基于社会媒体数据增强的交通态势感知研究及进展

陈苑文[1,2]，王晓[2,3]，李灵犀[3,4]，王飞跃[2,3]

（1. 厦门大学航空航天学院自动化系，福建 厦门 361100；
2. 中国科学院自动化研究所复杂系统管理与控制国家重点实验室，北京 100190；
3. 青岛智能产业技术研究院平行智能创新中心，山东 青岛 266114；
4. 美国印第安纳大学–普渡大学印第安纳波利斯分校电子与计算机工程系，美国 印第安纳州 IN 46204）

摘　要：交通态势感知是智能交通系统的重要研究方向。已有研究大多关注如何使用物理传感器感知当下交通态势并预测未来交通状况。然而，物理传感器性能易因天气影响、电磁干扰、能源限制等问题出现不稳定或失效情况，导致其采集的数据稀疏或缺失，使其对交通态势感知滞后且不准确。社会媒体数据为及时感知完善的交通态势信息提供了新的增强方式。面向当下异常交通情况频发的城市交通管控现状，社会传感与物理传感数据互为补充，可进一步满足城市交通高效管理需求。基于此，对基于社会媒体数据的交通事件检测和交通状况预测工作展开分析研究，探讨社会媒体数据增强的交通态势感知研究工作如何为交通管理部门提供决策支持，以合理规划、引导交通，缓解交通拥堵，最后提出社会媒体数据增强的交通态势感知还需进一步探索的方向。

关键词：交通态势感知；智能交通系统；社会感知；交通事件检测；交通状况预测
中图分类号：TP399
文献标志码：A
doi: 10.11959/j.issn.2096−6652.202220

Traffic situational awareness research and development enhanced by social media data: the state of the art and prospects

CHEN Yuanwen[1,2], WANG Xiao[2,3], LI Lingxi[3,4], WANG Fei-Yue[2,3]

1. Department of Automation, School of Aerospace Engineering, Xiamen University, Xiamen 361100, China
2. The State Key Laboratory for Management and Control of Complex Systems, Chinese Academy of Sciences, Beijing 100190, China
3. Parallel Intelligence Innovation Center, Qingdao Academy of Intelligent Industries, Qingdao 266114, China
4. Department of Electrical and Computer Engineering, Indiana University-Purdue University, Indianapolis IN 46204, USA

Abstract: Traffic situational awareness is an important research direction of intelligent transportation systems. Most of the existing research focused on how to use physical sensors to perceive the current traffic situation and predict the future traffic state. However, the performance of physical sensors is prone to instability or failure due to adverse weather, electromagnetic interference, energy limitation and other problems, resulting in sparse or missing collected data, which makes the perception of traffic situation lagging and inaccurate. Social media data provides a new and enhanced way of perceiving comprehensive traffic situation information in a timely manner. Facing with the current traffic situation where sudden abnormal traffic events occur frequently, social sensing and physical sensing data can complement with each other to further improve the efficiency of urban traffic management. The related work of traffic event detection and traffic state prediction enhancing based on social media data were analyzed, and how those research works provide decision support for the traffic management departments to plan and guide traffic reasonably and alleviate traffic congestion were explored. Finally, some future research directions of traffic situational awareness enhanced by social media data were proposed.

Key words: traffic situational awareness, intelligent transportation system, social perception, traffic event detection, traffic state prediction

收稿日期：2022−02−22；**修回日期：**2022−03−01
通信作者：王晓，x.wang@ia.ac.cn
基金项目：国家自然科学基金资助项目（No.62173329）
Foundation Item: The National Natural Science Foundation of China (No.62173329)

0　引言

近年来，随着我国社会经济的进步与繁荣，城市交通得到极大的发展。与此同时，人们的出行变得愈加频繁，出行工具与方式愈加多样，国内车辆数量增加，我国城市特别是大城市交通拥堵的问题普遍存在，并呈现出愈加严重的趋势[1-5]。由此产生的经济损失巨大，环境污染严重[6]，在紧急情况下，甚至会造成严重的生命财产损失。从供需角度看，当交通供给小于交通需求时便会产生交通拥堵。若能及时地感知交通状况，依此引导出行者的出行计划、出门时间、路线规划，就可以更好地分配交通资源，平衡交通供给和需求，减少交通拥堵。

目前，道路上安装的交通感知装备以物理传感器为主，如成像传感器、感应回路、磁传感器、声探测器、被动红外等。这些物理传感器的部署为交通管理部门提供了丰富的交通数据。然而，物理传感器数据成本高、空间覆盖范围有限、性能表现受天气影响严重、故障频发、数据可靠性难以保障。因此，交通管理部门往往无法仅依靠物理传感器获取交通状况的全貌，更无法根据其数据分析交通拥堵背后的原因。与此同时，由特殊事件（如体育赛事、音乐会、节日等）引发的异常交通拥堵成为常态，而这类事件受到多方面因素的影响。从供需角度看，影响交通供给的因素包括天气状况和交通事件（如交通事故、道路施工等），影响交通需求的因素包括通勤需要和特殊事件出席需要，这些因素均无法从物理交通数据中得到反映。

Yin J 等人[7]提出可利用社会媒体数据增强对交通状况的感知，并将交通态势感知分解为“理解当前状况”和“预测未来状况”两个方面。理解当前状况主要指了解当前是否发生交通拥堵、拥堵规模、原因等；预测未来状况以理解当前状况为前提，涉及的应用任务主要有交通速度预测、交通流预测、交通拥堵预测、交通事件预测等。Wang F Y[8]认为，基于社会媒体数据的交通分析和预测是社会交通的重要研究内容，而社会信号及其传感技术则是实现社会交通工程的主要手段。早在 2010 年左右，国内学者就提出了“社会传感网”和“社会传感器”[9-12]的概念，并将每个人作为一个智能传感器，通过他们在社会和自然环境中的自主移动来感知、解读和集成信息，这种感知方式将不再是仅对局部世界物理特性的感知，还有可能实现对大范围人类社会的全面感知。这种将人类作为传感器（又被称为“human as sensor”[9-12]或“citizenas sensor”[13]）的观点，随着移动设备及互联网技术的发展，几乎零成本、覆盖范围广泛且内容丰富的社会媒体平台成为触手可及的数据来源，已在城市及社会应急管理、舆情安全、民意调研中发挥了巨大作用[14]。

社会媒体数据亦可作为检测交通事件的信息源。除了对交通事件的描述，社会媒体数据中通常还记录了交通事件的可能原因、亲身参与者对当下交通状况的情绪反馈[15]、对交通基础设施规划和交通管理的建议等[16]，对这些信息的挖掘有利于交通管理部门更好地规划交通。此外，人们在社会媒体上讨论各种话题[17]，从社会媒体中可以挖掘出物理传感器无法反映的通勤需求和特殊事件出席需求。因此，使用社会媒体平台上与交通相关的数据进行数据集成、语义分析和理解，可弥补物理传感器数据的不足，两者相互补充，形成对交通态势更完整的感知。

为了更好地融合社会媒体数据和其他来源的数据以增强交通态势感知，提升城市交通管理效果，众多前沿技术已被应用于交通事件检测及预测，如 Wang S Z 等人[18]使用矩阵协同分解来补全交通拥堵矩阵。该方法融合了路网数据、社会媒体数据、GPS 探测数据，从中挖掘道路信息、事件信息和天气情况，使用耦合矩阵和张量分解来补全稀疏的交通拥堵矩阵。Nallaperuma D 等人[19]提出一种基于无监督在线增量机器学习、深度学习和深度强化学习的平台来集成物联网、智能传感器、社会媒体等异构大数据，检测概念漂移（concept drift），区分周期性和非周期性交通事件，并进行影响传播、交通流预测、通勤者情绪分析和优化交通控制决策。Rashid M T 等人[20]基于社会媒体检测交通事件，使用鲁棒真相发现（robust truth discovery，RTD）[21]算法来估计事件的真实性和置信度，用车载传感器数据进一步验证置信度低的事件。这些研究证实了在交通态势感知领域纳入社会媒体数据的有效性。

