软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(Suppl.(1)):87−99

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

基于路况相似性的城市公交车到站时间预测机制*

孙玉砚[1,2,3+], 刘 燕[4], 周新运[1], 孙利民[1]

[1](中国科学院 信息工程研究所 北京 100093)

[2](中国科学院 研究生院 北京 100049)

[3](信息安全国家重点实验室(中国科学院 信息工程研究所),北京 100093)

[4](北京大学 软件与微电子学院,北京 102600)

City Bus Arrival Time Prediction Based on Similarity of Road Conditions

SUN Yu-Yan[1,2,3+], LIU Yan[4], ZHOU Xin-Yun[1], SUN Li-Min[1]

[1](Institute of Information Engineering, The Chinese Academy of Sciences, Beijing 100093, China)

[2](Graduate University, The Chinese Academy of Sciences, Beijing 100049, China)

[3](State Key Laboratory of Information Security (Institute of Information Engineering, The Chinese Academy of Sciences), Beijing 100093, China)

[4](School of Software and Microelectronics, Peking University, Beijing 102600, China)

+ Corresponding author: E-mail: sunyuyan@iie.ac.cn

Sun YY, Liu Y, Zhou XY, Sun LM. City bus arrival time prediction based on similarity of road conditions. ***Journal of Software*, 2012,23(Suppl.(1)):87−99 (in Chinese).** http://www.jos.org.cn/1000-9825/12010.htm

Abstract: Currently, bus arrival time prediction studies have not yielded the best results under complex road conditions. A bus arrival time prediction mechanism based on the similarity of road conditions is proposed in this paper. By analyzing the past trip records of the bus, the historical traffic condition of each road section was estimated. The road condition for each trip records was extracted. In the operational phase, each road section's traffic condition was estimated by gathering bus locations continually. The mechanism predicted bus arrival time for the rest bus station by the most similar historical which was searched by the time, distance, and real-time road conditions. The evaluation result using historical data of real bus trips shows that the arrival time prediction of the mechanism is accurate—with an absolute percentage error around 5%.

Key words: arrival time prediction; road condition signature; *k*-nearest neighbor; similarity

摘 要: 目前已有多种公交车到站时间预测技术,但在城市复杂多变的道路交通环境下预测精确度不够理想,为此提出了一种基于路况相似性的城市公交车到站时间预测机制,通过分析公交车运行历史数据获取各个路段的历史路况评估值,提取历史行程记录的路况特征.在运行阶段,实时获取公交车运行数据并评估各个路段的路况,查找与当前时间、位置和路况特征最相似的历史行程记录,并据此预测公交车辆的到站时间.

关键词: 到站时间预测,路况特征,*k* 近邻,相似性

* 基金项目: 国家自然科学基金(60933011); 国家重点基础研究发展计划(973)(2011CB302902); 国家科技重大专项(2011ZX03005-006); 中国科学院先导专项课题(XDA06040100)

收稿时间: 2012-05-05; 定稿时间: 2012-08-17

公交车辆到站时间预测是实现城市公共交通系统信息化的重要内容,可对城市公共交通的发展起到积极的推动作用.精确的公交车到站时间预测信息可以减少乘客的候车时间,方便乘客换乘公交车,使乘客更合理地规划行程时间,提高公共交通的服务水平,吸引更多的出行者选择乘坐公交车.因此,公交车辆到站时间预测技术受到了国内外研究人员的极大关注.

目前在美国、英国、德国、日本等发达国家的城市,公交线路都有准确的运行时刻表,公交车一般能较为精确地控制好时间准时出发和到达公交站点.而在发展中国家,例如中国和印度的城市,存在市区人口密度大、乘客数量多、道路交通拥挤以及在城市化进程中常见的道路改建养护等问题.这些外在因素致使公交出行时间具有高度的不确定性,公众难以获得与出行密切相关的公交车辆实时运行信息,严重影响了公交客运的服务质量.

