网络(graph neural network, GNN)日益发展, 用GNN来处理位置预测的方法也逐渐显露头角[155-156]. 基于GNN的热度与重要性, 本节将基于GNN的方法单独进行介绍.

文献[157]提出了一种新的基于GNN并使用多方注意力机制的图神经网络(multiple-aspect attention graph neural network, MAGNN). MAGNN是一个多视图的用户地理位置(user geolocation, UG)模型, 为可解释的地理定位捕获语言和交互信息. 该模型将文本内容和社交网络统一起来, 其注意机制能够从多个来源的数据中获取多方面的信息, 这使得MAGNN易于归纳, 适应少数标签的场景. 为了解决过去方法进行地理位置检测时的网络表示学习中的位置不可知以及使用平面图嵌入时的噪声和不稳定的用户关系融合问题, 文献[158]提出了一种层次化图神经网络方法HGNN. 该方法将结构和区域特征使用层次图学习进行编码, 结合了用户的地理位置信息和区域聚集效应, 并且在保持其相对位置的同时捕捉了拓扑关系; 不仅利用了以前方法中无用的孤立节点的信息, 而且还捕获了未标记节点和标记子图之间的关系; 还提出了一种鲁棒的统计方法, 使其模型的行为和输出具有可解释性, 克服了以往的模型将用户地理定位看成一个“黑盒子”而无法解释的缺陷.

文献[159]提出了一种基于图卷积神经网络(graph convolutional network, GCN)[155]的模型, 并同时使用了文本特征和网络特征用于预测, 提升了预测精度. 文献[6]采用了图卷积神经网络来学习用户在一段时间内关于用户已知和未知地理位置的模型, 从而预测其位置. 文献[39]利用GCN中的卷积和池化操作来学习社交网络的结构表示, 从而实现了最佳的地理定位性能. 兴趣相似的异地用户也可以互相提及或关注, 大多数现有的方法都没有完全过滤与用户地理属性无关的用户关系, 而这些关系可能会影响位置推断的准确性. 因此, 文献[160]提出了一种基于表示学习和标签传播的Twitter用户位置预测方法. 该方法充分结合了异构关系, 有效过滤了与用户地理属性无关的关系, 并学习了准确表示用户地理属性特征的向量. 文献[1]提出了一种对Twitter用户地理位置进行预测的框架EDGE, 该框架对推文中提及的命名实体, 提出了一种基于GCN的实体扩散机制(entity diffusion mechanism), 对从用户自我中心网络上提取的称之为entity2vec的实体嵌入进行了平滑处理, 从而输出准确性高且可解释的预测结果. 文献[161]提出了一种在联合模型中兼顾个体偏好和群体决策的新方法预测模型. 该模型利用GCN和注意力机制构建了一个混合深度架构, 提取群体偏好和个人偏好之间的联系, 然后分别捕捉每个用户

对群体决策的影响.

文献 [162] 提出了一种基于层次结构化的 Transformer 网络，用来进行细粒度的空间事件预测，有研究者认为 Transformer 是一种图神经网络模型[163]，可克服传统方法带来的性能低问题.

总的来说，采用深度学习的方法，特别是神经嵌入的方法，由于非线性转换和注意机制，该方法提供了相对更好的性能. 虽然研究者可以定量地衡量各种模型的预测能力，甚至可以根据预测性能对模型进行定性的排序，但这些模型之间的数值差异不是很大. 另外，并不是所有的深度学习模型都能实现比传统方法更有效的预测结果.

5 现有方法总结对比

5.1 方法总结

为了与前述章节在逻辑上保持一致，本节分别从论文对应的预测粒度、输入数据类型、任务类型、数据集、数学模型和评价指标 6 个方面对每篇文献进行阐述. 对于预测粒度，本文使用 POI、坐标或城市等来表示文献所针对的粒度.

与文献 [12] 中的分类方法不同，本文将 OSN 用户位置预测划分为地名词典方法、传统机器学习的方法以及现在基于深度神经网络的方法，并且本文加入了最近几年新兴的模型和训练思路，例如，图卷积神经网络、多头注意力机制以及自然语言中最新的嵌入处理等.

在使用的模型中，最早使用的是基于地名词典的方法，使用地名词典的方法对语法的完全忽视可能会限制地理解析的性能. 随后兴起的是传统机器学习方法，其利用一系列概率方法和聚类技术在位置预测中取得了良好的结果.

在使用概率的方法中，第 1 种是最大似然估计，通过利用训练数据计算后验概率来选择一组选项中的一个位置，该算法依赖于文本处理技术来识别推文中提到的位置名称. 第 2 种是基于 EM 的位置预测技术，该方法额外利用事件、用户和地标之间的密切关系来抵抗信息扩散和人口异质性的影响. 第 3 种是基于滤波器的预测方法，该方法使用推文中的经纬度信息并将位置预测为一个点. 这种方法不需要训练、人工监督、文本分析或有关该区域的先验信息. 第 4 种是基于马尔可夫模型进行位置预测.FPMC 作为马尔可夫链的延伸，假设因素间存在很强的独立性，且 FPMC 不能用于建模多个组件之间的关系，所以难以捕捉到因素之间的联系.

在聚类技术中，信息特征的聚类是将区域建模为网格，预测结果表示为网格中的一个单元格. 该技术使用帖子与在一段时间内收集的坐标来生成空间特征，因此可以认为它对在线执行有一定的支持能力；另外，该技术还需要适度的监督来确定算法中的阈值，它不需要关于位置的广泛知识库，但应该事先知道感兴趣的区域并将其建模为网格. 另一种空间聚类技术 DBSCAN 需要适度的监督来设置阈值. 每次从在线的推文流中收到新推文时，运行 DBSCAN 可能不切实际，但 DBSCAN 可能会定期执行以实现最近实时的位置预测. 核密度估计通常是对帖子进行回顾性地分析，以获得推文在感兴趣区域分布的平滑可视化，根据兴趣区域和目标事件类型定义带宽θ和核函数 H.

近年来，基于地名词典的方法仍然被使用，如文献 [3] 使用基于地名词典的方法来推断 2019 年出现冠状病毒(COVID-19)的地理位置，从而可以统计各个地方疫情的严重程度，同时减少人们去疫情暴发地区的风险. 随着深度学习的流行，命名实体识别从最初的基于字典到方法到现在基于深度学习的方法得到了全新的发展，例如文献 [164] 使用 BiLSTM-CNN 的架构对实体进行提取，使命名实体识别的效果得到了很大的提升，BiLSTM-CNN 提取了地名实体之后首先要解决的是地理实体消歧，因为有的名词指代的可能是人名，也有可能是地名. 文献 [165] 利用 SVM 来联合优化提及地理位置的识别和歧义消除，识别特征(例如大小写)和消歧特征(例如实体流行度)都被集成来训练 SVM.另外，用户之间的友谊、上下文信息以及时间戳也可以用来消除歧义，文献 [166] 采用了一种增量消歧的方法，该方法以文献 [167] 为基础系统地对大量推文进行预处理，这种预处理能够预测基于 OSN 的用户感兴趣的实体. 文献 [72] 在研究中利用时间戳来消除歧义，并观察到工作日的凌晨 2 点至凌晨 5 点之间的 Twitter 帖子流量最低，当有多个候选位置(如悉尼、东京)时，会仔细选择一个地点以避免将时间戳放在低流量窗口中的时区. 同时对获取到的数据采用词嵌入的方式进行降维处理，能更好地提取到词与词之间不同的特征.