本文围绕基于社会媒体数据增强交通态势感知和预测方面展开分析，并详细介绍其中涉及的关键技术。首先总结了基于社会媒体增强交通态势感知的一般框架，对其中涉及的关键技术进行分析与讨论。其次分别详述基于统计学习的方法和基于机器学习的方法如何利用社会媒体数据增强交通预测。在此基础上，进一步探讨了该工作如何支撑交

通管理部门利用这些研究开展更有效的交通管控策略制定与管理规划工作。最后，列举了一些未来的研究方向。

1　基于社会媒体数据的交通事件检测流程及关键技术

基于社会媒体检测交通事件的主要任务是提取社会媒体数据的语义信息，检查道路上是否出现交通拥堵以及交通拥堵的原因（如交通事故、道路施工等）。研究表明，人们倾向于在发生事故或道路封闭时，通过社会媒体发布相关信息[22]。Zhang S 等人[23]将官方交通事件记录和推特（Twitter）数据 tweets 结合起来探索与交通相关的 tweets 的空间位置与交通事件位置的依赖关系，发现交通事件主题 tweets 倾向于在交通事件位置周围聚集成簇。Sinnott R O 等人[24]通过对比官方交通数据和社会媒体数据，发现社会媒体数据确实可以作为收集更多交通数据的来源。Steiger E 等人[25]对比社会媒体数据和官方数据，发现官方数据报道的特殊事件、交通事件之间的时空模式和与交通相关的社会媒体数据密切相关。这些研究验证了基于社会媒体数据挖掘交通事件的潜力。

基于社会媒体检测交通事件是从社会媒体中深入挖掘交通信息及交通预测的基础。Shen D Y 等人[6]从上报交通拥堵信息的 tweets 中提取拥堵路段和时间，提出 TC_Apriori（traffic congestion apriori）算法来发现路段拥堵共现模式。Zhang Y 等人[26]从社会媒体中提取交通事故信息，结合从在线地图中收集的卫星服务数据，使用多视图（multi-view）学习评估地区的交通危险程度。参考文献[18,27-29]从社会媒体中提取交通事件来改进交通预测。

社会媒体文本与一般文本相比存在以下特殊性：①文本中包含大量非正式的、不规则的表达、缩写词，存在拼写和语法错误；②文本中包含俚语和讽刺，如当收到石油涨价的消息时，有人发布“卖车去吧！”以表达不满情绪；③文本内容比较简短，缺乏上下文语境。针对这些特殊性，不同的研究者提出了不同的方法，但基本遵循以下框架：数据采集，数据预处理，数据表示，数据过滤，关联分析，事件位置提取，效果评估，事件描述及可视化。下面分别介绍其中使用的关键技术。

1.1　数据采集

从社会媒体上采集数据的方法主要有两种：①使用平台提供的应用程序接口（application programming interface，API）；②部署网络爬虫实时抓取网页。这两种方法需要结合研究需要及服务商提供的可用性和限制来灵活选择。以 Twitter 为例，它提供 REST API 和 stream API 供用户免费获取公开 tweets。REST API 使用户可以根据关键字、用户 ID、时间和地理边界框（由质心和半径指定）来获取 tweets。stream API 和 REST API 类似，但它不支持位置和关键词的联合查询。API 有一定的限制，如对于每个用户，每 15 min 只能查询 350 次。通过网络爬虫收集数据没有这样的限制，但后期需要复杂的预处理以抽取元数据。

社会媒体中包含各种各样的话题，与交通相关的数据占比不高，在获取数据时可以通过关键词过滤大部分与交通不相关的数据。关键词的设置需要尽可能地保守，以免漏掉一些重要数据。郑治豪等人[30]人为选择一些与交通相关的关键词，如“堵”“车祸”“剐蹭”“事故”等。Zhang Z H 等人[31]从报告交通事故的新闻中选择出现频率较高的词，将其作为搜索关键词。Fu K Q 等人[32]在收集数据的同时更新关键词。首先研究有影响力的、与交通相关的用户发布的 tweets，将这些 tweets 中的文档逆文档频率（term frequency-inverse document frequency，TF-IDF）值较高的 50 个词作为初始关键词。根据这些词搜索 tweets，将搜索到的 tweets 与之前的 tweets 合并，基于 TF-IDF 值重新选择关键词，由此不断迭代直至收敛。与此类似，Gu Y M 等人[33]提出一种自适应获取 tweets 的方法，分别计数查询结果中包含该词的、与交通相关的 tweets 数 N^+ 和与交通无关的 tweets 数 N^-，将 N^+较大的词作为与交通“正相关”的词，将 N^-较大的词作为与交通“负相关”的词，在查询中删去与交通“负相关”的词。基于单一关键词（如“车辆”）查询，会使得结果有较大噪声，加入词语组合能有效提高查询效率[32-33]。

社会媒体中包含大量社交机器人或垃圾邮件发送者，需要识别并过滤这样的用户。Yao W R 等人[29]将 tweets 位置范围小于 10 m 的用户看作可疑机器人，使用随机森林机器人分类器 Botometer API[34]选择出社交机器人。获取到的数据通常包含发布内容、用户信息、发布时间、位置标记（若用户开启了基于位置的服务），这些数据被转换为结构化的数据存储在数据库中供分析使用。

1.2 数据预处理

社会媒体文本包含大量非正式、不规则的表达，存在拼写和语法错误，在分析前需要一定的预处理，如大小写转换（对于英文）、纠正书写错误、俚语替换、噪声过滤、分词（对于中文）、停止词过滤、词干化等。原始文本中包含标点符号、标签、URL（uniform resource locator）链接等对信息提取无意义的符号，一般使用正则表达式过滤这些噪声符号。中文需要分词，目前已有多个开源的中文分词工具，如计算技术研究所汉语词汇分析系统（institute of computing technology Chinese lexical analysis system，ICTCLAS）、语言技术平台（language technology platform，LTP）[35]等。在英文中，大小写、同一词根派生出来的词（如“congested”和“congestion”）没有语义区别，因此需要对单词进行大小写转换和词干提取。波特（Porter）词干算法[36]是一种常用的词干提取算法[31,37]。

1.3 数据表示

采集到的数据通常包含发布内容、用户信息、发布时间、位置标记等。对于用户发布的内容，为了使计算机能够计算分析，需要将其表示为数值形式。最简单的表示方法是独热（one-hot）编码，每个词被表示为一个独热向量，特征是语料库中的单词，向量在该词的位置为 1，在其他位置为 0。独热编码具有高维稀疏性且无法表示词语间的语义相似性，词嵌入模型将每个词语表示为一个连续空间的向量，考虑了词语间的语义相似性。有两种常见的词嵌入模型[38]：CBOW（continuous bag of-word）和 Skip-Gram，两者互为镜像。上述方法使用向量来表示词语，文档被表示为矩阵。另一类方法将文档表示为向量：首先根据语料库构建词典，每个词对应一个特征，计算该词的某种重要性度量，并将其作为该特征的取值。常用的重要性度量为该词的 TF-IDF 值。上述方法纯粹基于单个词语，没有考虑词法和语序。*N*-gram 模型将连续的 *N* 个词语作为特征，以此纳入语序。除此之外，Zhang Z H 等人[31]指出只将单个词语作为特征是不够的，因为这只强调了标签与词语之间的相关性，而忽略了词语内部的相关性。例如，一个文本中，在“事故”出现的条件下，“车”的出现会增大与事故相关的概率，而在“维修”出现的条件下，“车”的出现可能会降低事故相关概率。因此一些研究[33,39]将词语组合作为特征，与 *N*-gram 模型不同，这里的词语组合不是相邻的，而是通过 Apriori 算法[39]或基于 MapReduce 框架抽取的[33]。

基于特征的表示方法没有考虑文本的上下文关系。主题建模方法（如隐性语义分析（latent semantic analysis，LSA）和隐狄利克雷分布（latent Dirichlet allocation，LDA））将文档映射到主题空间，降低了相似性度量的噪声，在一定程度上弥补了上述缺陷。

