基于社交媒体大数据的交通感知分析系统

郑治豪 [1] 吴文兵 [2] 陈 鑫 [3] 胡荣鑫 [1] 柳 鑫 [1] 王 璞 [1]

摘 要 社交媒体数据中蕴含了丰富的交通状态信息, 这些信息以人类语言为载体, 包含了大量对交通状态的因果分析与多角度描述, 可以为传统交通信息采集手段提供有力补充, 近年来已成为交通状态感知的重要信息来源. 本文以新浪微博为主要数据来源, 分别利用支持向量机算法、条件随机场算法以及事件提取模型完成微博的分类、命名实体识别与交通事件提取, 开发了基于社交媒体大数据的交通感知分析与可视化系统, 可以为交通管理部门及时提供交通舆情及突发交通事件的态势、影响范围、起因等信息. 在交通信息采集系统建设较为薄弱的地区, 本文建立的系统可以为交通管理提供信息补充.

关键词 社会交通, 机器学习, 文本分类, 命名实体识别, 数据可视化

引用格式 郑治豪, 吴文兵, 陈鑫, 胡荣鑫, 柳鑫, 王璞. 基于社交媒体大数据的交通感知分析系统. 自动化学报, 2018, **44**(4): 656−666

DOI 10.16383/j.aas.2017.c160537

A Traffic Sensing and Analyzing System Using Social Media Data

ZHENG Zhi-Hao[1] WU Wen-Bing[2] CHEN Xin[3] HU Rong-Xin[1] LIU Xin[1] WANG Pu[1]

Abstract Social media data, which encapsulate abundant traffic status information, have gradually become an important data source for sensing traffic status. The information recorded by human language contains a large amount of causality analysis and multi-angle descriptions of the traffic condition, acting as a powerful supplement to traditional traffic information collecting methods. Employing Sina Weibo as a main data source, we apply SVM algorithm, CRF algorithm and event extracting model for classification, named entity recognition and events extraction of microblogs. We develop a traffic sensing and visualizing system, which can collect public opinion, situations, scales and even origins of traffic incidents for transportation agency. Furthermore, this system can provide traffic information for the transportation department in the area which lack traffic detectors.

Key words Social transportation, machine learning, text classification, named entity recognition, data visualization

Citation Zheng Zhi-Hao, Wu Wen-Bing, Chen Xin, Hu Rong-Xin, Liu Xin, Wang Pu. A traffic sensing and analyzing system using social media data. *Acta Automatica Sinica*, 2018, **44**(4): 656−666

进入新世纪, 我国交通信息化建设快速推进, 公交车或出租车上的 GPS 轨迹数据[1−2]、磁感线圈数据[3]、视频监控数据[4−5] 大量涌现, 基于这些数据的交通状态感知与预测技术发展迅速. 翁剑成等[1] 基于浮动车 GPS 数据, 获取了路段区间运行速度与行程时间信息, 改善了传统交通检测方式高投入、精度低的缺点. 董均宇[2] 通过融合多类型车辆 GPS 轨迹数据与道路交通信息, 估计了城市路段的平均速度. 陶汉卿等[3] 基于分段序列相似度的分析方法对转弯车辆和直行车辆的感应数据进行联合分析, 获取了车辆的转弯信息, 提高了交通调查和交通信息采集的效率和准确性. 张佐等[4] 指出先进的视频技术将成为视频和无线传感器 "按需" 布设和多参数交通信息采集的基础, 利用视频处理技术可以发现混合交通流的新特性. 王川童[5] 运用视频检测技术获取的交通数据, 结合卡尔曼滤波跟踪与虚拟检测线法提取交通特征参数, 提高了车辆识别的精度. Li 等[6] 利用实测交通流数据, 提出了一种基于模糊集理论的交通流预测方法, 该方法还可以准确预测交通流变化的范围.

基于车辆 GPS 轨迹、磁感线圈、视频监控等数据的交通分析方法在智能交通系统的建设和发展中发挥了重要作用, 然而这些数据自身结构和特点也使它们在某些应用方面存在不足. Shang 等[7] 指出: 某些时刻很多路段上并没有出租车行驶, 浮动车数据一定程度上缺乏完整性; 感应线圈的埋置深度、性

收稿日期 2016-07-19 录用日期 2017-04-07
Manuscript received July 19, 2016; accepted April 7, 2017
国家自然科学基金面上项目 (61473320), 湖南省科技计划项目 (2015RS4011), 中南大学创新驱动计划项目 (2016CSX014) 资助
Supported by National Natural Science Foundation of China (61473320), Science and Technology Project of Hunan Province (2015RS4011), Innovation Driven Plan of Central South University (2016CSX014)
本文责任编委 王飞跃
Recommended by Associate Editor WANG Fei-Yue
1. 中南大学交通运输工程学院 长沙 410075 2. 中南大学软件工程学院 长沙 410075 3. 中南大学信息科学与工程学院 长沙 410075
1. School of Traffic and Transportation Engineering, Central South University, Changsha 410075 2. School of Software, Central South University, Changsha 410075 3. School of Information Science and Engineering, Central South University, Changsha 410075

能和寿命、线圈与导线接头的可靠性和防潮绝缘性能等均有待进一步完善和改进; 而视频检测设备在气象恶劣的情况和低光照强度下, 很难得到清晰可靠的图像. 陆锋[8] 指出基于移动目标速度感知方式的交通信息采集手段在运营成本和时空覆盖范围上仍然存在较大的局限性. Zhang 等[9] 发现节假日交通出行由于受到天气、旅游商业等特殊活动以及服务价格、交通事故等多种偶发、可变因素影响, 难以通过历史数据作出有效预测, 常常导致交通突发事件预报失当、应对失当.

进入 “互联网 +” 时代, 社交媒体已经成为人们生活的重要组成部分和人类语言的重要发布平台. 社交媒体中蕴含了大量与交通相关的语言描述, 在某些应用方面甚至比在物理空间中采集的交通信息更有优势. 复杂系统管理与控制国家重点实验室主任王飞跃教授认为社会信号是复杂系统平行管理与控制的重要一环[10], 并首次提出 “社会交通” 研究方向[11].