在深度学习的方法中，早期方法使用基于 RNN 的模型来预测地理位置[168-169]，但 RNN 由于短时记忆的特点而难以处理长序列信息，且存在消失梯度问题，因此研究人员开发了 RNN 变体，如 GRU, BRNN, LSTM, 并取得了一定的效果. 在实际情况下，OSN 帖

子中文本有短句的特点，而这 3 个 RNN 模型适用于顺序或长文本数据，因此有的学者利用 CNN 来学习输入数据的表示，因为好的表示学习对后续的位置预测更有利. 但由于卷积核滤波器的长度限制，CNN 在全局信息建模方面效果不佳. 为了捕获更全面的信息，一些研究者们开始采用注意力机制[149]来实现用户的位置预测. 同时，社交媒体数据中的非结构化特点以及 GNN 对非结构化数据处理的优势促使研究者基于 GNN 解决 OSN 中的位置预测问题. GNN 也有很多经过改进的模型，例如，简化的图神经网络（simplifying graph convolutional networks, SGC）、GCN 和 MAGNN 等被应用于位置预测研究中.

5.2 应用领域

从模型算法的应用角度，本文所综述的文章可以分为 5 个主要的应用领域：

1）基于位置的推荐及个性化广告推送

通过预测用户发出的帖子的位置，估计其喜好，从而对其进行个性化广告推送以及地点的推送，例如推荐的餐馆等. 文献 [32] 基于 LBSNs 上的用户发布的内容对其进行 POIs 推荐. 文献 [37] 提出对 Tweeter 时间线上的非地理标记推文进行位置推理，其结果可以为诸如位置感知推荐系统和本地具有影响力的用户搜索提供支持. 文献 [170] 中指出通过准确预测用户的精确场所，可提升基于位置广告推荐的准确程度. 文献 [171] 指出，在 LBSNs 中对用户的下一步位置进行预测可以提升定向营销的效果，例如用户下一步可能去电影院，那么向其推荐电影票将大概率会取得效果. 此类应用也是位置预测的主要应用场景.

2）事件动力学分析

通过对特定关键词或标记过的帖子信息进行研究，推断与之相关的位置来研究特定事件的传播过程，可为有效控制事件传播提供支持. 基于社交媒体内容的位置预测，是人类动力学分析研究的重要部分[133]. 文献 [1] 提出 EDGE 模型，并根据特定没有地理位置标签的推文（Twitter 用户发布的帖子）的混合概率分布预测其位置；通过预测一组不含地理位置标签的特定推文的可能位置，从而以实时的方式研究事件发展的动态性.

3）灾难应急响应的资源调度

通过对社交网络上的实时事件信息，如对地震、传染病流行[132]进行监测[88,134]，对推文相关的地理位置进行预测，管理机构可以在第一时间得知发生自然灾害或灾难事件的地理位置，从而在第一时间有目的性地对救灾资源进行调配[172-173]. 在这种情况下，往往不需要细粒度的预测方法. 文献 [174] 提出通过一个用户的多条推文的文本内容对用户进行城市级的位置预测，其预测结果可为灾难应急响应的物资调度提供帮助.

4）人类移动模式研究

文献 [150] 提出的 DeepMove 模型可以对长而稀疏的移动轨迹进行预测并提供了直观解释，有助于可解释地理解人类的移动模式预测，并提出其预测结果是智能交通、城市规划和资源管理等下游应用的重要基础. 文献 [128] 对社交网络的用户时空特性进行建模，达到理解个人或特定用户群组行为特征的目的，最终可以应用于基于 POI 的推荐系统中. 文献 [175] 提出，通过对一系列位置感知服务产生的数据（GPS 轨迹、WiFi 记录、移动电话日志等数据）进行挖掘，从而对用户行为预测，为研究人类移动行为带来了“前所未有”的可能. 文献 [96] 指出移动位置预测随着 LBSNs 的普及，成为了研究人类移动模式的重要手段；而这些人类行为模式，亦可应用于好友推荐、基于位置的广告、城市规划、公共交通系统规划管理和通用导航服务等领域.

5）本地化搜索及事件检测

文献 [27] 提出了在基于图片的社交媒体中对用户的地理位置进行预测，并指出此类任务对社交系统上“感知位置的图片搜索”（location-aware image search）至关重要. 文献 [176] 中提到地理位置预测是支持本地化搜索及本地化事件检测的重要应用，并提出了基于地点标识词检测的方法.

在本节中，对基于 OSN 数据的用户位置预测的现有方法进行了归纳整理，如表 1 所示.

6 未来发展趋势

本文回顾了目前提出的用于 OSN 用户位置预测技术的最新进展. 这些技术可以应用于许多领域，包括市场营销、消费者用户分析和广告定位等. 虽然现有的方法取得了一定的成果，但并未解决所有的问题. 未来可以在数据融合挖掘、复合模型、因果推断的结合 3 个方向进行改进.

6.1 数据融合挖掘

目前大多数的预测方法主要侧重于挖掘单一平台的单一类型数据，如只在 Twitter 上进行数据挖掘，或只针对文本数据或签到数据进行数据分析等. 而在真实世界中，LBSN 中的数据面临严重的数据稀疏

Table 1 Literatures Summary of User Location Prediction Based on OSN Data

表 1 基于 OSN 数据的用户位置预测的文献总结

文献	预测粒度	输入数据类型	任务类型	数据集	数学模型	评价指标
[2]	坐标	文本数据	Text 位置	Twitter	地名词典	*Top@k*
[10]	POI	签到数据	Future 位置	Foursquare, Gowalla	GNN	*Top@k*
[18]	街区	文本数据	Text 位置	Twitter	标签传播	精确率、召回率、*F*1
[24]	坐标	文本数据	Text 位置	Twitter	POI	签到数据
[27]	场所类型	签到数据、图像数据	Future 位置	Foursquare, Brightkite	矩阵分解	精确率、召回率
[28]	网格	文本数据	Text 位置	Twitter	核密度估计	*Acc*、*Top@k*、中值误差
[32]	POI	签到数据、评级数据	Future 位置	Foursquare	矩阵分解	*Top@k*
[37]	坐标	文本数据	Text 位置	Twitter	贝叶斯、LSTM	*Acc@d*
[39]	城市	文本数据	Text 位置	Weibo	聚类	消息文本、上下文
[52]	城市	文本数据	Home 位置	Tweets	地名词典	精确度
[53]	城市	文本数据	Text 位置	Twitter	注意力、Bi-LSTM	*Acc@d*、中值误差、平均误差
[72]	坐标	文本数据	Mentioned 位置	Twitter	地名词典	精确率、召回率
[80]	城市	签到数据	Home 位置	Foursquare	最大似然估计	*Top@k*
[88]	城市	文本数据	Home 位置	Twitter	概率方法	精确率、召回率、*F*1
[98]	POI	签到数据	Future 位置	Foursquare, Douban	核密度估计	*Top@k*
[103]	POI	签到数据	Text 位置	Foursquare, Gowalla	变分注意力、CNN	*Top@k*
[109]	坐标	文本数据	Text 位置	Twitter	贝叶斯	平均误差、中值误差
[122]	POI	签到数据	Future 位置	Yelp	矩阵分解	*MSE*
[128]	POI	签到数据	Future 位置	Gowalla,Yelp	协同过滤	精确率、召回率
[132]	网格	文本数据	Home 位置	Twitter	MLP	*Acc@d*、中值误差、平均误差
[133]	坐标	文本数据	Home 位置	Twitter	GMM	*Acc@d*、中值误差、平均误差
[134]	城市	文本数据	Home 位置	Twitter	注意力、RNN	*Acc*、中值误差
[136]	坐标	文本数据	Text 位置	Twitter	卷积混合密度网络	中值误差、平均误差
[138]	坐标	文本数据	Text 位置	Twitter	LSTM, NER	精确率、召回率、*F*1
[140]	坐标	文本数据	Text 位置	Twitter	CNN	中值误差、平均误差
[142]	坐标	文本数据	Home 位置	Tweets	深层神经网络	精确率、召回率
[143]	POI	文本数据	Text 位置	Twitter	注意力、Bi-LSTM	精确率、召回率、*F*1
[150]	POI	签到数据	Future 位置	Foursquare,Gowalla	注意力机制、RNN	*Top@k*
[151]	坐标	文本数据	Text 位置	Twitter	多头注意力	*Acc*、平均误差、中值误差
[157]	城市	文本数据	Home 位置	Twitter	注意力机制、GNN	*Acc@d*、中值误差、平均误差
[170]	POI	文本数据	Text 位置	Foursquare, Twitter	朴素贝叶斯	平均倒数秩
[171]	场所类型	签到数据	Future 位置	Brightkite, Foursquare	朴素贝叶斯	*Top@k*
[172]	坐标	文本数据	Home 位置	Twitter	标签传播	*Top@k*、平均误差、中值误差
[173]	坐标	文本数据	Home 位置	Twitter	标签传播	*Acc@d*、中值误差
[174]	城市	文本数据	Home 位置	Twitter	机器学习	*Acc*、*Top@k*、平均误差
[175]	场所类型	签到数据	Future 位置	Foursquare	矩阵分解	*Top @k*
[176]	城市	文本数据	Home 位置	Twitter	朴素贝叶斯	*Acc*、*Top@k*、平均误差、中值误差
[177]	POI	签到数据	Future 位置	Foursquare	矩阵分解	精确率、召回率
[178]	POI	签到数据	Future 位置	Gowalla	马尔可夫模型	精确率、召回率
[179]	区域	签到数据	Future 位置	Gowalla, Brightkite	聚类	*Top@k*
[180]	区域	签到数据	Future 位置	Geolife	LSTM	召回率
[181]	POI	签到数据	Future 位置	Instagram	矩阵分解	精确率、召回率