除了抽取文本的语义特征外，Khan S M 等人[40]还抽取了文本的情感分数、特定单词的出现分数、一些句法特征（如标签、问号、感叹号、大写字母的数量、文本长度）为特征。包含道路名的文本与交通相关的可能性更大，Gu Y M 等人[33]使用命名实体识别（named entity recognition，NER）技术识别道路名，将其标记为“道路名”，并作为特征输入分类器中。

很多方法将文本表示为高维向量，选择合理的特征或对向量降维可以提高分类效果和算法效率。D'Andrea E 等人[37]采用计算信息增益的方法，仅保留了信息增益大于 0 的特征。Zhang Z H 等人[31]基于 phi 系数[41]选择特征。Khan S M 等人[40]使用 LASSO（least absolute shrinkage and selection operator）特征选择方法选择特征，且对 TF-IDF 向量使用奇异值分解（singular value decomposition，SVD）来降低维度。

1.4 数据过滤

尽管在采集数据时已根据特定关键词搜索，但获得的数据仍十分嘈杂，还需要更精细地过滤。大部分研究采用基于监督学习的方法[33, 37, 40, 42-44]：首先人工标记大量数据，并将其作为训练真值，然后训练分类器。常见的分类算法有支持向量机（support vector machine，SVM）、朴素贝叶斯（naive Bayes，NB）、决策树（decision tree，DT）、*K* 最近邻（*K*-nearest neighbor，KNN）、神经网络（neural network，NN）等。Chen Y Y 等人[44]提出的长短时记忆−卷积神经网络（long short term memory-convolutional neural network，LSTM-CNN）将社会媒体文本分为与交通相关和与交通无关两类。LSTM 有潜力学习微博的上下文相关性，而 CNN 可以提取深层特征，将二者结合起来能得到比单独的 CNN、LSTM 效果更好的模型。

监督学习方法需要大量人工标记的数据，费时费力。Wang D 等人[45]提出了一种基于 LDA 的半监

督学习方法 tweet-LDA，以减少标记数据的需求量。Zhang S 等人[23]使用无监督学习方法选择与交通相关的文本，首先利用 LDA 将文本投影到主题空间中，然后在主题空间中使用层次聚类算法得到聚类簇，从中选出与交通相关的文本。

由于社会媒体数据容量大、变化快，计算时间是一个需要考虑的问题。Khan S M 等人[40]使用克莱姆森大学棕榈超级计算集群（Clemson University Palmetto Supercomputing Cluster）来支持并行计算，以最小化训练阶段的计算时间。

1.5　关联分析

为了进一步分析是否出现交通拥堵及挖掘交通拥堵背后的原因，还需要进一步分析选出的与交通相关的数据。郑治豪等人[30]通过关键词匹配法将文本分为路况正常、施工、封路、路况拥堵、车辆相撞、其他。与此类似，Gutiérrez C 等人[42]将文本分为交通拥堵、道路施工、货运、道路封闭、冰、雪、其他。关键词匹配法允许一个文本有多个类标签，但没有综合考虑整个文本的语义信息。Cui J 等人[43]采用基于 *N*-gram 的贝叶斯分类器将文本分为交通拥堵、交通事故、交通管制。Jain A K 等人[46]对比 NB、SVM、随机森林后，选择随机森林将文本分为拥堵、车辆故障、交通顺畅、阻塞、清除阻塞、其他。Gu Y M 等人[33]使用 sLDA（supervised latent Dirichlet allocation）将文本分为交通事故、道路施工、不利天气、特殊事件、车辆故障。D'Andrea E 等人[37]使用 SVM、NB、C4.5 决策树算法、KNN、PART 决策树算法[47]将文本分为外部事件交通流、交通拥堵、非交通，其中 SVM 表现最优。Zhang Z H 等人[31]使用深度信念网络（deep belief network，DBN）和 LSTM 模型将文本分为车辆碰撞、失灵、着火，对比 DBN、LSTM、ANN（artificial neural network）、SVM、sLDA，DBN 效果最好。Khan S M 等人[40]结合 L-LDA（labeled-LDA）和 SVM 将文本分为事件、拥堵、施工、特殊事件、其他事件。

多来源数据相互补充可以提高事件检测效果，Lu H 等人[48]融合微博和新闻来检测交通事件，使用无监督学习技术 LDA、w-LDA（weibo-LDA）分别从新闻、微博中检测话题词。由于新闻表达规范，而微博表达随意，因此将微博词汇和新闻词汇“对齐”以形成对事件的整体描述，结果表明将新闻和微博数据结合能有效提高事件检测的准确率。Alkouz B 等人[49]融合了 Twitter 和 Instagram 检测和预测道路交通拥堵，提出的方法能够分析多种语言，包括阿拉伯语、阿联酋当地语言和英语。

1.6　事件位置提取

位置是交通事件的重要特征之一，社会媒体的元数据中包含以经纬度表示的位置标记，然而具有位置标记的文本占比非常小且不一定是交通事件的位置，因此大部分情况下需要从文本本身提取位置信息。其主要环节包括文本实体识别[18,30,38]、位置对齐和消歧[21,28,30-31,34,40]。Wang S Z 等人[18]人工为一些文本进行命名实体标注，将其作为训练值输入条件随机场（conditional random field，CRF）模型中以识别命名实体；Gutiérrez C 等人[42]测试了 Alchemy、OpenCalais、Standford NER、NERD 4 种命名实体识别方法，NERD 效果最好；Tejaswin P 等人[50]使用正则表达式解析器生成候选实体。对于文本中的实体还需要结合背景知识来判断其是否属于一个位置。Khan S M 等人[40]将提取的位置名与街道名称字典进行匹配，使用基于编辑距离（Levenshtein distance）的相似度比来确定位置实体。Cui J 等人[43]采用线性参考方法（linear referencing method，LRM）[51]进行文本定位。Jain A K 等人[46]爬取多个网站获取研究城市的所有可能的地点名称列表，对于每一个文本，扫描列表以查看是否有匹配位置。Gu Y M 等人[33]使用两个地理解码器分别处理交叉路口名、高速公路名等地理信息和兴趣点名称，使用模糊匹配算法[52]解析与位置相关的词。Gutiérrez C 等人[42]使用 Geocoding、GeoNames 和 namation 这 3 个地理定位引擎来消除地理歧义，如果至少两个地理定位引擎的结果是一致的，那么使用这个结果，否则使用权威较高的 Geocoding 的结果。

从文本提取事件位置的一个问题在于有些文本只提及事件，缺少事件发生的位置。为此，Cui J 等人[43]设计了一个问答系统，当检测到不完整信息时主动询问用户，以完善事件信息。另外，对于单一文本信息不完整问题，可以通过分析多个描述同一事件但由不同用户发布的文本有效解决[42]。

一般而言，很难从社会媒体中直接获得准确的事件位置。Zheng Z H 等人[16]提出不需要从社会媒体中获取准确的事件位置，而使用出租车 GPS 数据来确定交通异常的时间和位置。该框架中社会媒体仅仅作为潜在关键交通事件的初步过滤以及交通事件的原因分析。首先从社会媒体文本上检测事件的大概位置，根据这个位置生成搜索区域，在此区

域内根据 GPS 数据检测异常路径，基于此异常路径定位事件位置。

1.7 效果评估

评估分类算法效果的指标有很多，如准确率、召回率等。但这些只能评估算法在网络空间中检测事件的效果，在物理空间中的效果仍需验证。Lu H 等人[48]咨询交通领域的专家以确定与测试数据集对应的交通事件。参考文献[31,33]将检测到的事件与交通事故日志对比来查看检测效果。然而因为官方事故日志本身并没有记录所有交通事故，那些在网络空间中检测到但不在官方事故日志中的交通事件不一定是系统误报的结果。

交通事件会引起交通异常，因此将社会媒体数据报告的交通事件与物理传感器的交通数据（如速度、流量、旅行时间等）对比[33]可以验证该交通事件是否真实发生。如果社会媒体报道的事件是真实的，那么事件位置附近的交通数据将会出现异常，如交通速度应该比正常速度慢，旅行时间应该比正常长。假设检验常被用来验证社会媒体中的交通事件是否真实发生，首先假设实际交通量与正常的交通量相同（即没有发生交通事件），然后基于该假设构建相应的假设检验统计量，当其落在拒绝域中时，认为交通事件的确发生了。假设检验法趋于保守，只有当证据充分时才认为交通事件真实发生。