社交媒体是社会交通研究的重要数据来源, 基于社交媒体大数据的交通研究与应用方兴未艾. Qiao 等[12] 指出微博消息可以作为线圈、视频等交通检测传感器的有效补充, 可以用于及时定位交通拥堵. Zeng 等[13] 指出社交媒体信息可以提供交通预警信号与路况信息预报. Wanichayapong 等[14] 开发了一个基于 Twitter 数据的交通信息采集与归类系统. Endarnoto 等[15] 开发了一个 Twitter 交通信息采集系统, 并设计了一款安卓手机软件用于显示交通信息. Balagapo 等[16] 开发了一款安卓手机软件, 采用信息众包的方式采集使用者记录, 分享公共交通出行数据. D′Andrea 等[17] 开发了一个基于 Twitter 信息流的交通信息实时监测系统, 该系统可以在新闻网站发布相同资讯之前监测到相关的交通信息. 张恒才等[18] 提出了一种从微博消息中快速获取和融合交通信息的技术方法, 他们首先对获取的微博进行分词和路网匹配, 然后用模糊 C 聚类方法对微博进行定量化结果分析, 从而获取各个路段置信度最高的交通状态描述. 张恒才等[19] 还提出了一种基于 D-S 证据理论的微博交通信息获取方法, 构建了微博文本中交通状态信息的评价体系, 作者定义了微博消息源的基本概率分布函数, 通过证据合成与证据决策, 实现微博消息中实时交通信息的甄别与融合. 崔健等[20] 开发出一套基于微博的节假日突发交通事件感知与分析系统, 旨在分析节假日交通状况, 评估个体出行者的情感状态等. 熊佳茜[21] 运用条件随机场模型对微博进行时间与地点词语的识别, 用于感知交通事件. Hasan 等[22] 利用社交媒体数据分析出行者活动模式. Gkiotsalitis 等利用社交媒体数据分析用户参加各类活动的出行意愿[23], 并针对休闲活动的出行随机特性进行分析[24]. Gu 等[25] 开发了一套基于社交媒体的交通事件探测系统, 并在两个城市得到应用. Kuflik 等[26] 提出了一个从社交媒体信息中抽取交通相关信息的框架. Rashidi 等[27] 探讨了社交媒体数据在挖掘人类出行行为方面的机遇与挑战. Cottrill 等[28] 讨论了社交媒体信息在大型事件发生时的交通信息传播策略. Xiong 等[29] 提出了一个基于信息–物理–社会系统的智能交通系统框架, 文中详细介绍了社会交通系统的运行机制.

车辆 GPS 轨迹数据、磁感线圈数据、视频监控数据等由物理空间的传感器采集, 具有量化、精确、客观的特点. 蕴涵交通信息的微博文本由社交媒体采集, 虽然在描述上存在一定模糊性、主观性, 但包含了人类的分析、推理和智慧. 物理空间与社会空间交通数据在交通分析、预测与应用中具有各自的优势. 目前物理空间交通数据的采集、处理方法比较成熟完善, 但社会空间交通数据方面的研究还较少, 尚缺乏系统性好、应用性强的社交媒体交通感知系统.

本文的结构组织如下: 第 1 节总述本文所构建系统的构架与要解决的重难点; 第 2 节阐述数据来源与数据处理; 第 3 节和第 4 节阐述两个难点的解决方案; 第 5 节展示数据可视化; 第 6 节讨论系统特色; 第 7 节是结论及下一步工作.

1　系统构架

基于社交媒体大数据的交通感知分析系统由以下几个模块组成: 1) 微博数据采集与预处理模块; 2) 微博分类模块; 3) 微博命名实体识别模块; 4) 交通事件归类与可视化模块. 系统构架图如图 1 所示.

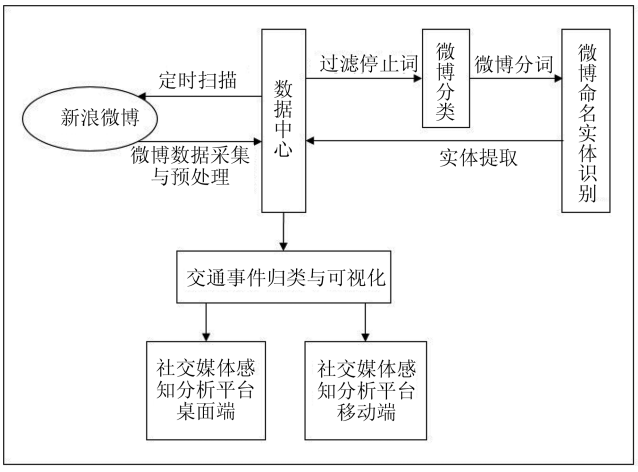


图 1　系统构架图

Fig. 1　Architecture of the system

建立基于社交媒体大数据的交通感知分析系统

需解决的难点如下:

1) 微博内容语义消歧与交通话题筛选. 中文具有一词多义的特点, 带有关键词的微博可能与交通无关, 且与交通相关的微博也不一定带有实际的交通信息, 如何进行语义消歧和交通话题筛选, 减少对无效微博的后续处理, 提高系统效率, 是本论文解决的第一个难点.

2) 微博数据中交通信息的有效识别与提取. 微博中包含的交通事件发生地点往往比普通的地点实体更复杂, 如何准确界定微博中的交通相关信息, 并选择相应的算法提取这些信息是本论文解决的第二个难点.

针对上述研究重点与难点问题, 我们使用数据挖掘、机器学习、自然语言处理的方法对社交媒体数据进行了大量的实验与测试, 最终选择出可靠有效、性能优良的方法, 建立了一个基于社交媒体大数据的交通感知分析系统. 该系统体现了 "社会交通" 的信息众包机制[30−31], 发挥了群体智慧的优势, 在突发事件的检测, 交通事件的原因分析、规模判断, 舆情采集等方面是现有交通检测方式的有力补充, 并且为未来 "社会交通" 研究提供基础数据与分析平台.

2 微博数据采集与预处理

首先运行网络爬虫, 通过设置好的关键词 (表 1) 随机收集 4 万条相关的原始微博数据.

表 1 关键词表

Table 1 Keywords list

堵	车祸	剐蹭	事故	绕行
路	追尾	相撞	塞车	高速

原始微博数据的每一条信息包含: 微博发布时间、官方标记 (是否源于认证的官方微博)、微博正文、微博定位地点. 原始微博正文中可能含有一些特定符号, 包括表情符号、话题标签 (##)、链接、转义字符、用户引用 (@ 符号) 以及多余的空格等, 这些内容没有实际含义与信息, 剔除后不影响全文语义表达.