表 1（续）

文献	预测粒度	输入数据类型	任务类型	数据集	数学模型	评价指标
[182]	POI	签到数据	Future 位置	Foursquare	矩阵分解	精确率、召回率
[183]	POI	签到数据	Future 位置	Foursquare, Gowalla	统计	精确率、召回率
[184]	POI	签到数据	Future 位置	Foursquare	最大似然估计	*Top@k*、召回率
[185]	场所类型	签到数据	Future 位置	Foursquare, Gowalla	矩阵分解	精确率、召回率

问题，导致位置预测效果不佳. 此外，基于位置的在线社交网络的用户生成数据是一种隐式反馈，这意味着在实践中可能只有正样本可用. 比如：用户在某些位置没有记录并不一定意味着他/她不喜欢这些位置，有可能是用户不知道该位置，这对从隐式数据中进行学习提出了挑战. 在最近的研究中，已有研究者对不同社交平台的同一用户数据或使用同一社交平台中进行同一用户的不同类型数据进行挖掘. 实验结果表明这种数据融合能缓解数据稀疏问题，而进一步挖掘隐式数据能提升预测性能.

1）跨平台数据融合

大多数研究只使用单一数据源（如 Twitter, Foursquare, Gowalla, Yelp 等）来挖掘用户行为模式. 不过随着多功能异构社交网站的普及，用户跨越多个社交网络[54]拥有账号的现象越来越普遍. 一般相同的用户在不同的社交平台上会展现出一致的行为模式，因此如果能够整合同一用户不同平台的多源数据，便可以很好地缓解数据稀疏性问题，并提升预测的准确度. 目前，越来越多的研究开始对跨平台的多源数据进行挖掘. 在文献 [186] 中，Foursquare 的签到数据和 Twitter 的每日状态被集合在一起，以增强用户生成内容的可用性.

2）同平台多类型数据融合

由于地理标记过的 OSN 数据只占所有数据的一小部分，除了利用跨平台的不同数据源以外，在同一社交网络内不同类型的数据（例如用户属性信息、所在时区、个人档案甚至图片等）也可以进行联合分析，补偿由于地理标记 OSN 数据稀疏性而导致的数据缺失，利用 DNN 来融合分析用户发布的内容，更精确地进行位置预测. 例如文献 [168] 利用 Node2Vec[83]来学习用户表示，并结合 Doc2Vec 学习到的文本表示来预测用户的位置. 文献 [187] 利用 GCN 中的卷积和池化操作来学习社交网络结构的表示，实现了当时最佳的地理定位性能.

在大数据时代，通过分析多源数据，可以很好地分析跨多个社交网站的用户属性，有效地揭示社交网络数据的潜在价值，研究者通过设计更具可解释性的模型，进一步提高用户位置预测的性能.

6.2 复合模型

目前的大多数方法假设数据是独立的特征，并将这些数据以线性方式组合，但没有将数据联合起来学习更有效的表示. 在实际的应用中，为了提高对用户位置预测的精确度与模型的实用性，很多预测方法都联合使用了多种模型. 已经有研究者进行多种方法相结合的尝试. 例如，一些学者结合用户历史轨迹与最近移动状态来研究用户位置预测问题，文献 [150] 使用 RNN 和注意力机制来共同对轨迹隐含的特征进行提取. 文献 [30] 在原始生成模型（包含 GRU、CNN 和注意力机制等模块）的基础上对数据进行增强，利用自监督的方式训练模型来预测用户的下一个可能被访问的 POI.

6.3 因果推断的结合

在现有的研究中，已经证明数据集之间存在着关联关系，通过对这种关联关系进行挖掘及利用，可以进一步提升位置预测的准确性. 而因果关系作为一种比关联更强的关系，普遍存在于社交网络的数据中. 例如，用户在社交网络上发布帖子，表示自己“饿了”，那么其下一步很有可能前往餐厅就餐；用户历史访问地点的位置序列（轨迹）的先后顺序也包含了一定的因果关系，未来访问的地点就可以利用这种因果关系进行推断. 如果能挖掘并利用这种因果关系，势必能弥补现有方法的不足，显著提升位置预测的准确度.

作者贡献声明：刘乐源设计论文框架，撰写及修订论文；代雨柔调研、整理文献，撰写论文；曹亚男对论文的组织结构和部分内容提供了重要的指导意见；周帆对本文选题、组织结构和文章写作提供了关键性的指导意见.

参 考 文 献

[1] Hui Bo, Chen Haiquan, Yan Da, et al. EDGE: Entity-diffusion Gaussian ensemble for interpretable tweet geolocation prediction[C]//

Proc of the 37th IEEE Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2021: 1092−1103

[2] Qazi U, Imran M, Ofli F. Geocov19: A dataset of hundreds of millions of multilingual Covid19 tweets with location information[J]. SIGSPATIAL Special, 2020, 12(1): 6−15

[3] Wang Shuihua, Govindaraj V V, Górriz J M, et al. Covid-19 classification by FGCNet with deep feature fusion from graph convolutional network and convolutional neural network[J]. Information Fusion, 2021, 67: 208−229

[4] Rill S, Reinel D, Scheidt J, et al. Politwi: Early detection of emerging political topics on Twitter and the impact on concept-level sentiment analysis[J]. Knowledge-Based Systems, 2014, 69: 24−33

[5] Unankard S, Li Xue, Sharaf M A. Emerging event detection in social networks with location sensitivity[J]. World Wide Web, 2015, 18(5): 1393−1417

[6] Luceri L, Andreoletti D, Giordano S. Infringement of tweets geo-location privacy: An approach based on graph convolutional neural networks[J]. arXiv preprint, arXiv: 1903. 11206, 2019

