基于 K-means 聚类算法的公交行程速度计算模型

高 曼 韩 勇 陈 戈 张小垒 李 洁
(中国海洋大学信息科学与工程学院 青岛 266100)

摘 要 定位与无线装置在公交系统中的广泛应用使得获取实时公交数据成为可能。为挖掘这些数据中蕴含的道路交通状况信息,提出了一种基于 K-means 聚类算法的数据融合模型,来计算相邻公交站点间的平均行程速度。首先对 K-means 聚类算法进行改进:(1)聚类数 K 不是预先设定的固定值,而是不重复样本数的平方根,不同路段不同时段 K 值不同;(2)初始聚类中心不是随机选取,而是根据 K 值按一定规则选取。其次利用改进的算法对样本数据进行聚类,然后对各类数据进行加权融合,计算出平均行程速度。最后通过折线图对青岛市 4 个城区的行程速度进行分析,挖掘交通流的演变规律。研究结果为交通管理、居民出行等提供了强有力的支持。

关键词 公共交通,平均行程速度,K-均值聚类算法,数据融合,数据挖掘
中图法分类号 TP301.6 文献标识码 A

Computational Model of Average Travel Speed Based on K-means Algorithms

GAO Man HAN Yong CHEN Ge ZHANG Xiao-lei LI Jie
(College of Information Science and Engineering, Ocean University of China, Qingdao 266100, China)

Abstract It is possible to retrieve real-time data using floating bus data acquisition system equipped with positioning and wireless communication apparatus. To explore traffic condition, a data fusion model based on the K-means clustering algorithm was put forward. The model was used to calculate the average travel speed between adjacent bus stops. At first, K-means clustering algorithm was improved: (1) the cluster number K is not predefined but the square root of non-identical sample size, and it is different at different sections and time; (2) the initial cluster center is not random but selected according to K. Then, the sample data were divided into K classes by the improved algorithm and the average travel speed was obtained by data fusion model. Finally, the average travel speed of four areas in Qingdao was shown by line charts to explore some evolution law of traffic flow. The research provides strong support for traffic management and residents travel.

Keywords Public transport, Average travel speed, K-means clustering algorithms, Data fusion, Data mining

1 引言

城市交通是城市社会生活、经济活动的纽带和动脉,对城市经济的可持续发展和人民生活水平的提高发挥着重要作用。公共交通作为城市交通一个重要的组成部分,也是城市居民正常出行的最重要的交通工具之一。近年来,城市规模不断扩大,机动车保有量迅速增长,造成了严重的道路拥堵,城市交通面临着日益严峻的挑战。行程速度和行程时间作为评价路段交通状况、拥堵水平的重要指标,是交通诱导、交通信息服务的重要基础[1-6]。

长期以来,关于行程速度和行程时间的研究多采用固定式交通采集数据(如感应线圈数据、微波检测数据、车牌自动识别数据)和移动式交通采集数据(如出租车 GPS 数据)[7]。文献[8]利用环形线圈检测数据结合随机服务系统相关理论建立了城市道路路段行程时间的动态计算模型;文献[9,10]分别利用线圈检测数据和出租车 GPS 数据建立不同模型来估计路段平均行程时间;文献[11]基于车牌识别数据研究城市道路行程时间的分布规律。文献[12]基于车牌识别数据提出计算道路行程速度的模型;文献[13]以微波检测数据为基础,构建了一种基于 K 近邻的非参数回归短时交通预测模型,实现了对路段行程速度的短时预测;文献[3,14]基于地点交通参数提出两种回归模型来估计主干路拥挤路段的行程速度,同时利用出租车 GPS 数据建立速度-时间积分与位置-时间插值改进模型来估计城市路段平均行程速度。然而,感应线圈数据、微波检测数据精度较低,并且作为截面数据难以估计路段区间的行程速度和行程时间;车牌自动识别数据虽然精度高,但是车牌自动识别系统目前只在我国个别城市中安装使用,不易于推广;出租车 GPS 数据受限于出租车运行的