文中使用 Python 的正则表达式模块对这些符号匹配剔除. 同时, 为了减小微博不准确信息和不真实信息经大量转发后的扩散影响, 在抓取微博时仅对原创微博进行抓取, 不使用转发微博. 数据预处理后, 得到了标准化的微博数据, 如表 2 所示.

3 微博分类

本文采用机器学习的方法进行微博分类, 解决微博内容语义消歧与交通话题筛选问题. 首先, 制定了有效微博交通信息的评判标准, 并以此为依据划分微博信息, 构建训练集; 其次, 利用不同的文本分类算法进行测试; 最后, 综合考虑各种因素选出最适合本系统的分类算法.

3.1 有效微博与无效微博

根据微博内容是否与交通信息有关进行评判, 本文将抽取到的微博分为有效微博与无效微博, 其定义如下:

定义 1. 有效微博

有效微博包含表 1 关键词, 所讨论的话题属于交通话题, 且描述实际交通情况. 例如:

"大鹏片区南西路沙坑农庄路段发生小车追尾事故, 民警正在现场处理事故, 疏导交通."

定义 2. 无效微博

无效微博包含表 1 关键词, 但其描述的话题与交通无关, 或者其虽然属于交通话题, 但并不描述实际交通情况. 例如:

"黄山再美都被人挤人的人群给淹没了还好下山不堵."

"交通管理部门要求: 1. 小汽车的司机和前排乘客必须系好安全带 — 这样可以防止惯性的危害; 2. 严禁车辆超载 — 不仅仅减小车辆对路面的破坏, 还有减小摩擦、惯性等; 3. 严禁车辆超速 — 防止急刹车时, 因反应距离和制动距离过长而造成车祸".

3.2 微博分类训练集

本文通过人工浏览标准化微博数据库中的 4 万条微博, 从中人工分类出 5 000 条有效微博与 5 000 条无效微博, 去除停止词后, 分别存入两个文档中, 其分类标签分别为 1 和 0.

在微博分类之前, 需要将文本向量化, 本文所构建的文本分类器使用隐性语义分析 (Latent semantic analysis, LSA) 进行向量化, 流程如图 2 所示.

表 2 标准化微博数据

Table 2 Standardized Weibo data

微博发布时间	官方标记	微博正文	微博定位地点 (缺省为 *)
201604022042	0	竟然能在一个地方堵车堵快 1 个小时了! 气得好多人中途下车了!	北京 · 北七家

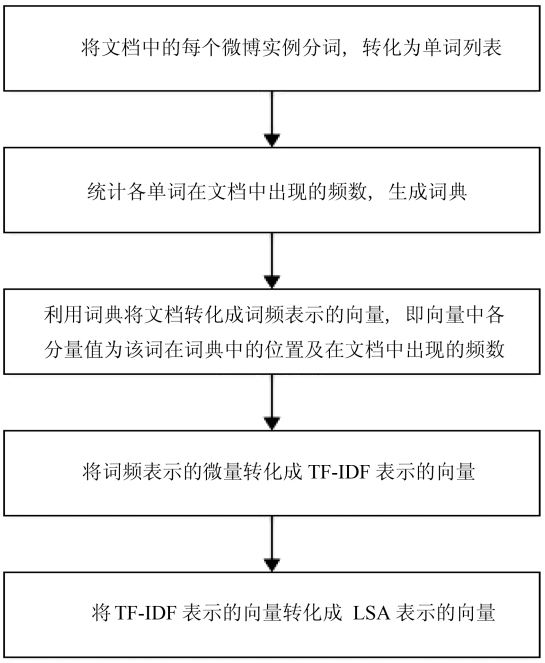


图 2　文本向量化流程图

Fig. 2　Flowchart of document vectorization

本文使用 LTP[32] 提供的中文停止词表去除微博正文中的停止词, 使用 Gensim[33] 工具包进行微博正文的向量化.

3.3　微博分类算法

文本分类算法则主要基于朴素贝叶斯 (Naive Bayes, NB)、k 最近邻 (k-nearest neighbor, KNN)、支持向量机 (Support vector machine, SVM)、决策树 (Decision tree, DT) 等算法. Scikit-learn[34] 是 Python 中的一个机器学习包, 提供了多种分类器算法. 其中, SVM 形参 kernel 的值代表其分类时所采用的核函数, 本测试取 linear、rbf、sigmoid、poly 四种核函数; KNN 方法中, k 表示分类决策时选取的最相似数据的个数, 测试选取 1NN、3NN、5NN; NB 方法中, 可以选择不同的模型训练, 本文选取高斯模型 (Gaussian NB) 和多项式模型 (Multinomial NB); DT 方法中, 形参 criterion 表示构造决策树时节点测试属性选取的标准, 测试选取信息熵 (Entropy) 和基尼不纯度 (Gini).

研究中使用第 3.2 节中得到的微博分类训练集训练分类模型. 在训练分类模型时, 采用十折交叉验证法, 对十次训练得到的模型评估参数取平均值作为最终评估模型的参数.

文中选择 MUC 会议制定的评估体系. 其评价模型性能的指标有准确率 (Precision)、召回率 (Recall) 和 F-score. 其中, 准确率是预测结果为有效微博中预测正确的比例, 召回率是预测结果为有效微博中预测正确的数量占全部人工标注的有效微博数量的比例, F-score 的计算公式如下:

$$\text{F-score} = \frac{(\lambda \times \lambda + 1) \times p \times r}{(\lambda \times \lambda \times p) + r} \tag{1}$$

其中, λ 是召回率相对于准确率的权重, 当 λ 取值小于 1 时, 结果偏向准确率; 大于 1 时, 结果偏向召回率. 在本次分类中, 准确率和召回率同等重要, λ 取值为 1.

在利用训练集对所有算法进行测试之后, 测试结果如表 3 所示.