[7] Eligüzel N, Çetinkaya C, Dereli T. Comparison of different machine learning techniques on location extraction by utilizing geo-tagged tweets: A case study[J]. Advanced Engineering Informatics, 2020, 46: 101151

[8] Suwaileh R, Imran M, Elsayed T, et al. Are we ready for this disaster? Towards location mention recognition from crisis tweets[C] //Proc of the 28th Int Conf on Computational Linguistics. Red Hook, NY: Curran Associates, 2020: 6252−6263

[9] Hu Yingjie, Wang Jimin. How do people describe locations during a natural disaster: An analysis of tweets from hurricane Harvey[C] //Proc of the 11th Int Conf on Geographic Information Science. Dagstuhl, Germany: Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021: 6: 1−6: 16

[10] Xie Min, Yin Hongzhi, Wang Hao, et al. Learning graph-based POI embedding for location-based recommendation[C] //Proc of the 25th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2016: 15−24

[11] Ajao O, Hong Jun, Liu Weirui. A survey of location inference techniques on Twitter[J]. Journal of Information Science, 2015, 41(6): 855−864

[12] Zheng Xin, Han Jialong, Sun Aixin. A survey of location prediction on Twitter[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(9): 1652−1671

[13] Xu Shuai, Fu Xiaoming, Cao Jiuxin, et al. Survey on user location prediction based on geo-social networking data[J]. World Wide Web, 2020, 23(3): 1621−1664

[14] Cheng Zhiyuan, Caverlee J, Lee K. A content-driven framework for geo-locating microblog users[J]. ACM Transactions on Intelligent Systems and Technology, 2013, 4(1): 1−27

[15] Roller S, Speriosu M, Rallapalli S et al. Supervised text based geolocation using language models on an adaptive grid[C] //Proc of the 2012 Joint Conf on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg, PA: ACL, 2012: 1500−1510

[16] Java A, Song Xiaodan, Finin T, et al. Why we Twitter: An analysis of a microblogging community[C] //Proc of the Int Workshop on Social Network Mining and Analysis. Berlin: Springer, 2007: 118−138

[17] Graham M, Hale S A, Gaffney D. Where in the world are you? Geolocation and language identification in Twitter[J]. The Professional Geographer, 2014, 66(4): 568−578

[18] Shen Wei, Wang Jianyong, Han Jiawei. Entity linking with a knowledge base: Issues, techniques, and solutions[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 27(2): 443−460

[19] MacEachren A M, Jaiswal A, Robinson A C, et al. Senseplace2: Geotwitter analytics support for situational awareness[C] //Proc of the IEEE Conf on Visual Analytics Science and Technology (VAST). Piscataway, NJ: IEEE, 2011: 181−190

[20] Feng Shanshan, Cong Gao, An Bo, et al. Poi2vec: Geographical latent representation for predicting future visitors[C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2017: 102−108

[21] Li Rui, Wang Shengjie, Deng Hongbo, et al. Towards social user profiling: Unified and discriminative influence model for inferring home locations[C]//Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2012: 1023−1031

[22] Schulz A, Hadjakos A, Paulheim H, et al. A multi-indicator approach for geo-localization of tweets[C] //Proc of the 7th Int AAAI Conf on Weblogs and Social Media. Palo Alto, CA: AAAI, 2013: 573−582

[23] Compton R, Jurgens D, Allen D. Geotagging one hundred million Twitter accounts with total variation minimization[C] //Proc of the IEEE Int Conf on Big Data, Palo Alto, CA: AAAI, 2014: 393-401

[24] Flatow D, Naaman M, Xie K E, et al. On the accuracy of hyper-local geotagging of social media content[C] //Proc of the 8th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2015: 127−136

[25] Ritter A, Clark S, Etzioni O, et al. Named entity recognition in tweets: An experimental study[C] //Proc of the 2011 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2011: 1524−1534

[26] Anantharam P, Barnaghi P, Thirunarayan K, et al. Extracting city traffic events from social streams[J]. ACM Transactions on Intelligent Systems and Technology, 2015, 6(4): 1−27

[27] Liu Bo, Yuan Quan, Cong Gao, et al. Where your photo is taken: Geolocation prediction for social images[J]. Journal of the Association for Information Science and Technology, 2014, 65(6): 1232−1243

[28] Ozdikis O, Ramampiaro H, Nørvåg K. Locality-adapted kernel densities of term co-occurrences for location prediction of tweets[J]. Information Processing & Management, 2019, 56(4): 1280−1299

[29] Young S D, Zhang Qingpeng, Zeng D D, et al. Social media images as an emerging tool to monitor adherence to COVID-19 public health guidelines: Content analysis[J]. Journal of Medical Internet Research, 2022, 24(3): e24787

[30] Zhou Fan, Dai Yurou, Gao Qiang, et al. Self-supervised human mobility learning for next location prediction and trajectory

classification[J]. Knowledge-Based Systems, 2021, 228: 107214

[31] Ye Jihang, Zhu Zhe, Cheng Hong. What's your next move: User activity prediction in location-based social networks[C] //Proc of the 2013 SIAM Int Conf on Data Mining. Philadelphia, PA: SIAM, 2013: 171−179

[32] Gao Huiji, Tang Jiliang, Hu Xia, et al. Content-aware point of interest recommendation on location-based social networks[C] // Proc of the 29th AAAI Conf on Artificial Intelligence, Palo Alto, CA: AAAI, 2015: 1721−1727

[33] Zhang Zhiyuan, Liu Yun, Zhang Zhenjiang, et al. Fused matrix factorization with multi-tag, social and geographical influences for POI recommendation[J]. World Wide Web, 2019, 22(3): 1135−1150

[34] Manotumruksa J, Macdonald C, Ounis I. Modelling user preferences using word embeddings for context-aware venue recommendation[J]. arXiv preprint, arXiv: 1606. 07828, 2016

[35] Giridhar P, Abdelzaher T, George J, et al. On quality of event localization from social network feeds[C] //Proc of the IEEE Int Conf on Pervasive Computing and Communication Workshops (PerCom Workshops). Piscataway, NJ: IEEE, 2015: 75−80

[36] Imran M, Castillo C, Diaz F, et al. Processing social media messages in mass emergency: A survey[J]. ACM Computing Surveys, 2015, 47(4): 1−38

[37] Li Pengfei, Lu Hua, Kanhabua N et al. Location inference for non-geotagged tweets in user timelines[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(6): 1150−1165

[38] Lin Yucheng, Lai Chunming, Chapman J W, et al. Geo-location identification of Facebook pages[C] //Proc of 2018 IEEE/ACM Int Conf on Advances in Social Networks Analysis and Mining (ASONAM). Piscataway, NJ: IEEE, 2018: 441−446

[39] Ao J, Zhang Peng, Cao Yanan. Estimating the locations of emergency events from Twitter streams[J]. Procedia Computer Science, 2014, 31: 731−739

[40] Sakaki T, Okazaki M Matsuo Y. Tweet analysis for real-time event detection and earthquake reporting system development[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 25(4): 919−931

[41] Lee R, Sumiya K. Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection[C] //Proc of the 2nd ACM SIGSPATIAL Int Workshop on Location Based Social Networks. New York: ACM, 2010: 1−10

[42] Wing B, Baldridge J. Simple supervised document geolocation with geodesic grids[C] //Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2011: 955−964

[43] Abrol S, Khan L. Tweethood: Agglomerative clustering on fuzzy k-closest friends with variable depth for location mining[C] //Proc of the 2nd IEEE Int Conf on Social Computing. Piscataway, NJ: IEEE, 2010: 153−160