随着信息技术的迅速发展,先进的定位、检测、监控和通信技术广泛应用在国内外的公交系统,能够实时获得大量的道路交通信息和公交车辆位置信息,为精确预测公交车到站时间提供了基础.本文提出了一种基于路况相似性(similarity of road conditions,简称 SORC)的预测技术,主要贡献包括:提出了一种基于 *k*-means 聚类算法的路段交通状况计算方法;提出了一种基于 Tanimoto 系数的路况多元组相似度计算方法;实现了基于路况多元组相似性的公交车辆到站时间预测方法.

本文第 1 节介绍公交车辆到站时间预测技术的研究进展;第 2 节对无锡市实际公交车辆运行数据进行分析;第 3 节提出了道路交通状况的评估模型,详细讨论基于路况信息多元组相似性的预测机制;第 4 节进行算法参数实验分析和算法性能评估实验分析, 进一步验证预测机制的有效性;第 5 节对全文的工作进行总结.

1 相关工作

1994 年美国的 GPS(global positioning system)系统开始全面运作,Ravi[1]等人首次利用公交车辆的 GPS 位置数据(automatic vehicle location data,简称 VAL)进行到站时间预测.此后公交车辆到站时间预测技术受到了许多研究人员的关注,特别是中国和印度等发展中国家,提出了大量的预测技术.这些公交车辆到站时间预测技术可以分为机器学习、历史数据相似性模型、回归模型和时间序列模型等.

机器学习机制是在海量历史数据基础上训练预测模型,由预测结果选择、训练并确定最佳的训练函数.早期的机器学习技术主要采用神经网络训练预测模型,典型的例如 Chen[2],Ran[3].近年来也有很多研究使用支持向量机(support vector machine,简称 SVM)实现预测技术,例如印度的 Vanajakshi 和 Rilett[4],中国的 Chen[5],Wang[6]等.其中 Ran[3]使用的三层前馈(three layer feed forward)神经网络预测模型是典型的机器学习预测技术,采用利文贝格一马夸特最优化算法(Levenberg-Marquardt,简称 LM)为训练函数,采用感知器(perceptron weight and bias)作为学习函数,建立了公交车辆行程时间预测模型.它的优点是利用经验尝试选择较优的训练函数、学习函数以及一些参数可以达到一定的预测精度,缺点是实现复杂度较高,同时很难实现实时在线的训练和动态预测.

历史数据相似性的预测模型以大量历史数据为基础,假设交通模式具有循环变化的规律,以历史相同时期公交车的行程时间均值预测当前的行程时间.目前大多数城市的电子站牌显示到站时间是基于这种简单模型进行预测的.例如爱尔兰的 Coffey[7]利用都柏林市的公交车辆 GPS 定位数据、公交时刻表等信息,提出基于典型范例的车辆到达时间预测模型.其他的还有美国的 Lin 和 Zeng[8],印度的 Manolis 和 Kwstis[9]等.

回归模型如 Li[10]和 Tetreault[11],假设公交车运行时间受路段长度、交叉口数量、天气情况、交通流量、路段平均速度等多个变量影响,构造包含多个自变量的回归函数方程并将回归方程作为预测模型,根据自变量在预测时段的变化对因变量公交车运行时间进行预测.

时间序列模型如 Ishak[12]和 Zhu[13],假设公交车辆未来的运行时间会按照历史运行时间序列的变化趋势继续延续下去,其预测的准确性取决于所预测的行程时间变化规律与历史规律的匹配度.

本文提出的基于路况相似性的到站时间预测机制,是对历史数据相似性预测模型的一种扩展.在查找历史记录中最相似行程的过程中,除了使用传统的时间和地点作为相似特征,创新性的增加了公交车运行线路的路

段交通状况特征,即路况信息多元组,并提出了路况信息多元组相似性计算方法.利用实际数据测试结果显示,本文提出的机制预测到站时间精确度较高.

2 实际数据分析

我们选择了无锡市的 3 条具有代表性的公交线路,这 3 条公交线路的站点都在市区内,而且都经过无锡市火车站和一些繁华的商业街.公交线路在本文中编号分别为 1,2,3 路.