表 3　不同分类算法的测试结果

Table 3　Test results of different algorithms

算法	Precision	Recall	F1-score
SVM (kernel = ‘linear’)	0.880	0.850	0.859
SVM (kernel = ‘rbf’)	0.747	0.574	0.504
SVM (kernel = ‘sigmoid’)	0.799	0.524	0.419
SVM (kernel = ‘poly’)	0.234	0.500	0.318
1NN	0.693	0.685	0.683
3NN	0.725	0.699	0.692
5NN	0.727	0717	0.717
Gaussian NB	0.645	0.626	0.618
Multinomial NB	0.766	0.768	0.767
DT (criterion = ‘entropy’)	0.676	0.687	0.676
DT (criterion = ‘gini’)	0.674	0.677	0.672

由表 3 结果可以看出, SVM 算法总体表现优异, 采用的各种核函数中, 线性核表现最优, 表明文本向量化得到的数据是线性可分的; KNN 算法整体的表现不佳, 这与 KNN 算法的归纳偏置密切相关: 一个新数据的分类标签总是与其在欧氏空间中若干个临近数据的多数标签相同. 在算法应用的过程中, 数据间的距离是根据数据的所有属性计算的, 近邻间的距离往往会被大量的不相关属性所主导, 从而降低 KNN 算法的分类性能. 对比不同 k 值的 KNN 算法可以看出, 当 k 增大时, 分类性能有所提升, 说明在一定范围内 k 值增大能够更好地排除错误数据与噪声的影响, 提高分类性能; 朴素贝叶斯分类器采用不同的模型时, 分类性能差异较大. 高斯分布的朴素贝叶斯分类器的性能明显低于多项式分布的朴素贝叶斯分类器. 其原因在于, Gaussian NB 假定训练集中的各样本特征值服从高斯分布, 而这一假定并不一定符合微博语料的实际情况. Multinomial NB 以文档中的单词作为特征, 对应的特征值是单词在文档中出现的次数, 是典型的词袋模型, 适用于文本分类; 决策树算法在测试中表现较差. 构造决策树时节点测试属性选取标准的不同, 并不会对最终

的分类性能产生明显的影响.

综上, 在本系统中选择性能最优的 SVM 算法进行微博分类, 为解决有效、无效微博分类提供了一个可行的方案, 解决了第 1 节所述难点 1.

4 微博命名实体识别

本文同样采用机器学习的方法进行微博命名实体识别, 解决微博内容中交通信息的有效识别与提取问题. 首先, 我们对微博蕴含交通信息的实体名词进行定义; 其次, 我们讨论了不同实体名词标注方案的优劣, 建立了微博交通信息实体的界定方法; 最后, 我们讨论并确定了最适合本系统的实体识别算法.

4.1 时间实体与地点实体

在对微博分类后, 我们使用命名实体识别 (Named entity recognition, NER) 对有效标准微博数据进行时间实体和地点实体的识别 (如图 3 所示).

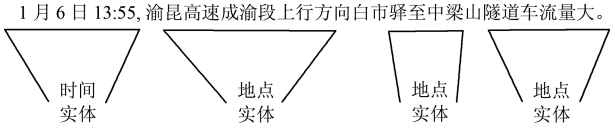


图 3 时间实体与地点实体示例

Fig. 3 An example of time entity and location entity

最常见的两种命名实体识别方法为基于语法规则的方法和基于机器学习的方法. 前者在所制定的规则适应于相对应文本情景的情况下具有良好的表现, 但在面对陌生随机文本时表现不佳[35]. 后者的优点在于它可以利用标记文本反复训练, 适应性强, 维护成本远小于基于语法规则的方法[36]. 基于机器学习的方法又分为有监督、半监督和无监督方法. 由于后者无需太多的语言学知识, 且有监督的机器学习方法只需通过训练模板设定待考察的特征, 并用算法对人工标注真值的训练集进行训练, 便可得出相应的模型文件用于实体识别, 简单易用, 对随机文本适应性强. 所以, 文中选择基于有监督的机器学习算法完成微博命名实体识别的工作.

由于命名实体识别需要基于词序列进行建模, 文中使用 LTP[32] 分词工具将每一条微博文本切分为词序列并标注词性后进行序列标注, 如表 4 所示.

4.2 命名实体识别训练集

文中选取分类阶段中筛选出的 5 000 条未过滤停止词的有效微博进行分词序列化处理及词性标注, 并采用文献 [21] 提出的方法进行人工命名实体标注, 作为训练真值. 标注规则与示例如表 5 所示.

表 4 微博的词序列示例

Table 4 An example of a sequence of Weibo word

微博词序列示例	词性符号	词性	
1 月	nt	nt	temporal noun
6 日	nt		
13:55	m	m	number
,	wp		
愈	j	wp	punctuation
昆	j		
高速	d	j	abbreviation
成	v		
渝段	n	d	adverb
上行	v		
方向	n	v	verb
白市驿	ns		
至	p	n	general noun
中梁山	ns		
隧道	n	ns	geographical name
车流量	n		
大	a	p	preposition

在标注命名实体的过程中, 我们发现, 较长的交通地点实体常常占据 5 ∼ 7 个窗口, 且由多个短地点实体组成, 导致不同的人对同一个地点实体的标注会有不同的结果 (如图 4 (a) 和 4 (b) 所示).

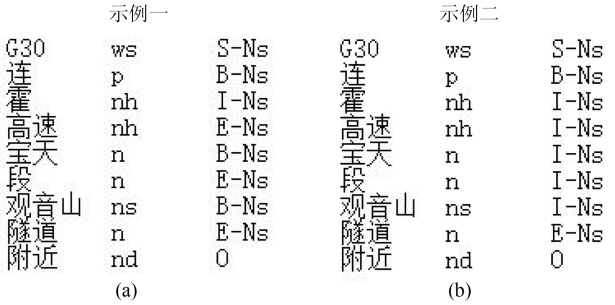


图 4 命名实体标注示例

Fig. 4 Examples of NER labels

从图 4 可以看出, “G30 连霍高速宝天段观音山隧道” 描述的是一个具体交通事件发生的位置, 在这个位置中包含了多个可以作为命名实体的地点, 例如 “连霍高速”、“观音山隧道”.

从词义角度分析, 描述一个交通事件发生地点通常是由高级地名向低级地名递减. 示例二将这种地理描述完整的标记出来, 作为一个地点实体. 而示例一则将一个地理描述中的多个地名作为单独的地点实体.

从应用角度分析, 如果利用示例一的标注方法, 多个地点实体之间的从属性较难判断, 造成定位困难. 而示例二则避免了这个问题, 降低了定位难度.