[44] Backstrom L, Sun E, Marlow C. Find me if you can: Improving geographical prediction with social and spatial proximity[C] //Proc of the 19th Int Conf on World Wide Web. New York: ACM, 2010: 61−70

[45] McGee J, Caverlee J, Cheng Zhiyuan. Location prediction in social media based on tie strength[C] //Proc of the 22nd ACM Int Conf on Information & Knowledge Management. New York: ACM, 2013: 459−468

[46] Jurgens D. That's what friends are for: Inferring location in online social media platforms based on social relationships[C] //Proc of the 7th Int AAAI Conf on Weblogs and Social Media. Palo Alto, CA: AAAI, 2013: 273−282

[47] Kossinets G, Watts D J. Empirical analysis of an evolving social network[J]. Science, 2006, 311(5757): 88−90

[48] McGee J, Caverlee J A, Cheng Zhiyuan. A geographic study of tie strength in social media[C] //Proc of the 20th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2011: 2333−2336

[49] Hecht B, Hong L, Suh B, et al. Tweets from justin bieber's heart: The dynamics of the location field in user profiles[C] //Proc of the SIGCHI Conf on Human Factors in Computing Systems. New York: ACM, 2011: 237−246

[50] Achrekar H, Gandhe A, Lazarus R, et al. Online social networks flu trend tracker: A novel sensory approach to predict flu trends[C] //Proc of the Int Joint Conf on Biomedical Engineering Systems and Technologies. Berlin: Springer, 2012: 353−368

[51] Yin Jie, Karimi S, Lampert A, et al. Using social media to enhance emergency situation awareness[C] //Proc of the 24th Int Joint Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2015: 4234−4239

[52] Mahmud J, Nichols J, Drews C. Home location identification of twitter users[J]. ACM Transactions on Intelligent Systems and Technology, 2014, 5(3): 1−21

[53] Huang Binxuan, Carley K M. A hierarchical location prediction neural network for twitter user geolocation[C] //Proc of the 2019 Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing. Stroudsburg, PA: ACL, 2019: 4731−4741

[54] Van Der Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2579−2605

[55] Li Chenliang, Sun Aixin. Fine-grained location extraction from tweets with temporal awareness[C] //Proc of the 37th Int ACM SIGIR Conf on Research & Development in Information Retrieval. New York: ACM, 2014: 43−52

[56] Ikawa Y, Enoki M, Tatsubori M. Location inference using microblog messages[C] //Proc of the 21st Int Conf on World Wide Web. New York: ACM, 2012: 687−690

[57] Sankaranarayanan J, Samet H, Teitler B E, et al. Twitterstand: News in tweets[C] //Proc of the 17th ACM SIGSPATIAL Int Conf on Advances in Geographic Information Systems. New York: ACM, 2009: 42−51

[58] Teitler B E, Lieberman M D, Panozzo D, et al. NewsStand: A new view on news[C/OL] //Proc of the 16th ACM SIGSPATIAL Int Conf on Advances in Geographic Information Systems. New York: ACM, 2008 [2022-11-11].https://dl.acm.org/doi/10.1145/1463434.1463458

[59] Chauhan A, Kummamuru K, Toshniwal D. Prediction of places of visit using tweets[J]. Knowledge and Information Systems, 2017, 50(1): 145−166

[60] Manotumruksa J, Macdonald C, Ounis I. Matrix factorisation with word embeddings for rating prediction on location-based social networks[C] //Proc of the European Conf on Information Retrieval. Berlin: Springer, 2017: 647−654

[61] Lim W L, Ho C C, Ting C Y. Sentiment analysis by fusing text and location features of geo-tagged tweets[J]. IEEE Access, 2020, 8: 181014−181027

[62] Abdelhaq H, Sengstock C, Gertz M. Eventweet: Online localized event detection from Twitter[J]. Proceedings of the VLDB Endowment, 2013, 6(12): 1326−1329

[63] Samuel A, Sharma D K. Location estimation of non-geo-tagged tweets[J]. Evolutionary Intelligence, 2021, 14(2): 205−216

[64] Ozdikis O, Oguztuzun H, Karagoz P. Evidential location estimation for events detected in Twitter[C] //Proc of the 7th Workshop on Geographic Information Retrieval. New York: ACM, 2013: 9−16

[65] Ozdikis O, Oğuztüzün H, Karagoz P. Evidential estimation of event locations in microblogs using the Dempster–Shafer theory[J]. Information Processing & Management, 2016, 52(6): 1227−1246

[66] Nadeau D, Sekine S. A survey of named entity recognition and classification[J]. Lingvisticae Investigationes, 2007, 30(1): 3−26

[67] Roick O, Heuser S. Location based social networks-definition, current state of the art and research agenda[J]. Transactions in GIS, 2013, 17(5): 763−784

[68] Gelernter J, Mushegian N. Geo-parsing messages from microtext[J]. Transactions in GIS, 2011, 15(6): 753−773

[69] Lingad J, Karimi S, Yin Jie. Location extraction from disaster-related microblogs[C] //Proc of the 22nd Int Conf on World Wide Web. New York: ACM, 2013: 1017−1020

[70] Hinduja S, Patchin J W. Bullying, cyberbullying, and suicide[J]. Archives of Suicide Research, 2010, 14(3): 206−221

[71] Malmasi S, Dras M. Location mention detection in tweets and microblogs[C] //Proc of the Conf of the Pacific Association for Computational Linguistics. Berlin: Springer, 2015: 123−134

[72] Zhang Wei, Gelernter Judith. Geocoding location expressions in Twitter messages: A preference learning method[J]. Journal of Spatial Information Science, 2014, 2014(9): 37−70

[73] Paradesi S M. Geotagging tweets using their content[C] //Proc of the 24th Int FLAIRS Conf. Palo Alto, CA: AAAI, 2011: 355−356

[74] Vieweg S, Hughes A L, Starbird K, et al. Microblogging during two natural hazards events: What Twitter may contribute to situational awareness[C] //Proc of the SIGCHI Conf on Human Factors in Computing Systems. New York: ACM, 2010: 1079−1088

[75] Pennacchiotti M, Popescu A. A machine learning approach to Twitter user classification[C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media. Palo Alto, CA: AAAI, 2011: 281−288

[76] Gültekin G, Bayat O. A Naïve Bayes prediction model on location-based recommendation by integrating multi-dimensional contextual information[J]. Multimedia Tools and Applications, 2022, 81(5): 6957−6978

[77] Stock K, Jones C B, Russell S, et al. Detecting geospatial location descriptions in natural language text[J]. International Journal of Geographical Information Science, 2022, 36((3):): 547−584

[78] Giridhar P, Wang Shiguang, Abdelzaher T F, et al. Joint localization of events and sources in social networks[C] //Proc of the Int Conf on Distributed Computing in Sensor Systems. Piscataway, NJ: IEEE, 2015: 179−188

[79] Cheng Zhiyuan, Caverlee J, Lee K. You are where you tweet: A content-based approach to geo-locating Twitter users[C] //Proc of the 19th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2010: 759−768

[80] Huang Chao, Wang Dong, Zhu Shenglong. Where are you from: Home location profiling of crowd sensors from noisy and sparse crowdsourcing data[C/OL] //Proc of the IEEE Conf on Computer Communications (INFOCOM 2017). Piscataway, NJ: IEEE, 2017 [2023-03-29].https://ieeexplore.ieee.org/abstract/document/8057112

[81] Eisenstein Ja, O'Connor B, Smith N A, et al. A latent variable model for geographic lexical variation[C] //Proc of the 2010 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2010: 1277−1287