其中 1 路公交车有 44 个站点,2 路公交车有 46 个站点,3 路公交车有 36 个站点.每条公交线路每天平均有 20 辆公交车在运行,每辆公交车在线路始发站和终点站之间往返运行 4 次,其中从始发站开往终点站的方向称为下行线,从终点站开往始发站的方向称为上行线.我们获取了 2011 年下半年 3 条线路公交车辆的运行记录日志,公交车辆上搭载的信息采集设备每 15s 发送包含 GPS 定位信息的汇报数据见表 1.

Table 1 Report data of buses

表 1 公交车辆汇报数据

线路	1 路
车辆编号	苏 B-xxxxx
时间	2011/11/5 09:33:20.163
经度	31.52217
纬度	120.39561
速度	20
方位角	186
方向	上行

每个公交线路分别按上行和下行线路顺序依次排列公交站点,公交站点位置数据包括站名、线路、线路编号、序号、方向、距起点里程、经度纬度等.

运行日志数据库中错误的记录需要被预先去除,例如由于 GPS 定位错误导致其位置不在无锡市范围内的记录等.同时借助公交车内摄像头保存的视频记录校准了一些公交车辆的到站时间.

通过对获取的数据进行分析,本文得出了下列结论:

(1) 公交车辆在相同路段行驶耗费的时间在一天不同时段变化很大,同一个时间段运行耗费时间的最大值和最小值差值也很大.

图 1 是 2 路公交车在无锡市长江南路站到黄山路站路段运行耗费时间的平均值、最大值和最小值,横轴表示每天不同时间段.其他路段的运行耗费时间变化特征与该路段基本相同.明显可以看出上午 7 点左右和下午 17 点左右是交通高峰期,公交车辆运行时间比其他时间段多出很多.同一个时间段运行耗费时间的最大值和最小值之间的差值基本在 1 分钟以上,特别是高峰期差值通常有 3 分钟左右.

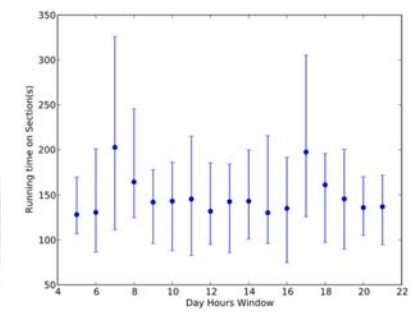

Fig.1 The running time on section between Changjiangnan-Road and Huangshan-Road

图 1 长江南路站至黄山路站路段公车运行耗费时间

(2) 在相同路段上连续运行的公交车运行耗费时间有密切关系,前方公交车的运行时间与后方公交车的运行时间相似的概率很大.

图 2 是前后连续两辆公交车运行在同一路段,运行耗费时间差值百分比的概率分布.横轴表示前车与后车运行时间差值的百分比,每个柱状图代表差值 15%区间的分布概率.前后连续的两辆公交车在同一个路段,发车的时间间隔分别在 0~10 分钟、10~20 分钟、20~30 分钟和 30~40 分钟共 4 个时间段内.图 2 中显示的耗费时间差值百分比分布呈现典型的瑞利分布,下一时刻公交车的运行时间与当前的运行时间相同的概率很大.

另外在图 2 中可以看到运行时间差值百分比在±10%~30%左右的概率也较高,出现这种情况的一个主要原因是由于任意两个公交站点之间的路段中基本都有红绿交通信号灯,等待交通信号变化的时间容易造成前后两辆公交车辆运行时间差几十秒.