表 5 命名实体标注方案
Table 5 Method of NER labelling

类别	标注符号	说明	词序列示例		标注示例
	B-Ns	地点词的起始	1 月	nt	B-Nm
			6 日	nt	I-Nm
地	I-Ns	地点词的中部	13:55	m	E-Nm
点			渝	wp	B-Ns
实	E-Ns	地点词的结尾	昆	j	I-Ns
体			高速	j	I-Ns
	S-Ns	完整的地点词	成	d	I-Ns
			渝段	v	E-Ns
	B-Nm	时间词的起始	上行	n	O
			方向	v	O
时	I-Nm	时间词的中部	白市驿	n	S-Ns
间			至	ns	O
实	E-Nm	时间词的结尾	中梁山	nt	B-Ns
体			隧道	n	E-Ns
	S-Nm	完整的时间词	车流量	n	O
			大	wp	O

综上, 文中采用示例二所示的标注方法对微博命名实体训练集进行标注, 该方法为: 在连续的地理位置描述中, 以两个相同等级的地名为地点实体分隔点, 每个地点实体由最高等级地名开始至最低等级地名结束. 例如 “G30 连霍高速宝天段观音山隧道” 这一描述中, “G30” 是 “连霍高速” 的代号, 故二者属于平行关系, 我们将 “G30” 作为单独的地点实体. “连霍高速” 和 “观音山隧道” 分别是该描述中最高级和最低级的地名, 故我们将 “连霍高速宝天段观音山隧道” 标注为一个地点实体. “附近” 一词不具有定位意义, 不作标注. 该方法能够清晰地标定微博文本中的交通地点实体, 减少判定尺度不一致带来的误差, 为解决微博交通信息提取提供了可行方案, 解决了第 1 节所述难点 2.

4.3 测试分析

文献 [37] 指出, 较常用的用于命名实体识别的序列标注算法有: 最大熵马尔科夫模型 (Maximum-entropy markov model, MEMM)、隐性马尔科夫模型 (Hidden markov model, HMM)、条件随机场模型 (Conditional random field, CRF) 以及支持向量机模型 (Support vector machine, SVM). 对于序列标注问题, 隐性马尔科夫模型的识别速度快[38], 但对观察序列的多个非独立特征建模存在困难[39]. 支持向量机模型则需要进行两步操作, 先对各行独立分配标签, 再进行调整, 这种方式忽略了状态转移和观察之间的紧密关系[39]. 最大熵马尔科夫模型虽然克服了 HMM 模型输出独立性假设的缺点, 但只在局部统计归一化概率, 且会产生标注偏置的问题. 条件随机场模型汲取了 HMM 和 SVM 的优点, 特征设计灵活, 可以容纳任意的上下文信息, 被广泛运用于诸如命名实体识别等多种自然语言处理任务中[39]. 而 CRF 与 MEMM 相比, CRF 模型计算的是全局最优输出节点的条件概率, 也克服了标注偏置的问题. 虽然 CRF 复杂度高, 训练代价大, 但在使用时速度满足本系统的使用要求. 所以, 我们拟运用 CRF++[40] 工具包对 CRF[41] 算法的性能进行测试.

在测试 CRF 算法时, 为了得到最准确的模板, 我们采用了 6 套适合我们数据结构的模板进行实验, 以期得到一个准确率和召回率最高的模板. 在此过程中, 同样采用第 3.3 节中使用的评价体系. 模板的设定方式和性能如表 6 所示, 表中用 a 代表分词结果, b 代表词性.

表 6 CRF 不同模板的设置方案与测试结果
Table 6 Settings of different CRF templates and test results

方案	窗口大小	考虑的列	考虑的相对关系	Precision	Recall	F1-score
一	3	a	N/A	0.790	0.665	0.72
二	3	a, b	N/A	0.798	0.743	0.769
三	3	a, b	a, b	0.794	0.754	0.773
四	5	a	N/A	0.787	0.639	0.703
五	5	a, b	N/A	0.788	0.735	0.760
六	5	a, b	a, b	0.791	0.741	0.764

根据测试结果, 方案三的 F1 值最高, 在准确率和召回率上都有良好的表现, 故本文采用方案三作为训练模板.

4.4 标注信息的提取

系统运用训练好的 CRF 模型对词序列进行标注, 逐行遍历标注结果并提取出相关的词语并将其组合起来, 如图 5 所示.

系统通过标签尾部的 Ns 和 Nm 标识判断该词是一个交通地点实体, 还是一个交通时间实体的组成部分, 再通过标签前部的 B、I、E、S 标识判断该词属于该实体的哪一部分. 若是 S 标签, 该词即为一个完整的实体; 若是 B 标签, 则读取至下一个 E 标签处, 将这两个标签之间对应的词组合起来作为一个实体.

在获取了微博中的交通时间实体和交通地点实体后, 我们不能直接将其作为交通事件的发生时间和地点. 因为我们在采集微博时获得了微博的发布时间, 所以我们通过系统将交通时间实体数字化后,

选取两个时间中较早的时间作为事件发生时间. 同时, 在微博定位地点不缺省时, 文中优先选择微博定位地点作为事件发生地点. 最后, 使用百度地址解析 API[42] 将其转化为 GPS 坐标供可视化模块调用.

1月	nt	B-Nm
6日	nt	I-Nm
13:55	m	E-Nm
,	wp	O
渝	j	B-Ns
昆	j	I-Ns
高速	d	I-Ns
成	v	I-Ns
渝段	n	E-Ns
上行	v	O
方向	n	O
白市驿	ns	S-Ns
至	p	O
中梁山	ns	B-Ns
隧道	n	E-Ns
车流量	n	O
大	a	O

图 5 微博命名实体标注结果

Fig. 5 Weibo NER labelling results

5 交通事件归类与可视化

5.1 交通事件归类

在这个部分我们用关键词对采集的微博交通事件作简要归类, 实现可视化模块中信息分类浏览的功能. 交通事件类别如表 7 所示.

我们人工将第 3 节中的有效微博归为表 7 所示 6 类, 统计每一类中出现频率最高的词, 从高频率词表中选取具有代表性的且与交通相关的词语作为该类别对应的关键词库. 在进行微博事件归类的过程中, 我们用每一个关键词库中的词语对微博进行匹配, 若微博中含有该词语, 则我们将该微博贴上相应类别标签. 例如涉及车辆相撞等事故的的微博中, 可能出现" 撞"、" 追尾"、" 剐蹭" 等词语, 我们将这些词语作为车辆相撞类别的关键词库. 由于交通事件之间常具有一些因果关联, 如事故可能导致路段拥堵, 所以每一条微博可能同时具有多个类别标签. 值得注意的是, 由于本环节处理的微博已是有效微博, 所以不需考虑一词多义等问题.