[82] Hong Liangjie, Ahmed A, Gurumurthy S, et al. Discovering geographical topics in the Twitter stream[C] //Proc of the 21st Int Conf on World Wide Web. New York: ACM, 2012: 769−778

[83] Grover A, Leskovec J. Node2Vec: Scalable feature learning for networks[C] //Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 855864

[84] Chang H, Lee D, Eltaher M, et al. @Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage[C] //Proc of the IEEE/ACM Int Conf on Advances in Social Networks Analysis and Mining. New York: ACM, 2012: 111−118

[85] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes twitter users: Real-time event detection by social sensors[C] //Proc of the 19th Int Conf on World Wide Web. New York: ACM, 2010: 851−860

[86] Arulampalam M S, Maskell S, Gordon N, et al. A tutorial on particle filters for online nonlinear/nongaussian Bayesian tracking[J]. IEEE Transactions on Signal Processing, 2002, 50(2): 174−188

[87] Li Wen, Eickhoff C, de Vries A P. Want a coffee? Predicting users' trails[C] //Proc of the 35th Int ACM SIGIR Conf on Research and Development In Information Retrieval. New York: ACM, 2012: 1171−1172

[88] Rodrigues E, Assunção R, Pappa G L, et al. Exploring multiple evidence to infer users'location in Twitter[J]. Neuro Computing, 2016, 171: 30−38

[89] Zhang Jiadong, Chow C Y, Li Yanhua. Lore: Exploiting sequential influence for location recommendations[C] //Proc of the 22nd ACM SIGSPATIAL Int Conf on Advances in Geographic Information Systems. New York: ACM, 2014: 103−112

[90] Lian Defu, Xie Xing, Zheng Vincent W, et al. CEPR: A collaborative exploration and periodically returning model for location prediction[J]. ACM Transactions on Intelligent Systems and Technology, 2015, 6(1): 1−27

[91] Cheng Chen, Yang Haiqin, Lyu M R, et al. Where you like to go next: Successive point-of-interest recommendation[C] //Proc of the

23rd Int Joint Conf on Artificial Intelligence. Los Altos, CA: IJCAI, 2013: 2605−2611

[92] Feng Shanshan, Li Xutao, Zeng Yifeng, et al. Personalized ranking metric embedding for next new POI recommendation[C] //Proc of the 24th Int Joint Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2015: 2069−2075

[93] Chen Sijia, Zhang Jian, Meng Fanwei, et al. A Markov chain position prediction model based on multidimensional correction[J]. Complexity, 2021, 2021: 6677132

[94] Good I J. The population frequencies of species and the estimation of population parameters[J]. Biometrika, 1953, 40(34): 237−264

[95] Rendle S, Freudenthaler C SchmidtThieme L. Factorizing personalized Markov chains for nextbasket recommendation[C] //Proc of the 19th Int Conf on World Wide Web. New York: ACM, 2010: 811−820

[96] Lu Jing, Anjiri S N, Liu Cong. Compredict: A location prediction method by mining semantic trajectories and mobile intention[J/OL]. Available at SSRN, 2022 [2022-05-06].https://ssrn.com/abstract=4034463

[97] Han Jiawei, Kamber M, Tung A. Spatial clustering methods in data mining: A survey[J]. Geographic Data Mining and Knowledge Discovery, 2001, 1: 188−217

[98] Yin Hongzhi, Hu Zhiting, Zhou Xiaofang, et al. Discovering interpretable geosocial communities for user behavior prediction [C]//Proc of the IEEE 32nd Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2016: 942−953

[99] Kong Longbo, Liu Zhi, Huang Yan. Spot: Locating social media users based on social network context[J]. Proceedings of the VLDB Endowment, 2014, 7(13): 1681−1684

[100] Poulston A, Stevenson M, Bontcheva K. Hyperlocal home location identification of Twitter profiles[C]//Proc of the 28th ACM Conf on Hypertext and Social Media. New York: ACM, 2017: 45−54

[101] Alaskar H, Vaiyapuri T, Sbai Z. Twitter analytics for discovering socially important locations for business improvement[C/OL] //Proc of the IEEE Int Symp on Signal Processing and Information Technology (ISSPIT). Piscataway, NJ: IEEE, 2019 [2023-03-29].https://ieeexplore.ieee.org/abstract/document/9001873

[102] Senefonte H C M, Delgado M R, Lüders R, et al. PredicTour: Predicting mobility patterns of tourists based on social media user's profiles[J]. IEEE Access, 2022, 10: 9257−9270

[103] Gao Qiang, Zhou Fan, Trajcevski G, et al. Predicting human mobility via variational attention[C] //Proc of the World Wide Web Conf. New York: ACM, 2019: 2750−2756

[104] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C] //Proc of the 2nd Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 1996: 226−231

[105] Boettcher A, Lee D. Eventradar: A real-time local event detection scheme using Twitter stream[C] //Proc of the IEEE Int Conf on Green Computing and Communications. Piscataway, NJ: IEEE, 2012: 358−367

[106] Tamura K, Ichimura T. Density-based spatiotemporal clustering algorithm for extracting bursty areas from georeferenced documents[C] //Proc of the IEEE Int Conf on Systems, Man, and Cybernetics. Piscataway, NJ: IEEE, 2013: 2079−2084

[107] Sakai T, Tamura K. Identifying bursty areas of emergency topics in geotagged tweets using density-based spatiotemporal clustering algorithm[C] //Proc of the 7th IEEE Int Workshop on Computational Intelligence and Applications (IWCIA). Piscataway, NJ: IEEE, 2014: 95−100

[108] Ebrahimi M, ShafieiBavani E, Wong R, et al. Exploring celebrities on inferring user geolocation in Twitter[C] //Proc of the Pacific-Asia Conf on Knowledge Discovery and Data Mining. Piscataway, NJ: IEEE, 2017: 395−406

[109] Hulden M, Silfverberg M, Francom J. Kernel density estimation for text-based geolocation[C] //Proc of the 29th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2015: 145-150

[110] Yin Zhijun, Cao Liangliang, Han Jiawei, et al. Geographical topic discovery and comparison[C] //Proc of the 20th Int Conf on World Wide Web. New York: ACM, 2011: 247−256

[111] Van L O, Quinn J, Schockaert S, et al. Spatially aware term selection for geotagging[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 26(1): 221−234

[112] Priedhorsky R, Culotta A, Del V S Y. Inferring the origin locations of tweets with quantitative confidence[C] //Proc of the 17th ACM Conf on Computer Supported Cooperative Work & Social Computing. New York: ACM, 2014: 1523−1536

[113] Bakerman J, Pazdernik K, Wilson A, et al. Twitter geolocation: A hybrid approach[J]. ACM Transactions on Knowledge Discovery from Data, 2018, 12(3): 1−17

[114] Lu Yafeng, Hu Xia, Wang Feng, et al. Visualizing social media sentiment in disaster scenarios[C] //Proc of the 24th Int Conf on World Wide Web. New York: ACM, 2015: 1211−1215

[115] Lichman M, Smyth P. Modeling human location data with mixtures of kernel densities[C] //Proc of the 20th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2014: 35−44

[116] Zhang Jiadong, Chow C Y. iGSLR: Personalized geo social location recommendation: A kernel density estimation approach[C] //Proc of the 21st ACM SIGSPATIAL Int Conf on Advances in Geographic Information Systems. New York: ACM, 2013: 334−343

[117] Botev Z I, Grotowski J F, Kroese D P. Kernel density estimation via diffusion[J]. The Annals of Statistics, 2010, 38(5): 2916−2957

[118] Chen Wei, Wang Weiqing, Yin Hongzhi, et al. HFUL: A hybrid framework for user account linkage across location-aware social networks[J]. The VLDB Journal, 2023, 32(1): 1−22