1.8 事件描述及可视化

对异常事件进行准确描述及可视化有助于帮助公众及交通管理部门及时关注检测到的异常。事件描述及可视化的主要方式是文本报告及图形图像。Fu K Q 等人[32]使用 LexRank 方法[53]从大量与交通相关的文本中选出重要的文本供用户阅读。Pan B 等人[54]从社会媒体中挖掘异常发生时频繁出现但在其他情况很少出现的代表性术语来描述异常。Lu H 等人[48]将所有交通事件映射到城市道路上。郑治豪等人[30]分别开发了电脑和安卓端的可视化模块，将事件用不同颜色在地图上标注出来。Cui J 等人[43]开发了一个基于安卓的应用程序，在百度地图二次开发的基础上显示特定位置的交通事件。Jain A K 等人[46]开发了一个 Web 应用程序，显示地图上的拥堵数据趋势分析。

1.9 小结

基于社会媒体数据的交通事件检测，本质上是利用众包模式调动每一位交通参与者提供的数据，让每一位城市公民都能够参与到城市交通的检测、管理与评估中，进一步提升城市交通管理的质量和效率。近年来，基于众包的交通数据采集、交通管理、交通衍生服务等[43,55-56]已经渗透到人们生活的方方面面，极大地提升了人民生活的满意度。

2 基于社会媒体数据的交通预测增强方法

交通预测是智能交通系统的重要组成部分，准确的交通预测可以辅助路线规划，指导车辆调度，缓解交通拥堵[57]。目前已有很多先进的交通预测方法，包括使用概率模型来建模交通数据[58-59]，使用时间序列模型来预测交通状况[60]，使用扩展卡尔曼滤波器来预测车速[61]，使用神经网络和深度学习来预测交通[62-72]。这些方法在常规情况下效果良好，但当交通事件（如交通事故）、特殊事件（如运动会、音乐会、节日、集会）、恶劣天气等突发情况发生时，由于使用的数据源中没有反映这些事件，预测往往失效。社会媒体中包含丰富的信息，社会媒体是检测交通事件的一大数据来源。公众倾向于在社会媒体上发表对特殊事件的出席意愿，在社会媒体上查看并讨论天气。因此从社会媒体中挖掘影响交通供需的多方面因素，并将其作为物理交通数据的补充可以有效改善交通预测。

交通预测涉及多个应用任务，包括交通速度预测、交通流量预测、交通拥堵预测等。路网中不同区域之间的交通量存在复杂的空间依赖性、动态的时间依赖性[57]，在预测时需要充分考虑交通量间的时空交互。因此，不同于基于社会媒体数据的交通事件检测，基于社会媒体的交通预测增强难以抽取一般框架。本文主要关注基于统计学习的方法和基于机器学习的方法。基于统计学习的方法首先对数据提出一些合理的假设，建立数据的分布模型，使用参数估计方法（如极大似然估计（maximum likelihood estimate，MLE）、最大后验概率（max a posterior，MAP））估计模型参数，然后推断预测。该方法的优点在于可以对不同来源的数据提出不同的概率模型，能方便地融合多来源数据，但效果非常依赖于假设与真实世界的吻合度，并且由于模型中涉及大量随机变量，往往存在计算困难问题。基于机器学习的方法需要考虑影响交通的多方面因素，抽取复杂的内外特征，将这些特征转换为有意义的数值表示，并进行时空聚合。

2.1　基于统计学习的方法

基于统计学习的方法将预测量及一些已观测量看作随机变量，建立其概率分布模型，通过已观测数据估计模型参数，计算变量的后验分布来预测。该方法可以方便地融合多个来源的数据，如参考文献[27,73-74]集成了社会媒体数据、GPS 数据和交通速度数据来改进交通预测；也可以方便地纳入多种影响因素，如 Zhou T 等人[75]从社会媒体中提取交通状况及天气、地方事件（如体育赛事）、特殊日子（风暴日、节日等）等因素，使用层次贝叶斯网络来建模这些变量间的交互，Chen P T 等人[28]将马尔可夫随机场（Markov random field，MRF）纳入先验规则。在基于统计学习的方法中，各链路的交通速度通常被假设为服从联合高斯分布，分布的均值描述了该路段正常行驶的速度水平，协方差矩阵描述了交通速度间的空间相关性。使用高斯分布具有一大优势：对于无法观测速度的路段，其速度的后验均值可由先验均值和观测值与先验均值之间的偏差计算，这可以有效解决数据稀疏性问题。

社会媒体数据、GPS 数据和交通速度数据三者相互印证、相互补充，为交通状况提供更加准确的推断。当交通速度缓慢时，社会媒体数据应该反映道路拥堵，GPS 数据应该反映旅行时间变长，3 种数据的生成具有一定的依赖关系。Lin L 等人[73]认为交通速度决定旅行时间和交通拥堵状况，而交通拥堵状况决定 tweets 主题。因此对于社会媒体数据，根据交通速度选择道路拥堵状况，根据道路拥堵状况选择 tweets 主题，根据主题生成词语。Liu S Y 等人[74]通过一个函数将链路、tweets 内容和 tweets 主题映射到交通速度，并假设该函数服从高斯分布，提出如下主题模型来计算 tweets 主题：从依赖于用户的分布、依赖于空间的分布和依赖于时间的分布中选择一个作为 tweets 的主题分布，基于该主题从背景词分布和主题特定的词分布中选择一个作为词分布。

概率图模型使用图来表示变量间的依赖关系，可以方便地表示多种变量间的联合概率分布并纳入先验规则。Wang S Z 等人[27]从社会媒体中提取事件信息，扩展了耦合隐马尔可夫模型（coupled hidden Markov model，CHMM）来建模交通状况、交通事件和交通速度间的关系，认为交通状况是隐状态，交通事件和交通速度都以该隐状态为条件。在链路速度条件独立性假设、交通事件条件独立性假设、交通状态条件独立性假设的基础上建立交通概率模型。Chen P T 等人[28]使用马尔可夫随机场来建模交通状况间的复杂交互，定义了 11 个概率软逻辑（probabilistic soft logic，PSL）规则以刻画交通事件和交通状况之间、交通状况之间、社会媒体文本和交通状况之间的相互关系，通过这些规则生成 MRF 的特征函数。Zhou T 等人[75]使用层次贝叶斯网络结合线上线下信息来预测交通状况。层次贝叶斯网络将随机变量分层表示，用图中的节点表示交通状况、天气、地方事件、特殊日子等随机变量，而节点间的边表示变量间的因果依赖关系，使用贝叶斯分数[76]来衡量提出的网络与实际数据的吻合度。

以上方法建立了变量间的概率分布模型，接下来需要根据观测数据估计模型参数，然后计算变量的后验概率进行预测。常用的估计模型参数的方法有 MLE 和 MAP。然而由于模型中涉及大量变量间的交互，直接计算往往存在困难。Lin L 等人[73]使用 MLE 来估计模型参数，使用变分推断[77]来计算后验概率和对数似然函数下界，并使用期望最大化（expectation-maximization，EM ）算法迭代更新模型参数。Chen P T 等人[28]通过 MAP 来推断模型参数，用期望最大化（maximization expectation，ME）算法和交替方向乘子法（alternating direction method of multipliers，ADMM）解决复杂问题。Liu S Y 等人[74]在估计交通速度的均值和协方差矩阵时，从历史数据中进行子抽样并舍去非常旧的数据以解决随着观测数据增多计算量急剧增加的问题。Wang S Z 等人[27]提出一种并行粒子滤波方法来减少计算时间。寻找最优的贝叶斯网络是一个 NP 难问题[78]，Zhou T 等人[75]提出了一种启发式算法，利用网络的子结构特性和层次特性，将网络划分为更小的网络，对每个小网络进行贪婪搜索，该算法可以在多项式时间内找到可接受的结构，并且使用信念传播算法[79]进行预测。