表 7 交通事件归类

Table 7 Classification of traffic events

路况正常	施工	封路
路况拥堵	车辆相撞	其他

5.2 数据可视化模块

本系统的可视化模块桌面端基于 Web 平台构建, 采用 PHP 语言编写. 可视化模块移动端基于安卓平台构建, 采用 Java 语言编写. 系统对获取到的原始微博信息进行处理后, 获得了交通事件发生的时间、地理坐标以及事件类型等信息, 可视化模块读取上述格式的数据后, 根据不同的事件类型用不同颜色的图标在地图上进行可视化标记, 点击该图标, 会弹出具体的事件信息. 对于含有多个类别标签的数据, 我们以封路、施工、车辆相撞、路况拥堵、路况正常、其他的优先级顺序显示标记的颜色. (如图 6 (a) 和图 6 (b) 所示).

(a) (b)

图 6 可视化模块

Fig. 6 Visualization module

6　系统应用评估

由于实时数据量巨大, 系统在实际运用中采取少量高频采集、采集与处理同时进行的方式采集和处理数据, 以保证系统的实时性与高效性.

在进行性能与可靠性测试期间, 系统于 2016 年 4 月 2 日下午全程跟踪监测到了沪蓉高速常州段两辆大货车相撞翻车, 事故发生后几分钟内, 车辆连续追尾, 造成重大交通事故.

根据央视等官方权威媒体事后的报道, 该事件发生于 4 月 2 日下午 13 时 20 分左右, 最终导致约 56 辆汽车追尾, 本系统于事故发生前就监测到多条微博信息反应该路段拥堵, 而在事故发生 14 分钟之后, 本系统即监测到该路段交通中断, 而事故发生 35 分钟之后, 系统即报告了该事故的严重程度. 图 7 显示的是 13:55 分系统在该路段监测到的数据量.

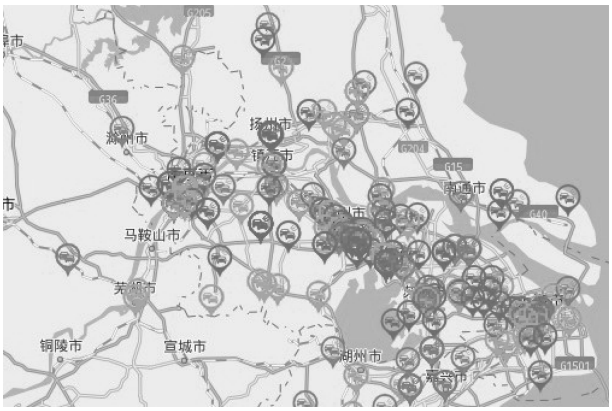


图 7　13:55 系统在相关路段的监测截图

Fig. 7　A system screenshot at 13:55

官方新闻最早播报这起事故是中国广播网于 15:00 发出了一条新闻, 这比本系统首次监测到该路段交通中断延迟了 1 小时 26 分钟. 分析其原因在于, 本系统利用信息众包的思想, 信息来源更加广泛, 而传统的新闻媒体由于其工作性质, 需要对信息反复沟通确认后才会发布信息, 这使得社交媒体在突发情况的信息传播方面通常比传统媒体更加快速, 也更能反映普通民众针对该事件的舆情导向. 除此之外, 相较于车辆 GPS 轨迹、感应线圈、微波等常用的交通检测手段, 社交媒体数据中蕴含着因果关联与对事件的文字描述, 能够直接反应出事件的原因、规模、影响程度等信息, 与视频监控等方式相比又具有成本低廉的特点, 是一种行之有效的交通检测辅助手段.

当然, 社交媒体数据也有不足之处. 首先, 社交媒体数据的置信度有待进一步考量. 虽然我们在采集信息时已排除转发信息的影响, 但系统采集到的信息中仍含有部分不准确的信息, 这一定程度上是人们在主观上对同一事件的不同程度的判断所导致的. 如图 8 所示数据, 系统在采集信息的过程中也采集到一些过分夸张的信息, 这些信息并不是真实的, 图中微博显示该路段有百车相撞, 但实际上这只是微博用户对于现场情况的夸大估计. 此外, 在少数情况下, 微博中也存在部分虚假信息.


```
"time":"2016年04月02日15时53分",
"context":"常州段百车相撞，伤亡不明.",
"place":"苏州·姑苏区",
"piece":"2",
"res":"车辆相撞"
```


图 8　偏差数据示例

Fig. 8　An example of bias

其次, 社交媒体数据具有一定的地理模糊性. 我们从社交媒体数据中获得的地理位置信息来源于原文或发布者的地理定位, 部分位置信息在进行地理坐标解析时, 难以在地图上找到准确的位置, 仍需进一步研究解决.

7　结论

本文开发了一个基于社交媒体大数据的交通感知分析系统. 该系统能够自动采集、分类、提取微博中的有效交通信息并在地图上进行可视化标注. 系统充分利用了社交媒体上人们对于交通事件的最新信息分享、原因分析和程度描述, 相比与传统交通检测设备所采集的数据, 本系统所采用的数据包含了更多角度的信息, 且空间分布不受限制, 不需要布设、维护地面传感设备, 具有明显的经济优势, 可以为交通数据的采集提供有力支持.

我们将当前最有效的软件工具和自然语言处理技术引入到社会交通领域, 并对比、分析了多种算法在微博交通信息处理方面的技术性能, 将其中表现最优异的算法整合到整个系统中.

随着社交媒体交通信息数据库的逐步完善, 数据的实时性、准确性逐步提升, 数据量不断扩大, 可以为从事社会交通研究的学者提供必要的数据资源和可视化平台.

本文工作也存在一些值得进一步研究的问题. 由于社交媒体数据中人们对于同一个交通事件的描述是不同的, 其中也可能包含有一些不真实的信息, 如何将这些类似的信息融合, 是我们今后需要继续研究的问题. 此外, 社交媒体交通数据与传统交通数据的交叉验证与多元信息融合也可能成为剔除不真实社交媒体数据的重要手段, 值得进一步研究与探索.