[119] Cai Liangqi, Wen Wen, Wu Biao, et al. A coarse-to-fine user preferences prediction method for point-of-interest recommendation[J]. Neurocomputing, 2021, 422: 1−11

[120] Lian Defu, Zhao Cong, Xie Xing, et al. GeoMF: Joint geographical modeling and matrix factorization for point-of-interest recommendation[C] //Proc of the 20th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2014: 831−840

[121] Lian Defu, Zheng Kai, Ge Yong, et al. GeoMF++ scalable location

recommendation via joint geographical modeling and matrix factorization[J]. ACM Transactions on Information Systems, 2018, 36(3): 1−29

[122] Manotumruksa J, Macdonald C, Ounis I. Regularising factorised models for venue recommendation using friends and their comments[C] //Proc of the 25th ACM Int on Conf on Information and Knowledge Management. New York: ACM, 2016: 1981−1984

[123] Zhao Shenglin, Zhao Tong, Yang Haiqin, et al. STELLAR: Spatial-temporal latent ranking for successive point-of-interest recommendation[C] //Proc of the 30th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2016: 315−322

[124] Liu Bin, Xiong Hui, Papadimitriou S, et al. A general geographical probabilistic factor model for point of interest recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 27(5): 1167−1179

[125] Liu Bin, Fu Yanjie, Yao Zijun, et al. Learning geographical preferences for point-of-interest recommendation[C] //Proc of the 19th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2013: 1043−1051

[126] Zhao Pengpeng, Xu Xiefeng, Liu Yanchi, et al. Exploiting hierarchical structures for POI recommendation[C] //Proc of the 2017 IEEE Int Conf on Data Mining (ICDM). Piscataway, NJ: IEEE, 2017: 655−664

[127] Li Xin, Jiang Mingming, Hong Huiting, et al. A time-aware personalized point-of-interest recommendation via high-order tensor factorization[J]. ACM Transactions on Information Systems, 2017, 35(4): 1−23

[128] Yang Dingqi, Zhang Daqing, Zheng V W, et al. Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2014, 45(1): 129−142

[129] Zhang Chenyi, Wang Ke. POI recommendation through cross-region collaborative filtering[J]. Knowledge and Information Systems, 2016, 46(2): 369−387

[130] Liu Qiang, Wu Shu, Wang Liang, et al. Predicting the next location: A recurrent model with spatial and temporal contexts[C] //Proc of the 30th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2016: 194−200

[131] Miura Y, Taniguchi M, Taniguchi T, et al. A simple scalable neural networks based model for geolocation prediction in Twitter[C] //Proc of the 2nd Workshop on Noisy User-generated Text (WNUT). New York: ACM, 2016: 235−239

[132] Rahimi A, Cohn T, Baldwin T. A neural model for user geolocation and lexical dialectology[C] //Proc of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2017: 209−216

[133] Rahimi A, Baldwin T, Cohn T. Continuous representation of location for geolocation and lexical dialectology using mixture density networks[C] //Procs of the 2017 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2017: 167−176

[134] Miura Y, Taniguchi M, Taniguchi T, et al. Unifying text, metadata, and user network representations with a neural network for geolocation prediction[C] //Proc of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2017: 1260−1272

[135] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12: 2493−2537

[136] Iso H Wakamiya S, Aramaki E. Density estimation for geolocation via convolutional mixture density network[J]. arXiv preprint, arXiv: 1705. 02750, 2017

[137] Mousset P, Pitarch Y, Tamine L. End-to-end neural matching for semantic location prediction of tweets[J]. ACM Transactions on Information Systems, 2020, 39(1): 1−35

[138] Chen Zi, Pokharel B, Li Bingnan, et al. Location extraction from twitter messages using a bidirectional long short-term memory neural network with conditional random field model[C] //Proc of the Int Conf on Geographical Information Systems Theory, Applications and Management. Berlin: Springer, 2020: 18−30

[139] Ajao O, Bhowmik D, Zargari S. Content-aware tweet location inference using quadtree spatial partitioning and Jaccard-cosine word embedding[C] //Proc of the IEEE/ACM Int Conf on Advances in Social Networks Analysis and Mining (ASONAM). New York: ACM, 2018: 1116−1123

[140] Izbicki M, Papalexakis V, Tsotras V. Geolocating tweets in any language at any location[C]//Proc of the 28th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2019: 89−98

[141] Xu Shuai, Cao Jiuxin, Legg P, et al. Venue2vec: An efficient embedding model for fine-grained user location prediction in geosocial networks[J]. IEEE Systems Journal, 2019, 14((2):): 1740−1751

[142] Ghaffari M, Srinivasan A, Liu Xiuwen. High resolution home location prediction from tweets using deep learning with dynamic structure[C] //Proc of the 2019 IEEE/ACM Int Conf on Advances in Social Networks Analysis and Mining. Piscataway, NJ: IEEE, 2019: 540−542

[143] Xu Canwen, Li Jing, Luo Xiangyang, et al. Dlocrl: A deep learning pipeline for fine grained location recognition and linking in tweets[C] //Proc of the World Wide Web Conf. New York: ACM, 2019: 3391−3397

[144] Bao Yi, Huang Zhou, Li Linna, et al. A BiLSTM-CNN model for predicting users' next locations based on geotagged social media[J]. International Journal of Geographical Information Science, 2021, 35(4): 639−660

[145] Dat V A D, Yoon S. A human location prediction-based routing protocol in mobile crowdsensing-based urban sensor networks[J]. Applied Sciences, 2022, 12(8): 3898

[146] Chiu J P C, Nichols E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357−370

[147] Kumar A, Singh J P. Location reference identification from tweets during emergencies: A deep learning approach[J]. Internatinal Journal of Disaster Risk Reduction, 2019, 33: 365−375

[148] Huang Binxuan, Carley K M. On predicting geolocation of tweets

using convolutional neural networks[C] //Proc of the Int Conf on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Berlin: Springer, 2017: 281−291

[149] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C] //Proc of the Advances in Neural Information Processing Systems. Piscataway, NJ: IEEE, 2017: 5998−6008

[150] Feng Jie, Li Yong, Zhang Chao, et al. Deepmove: Predicting human mobility with attentional recurrent networks[C] //Proc of the 2018 World Wide Web Conf. New York: ACM, 2018: 1459−1468

[151] Huang Chieh Yang, Tong Hanghang, He Jingrui, et al. Location prediction for tweets[J]. Frontiers in Big Data, 2019, 2: 5

[152] Wang Zhixiao, Yan Wenyao, Gao Ang. Fine-grained user location prediction using meta-path context with attention mechanism[J]. Journal of Web Engineering, 2021, 20(3): 597−614

[153] Wang Shuang, Li Anlian, Xie Shuai, et al. A spatial-temporal self-attention network for location prediction[J/OL]. Complexity, 2021 [2022-11-11]. https://downloads.hindawi.com/journals/complexity/2021/6692313.pdf

[154] Liu Yuwen, Song Zuolong, Xu Xiaolong, et al. Bidirectional GRU networks-based next POI category prediction for healthcare[J]. International Journal of Intelligent Systems, 2022, 37(7): 4020−4040

[155] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[C/OL] //Proc of the 5th Int Conf on Learning Representations. 2017 [2022-05-05].https://openreview.net/pdf?id=SJU4ayYgl

[156] Velickovic P, Cucurull G, Casanova A, et al. Graph attention networks[C/OL] //Proc of the 6th Int Conf on Learning Representations. 2018 [2022-05-05].https://openreview.net/pdf?id=rJXMpikCZ