2.2　基于机器学习的方法

不同于基于统计学习的方法，基于机器学习的方法不需要事先假设数据的分布，而是直接抽取可能影响交通的量，并将其作为特征构建预测模型。基于机器学习的方法将交通预测归结为时间序列预测。一个时间序列可分为 3 个部分[80]：季节（season）、趋势（trend）、噪声（noise）。从历史数据中可以挖掘季节和趋势，噪声需要从其他外界因

素（如交通事件、恶劣天气、特殊事件）中提取。

社会媒体的一大优势在于其能反映特殊事件下的交通需求。研究表明，事件出席率和与事件相关的社会媒体量存在一定程度的正相关[80]，Twitter 背后反映的公共活动可能在特定的时间地点造成流量激增[81]。因此与特殊事件相关的社会媒体量和情绪特征表达了特殊事件下的交通需求。Ni M 等人[82]基于话题标签（hashtag）从社会媒体中检测特殊事件，将与事件相关的 tweets 数量和不同用户数量作为事件特征。Ni M 等人[83]从社会媒体中抽取 5 个数量特征和 5 个情绪特征作为事件特征。其中，5 个数量特征为与比赛相关的 tweets 数量、发送这些 tweets 的独立用户数、这些 tweets 中的话题标签数、提到其他用户的 tweets 数、有 URL 的 tweets 数。5 个情绪特征由 Lydia/TextMap 系统[84-85]抽取。Yao W R 等人[29]认为 tweets 的异常发布频率和 tweets 情绪可以用来衡量特殊事件是否发生，将 tweets 数量以及中性情绪 tweets 在所有情绪中的比例作为事件特征。Xue G 等人[80]计算每个时间窗内与事件相关的帖子数量，将其聚合为时间序列数据，使用 SAE（stacked autoencoder）进行特征压缩，用深度神经网络（deep neural network，DNN）训练得到衡量特殊事件扰动的输出 y。

除了交通需求，社会媒体中的其他信息也被用来改进交通预测。Essien A 等人[86]将社会媒体数据和天气数据结合，从两个道路交通用户中搜集 tweets，将 tweets 的时间戳和天气数据合并，采用重叠滑动窗口方法[87]将输入的多元时间序列数据集重构为监督学习格式。Liu X Y 等人[88]在预测模型中加入社会媒体的语义特征。社会媒体数据在细粒度上具有稀疏性，即在较小的时间窗口、较小的区域内只生成少量的社会媒体内容，因此他们将每个时间窗口 t 和区域 g 内的 tweets 数据聚集在一起，删除停止词后，将它们投射到词根空间，将得到的向量作为社会媒体的语义向量加入预测模型中。

在交通拥堵预测方面，Wongcharoen S 等人[89]认为 tweets 数量表明特定区域存在大量人群，他们使用 tweets 密度来预测交通拥堵的严重程度。他们收集了带有地理标记的 tweets、选定的交通账号的 tweets、包含交通关键词的 tweets，将所有 tweets 匹配到附近的道路并按每 30 min 为一组计算其数量，给 3 种 tweets 分配不同的权重，加权求和得到 tweets 密度。Yao W R 等人[29]在预测模型中纳入需求和供给两方的因素来预测第二天早晨的交通拥堵状况。需求包含常规的通勤需要和非常规的特殊事件需要，影响供给的因素包括交通事件和天气。一个人睡得越早，越有可能需要早起通勤，可以从 tweets 中提取睡眠–清醒模式来表达这种交通需求。收集的 tweets 均在晚上 9 点到凌晨 5 点发布，选择在晚上 9 点到凌晨 3 点之间发布的最后一条时间线 tweets，以及凌晨 3 点到 5 点之间发布的第一条时间线 tweets，按时空聚合计算其数量并归一化得到两个特征向量，将此作为早起通勤的出行需求表示。使用 tweets 数量以及中性情绪 tweets 在所有情绪中的比例表示特殊事件下的交通需求。交通事件从与交通事件相关的 tweets 和交通管理部门数据中解析，考虑 3 个方面的事件影响：①车道关闭类型；②事件位置；③事件时间窗。将交通事件表示为两个向量：部分闭合影响向量和完全闭合影响向量。

影响交通的因素非常多，可能抽取出来的特征非常高维，然而不是所有特征对预测都有帮助，且维数过高将降低计算效率，因此不少研究者使用特征选择技术或降维技术来降低特征维数。Ni M 等人[83]使用了带有 L1 范式正则项的最小平方优化进行特征选择，选择系数项非 0 的特征。Cui Y 等人[90]使用弹性网（elastic net），其结合了 LASSO 和岭回归（ridge regression）进行特征选择。Roy K C 等人[91]使用排序重要性来计算特征的重要性[92]。Essien A 等人[86]使用自动编码器作为降维组件，使输入向量降到更小的维空间[93]。Xue G 等人[80]将抽取的特征按时间序列排列为矩阵，将矩阵转为图像，利用基于注意力模块的 CNN 对其进行训练，将矩阵压缩为一个输出值。

基于选择出的特征，需要构建模型来实现预测。常用的模型有线性回归、支持向量回归、K 最近邻回归、xGBoost 回归、随机森林回归等。近年来，更复杂的模型被提出以学习数据间的复杂交互。Essien A 等人[86]提出深层双向 LSTM 模型。双向 LSTM 结构由两个 LSTM 反向叠加组成，而深层双向 LSTM 将多层双向 LSTM 垂直叠加。Ni M 等人[82]使用季节性差分自回归移动平均（seasonal autoregressive integrated moving average，SARIMA）模型和线性回归模型分别从历史交通数据和社会媒体数据中学习，提出一种基于参数和凸优化的方法 OPL（optimization and prediction with hybrid loss function）将两者融合起来。用堆叠（stacking）策

略将 SVR 和 OPL 结合起来，得到比单独的 SVR 或 OPL 更优的结果。Xue G 等人[80]认为预测变量含有规则模式和不规则模式，使用深度神经网络从外界特征中学习不规则模式，将预测变量真值减去不规则模式后输入深度神经网络中学习规则模式。将规则模式和不规则模式相加作为预测值。Yao W R 等人[29]开发一个聚类学习管道（clustered learning pipeline）来预测交通拥堵。将道路根据交叉路口分段，认为各路段的拥堵状况受到所属道路拥堵状况的影响，因此首先预测道路级的交通拥堵模式，基于此模式预测各路段的拥堵状况。道路拥堵模式用拥堵聚类指数来衡量：首先生成旅行时间指数（travel time index，TTI），在第 d 天，将道路上所有路段从 5 点到 11 点的 TTI 连接，得到该道路的日交通拥堵情况。使用主成分分析（principal component analysis，PCA）降维后，用 K-means 聚类分析识别各道路的典型日拥堵模式。

交通数据存在复杂的时空交互，在预测模型中往往需要加入当前位置的历史数据、周边位置的历史数据。然而对于那些没有安装物理传感器的区域，这些数据无法获取。Liu X Y 等人[88]提出集体交通预测（collective traffic prediction，CTP）算法，首先利用已知数据训练基学习器得到局部模型，将未观测数据初始化为 0，利用局部模型对未来的所有区域进行初始预测，根据预测更新未观测数据，再次应用局部模型更新预测，如此迭代直到收敛。

3　结束语

交通态势感知是智能交通领域的一个重要研究方向，包括了解当前的交通状况和预测未来的交通状况。由于传统物理数据存在稀疏性，缺乏对交通信息的理解和分析，无法反映某些突发情况，因此不少研究将社会媒体纳入感知交通态势的信息来源。本文针对社会媒体数据增强的交通态势感知展开分析，并着重对基于社会媒体的交通事件检测以及与基于社会媒体增强交通预测相关的研究进行分析详述，未来进一步深化社会传感数据与物理传感数据的融合，为城市交通管理部门做出更加符合社情民意的决策提供有效数据支持，并产生新的衍生服务。