References

1 Weng Jian-Cheng, Rong Jian, Yu Quan, Ren Fu-Tian. Optimization on estimation algorithms of travel speed based on the real-time floating car data. *Journal of Beijing University of Technology*, 2007, **33**(5): 459−464
(翁剑成, 荣建, 于泉, 任福田. 基于浮动车数据的行程速度估计算法及优化. 北京工业大学学报, 2007, **33**(5): 459−464)

2 Dong Jun-Yu. Study on Link Speed Estimation in Urban Arteries Based on GPS Equipped Floating Vehicle [Master thesis], Chongqing University, China, 2006.
(董均宇. 基于 GPS 浮动车的城市路段平均速度估计技术研究 [硕士学位论文], 重庆大学, 中国, 2006.)

3 Tao Han-Qing, Li Wen-Yong. Acquisition of turning vehicles information based on induction loop detector. *Journal of Guilin University of Electronic Technology*, 2008, **28**(5): 387−391
(陶汉卿, 李文勇. 基于感应线圈车辆检测器的车辆转弯信息获取. 桂林电子科技大学学报, 2008, **28**(5): 387−391)

4 Zhang Z, Yao D Y, Zhang Y, Hu J M. Mixed urban traffic data collection and processing with advanced information technologies. In: Proceedings of the 3rd China Annual Conference on ITS. Nanjing, China: Southeast University Press, 2007. 474−479

5 Wang Chuan-Tong. Study on Video-based Traffic Congestion Identification Technology of City Road [Master thesis], Chongqing University, China, 2010.
(王川童. 基于视频处理的城市道路交通拥堵判别技术研究 [硕士学位论文], 重庆大学, 中国, 2010.)

6 Li R M, Jiang C Y, Zhu F H, Chen X L. Traffic flow data forecasting based on interval type-2 fuzzy sets theory. *IEEE/CAA Journal of Automatica Sinica*, 2016, **3**(2): 141−148)

7 Shang J B, Zheng Y, Tong W Z, Chang E, Yu Y. Inferring gas consumption and pollution emission of vehicles throughout a city. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2014. 1027−1036

8 Lu Feng, Zheng Nian-Bo, Duan Ying-Ying, Zhang Jian-Qin. Travel information services: state of the art and discussion on crucial technologies. *Journal of Image and Graphics*, 2009, **14**(7): 1219−1229
(陆锋, 郑年波, 段滢滢, 张健钦. 出行信息服务关键技术研究进展与问题探讨. 中国图像图形学报, 2009, **14**(7): 1219−1229)

9 Zhang J P, Wang F Y, Wang K F, Lin W H, Xu X, Chen C. Data-driven intelligent transportation systems: a survey. *IEEE Transactions on Intelligent Transportation Systems*, 2011, **12**(4): 1624−1639

10 Wang F Y, Zhang J J, Zheng X H, Wang X, Yuan Y, Dai X X, Zhang J, Yang L Q. Where does AlphaGo go: from church-turing thesis to AlphaGo thesis and beyond. *IEEE/CAA Journal of Automatica Sinica*, 2016, **3**(2): 113−120

11 Wang F Y. Scanning the issue and beyond: crowdsourcing for field transportation studies and services. *IEEE Transactions on Intelligent Transportation Systems*, 2015, **16**(1): 1−8

12 Qiao F X, Zhu Q, Yu L. Social media applications to publish dynamic transportation information on campus. In: Proceedings of the 11th International Conference of Chinese Transportation Professionals. Nanjing, China: Southeast University Press, 2011. 4318−4329

13 Zeng K, Liu W L, Wang X, Chen S H. Traffic congestion and social media in China. *IEEE Intelligent Systems*, 2013, **28**(1): 72−77

14 Wanichayapong N, Pruthipunyaskul W, Pattara-Atikom W, Chaovalit P. Social-based traffic information extraction and classification. In: Proceedings of the 11th International Conference on ITS Telecommunications. St. Petersburg, Russia: IEEE, 2011. 107−112

15 Endarnoto S K, Pradipta S, Nugroho A S, Purnama J. Traffic condition information extraction & visualization from social media Twitter for Android mobile application. In: Proceedings of the 2011 International Conference on Electrical Engineering and Informatics. Bandung, Indonesia: IEEE, 2011. 1−4

16 Balagapo J, Sabidong J, Caro J. Data crowdsourcing and traffic sensitive routing for a mixed mode public transit system. In: Proceedings of the 5th International Conference on Information, Intelligence, Systems and Applications. Chania, Crete, Greece: IEEE, 2014. 1−6

17 D'Andrea E, Ducange P, Lazzerini B, Marcelloni F. Real-time detection of traffic from twitter stream analysis. *IEEE Transactions on Intelligent Transportation Systems*, 2015, **16**(4): 2269−2283

18 Zhang Heng-Cai, Lu Feng, Chen Jie. Extracting traffic information from massive micro-blog messages. *Journal of Image and Graphics*, 2013, **18**(1): 123−129
(张恒才, 陆锋, 陈洁. 微博客蕴含交通信息的提取. 中国图象图形学报, 2013, **18**(1): 123−129)

19 Zhang Heng-Cai, Lu Feng, Qiu Pei-Yuan. Extracting traffic information from micro-blog based on D-S evidence theory. *Journal of Chinese Information Processing*, 2015, **29**(2): 170−178
(张恒才, 陆锋, 仇培元. 基于 D-S 证据理论的微博客蕴含交通信息提取方法. 中文信息学报, 2015, **29**(2): 170−178)

20 Cui Jian, Feng Xuan, Zhang Zuo. Extraction and analysis system of traffic incident based on microblog. *Journal of Transport Information and Safety*, 2013, **31**(6): 132−135
(崔健, 冯璇, 张佐. 基于微博的交通事件提取与文本分析系统. 交通信息与安全, 2013, **31**(6): 132−135)

21 Xiong Jia-Xi. Civil Transportation Event Extraction from Chinese Microblogs Based on CRF [Master thesis], Shanghai Jiao Tong University, China, 2014.
(熊佳茜. 基于 CRF 的中文微博交通信息事件抽取 [硕士学位论文], 上海交通大学, 中国, 2014.)