本文受青岛市科技发展计划(13-1-3-117-nsh)资助。

高 曼(1988—),女,硕士生,主要研究方向为城市交通数据挖掘,E-mail:gaoman2014@163.com(通信作者);韩 勇(1969—),男,博士,教授,主要研究方向为虚拟地理环境及海洋地理信息系统;陈 戈(1965—),男,博士,教授,博士生导师,主要研究方向为卫星海洋遥感、海洋地理信息系统、虚拟现实;张小垒 男,博士生,主要研究方向为复杂网络的理论研究与实践探索;李 洁(1991—),女,硕士生,主要研究方向为 Web 前端可视化。

随意性，不能很好地反映固定路段交通流状态的周期性特征。

当前，大中型城市公交车都装有 GPS，因而能够获取公交车的实时运行数据。公交线路长期固定的强大优势使得公交车 GPS 数据具有明显的周期性，并且公交线路的设置遵循覆盖主要客流走廊、发挥公交主导作用的原则，因而能够更好地挖掘城市道路公交车流运行的固定模式。本文基于青岛市公交 GPS 数据进行研究，建立了一种基于 K-means 聚类算法的数据融合模型来计算公交平均行程速度，并对青岛市四大城区的交通状况做了分析，分析结果可为交管部门提供决策依据。

2 数据处理方法

2.1 行程时间计算

本文利用从相关单位获取的公交车 GPS 数据进行研究。数据获取的时间范围是 2014 年 7 月至 2015 年 7 月，包括 200 多条公交线路、4372 辆公交车，数据量约为 10 亿条，数据的属性结构如表 1 所列。

表 1 公交车 GPS 数据属性结构

公交线路名称	线路终点站名称	到站名称	到站时间	距离终点站数	公交车编号
26 路	南京路	栈桥	2014-07-14 20:00:07	19	DD102
28 路	台东一路	错埠岭	2014-07-14 20:00:10	7	DD110
…	…	…	…	…	…

由于数据量巨大，首先将数据按照公交线路名称建立索引，以便后续对数据的存取操作。路段划分是进行路段平均行程速度估计的前提，若将整条道路作为一个整体进行估计，则不能准确反映道路的交通状况。以公交站点作为路段划分的依据，计算经过相邻站点的公交车行程时间。计算流程如图 1 所示。