据笔者分析，基于社会媒体数据的交通态势感知增强将至少从以下 4 个方面提升城市交通管理的满意度。

（1）有效提升交通态势感知和交通预测的准确度，为交通管理部门及时地采取合理的交通管控措施提供支持。目前已有多项研究关注智能交通管理系统，如 Hashemi H 等人[94]开发了一种模拟实时操作的交通管理系统，集成了交通状况估计、预测及决策支持功能；Abdelghany K 等人[95]提出了主动−鲁棒交通网络管理决策支持系统；Hashemi H 等人[96]提出了一种使用端到端深度学习方法的新型实时交通网络管理系统。这些交通管理系统为交通管理部门提供特定场景下的最优交通管理措施。然而，交通决策需与实际的交通状况相适应才能发挥效果，因此这些交通管理系统的性能十分依赖于交通监测和预测的准确性。社会媒体的加入提高了交通监测和预测的准确率，使得这些系统生成与实际交通状况相适应的交通管理方案和应急预案。

（2）从多个维度观察并增强交通管理服务能力。社会媒体提供了交通事件的规模和原因描述，据此交通管理部门可以采取合理的措施来降低交通事件的影响，如动态调控交通信号灯、根据需要改变车道的车流方向、封锁道路禁止车辆进入等。一个缓解交通拥堵的自然想法是扩建道路，然而受地理位置的限制，这在有些情况下是不可行的。况且，交通需求日益膨胀，不可能无限制地扩建道路。由特殊事件导致的交通拥堵属于非经常性拥堵，可以通过提前规划避免。社会媒体中反映了公众对特殊事件的出席需求，据此，交通管理部门可以提前采取措施规避交通拥堵，如调度公共交通、改变速度限制等。导航软件也可以在计算路线时考虑这些需求，从而更有效地引导公众出行。

（3）实时发布交通现状，让公众能够及时准确地了解当前交通状况。社会媒体中包含了交通事件的时间、地点、规模、原因等详细描述，挖掘这些信息并通过道路上的警告屏幕、广播、社会媒体平台实时播报，有利于公众合理选择出行路线、规划出行时间。

（4）及时感知社会媒体上发布的相关信息，增强管理部门与交通参与者的互动，提升民众满意度。除了交通事件描述和交通需求反应，社会媒体中还包含用户对交通的需求、意见等丰富的信息，这为交通管理部门制定交通政策提供了支持，如根据用户需求创建新的交通服务、根据用户意见改善交通服务、更好地规划道路扩建、路标和限速设置等。

然而，社会媒体数据本身存在着一些问题：

①数据量大；②类型具有多样性和异质性；③可靠性存在问题[97-99]。因此还有很多方向值得探索，举例如下。

（1）探索更加有效的短文本语言分析模型。社会媒体数据一般都比较简短，缺乏上下文，传统的文本挖掘技术在分析社会媒体文本时效果并不好。

（2）联合分析文本、视频、图片，提升事件检测精度。除了文本，社会媒体中还包含大量图片、视频，这些数据中也包含丰富的信息，但其价值尚待挖掘。

（3）融合多来源数据提升交通态势感知。引入多渠道社会信号（如 Instagram、脸书（Facebook）、Quora、新闻等）或将社会数据与物理数据融合，可以解决单一数据造成的数据稀疏和不可靠问题。

（4）纳入半监督学习或无监督学习、增量学习和深度学习。这能减少对有标记数据的需求量，适应实时变化的数据，提高预测精度。

（5）知识图融合。交通知识图的构建、学习和深度知识搜索有助于挖掘更深层次的交通语义信息[57]，以进行交通现象原因分析和推理。

未来，笔者将进一步围绕如何高效地利用社会媒体数据增强交通态势的感知和预测逐步展开研究探索，将社会媒体数据与知识图谱融合起来，结合深度学习、半监督学习等技术，预期从社会媒体中挖掘更多、更准确的交通信息，并与其他来源的数据融合以进行交通状况推理和预测。

参考文献：

[1] 高德地图. 2021 年度中国主要城市交通分析报告[R]. 2021.
Amap. Traffic analysis report of China’s major cities in 2021[R]. 2021.

[2] 高德地图. 2020 年度中国主要城市交通分析报告[R]. 2020.
Amap. Traffic analysis report of China’s major cities in 2020[R]. 2020.

[3] 高德地图. 2019 年度中国主要城市交通分析报告[R]. 2019.
Amap. Traffic analysis report of China’s major cities in 2019[R]. 2019.

[4] 高德地图. 2018 年度中国主要城市交通分析报告[R]. 2018.
Amap. Traffic analysis report of China’s major cities in 2018[R]. 2018.

[5] 高德地图. 2017 年度中国主要城市交通分析报告[R]. 2017.
Amap. Traffic Analysis report of China’s Major Cities in 2017[R]. 2017.

[6] SHEN D Y, ZHANG L F, CAO J P, et al. Forecasting citywide traffic congestion based on social media[J]. Wireless Personal Communications, 2018, 103(1): 1037-1057.

[7] YIN J, LAMPERT A, CAMERON M, et al. Using social media to enhance emergency situation awareness[J]. IEEE Intelligent Systems, 2012, 27(6): 52-59.

[8] WANG F Y. Scanning the issue[J]. IEEE Transactions on Intelligent Transportation Systems, 2014, 15(3): 909-914.

[9] WANG D, AMIN M T, LI S, et al. Using humans as sensors: an estimation-theoretic perspective[C]//Proceedings of the 13th International Symposium on Information Processing in Sensor Networks. Piscataway: IEEE Press, 2014: 35-46.

[10] WANG D, KAPLAN L, LE H, et al. On truth discovery in social sensing: a maximum likelihood estimation approach[C]//Proceedings of 2012 ACM/IEEE 11th International Conference on Information Processing in Sensor Networks. Piscataway: IEEE Press, 2012: 233-244.

[11] 王飞跃. 万维社会媒体在防灾应急中的作用[J]. 科技导报, 2008, 26(10): 30-31.
WANG F Y. Web social media in disaster reduction and emergence management[J]. Science & Technology Review, 2008, 26(10): 30-31.

[12] 王晖, 姜志宏, 李沛, 等. 基于 Web 社会媒体的社会传感器网络[C]//第二届全国社会计算会议. 北京: 中国自动化学会, 2010: 1-5.
WANG H , JIANG Z H, LI P, et al. Social sensor network based on Web social media[C]//Proceedings of the 2nd Chinese Conference on Social Computing. Beijing: Chinese Association of Automation, 2010: 1-5.

[13] SHETH A. Citizen sensing, social signals, and enriching human experience[J]. IEEE Internet Computing, 2009, 13(4): 87-92.

[14] 王飞跃. 社会信号处理与分析的基本框架: 从社会传感网络到计算辩证解析方法[J]. 中国科学: 信息科学, 2013, 43(12): 1598-1611.
WANG F Y. A framework for social signal processing and analysis: from social sensing networks to computational dialectical analytics[J]. Scientia Sinica (Informationis), 2013, 43(12): 1598-1611.

[15] WANG X, ZENG K, ZHAO X L, et al. Using Web data to enhance traffic situation awareness[C]//Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems. Piscataway: IEEE Press, 2014: 195-199.

[16] ZHENG Z H, WANG C C, WANG P, et al. Framework for fusing traffic information from social and physical transportation data[J]. PLoS One, 2018, 13(8): e0201531.

[17] ZENG K, LIU W L, WANG X, et al. Traffic congestion and social media in China[J]. IEEE Intelligent Systems, 2013, 28(1): 72-77.

[18] WANG S Z, ZHANG X M, CAO J P, et al. Computing urban traffic congestions by incorporating sparse GPS probe data and social media data[J]. ACM Transactions on Information Systems, 2017, 35(4): 1-30.

[19] NALLAPERUMA D, NAWARATNE R, BANDARAGODA T, et al. Online incremental machine learning platform for big data-driven smart traffic management[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(12): 4679-4690.

[20] RASHID M T, ZHANG D, WANG D. DASC: Towards a road damage-aware social-media-driven car sensing framework for disaster response applications[J]. Pervasive and Mobile Computing, 2020, 67: 101207.