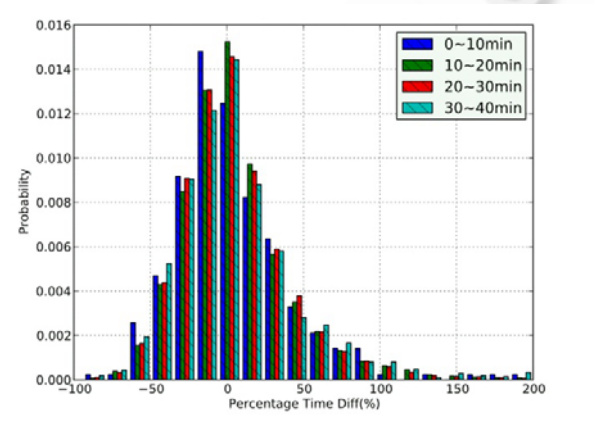

Fig.2 The probability distribution of the Percentage time diff between Continuous bus

图 2 连续公交车耗费时间差值百分比的概率分布

(3) 每天的相近时刻公交车辆在同一路段运行耗费的时间很相近.

图 3 统计了两辆公交车辆的发车时间间隔与相应到达终点站耗费时间的关系分布.横轴表示公交车辆发车时间的间隔,单位是小时;纵轴是在路段运行耗费的时间,单位是秒.两个公交车行程记录分别取自两个星期的数据,从始发站发车的时间不在同一天但是星期数相同.

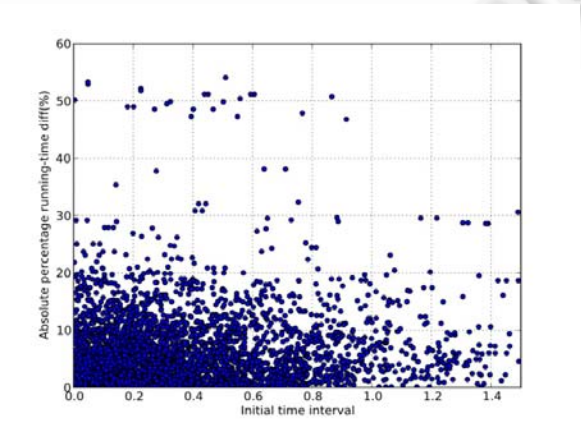

Fig.3 The running-time distribution of different initial time on the same road section

图 3 同一路段不同时刻运行耗费时间分布

在图 3 中可以看到绝大部分点都分布在左下角,发车时间间隔越小,其最终到达终点站的时间差值越小.每天相近时刻公交车辆在同一路段运行耗费的时间变化很相近,可以利用这种周期变化的特性,以历史最相近时期公交车的记录预测当前的行程时间.

(4) 每天相近时刻发车的两辆公交车辆,在发车时刻,其前方路段的交通状况越相近,则两辆公交车辆运行耗费时间也越相近.

图 4 是在不同周期相近时刻从始发站发车的公交车辆运行记录,列举了其前方路段交通状况相似度与公交车到达终点运行耗费时间的分布.纵坐标表示两辆公交车到达终点总共的运行时间差值百分比;横坐标表示公交车始发时刻该线路所有路段的交通状况相似度.

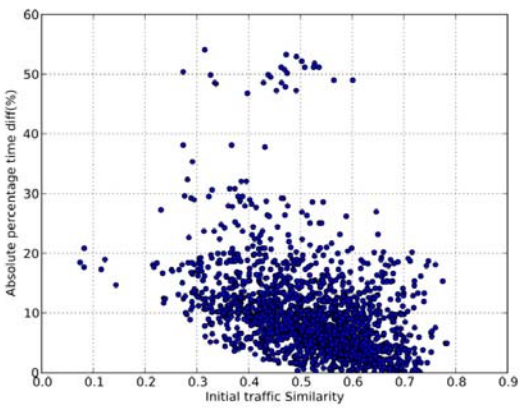

Fig.4 The running-time diff distribution of different initial traffic similarity

图 4 始发交通状况相似度与行程运行时间关系

两个公交车行程记录分别取自两个星期的数据,从始发站发车的时间不在同一周但是星期数相同,并且发车时刻的周期间隔在 1 分钟之内,路段交通状况相似度用 Taminoto 系数[14]计算公交车辆的运行速度计算,其计算公式如下,其中 x_i 和 y_i 分别是在 x 时刻和 y 时刻,公交线路第 i 站点到第 i+1 站点路段公交车辆平均运行耗费时间.

$$T(x,y)=\frac{\sum x_i y_i}{\sum x_i^2+\sum y_i^2-\sum x_i y_i} \tag{1}$$

从图 4 可以看到相似度超过 50%的行程运行时间差值百分比大部分分布在 10%以下.两辆公交车辆在相似时刻始发并且前方交通状况也相近,两辆公交车运行的初始环境更接近,最终运行过程中的交通状况变化也更相似,相应到站时间差别较小.在此假设道路的交通状况变化情况也具有历史循环的特性.