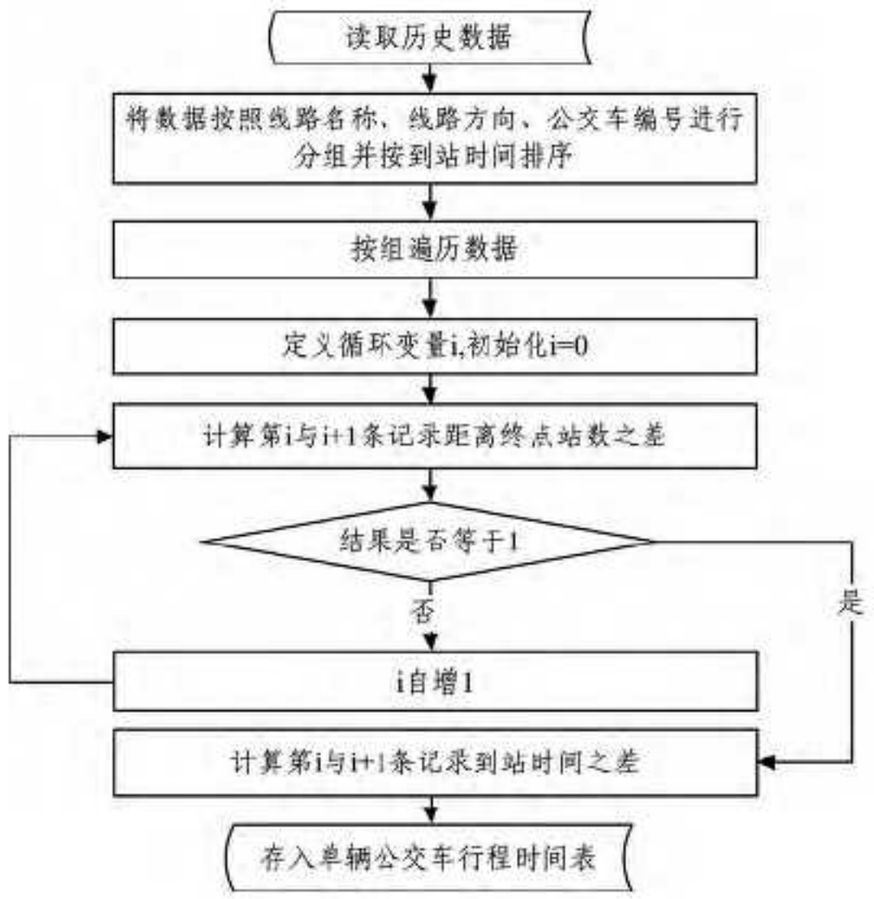

图 1 行程时间计算流程

计算步骤如下：

步骤 1 按照公交车编号对数据进行分组，将每组中的数据记录按公交线路、线路方向、到站时间多字段排序；

步骤 2 按组遍历数据，判断每组中相邻两条记录距离终点站数字段之差是否为 1，若是，则计算其到站时间差并存入单辆公交车行程时间数据表中；否则，继续往下遍历数据，直到遍历完数据；

步骤 3 计算完成，得到各路段的行程时间序列，为 3.2 节数据聚类做准备。

2.2 剔除噪声数据

计算得到的行程时间序列中存在很多的“噪声”数据，具体原因包括：

1）数据传输错误，由于 GPS 接收器工作异常导致数据接收异常；

2）数据异常值，由于道路事故、车辆故障、天气等外界因素导致数据明显偏离正常水平[11]。

文献[15]提出基于四分位数筛选法的异常数据值处理方法，将超出有效数据区间的数据作为异常值剔除。

$$Z=[Q_{0.25}-1.5R,Q_{0.75}+1.5R] \tag{1}$$

$$R=Q_{0.75}-Q_{0.25} \tag{2}$$

式中，Z 表示有效数据区间，$Q_{0.75}$、$Q_{0.25}$ 分别为上、下四分位值，R 表示四分位极差。

3 模型建立

3.1 历史数据的构建与更新

在不同时刻、同一道路上和同一时刻、不同道路上的交通流是不断变化的，此过程是一个复杂的动态选择过程[16]。但是，特定区域一般具有比较稳定的社会经济活动模式即上班、上学等活动在时间和空间上的分布具有一定的规律性，导致交通流仍然存在着很强的内在规律性，即同一空间位置、相同时段的行程时间序列具有极大的相似性[17]。

本文以周为单元，对每个单元（周一，周二，…，周日）的历史数据进行构建和更新，历史数据的时间范围选取最近的两个月。为了保持历史数据集合的时效性，当任意一天结束后，将该天数据补充为新的历史数据，同时剔除原来历史数据集合中最早一天的数据。将实时检测的交通数据整合为 10min 一个状态分析段，用 10min 交通参数平均值表示路段的平均交通状态[10]，本文选取人群活动集中的时间段即 6:00～22:00 进行研究。

3.2 K-means 聚类算法改进

K-means 聚类是 Mac Queen 提出的一种非监督实时聚类算法[18]。该算法理论严密，实现简单，能对数据集进行高效分类，但其也有自身无法克服的缺点，即聚类类数的主观性和对初始聚类中心的依赖性，主观设定聚类类数以及随机选取初始聚类中心都容易导致聚类结果不稳定。

本文的基本思路是在最小化误差函数的基础上将行程时间数据划分为 K 类，使得处于相同水平的行程时间归为同一类，减小过大或者过小的数值在后续加权计算中的权重，从而提高数据融合计算的精度。

为了减小聚类误差，本文对 K-means 聚类算法进行了改进，主要表现在以下两个方面。

1）改进聚类类数的设定方法

由于行程时间序列在不同路段不同时段的样本量大小不同，统一将其设为固定的类数可能会造成数据分类不均匀。为了避免上述问题，聚类类数根据样本量设定，将 K 设为 $\sqrt{m}$ 取整[19]（m 表示每个路段每个时段不重复样本 $T\{t_1,t_2,\cdots,t_n\}$ 的个数）；