[157] Zhong Ting, Wang Tianliang, Wang Jiahao, et al. Multiple-aspect attentional graph neural networks for online social network user localization[J]. IEEE Access, 2020, 8: 95223−95234

[158] Zhou Fan, Wang Tianliang, Zhong Ting, et al. Identifying user geolocation with hierarchical graph neural networks and explainable fusion[J/OL]. Information Fusion, 2022 [2023-03-29].https://www.sciencedirect.com/science/article/pii/S1566253521002293

[159] Miyazaki T, Rahimi A, Cohn T, et al. Twitter geolocation using knowledge-based methods[C] //Proc of the 4th Workshop on Noisy User-generated Text. Stroudsburg, PA: ACL, 2018: 7−16

[160] Tian Hechan, Zhang Meng, Luo Xiangyang, et al. Twitter user location inference based on representation learning and label propagation[C] //Proc of the Web Conf. New York: ACM, 2020: 2648−2654

[161] Elmi S, Tan K L. Influence-based deep network for next POIs prediction[C] //Proc of the 44th European Conf on Information Retrieval. Berlin: Springer, 2022: 170−183

[162] Wu Xian, Huang Chao, Zhang Chuxu, et al. Hierarchically structured transformer networks for fine-grained spatial event forecasting[C] //Proc of the Web Conf 2020. New York ACM, 2020: 2320–2330

[163] Chaitanya J. Transformers are graph neural networks [EB/OL]. 2020 [2022-05-05].https://graphdeeplearning.github.io/post/transformers-are-gnns/

[164] Tsochantaridis I, Joachims T, Hofmann T, et al. Large margin methods for structured and interdependent output variables[J]. Journal of Machine Learning Research, 2005, 6(9): 1453−1484

[165] Guo S, Chang Mingwei, Kiciman E. To link or not to link? A study on end-to-end tweet entity linking[C] //Proc of the 2013 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2013: 1020−1030

[166] Hua Wen, Zheng Kai, Zhou Xiaofang. Microblog entity linking with social temporal context[C] //Proc of the 2015 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2015: 1761−1775

[167] Shen Wei, Wang Jianyong, Luo Ping, et al. Linking named entities in tweets with knowledge base via user interest modeling[C] //Proc of the 19th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2013: 68−76

[168] Do T H, Nguyen D M, Tsiligianni E, et al. Multiview deep learning for predicting Twitter users' location[J]. arXiv preprint, arXiv: 1712.08091, 2017

[169] Wang Dong, Al Amin M T , Abdelzaher T, et al. Provenance-assisted classification in social networks[J]. IEEE Journal of Selected Topics in Signal Processing, 2014, 8(4): 624−637

[170] Chong W H, Lim E P. Tweet geolocation: Leveraging location, user and peer signals[C] //Proc of the 2017 ACM on Conf on Information and Knowledge Management. New York: ACM, 2017: 1279−1288

[171] Likhyani A, Padmanabhan D, Bedathur S, et al. Inferring and exploiting categories for next location prediction[C] //Proc of the 24th Int Conf on World Wide Web. New York: ACM, 2015: 65−66

[172] Rahimi A, Cohn T, Baldwin T. Twitter user geolocation using a unified text and network prediction model[C] //Proc of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int Joint Conf on Natural Language Processing of the Asian Federation of Natural Language Processing. Stroudsburg, PA: ACL, 2015: 630−636

[173] Rahimi A, Vu D, Cohn T, et al. Exploiting text and network context for geolocation of social media users[C] //Proc of the 2015 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2015: 1362−1367

[174] Han B, Cook P, Baldwin T. Text-based Twitter user geolocation prediction[J]. Journal of Artificial Intelligence Research, 2014, 49: 451−500

[175] Liao Dongliang, Zhong Yuan, Li Jing. Location prediction through activity purpose: Integrating temporal and sequential models[C] //Proc of Pacific-Asia Conf on Knowledge Discovery and Data Mining. Berlin: Springer, 2017: 711−723

[176] Han B, Cook P, Baldwin T. Geolocation prediction in social media data by finding location indicative words[C]//Proc of the Int Conf on Computational Linguistics 2012. New York: ACM, 2012: 1045−1062

[177] Gao Huiji, Tang Jiliang, Liu Huan. Exploring social-historical ties on location-based social networks[C] //Proc of the 6th Int AAAI Conf on Weblogs and Social Media. Palo Alto, CA: AAAI, 2012: 114−121

[178] Mazumdar P, Patra B K, Babu K S, et al. Hidden location prediction using check-in patterns in location-based social networks[J]. Knowledge and Information Systems, 2018, 57(3): 571−601

[179] Hai N T, Nguyen H H, Thai-Nghe N. A mobility prediction model for location-based social networks[C] //Proc of the Asian Conf on Intelligent Information and Database Systems. Berlin: Springer, 2016: 106−115

[180] Wong M H, Tseng V S, Tseng J C, et al. Long-term user location prediction using deep learning and periodic pattern mining[C] //Proc of the Int Conf on Advanced Data Mining and Applications. Berlin: Springer, 2017: 582−594

[181] Wang Yingzi, Yuan N J, Lian Defu, et al. Regularity and conformity: Location prediction using heterogeneous mobility data[C] //Proc of the 21st ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2015: 1275−1284

[182] Gao Huiji, Tang Jiliang, Hu Xia, et al. Modeling temporal effects of human mobile behavior on location-based social networks[C] //Proc of the 22nd ACM Int Conf on Information & Knowledge Management. New York: ACM, 2013: 1673−1678

[183] Li Xutao, Cong Gao, Li Xiaoli, et al. Rank-geofm: A ranking based geographical factorization method for point of interest recommendation[C] //Proc of the 38th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2015: 433−442

[184] Pang Jun, Zhang Yang. Exploring communities for effective location prediction[C] //Proc of the 24th Int Conf on World Wide Web. New York: ACM, 2015: 87−88

[185] Yang Dingqi, Qu Bingqing, Yang Jie, et al. Revisiting user mobility and social relationships in lbsns: A hypergraph embedding approach[C] //Proc of the World Wide Web Conf. New York: ACM, 2019: 2147−2157

[186] Hristova D, Williams M J, Musolesi M, et al. Measuring urban social diversity using interconnected geo-social networks[C]//Proc of the 25th Int Conf on World Wide Web. New York: ACM, 2016: 21−30

[187] Rahimi A, Cohn T, Baldwin T. Semi-supervised user geolocation via graph convolutional networks[C] //Proc of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 2009−2019



Liu Leyuan, born in 1982. PhD, assistant professor. His main research interests include machine learning, graph learning, and social network data mining.

刘乐源, 1982 年生. 博士, 助理研究员. 主要研究方向为机器学习、图学习、社交网络数据挖掘.



Dai Yurou, born in 1997. Master, research assistant. Her main research interests include data mining, machine learning, computer vision, and autonomous driving.

代雨柔, 1997 年生. 硕士, 研究助理. 主要研究方向为数据挖掘、机器学习、计算机视觉、自动驾驶.



Cao Yanan, born in 1985. PhD, professor. Member of CCF. Her main research interests include natural language processing, social network analysis, and machine learning.

曹亚男, 1985 年生. 博士, 研究员. CCF 会员. 主要研究方向为自然语言处理、社交网络分析、机器学习.



Zhou Fan, born in 1981. PhD, professor. His main research interests include machine learning, neural networks, spatial-temporal data analysis, graph learning, recommender systems, and social networks data mining.

周　帆, 1981 年生. 博士, 教授. 主要研究方向为机器学习、神经网络、时空数据分析、图学习、推荐系统、社交网络数据挖掘.