通过分析无锡市三条公交线路公交车的历史数据,我们可以看出影响公交车到站时间的因素包括时间段和各个路段的交通拥塞情况.每星期同一天,公交发车时间越相近,运行到终点耗费时间越相近;每星期同一天相近时刻,公交发车时刻前方路段交通状况相似性越高,运行到终点耗费时间越相近.

3 基于路况相似性的到站时间预测机制

基于路况相似性 SORC 的公交车到站时间预测机制,其基本思想是在时间段和地点最接近的历史行程记录中,查找道路交通状况最近似的行程记录,最后用最相似历史行程的到站时间预测当前的公交车辆的到站时间.

基于 SORC 的公交车到站时间预测技术分为初始化阶段和运行阶段两部分.在初始化阶段,利用 k-means 聚类算法评估公交线路所有路段在历史不同时间段的交通状况,为历史行程数据库中的每一条行程记录生成一个对应的路况元组;在运行阶段,根据历史路段交通状况评估结果生成实时路况评估标准,实时计算公交线路的各个路段的交通状况,利用 k-近邻算法查找历史数据中最相似时间段、距离的候选行程记录,并在候选行程记录中挑选路况元组与实时路况元组相似性最高的行程记录,利用最相似行程记录的到站时间均值预测公交

车辆到站时间,如图 5 所示.

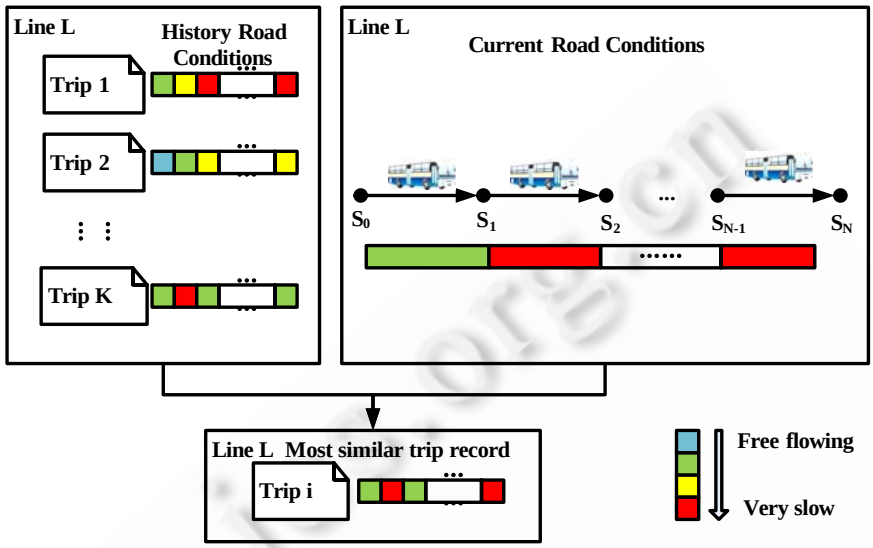

Fig.5 The bus arrival time prediction based on similarity of road conditions

图 5 基于路况相似性的到站时间预测机制示意图

3.1 历史路段交通状况评估

根据公交车路线数据,每一组相邻站点(S_i,S_j)划分为一个独立路段 $R(i,j)$.并查找所有在两个站点之间运行的所有公交线路[L_a,L_b…],查找到的公交线路集合称为复用线路集合,建立路段合和复用线路的映射关系〈$R(I,j)$,[L_a,L_b…]〉.