2）改进初始聚类中心的选择方法

将样本数据按大小升序排列，平均分为 K 段，取 $T\{t_1, t_2, \cdots, t_n\}$ 中位于第 $0*\frac{m-1}{K}+1, 1*\frac{m-1}{K}+1, \cdots, (K-1)*\frac{m-1}{K}+1$ 处的数据为初始聚类中心，其目的是使初始聚类中心更大程度上接近真实聚类中心，减少迭代次数，克服了随机选取可能造成的聚类中心过于集中的问题。

例如，一个行程时间序列{113,167,191,113,168,194,81,114,168,196,246,81,128,169,201,252,82,141,169,201,252,82,141,169,201,260,100,142,173,206,261,103,143,173,210,266,107,145,174,212,275,145,174,216,279,145,175,216,149,220}，根据改进算法计算，K 值为 5，初始聚类中心为 82、128、167、191、212，只需迭代 3 次即可确定最终聚类中心 92、134、171、205、262。若初始聚类中心随机选取，所选数据大小可能会相近，如 128、261、145、206、220，则导致迭代增加，使得最终聚类中心计算的时间复杂度增加。

算法改进后的聚类结果示意图如图 2 所示。

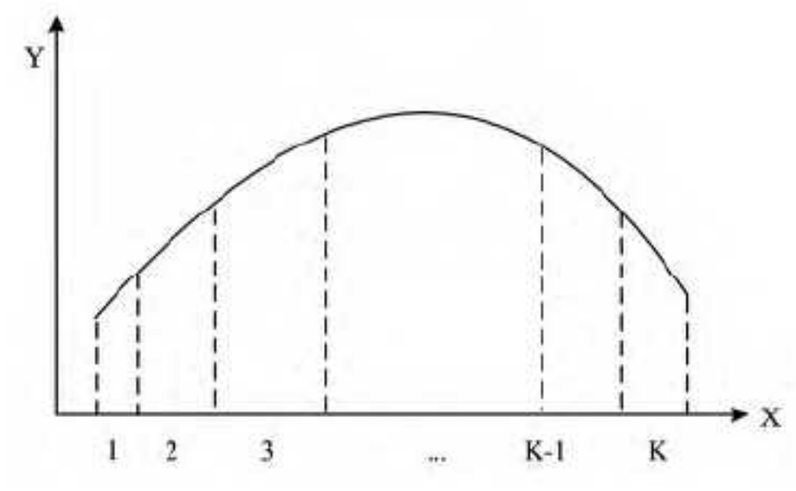

图 2　聚类结果示意图

X 轴表示数据的聚类类数 K，可统计出各类中包含样本的数量，K 是动态变化的，其大小由样本量决定；Y 轴表示落入每一类的样本值。改进聚类算法能够使大概率事件的数据相对集中，而使小概率事件的数据相对分散。

3.3　数据融合模型的建立

模型的核心思想是利用 K-means 聚类算法将同路段同时段的公交车行程时间序列分成预先设定的 K 类，然后采用频数加权算法对不同类的数据进行加权融合，得到平均行程时间，最后根据已知的站点间距离计算出平均行程速度。

频数加权法是在分类的基础上把所有样本数据按类归并，形成样本数据在各类中的分布，然后根据各组的样本数量确定该类数据的权重。本方法基于以下事实：大概率事件可靠度更高，因此出现频率越高的数据越接近于真实数据，所占权重应更大[20]。具体处理流程如下。

1）由 K-均值聚类算法对样本数据进行分类，根据聚类中心计算区间边界值：

$$(0, Z_0+\frac{Z_1-Z_0}{2}), [Z_0+\frac{Z_1-Z_0}{2}, Z_1+\frac{Z_2-Z_1}{2}), \cdots, [Z_{k-2}+\frac{Z_{k-1}-Z_{k-2}}{2}, Z_{k-1}+\frac{Z_k-Z_{k-1}}{2}), [Z_{k-1}+\frac{Z_k-Z_{k-1}}{2}, t_{\max}] \quad (3)$$