[21] ZHANG D Y, HAN R G, WANG D, et al. On robust truth discovery in sparse social media sensing[C]//Proceedings of 2016 IEEE International Conference on Big Data. Piscataway: IEEE Press, 2016: 1076-1081.

[22] XU S S, LI S N, WEN R. Sensing and detecting traffic events using geosocial media data: a review[J]. Computers, Environment and Urban Systems, 2018, 72: 146-160.

[23] ZHANG S, TANG J J, WANG H, et al. Enhancing traffic incident detection by using spatial point pattern analysis on social media[J]. Transportation Research Record: Journal of the Transportation Research Board, 2015, 2528(1): 69-77.

[24] SINNOTT R O, GONG Y K, CHEN S P, et al. Urban traffic analysis using social media data on the cloud[C]//Proceedings of 2018 IEEE/ACM International Conference on Utility and Cloud Computing Companion. Piscataway: IEEE Press, 2018: 134-141.

[25] STEIGER E, RESCH B, DE ALBUQUERQUE J P, et al. Mining and correlating traffic events from human sensor observations with official transport data using self-organizing-maps[J]. Transportation Research Part C: Emerging Technologies, 2016, 73: 91-104.

[26] ZHANG Y, LU Y W, ZHANG D, et al. RiskSens: a multi-view learning approach to identifying risky traffic locations in intelligent transportation systems using social and remote sensing[C]//Proceedings of 2018 IEEE International Conference on Big Data. Piscataway: IEEE Press, 2018: 1544-1553.

[27] WANG S Z, ZHANG X M, LI F X, et al. Efficient traffic estimation with multi-sourced data by parallel coupled hidden Markov model[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(8): 3010-3023.

[28] CHEN P T, CHEN F, QIAN Z. Road traffic congestion monitoring in social media with hinge-loss Markov random fields[C]//Proceedings of 2014 IEEE International Conference on Data Mining. Piscataway: IEEE Press, 2014: 80-89.

[29] YAO W R, QIAN S A. From Twitter to traffic predictor: next-day morning traffic prediction using social media data[J]. Transportation Research Part C: Emerging Technologies, 2021, 124: 102938.

[30] 郑治豪, 吴文兵, 陈鑫, 等. 基于社交媒体大数据的交通感知分析系统[J]. 自动化学报, 2018, 44(4): 656-666.
ZHENG Z H, WU W B, CHEN X, et al. A traffic sensing and analyzing system using social media data[J]. Acta Automatica Sinica, 2018, 44(4): 656-666.

[31] ZHANG Z H, HE Q, GAO J, et al. A deep learning approach for detecting traffic accidents from social media data[J]. Transportation Research Part C: Emerging Technologies, 2018, 86: 580-596.

[32] FU K Q, LU C T, NUNE R, et al. Steds: social media based transportation event detection with text summarization[C]//Proceedings of 2015 IEEE 18th International Conference on Intelligent Transportation Systems. Piscataway: IEEE Press, 2015: 1952-1957.

[33] GU Y M, QIAN Z S, CHEN F. From Twitter to detector: real-time traffic incident detection using social media data[J]. Transportation Research Part C: Emerging Technologies, 2016, 67: 321-342.

[34] DAVIS C A, VAROL O, FERRARA E, et al. BotOrNot: a system to evaluate social bots[C]//Proceedings of the 25th International Conference Companion on World Wide Web. New York: ACM Press, 2016: 273-274.

[35] CHE W X, LI Z H, LIU T. LTP: a Chinese language technology platform[C]//Proceedings of Coling 2010: Demonstrations. [S.l.:s.n.], 2010:13-16

[36] PORTER M F. An algorithm for suffix stripping[J]. Program, 1980, 14(3): 130-137.

[37] D'ANDREA E, DUCANGE P, LAZZERINI B, et al. Real-time detection of traffic from twitter stream analysis[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(4): 2269-2283.

[38] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint, 2013, arXiv: 1301.3781.

[39] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules[C]//Proceedings of Very Large Data Bases Conferences. [S.l.:s.n.], 1994: 487-499

[40] KHAN S M, CHOWDHURY M, NGO L B, et al. Multi-class twitter data categorization and geocoding with a novel computing framework[J]. Cities, 2020, 96: 102410.

[41] CRAMÉR H. Mathematical methods of statistics (PMS-9)[M]. [S.l]: Princeton University Press, 1946.

[42] GUTIÉRREZ C, FIGUERIAS P, OLIVEIRA P, et al. Twitter mining for traffic events detection[C]//Proceedings of 2015 Science and Information Conference. Piscataway: IEEE Press, 2015: 371-378.

[43] CUI J, FU R, DONG C H, et al. Extraction of traffic information from social media interactions: methods and experiments[C]//Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems. Piscataway: IEEE Press, 2014: 1549-1554.

[44] CHEN Y Y, LYU Y S, WANG X, et al. Detecting traffic information from social media texts with deep learning approaches[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(8): 3049-3058.

[45] WANG D, AL-RUBAIE A, DAVIES J, et al. Real time road traffic monitoring alert based on incremental learning from tweets[C]//Proceedings of 2014 IEEE Symposium on Evolving and Autonomous Learning Systems. Piscataway: IEEE Press, 2014: 50-57.

[46] JAIN A K, KUMAR A, GARG J, et al. TraffTrend: real time traffic updates and traffic trends using social media analytics[C]//Proceedings of the 2nd IKDD Conference on Data Sciences. [S.l.:s.n.], 2015: 1-4.

[47] FRANK E, WITTEN I. Generating accurate rule sets without global optimization[C]//Proceedings of the 15th International Conference on Machine Learning. [S.l.:s.n.], 1998.

[48] LU H, SHI K Z, ZHU Y F, et al. Sensing urban transportation events from multi-channel social signals with the Word2Vec fusion model[J]. Sensors (Basel, Switzerland), 2018, 18(12): 4093.

[49] ALKOUZ B, AL AGHBARI Z. SNSJam: road traffic analysis and prediction by fusing data from multiple social networks[J]. Information Processing & Management, 2020, 57(1): 102139.

[50] TEJASWIN P, KUMAR R, GUPTA S. Tweeting traffic: analyzing Twitter for generating real-time city traffic insights and predictions[C]//Proceedings of the 2nd IKDD Conference on Data Sciences. New York: ACM Press, 2015.

[51] SCARPONCINI P. Generalized model for linear referencing[C]// Proceedings of the 7th ACM International Symposium on Advances in

Geographic Information Systems. New York: ACM Press, 1999.

[52] GELERNTER J, BALAJI S. An algorithm for local geoparsing of microtext[J]. GeoInformatica, 2013, 17(4): 635-667.

[53] ERKAN G, RADEV D R. LexRank: graph-based lexical centrality as salience in text summarization[J]. Journal of Artificial Intelligence Research, 2004, 22: 457-479.

[54] PAN B, ZHENG Y, WILKIE D, et al. Crowd sensing of traffic anomalies based on human mobility and social media[C]//Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM Press, 2013: 344-353.

[55] WANG F Y. Scanning the issue[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(1): 1-8.

[56] WANG X, ZHENG X H, ZHANG Q P, et al. Crowdsourcing in ITS: the state of the work and the networking[J]. IEEE Transactions on Intelligent Transportation Systems, 2016, 17(6): 1596-1605.

[57] YIN X Y, WU G Z, WEI J Z, et al. Deep learning on traffic prediction: methods, analysis and future directions[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 4840(99): 1-17.

[58] LIU S Y, YUE Y S, KRISHNAN R. Non-myopic adaptive route planning in uncertain congestion environments[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(9): 2438-2451.

[59] ALEMAZKOOR N, MEIDANI H. A data-driven multi-fidelity approach for traffic state estimation using data from multiple sources[J]. IEEE Access, 2021, 9: 78128-78137.

[60] GHOSH B, BASU B, O'MAHONY M. Multivariate short-term traffic flow forecasting using time-series analysis[J]. IEEE Transactions on Intelligent Transportation Systems, 2009, 10(2): 246-254.

[61] HUANG Y P, QIAN L P, FENG A Q, et al. RFID data-driven vehicle speed prediction via adaptive extended Kalman filter[J]. Sensors (Basel, Switzerland), 2018, 18(9): 2787.