根据获取的公交车运行位置记录日志和公交线路信息,我们可以生成每个路段的运行历史信息,以一个路段 $R(i,j)$为例每一条运行历史信息包括到达站点 i 的时间 t_i,离开时间即到达站点 j 的时间 t,在此期间的位置记录 P_j,公交线路编号 L 等,其中 L 属于复用线路集合.路段的历史交通状况计算是面向所有路段复用线路集合,并不针对单独一条公交线路.

利用获取的交通状况特征计算路段的历史交通状况可以使用聚类算法进行划分,本文采用了比较常用的 k-means 聚类算法[15].k-means 算法输入聚类个数 k,以及包含 n 个数据对象的数据库,输出满足方差最小标准的 k 个聚类.k-means 聚类算法具有收敛较快的特点,各聚类本身尽可能的紧凑,而各聚类之间尽可能的分开.

目前从一个路段的公交车运行历史信息可以简单获取的交通状况特征是公交车在该路段运行过程中 GPS 采样的瞬时速度,以及公交车运行耗费时间.k-means 聚类算法根据平均瞬时速度和运行耗费时间这两个特征对历史运行数据进行聚类划分.当 k 取值为 4 时,长江南路站至黄山路站路段的运行历史信息聚类结果如图 6 所示,横轴表示运行耗费时间,纵轴表示公交车在此期间 GPS 设备采样的所有瞬时速度均值,每个类的中心点用圆点表示.

根据 k-mean 聚类的结果将历史数据分为 k 个聚类之后,按照中心点的运行耗费时间从小到大顺序排列 k 个聚类,然后每个聚类按照顺序赋权值从 1~k.每个聚类的权值就是该聚类中每条路段运行历史记录对应的交通状况评估值.最后以路段运行历史信息中的离开时间 tj 为索引建立该路段的历史交通状况评估值数据库.

根据获取的公交车运行位置记录日志和公交线路信息,为每一条公交线路生成历史行程数据库.首先获取一个从线路始发站出发,到达终点站结束的完整的行程记录,该记录包含了到达每一个线路站点的时间.然后将这个一个完整行程记录扩展为多个行程记录,以该线路中不同的站点作为起点到达线路终点站作为扩展的行程记录.最终生成一个以时间和起始站点为索引的历史行程数据库.

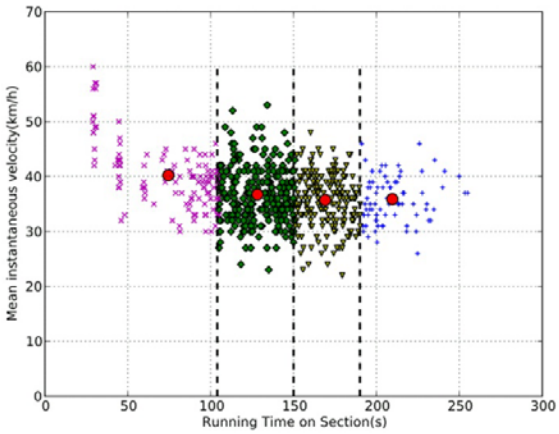

Fig.6 *k*-Mean clustering results on historical traffic conditions of same section

图 6 路段历史交通状况 *k*-mean 聚类结果

最后需要为历史行程数据库中的每一条记录生成一个对应初始的交通状况信息,以方便实时运行阶段查找.每个历史行程数据库的记录都是以起始站点和起始时间作为索引,则初始的交通状况信息通过查询该线路路段的历史交通状况评估值数据库获得该起始时间的交通状况评估值元组.

假设公交线路 L 一共有 N+1 个站点,N 个路段,则以 N 个路段的交通评估值组成的 N 维元组[RC_{1x}, RC_{2x},…,RC_{Nx}]表示始发时间为 x 的行程记录其始发时刻的道路交通情况.在 x 时刻每个路段的路况通过查询该路段的历史交通状况评估值数据库获得.