式中，$Z_0, Z_1, \cdots, Z_k$ 表示聚类中心，$t_{\max}$ 表示样本中最大数值。

2）确定落入每个区间的样本个数，计算各类所占权重：

$$W_x=\frac{N_x}{\sum_{x=1}^{k} N_x^2} \quad (4)$$

式中，N_x 表示落入区间 X 的样本数。

3）计算某一时间段的平均行程时间：

$$T_{avg}=\sum_{x=1}^{k} W_x \sum_{m=1}^{N_x} x_m \quad (5)$$

式中，x_m 表示落入区间 X 的各样本值。

4）因相邻公交站点之间的距离已知，故可以计算路段某时段的平均行程速度：

$$V_{avg}=\frac{S}{T_{avg}} \quad (6)$$

式中，S 表示相邻公交站点间距离。

4　结果分析

本文选取公交站点位于青岛市南、市北、李沧、崂山 4 个城区的行程速度数据，按照 1 小时间隔对数据进行采样，分别计算出工作日（周一到周五）与非工作日（周六与周日）各个城区的平均行程速度，绘制成时间序列图，如图 3、图 4 所示。

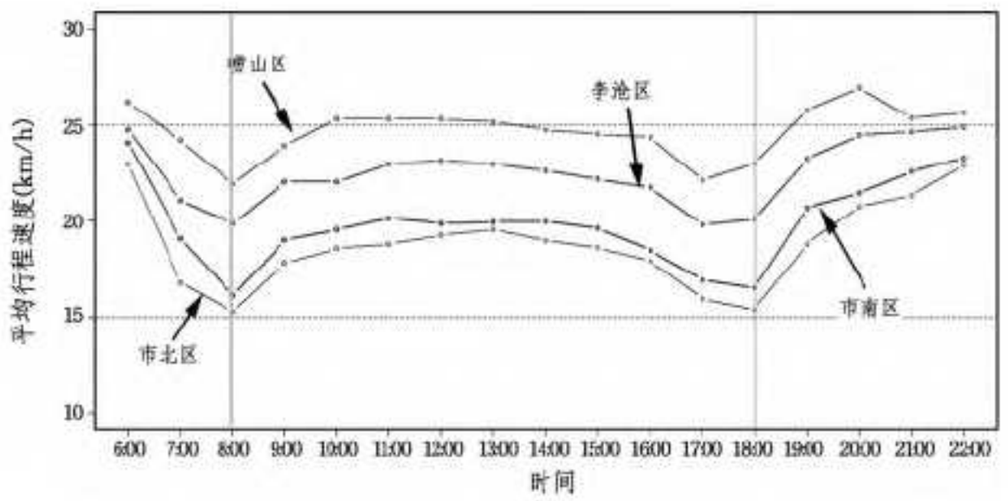

图 3　4 大城区平均行程速度时间序列图（工作日）

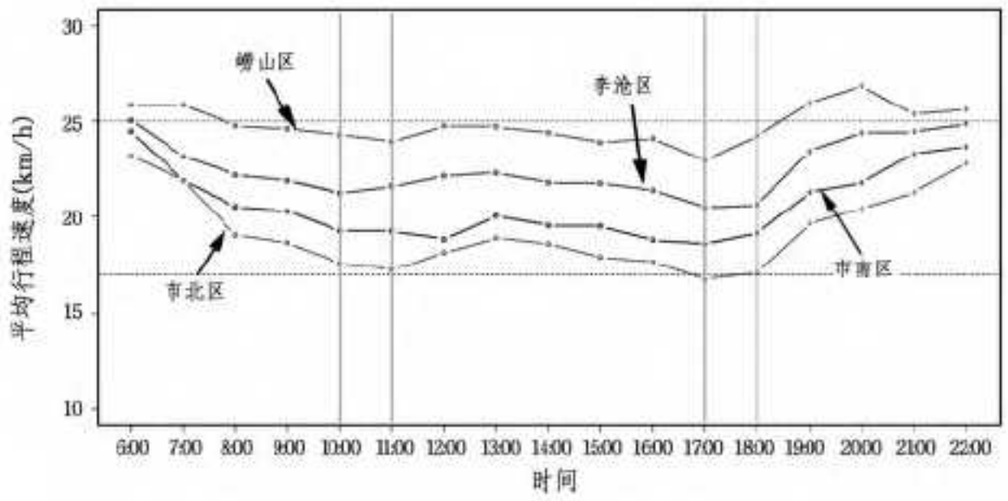

图 4　4 大城区平均行程速度时间序列图（非工作日）

由图可以发现以下规律：

1）不管是工作日还是非工作日，青岛四区中崂山、李沧、市南、市北的平均行程速度均依次减小，表明崂山区整体交通状况最好，市北区最差。

2）其中，在工作日，4 个城区都有明显的早晚高峰，8:00 左右达到早高峰，18:00 左右达到晚高峰，此时段平均行程速度最小；在非工作日，早晚高峰都不明显，相对于其它时段也没有明显拥堵。这是由于在工作日，人们的生活模式比较固定，固定的上下班、上下学时间导致在 8:00 及 18:00 道路车辆数量达到顶峰，从而造成交通拥堵的现象；而在非工作日，人们的活动丰富，车辆在道路上的时间相对分散，从而不会产生明显的早晚高峰现象。

3）在工作日，4 个城区公交车的平均行程速度多处于 15～25km/h 的范围；在非工作日，则处于 18～25km/h 的范围。由此可见，非工作日的交通状况整体要好于工作日的。

交通状况的影响因素可归结为城市道路设计和人员密度两方面。市北区、市南区是青岛市老城区，其道路设计狭窄，畸形路口较多，作为青岛市最初的经济中心，商业、娱乐活动

（下转第 439 页）

较句分析的难点。统计模型与规则分析方法相融合有可能提高比较句及其倾向性分析的结果和性能。

参 考 文 献

[1] 宋锐,林鸿飞,常富洋. 中文比较句识别及比较关系抽取[J]. 中文信息学报,2009,23(2):102-107,122

[2] 陈珺,周小兵. 比较句语法项目的选取和排序[J]. 语言教学与研究,2005(2):22-33

[3] 许国萍. 现代汉语差比范畴研究[M]. 上海:学林出版社,2007

[4] 车竞. 现代汉语比较句论略[J]. 湖北师范学院学报,2005(3):60-63