22 Hasan S, Ukkusuri S V. Urban activity pattern classification using topic models from online geo-location data. *Transportation Research Part C: Emerging Technologies*, 2014, **44**: 363−381

23 Gkiotsalitis K, Stathopoulos A. A utility-maximization model for retrieving users′ willingness to travel for participating in activities from big-data. *Transportation Research Part C: Emerging Technologies*, 2015, **58**: 265−277

24 Gkiotsalitis K, Stathopoulos A. Joint leisure travel optimization with user-generated data via perceived utility maximization. *Transportation Research Part C: Emerging Technologies*, 2016, **68**: 532−548

25 Gu Y M, Qian Z, Chen F. From Twitter to detector: real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies*, 2016, **67**: 321−342

26 Kuflik T, Minkov E, Nocera S, Grant-Muller S, Gal-Tzur A, Shoor I. Automating a framework to extract and analyse transport related social media content: the potential and the challenges. *Transportation Research Part C: Emerging Technologies*, 2017, **77**: 275−291

27 Rashidi T H, Abbasi A, Maghrebi M, Hasan S, Waller T S. Exploring the capacity of social media data for modelling travel behaviour: opportunities and challenges. *Transportation Research Part C: Emerging Technologies*, 2017, **75**: 197−211

28 Cottrill C, Gault P, Yeboah G, Nelson J D, Anable J, Budd T. Tweeting Transit: an examination of social media strategies for transport information management during a large event. *Transportation Research Part C: Emerging Technologies*, 2017, **77**: 421−432

29 Xiong G, Zhu F H, Liu X W, Dong X S, Huang W L, Chen S H, Zhao K. Cyber-physical-social system in intelligent transportation. *IEEE/CAA Journal of Automatica Sinica*, 2015, **2**(3): 320−333

30 Wang F Y. Scanning the issue and beyond: real-time social transportation with online social signals. *IEEE Transactions on Intelligent Transportation Systems*, 2014, **15**(3): 909−914

31 Wang X, Zheng X H, Zhang Q P, Wang T, Shen D Y. Crowdsourcing in ITS: the state of the work and the networking. *IEEE Transactions on Intelligent Transportation Systems*, 2016, **17**(6): 1596−1605

32 HIT-SCIR. LTP [Online], available: http://ltp.readthedocs.io/zh_CN/latest/, July 12, 2016.

33 Řehuřek R, Sojka P. Software framework for topic modelling with large corpora. In: Proceedings of LREC 2010 Workshop New Challenges for NLP Frameworks. Valletta, Malta: University of Malta, 2010. 45−50

34 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. *The Journal of Machine Learning Research*, 2011, **12**: 2825−2830

35 Pan S J, Toh Z Q, Su J. Transfer joint embedding for cross-domain named entity recognition. *ACM Transactions on Information Systems*, 2013, **31**(2): Article No. 7

36 Zhou G D, Su J. Named entity recognition using an HMM-based chunk tagger. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002. 473−480

37 Morwal S, Jahan N, Chopra D. Named entity recognition using hidden Markov model (HMM). *International Journal on Natural Language Computing*, 2012, **1**(4): 15−23

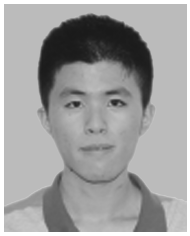
38 Wang Dan, Fan Xing-Hua. Named entity recognition for short text. *Journal of Computer Applications*, 2009, **29**(1): 143−145
(王丹, 樊兴华. 面向短文本的命名实体识别. 计算机应用, 2009, **29**(1): 143−145)

39 Peng F C, McCallum A. Information extraction from research papers using conditional random fields. *Information Processing & Management*, 2006, **42**(4): 963−79

40 Taku-ku. CRF++ [Online], available: http://sourceforge.net/projects/crfpp/files/, July 12, 2016.

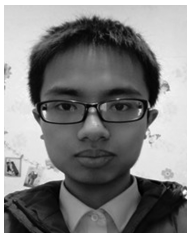
41 Lafferty J D, McCallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 2001. 282−289

42 Baidu. Baidu map API [Online], available: http://lbsyun.baidu.com, October 12, 2016.



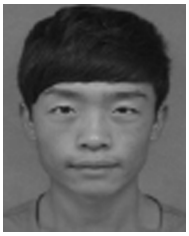
郑治豪 中南大学交通运输工程学院本科生. 主要研究方向为交通大数据.
E-mail: vincentzheng@csu.edu.cn
(**ZHENG Zhi-Hao** Undergraduate at the School of Traffic and Transportation Engineering, Central South University. His research interest covers transportation big data analysis.)

吴文兵 中南大学软件学院本科生. 主要研究方向为机器学习.
E-mail: SoundsOfLife@163.com
(**WU Wen-Bing** Undergraduate at the School of Software, Central South University. His main research interest is machine learning.)

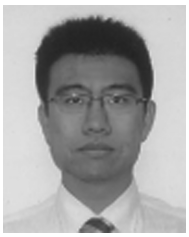


陈 鑫 中南大学信息科学与工程学院本科生. 主要研究方向为网络大数据挖掘与分析.
E-mail: 1774885528@qq.com
(**CHEN Xin** Undergraduate at the School of Information and Science and Engineering, Central South University. His research interest covers network big data mining and analysis.)

胡荣鑫 中南大学交通运输工程学院本科生. 主要研究方向为物流与电子商务.
E-mail: hurongxin@csu.edu.cn
(**HU Rong-Xin** Undergraduate at the School of Traffic and Transportation Engineering, Central South University. His research interest covers logistics and e-commerce.)



柳 鑫 中南大学交通运输工程学院本科生. 主要研究方向为城市公共交通规划、运营与管理.
E-mail: 1104130901@csu.edu.cn
(**LIU Xin** Undergraduate at the School of Traffic and Transportation Engineering, Central South University. His research interest covers urban public transport planning, operation, and management.)



王 璞 中南大学交通运输工程学院教授. 2010 年 5 月在美国圣母大学获得博士学位, 2010～2011 年于美国麻省理工学院进行博士后研究工作. 主要研究方向为交通大数据, 社会交通, 复杂网络. 担任 *IEEE Transactions on Intelligent Transportation Systems* 副主编, IEEE 智能交通系统学会–社会交通系统技术委员会 Co-Chair. 本文通信作者.
E-mail: wangpu@csu.edu.cn
(**WANG Pu** Professor at the School of Traffic and Transportation Engineering in Central South University. He received his Ph. D. degree in Physics from University of Notre Dame in 2010. From 2010 to 2011, he worked as a postdoctor researcher in the Department of Civil and Environmental Engineering in MIT. His research interest covers transportation big data, social transportation and complex networks. He is an associate editor of IEEE Transactions on Intelligent Transportation Systems and the Co-Chair of IEEE ITSS Social Transportation Systems Technical Committee. Corresponding author of this paper.)