3.2 实时路段交通状况评估

利用路段的历史交通状况评估结果,可以获取交通状况的评估标准,为接下来的实时交通状况评估提供基础.在图 5 中可以直观看出分析路段的分类主要依据是运行耗费时间,在图中用 3 条虚线划分了 4 个区间,每个区间大致分别包含了一个聚类.

实时路况评估标准就是根据路段的运行耗费时间划分为 k 个区间,其中每个区间的上界是该区间聚类中运行耗费时间最大值与下一个区间运行耗费时间最小值的均值,区间下界是该区间聚类中运行耗费时间最小值与上一个区间运行耗费时间最大值的均值.第 i 个聚类 C_i 对应的区间[$left_i$,$right_i$]计算计算公式如下,其中 T_r 是运行耗费时间:

$$left_i=\frac{\max\{T_r \mid T_r\in C_{i-1}\}+\min\{T_r \mid T_r\in C_i\}}{2}$$
$$right_i=\frac{\max\{T_r \mid T_r\in C_i\}+\min\{T_r \mid T_r\in C_{i+1}\}}{2} \quad (2)$$

在运行阶段首先建立公交线路所有路段的实时交通状况信息数据库.当接收到公交车到站信息,则立刻更新该公交车刚刚行驶过的路段的交通状况.获取该路段的交通状况评估区间信息,根据该公车在路段上运行耗费时间判断属于哪一个区间 $i(i\in K)$,则该路段的交通状况就更新为该区间的权值.

算法. 实时道路交通状况评估算法的伪码描述.

Real-Time traffic evaluation

输入:

L:公交线路编号

R:当前路段

t:当前时间

TR:当前公交车在路段运行耗费时间

输出:路段 R 当前时刻道路交通状况评估值

过程:

1: 根据当前时间 t 和路段 R 在关系数据库中查找交通状况评估标准集合 S

2: 如果集合 S 是空的,返回 0

3: 否则从集合 S 中按顺序选择一个区间标准 C_i

4: 获取 C_i 的运行时间区间最小值 $left_i$ 和最大值 $right_i$,以及对应的交通状况评估值 I

5: 如果 $left_i<Tr<right_i$,返回 i

6: 返回 0

3.3 基于SORC的到站时间预测算法

在运行阶段会实时接收到各路公交车的汇报信息,对这些信息进行处理之后如果发现有公交车到站,则除了更新上一路段的交通状况之外,启动到站时间预测机制,预测公交车辆到达该公交线路 L 后续站点的时间.

预测算法首先使用 K 近邻算法(K-Nearest Neighbor,简称 KNN)[15]以二元组$\langle t,d\rangle$在历史行程数据库中挑选最相似的行程记录,其中 t 是当前时间,d 是公交车到达的站点序号.目前有很多种数据结构可以实现 K 近邻算法,例如 kd 树、B 树和 R 树等,在本文中使用 R 树[16]作为 k-近邻算法的数据结构,索引值分别是行程记录的时间和距离.针对当前行程的时间 t 和站点 d 序号在 R 树中选择 k 个最相似的候选历史行程.时间 t 以星期为周期,其计算公式如下,其中 Weekday 是星期数,Hour,Minute 和 Sencond 分别是当时的小时、分钟和秒数.

$$t = Weekday \times 24 + Hour + \frac{Minute}{60} + \frac{Second}{3600} \tag{3}$$

其次从实时交通状况信息数据库中获取公交线路 L 的路况元组[$RC1t$,$RC2t$,…,$RCNt$],以 $RC(L,t)$表示.分别获取 k-近邻算法获取的 k 个候选历史行程的历史路况元组,采用 Tanimoto 相似性系数[17]计算获选历史行程的路况元组与当前路况元组的相似性,其计算公式同公式(1).根据计算的相似度选择 m 个路况元组相似度最高的行程.

最后计算 m 个最相似行程中记录的到达公交线路 L 每个站点时间的平均值,以此作为到达该站点时间的预测结果.

基于路况相似度 SORC 的历史趋势预测技术完整算法如下所示.