[5] 刘焱. 现代汉语比较范畴的语义认知基础[M]. 上海:学林出版社,2004

[6] 黄小江,万小军,杨建武,等. 汉语比较句识别研究[J]. 中文信息学报,2008,22(5):30-37

[7] 黄高辉,姚天昉,刘全升. 基于 CRF 算法的汉语比较句识别和关系抽取[J]. 计算机应用研究,2010,27(6):2061-2064

[8] 李建军. 比较句与比较关系识别研究及其应用[D]. 重庆:重庆大学,2011

[9] 杜文韬,刘培玉,费绍栋,等. 基于关联特征词表的中文比较句识别[J]. 计算机应用,2013,33(6):1591-1594

[10] 张辰,冯冲,刘全超,等. 基于多特征融合的中文比较句识别算法[J]. 中文信息学报,2013,27(6):110-116

[11] 王素格,王凤霞,宋雅. 基于序列模式的汉语比较句识别方法[J]. 山西大学学报(自然科学版),2013,36(2):172-179

[12] 王凤霞. 比较句识别及观点要素抽取方法研究[D]. 太原:山西大学,2013

[13] 周红照,侯明午,侯敏,等. 基于语义分类的比较句识别与比较要素抽取研究[J]. 中文信息学报,2014,28(3):136-141,149

[14] Jindal N, Liu Bing. Identifying comparative sentences in text documents[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006. New York: ACM Press, 2006: 244-251

[15] Jindal N, Liu Bing. Mining comparative sentences and relations [C]// Proceedings of the 21st National Conference on Artificial Intelligence, 2006. Boston: AAAI Press, 2006: 1331-1336

[16] Feldman R, Fresko M, Goldenberg J. Extracting product comparisons from discussion boards[C]// Proceedings of the 7th IEEE International Conference on Data Mining, 2007. Washington DC: IEEE Computer Society, 2007: 469-474

[17] Sun Jian-tao, Wang Xuan-hui, Shen Dou, et al. CWS: a comparative web search system[C]// Proceedings of the 15th International Conference on World Wide Web, 2006. New York: ACM Press, 2006: 467-476