[62] LU Z L, LYU W F, CAO Y B, et al. LSTM variants meet graph neural networks for road speed prediction[J]. Neurocomputing, 2020, 400: 34-45.

[63] JIA H W, LUO H Y, WANG H, et al. ADST: forecasting metro flow using attention-based deep spatial-temporal networks with multi-task learning[J]. Sensors, 2020, 20(16): 4574.

[64] PENG H, WANG H F, DU B W, et al. Spatial temporal incidence dynamic graph neural networks for traffic flow forecasting[J]. Information Sciences, 2020, 521: 277-290.

[65] DU B W, PENG H, WANG S Z, et al. Deep irregular convolutional residual LSTM for urban traffic passenger flows prediction[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 21(3): 972-985.

[66] GEORGE S, SANTRA A K. An improved long short-term memory networks with Takagi-Sugeno fuzzy for traffic speed prediction considering abnormal traffic situation[J]. Computational Intelligence, 2020, 36(3): 964-993.

[67] YANG X, YUAN Y, LIU Z Y. Short-term traffic speed prediction of urban road with multi-source data[J]. IEEE Access, 2020, 8: 87541-87551.

[68] ALI A, ZHU Y M, ZAKARYA M. Exploiting dynamic spatio-temporal correlations for citywide traffic flow prediction using attention based neural networks[J]. Information Sciences, 2021, 577: 852-870.

[69] YANG X, ZHU Q, LI P H, et al. Fine-grained predicting urban crowd flows with adaptive spatio-temporal graph convolutional network[J]. Neurocomputing, 2021, 446: 95-105.

[70] ZHANG J B, ZHENG Y, SUN J K, et al. Flow prediction in spatio-temporal networks based on multitask deep learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 32(3): 468-478.

[71] LI W, TAO W, QIU J Y, et al. Densely connected convolutional networks with attention LSTM for crowd flows prediction[J]. IEEE Access, 2019, 7: 140488-140498.

[72] LYU Y S, DUAN Y J, KANG W W, et al. Traffic flow prediction with big data: a deep learning approach[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(2): 865-873.

[73] LIN L, LI J X, CHEN F, et al. Road traffic speed prediction: a probabilistic model fusing multi-source data[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(7): 1310-1323.

[74] LIU S Y, QU Q. Dynamic collective routing using crowdsourcing data[J]. Transportation Research Part B: Methodological, 2016, 93: 450-469.

[75] ZHOU T, GAO L X, NI D H. Road traffic prediction by incorporating online information[C]//Proceedings of the 23rd International Conference on World Wide Web. New York: ACM Press, 2014: 1235-1240.

[76] FRIEDMAN J H. On bias, variance, 0/1—loss, and the curse-of-dimensionality[J]. Data mining and knowledge discovery, 1997, 1(1): 55-77.

[77] JORDAN M I, GHAHRAMANI Z, JAAKKOLA T S, et al. An introduction to variational methods for graphical models[J]. Machine Learning, 1999, 37(2): 183-233.

[78] PARSONS S. Probabilistic graphical models: principles and techniques[J]. The Knowledge Engineering Review, 2011, 26(2): 237-238.

[79] PEARL J. Reverend Bayes on inference engines: a distributed hierarchical approach[M]. [S.l.:s.n.], 1982.

[80] XUE G, LIU S F, REN L, et al. Forecasting the subway passenger flow under event occurrences with multivariate disturbances[J]. Expert Systems With Applications, 2022, 188: 116057.

[81] ZHANG Z H, NI M, HE Q, et al. Exploratory study on correlation between twitter concentration and traffic surges[J]. Transportation Research Record: Journal of the Transportation Research Board, 2016, 2553(1): 90-98.

[82] NI M, HE Q, GAO J. Forecasting the subway passenger flow under event occurrences with social media[J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 18(6): 1623-1632.

[83] NI M, HE Q, GAO J. Using social media to predict traffic flow under special event conditions[C]//Proceedings of the 93rd Annual Meeting of Transportation Research Board. [S.l.:s.n.], 2014.

[84] GODBOLE N, SRINIVASAIAH M, SKIENA S. Large-scale sentiment analysis for news and blogs (system demonstration)[J]. Icwsm, 2007, 7(21): 219-222.

[85] NAGARAJAN M, GAMON M. Automating analysis of social media communication: insights from CMDA[C]//Proceedings of the Work-

shop on Languages in Social Media. Oregon: Association for Computational Linguistics, 2011: 1.

[86] ESSIEN A, PETROUNIAS I, SAMPAIO P, et al. A deep-learning model for urban traffic flow prediction with traffic events mined from twitter[J]. World Wide Web, 2021, 24(4): 1345-1368.

[87] ESSIEN A, PETROUNIAS I, SAMPAIO P, et al. Improving urban traffic speed prediction using data source fusion and deep learning[C]//Proceedings of 2019 IEEE International Conference on Big Data and Smart Computing. Piscataway: IEEE Press, 2019: 1-8.

[88] LIU X Y, KONG X N, LI Y H. Collective traffic prediction with partially observed traffic history using location-based social media[C]//Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. New York: ACM Press, 2016: 2179-2184.

[89] WONGCHAROEN S, SENIVONGSE T. Twitter analysis of road traffic congestion severity estimation[C]//Proceedings of 2016 13th International Joint Conference on Computer Science and Software Engineering. Piscataway: IEEE Press, 2016: 1-6.

[90] CUI Y, MENG C S, HE Q, et al. Forecasting current and next trip purpose with social media data and Google places[J]. Transportation Research Part C: Emerging Technologies, 2018, 97: 159-174.

[91] ROY K C, HASAN S, CULOTTA A, et al. Predicting traffic demand during hurricane evacuation using real-time data from transportation systems and social media[J]. Transportation Research Part C: Emerging Technologies, 2021, 131: 103339.

[92] ALTMANN A, TOLOŞI L, SANDER O, et al. Permutation importance: a corrected feature importance measure[J]. Bioinformatics, 2010, 26(10): 1340-1347.

[93] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning[M]. [S.l.]: MIT press, 2016.

[94] HASHEMI H, ABDELGHANY K. Real-time traffic network state prediction for proactive traffic management[J]. Transportation Research Record: Journal of the Transportation Research Board, 2015, 2491(1): 22-31.

[95] ABDELGHANY K, HASHEMI H, KHODAYAR M E. A decision support system for proactive-robust traffic network management[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(1): 297-312.

[96] HASHEMI H, ABDELGHANY K. End-to-end deep learning methodology for real-time traffic network management[J]. Computer-Aided Civil and Infrastructure Engineering, 2018, 33(10): 849-863.

[97] WANG D, 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.

[98] WANG D, ABDELZAHER T, KAPLAN L. Surrogate mobile sensing[J]. IEEE Communications Magazine, 2014, 52(8): 36-41.

[99] ZHANG D, VANCE N, WANG D. When social sensing meets edge computing: vision and challenges[C]//Proceedings of 2019 28th International Conference on Computer Communication and Networks. Piscataway: IEEE Press, 2019: 1-9.

[作者简介]

陈苑文（2000− ），女，厦门大学航空航天学院自动化系在读，主要研究方向为基于社会媒体数据增强的交通态势感知及状态推理。

王晓（1988− ），女，博士，中国科学院自动化研究所复杂系统管理与控制国家重点实验室副研究员，青岛智能产业技术研究院院长。主要研究方向为平行智能、社会交通、动态网群组织行为和社交网络分析。

李灵犀（1977− ），男，博士，美国印第安纳大学−普渡大学印第安纳波利斯分校电子与计算机工程系教授，主要研究方向为复杂系统的建模、分析、控制与优化，智能交通系统，智能网联汽车，驾驶辅助系统与人因学。

王飞跃（1961− ），中国科学院自动化研究所研究员，复杂系统管理与控制国家重点实验室主任，中国科学院大学中国经济与社会安全研究中心主任。主要研究方向为平行系统的方法与应用、社会计算、平行智能以及知识自动化。