[18] 苗传江. HNC(概念层次网络理论)导论[M]. 北京:清华大学出版社,2005

(上接第 424 页)

频繁,并且靠海的优越条件使得其成为游客集中的区域,故交通状况相对较差,李沧区、崂山区是新兴城区,其道路依据现代城市规格设计,相对宽阔,商业中心处于建设完善期,规划相对合理,故交通状况普遍良好。

结束语 公共交通是保障大多数市民交通出行的最佳方式,行程速度作为评价公共交通的重要指标,是实现公共交通调度的关键。本文采用公交车 GPS 数据,首先从聚类类数的确定与初始聚类中心的选择两方面对 K-means 聚类算法进行改进,然后提出了基于 K-means 聚类算法计算平均行程速度的数据融合模型,并以青岛市 4 个城区的实例进行了验证。结果表明应用该模型分析公共交通状况可以达到很好的效果,能够为人们的出行和交管部门的决策提供帮助。

参 考 文 献

[1] 赖云波. 公交浮动车到达时间实时预测研究[D]. 重庆:重庆大学,2011

[2] 沙云飞,曹瑾鑫,史其信. 基于 GPS 的路段旅行时间和速度估计算法研究[C]// 第一届中国智能交通年会论文集. 2005

[3] 姜桂艳,常安德,李琦,等. 基于出租车 GPS 数据的路段平均速度估计模型[J]. 西南交通大学学报,2011,46(4):638-645

[4] 杨兆升. 关于智能运输系统的关键理论——综合路段行程时间预测的研究[J]. 交通运输工程学报,2001,1(1):65-67

[5] Rice J, van Zwet E. A. Simple and effective method for predicting travel times on freeways [J]. IEEE Transactions on Intelligent Transportation System, 2004, 5(3): 200-207

[6] Kerner B, Demir C, Herrtwichr, et al. Traffic state detection with floating car data in road networks[C]// Proceedings of IEEE Conference on ITS. Vienna, Austria: IEEE, 2005: 700-705

[7] 聂庆慧,夏井新,张韦华. 基于多源 ITS 数据的行程时间预测体系框架及核心技术[J]. 东南大学学报(自然科学版),2011,41(1):199-205

[8] 姚丽亚,关宏志,魏连雨,等. 基于实时交通信息的行程时间估算及路径选择分析[J]. 公路交通科技,2006,23(11):86-90

[9] 张和生,张毅,胡东成. 路段平均行程时间估计方法[J]. 交通运输工程学报,2008,8(1):89-97

[10] 张和生,张毅,温慧敏. 利用 GPS 数据估计路段的平均行程时间[J]. 吉林大学学报(工学版),2007,37(3):533-537

[11] 柴华骏,李瑞敏,郭敏. 基于车牌识别数据的城市道路旅行时间分布规律及估计方法研究[J]. 交通运输系统工程与信息,2012,12(6):41-48

[12] 徐巍,黄浩斌,林建华,等. 基于智能卡口系统的道路行程速度计算与实际应用[C]// 第八届中国智能交通年会论文集. 2013

[13] 翁建成,荣建,任福田,等. 基于非参数回归的快速路行程速度短期预测算法[J]. 公路交通科技,2007,24(3):93-98

[14] 姜桂艳,李继伟,张春勤. 城市主干路拥挤路段基于地点交通参数的行程速度估计[J]. 吉林大学学报(工学版),2010,40(5):1203-1209

[15] 朱健梅. 竞争性运输通道选择的博弈模型研究[J]. 西南交通大学学报,2003,38(3):336-340

[16] 王殿海,曲昭伟. 对交通流理论的再认识[J]. 交通运输工程学报,2001,1(4):55-59

[17] 李明涛. 快速路短时间尺度地点交通参数多部预测方法研究[D]. 吉林:吉林大学,2009

[18] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报,2008,19(1):48-61

[19] 周世兵,徐振源,唐旭清. 新的 K-均值算法最佳聚类数确定方法[J]. 计算机工程与应用,2010,46(16):27-31

[20] 张原. 公交路网旅行速度估计方法[D]. 北京:北京交通大学,2012