算法. 基于路况相似度到站时间预测算法的伪码描述.

Arrival time prediction based on SORC

输入:

T:行程历史数据

L:公交线路编号

k:候选相似行程的数目

m:相似行程的数目

d:当前站点序号

t:当前时间

V:当前路况多元组

输出:预测的到达站点时间列表过程:

1:根据 L 确定当前预测需要的 R 树

2: 如果 R 树是空的，返回失败

3: 否则利用 R 树的 KNN 搜索算法获取与当前行程 d 和 t 相近的 k 个最相似的候选行程

4: 获取当前线路 L 的路况元组 $RC(L,T)$

5: 获取 k 个候选行程的路况多元组记录集合 S

6: 在 S 中选择 m 个与当前路况元组 V 相似度最高的行程

7: 计算 m 个最相似行程的每个站点到站时间均值

8: 返回

4 实验及结果分析

本文设计和实现了公交车辆到达时间实时预测系统,采用了基于 SORC 的到站时间预测算法.系统包括在公交车辆上的智能采集设备、处理中心、显示公交车辆到达时间的智能公交站牌和手机注册用户,系统的结构示意图如图 7 所示.

处理中心负责接收到城市内所有公交车辆智能采集设备发送的公交车辆信息和位置信息,同时运行基于SORC 的到站时间预测算法预测公交车辆到达公交站点的时间,并将预测结果推送到相应的智能公交站牌和手机短信注册用户.智能公交站牌负责显示接收到的公交车辆到站时间.

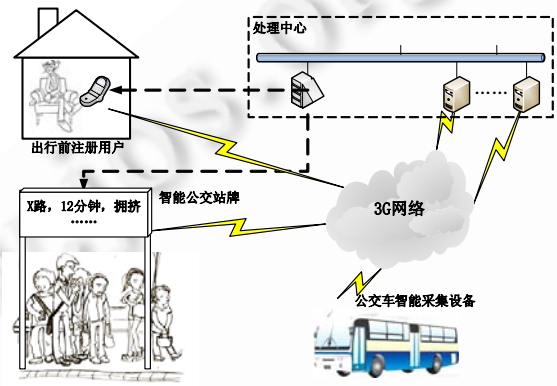

Fig.7 Real-Time bus arrival time prediction system

图 7 公交车辆到站时间实时预测系统

本实验将从无锡公交公司获取的 3 个公交线路公交车行程记录日志按照时间分为两个部分,前一段时间的记录做为算法使用的历史参考数据,后一段时间的记录作为检验预测精度的比较数据.

公交车到站时间预测精度评估采用绝对百分误差(absolute percentage error,简称 APE)和平均累计绝对百分误差(mean absolute percentage error,简称 MAPE),其计算公式为:

$$APE=\frac{\left|\sum_{i=1}^{N}E(i)-\sum_{i=1}^{N}T(i)\right|}{\sum_{i=1}^{N}T(i)} \tag{4}$$

$$MAPE=\frac{\sum_{i}^{N}\frac{\left|E(i)-T(i)\right|}{T(i)}}{N} \tag{5}$$

其中,N 是公交线路 L 的路段总数目,$T(i)$是公交车在第 i 个路段的实际运行时间,$E(i)$是算法预测在路段 i 运行需要的时间.绝对百分误差 APE 衡量公交车辆到达终点站的时间预测精度,而平均累计绝对百分误差 MAPE 衡量公交车到达线路每一站的综合时间预测精度.

4.1 算法参数设置

路段交通状况评估采用的 k-means 算法采用不同的 k 值其预测精度会有变化,如图 8 所示.在预测算法其他参数不变的情况下,当 k 取值为 2 时 APE 值明显偏高,而当 k 取值从 3~7 逐渐增大,预测精度 APE 值相差不大.

由图 8 可以看出,当路段交通状况划分为 2 个级别时,计算得到的评估值元组并没有很好的体现当时的路况特征,基于路况相似性的预测算法效果不佳;而如果路况划分比较细,例如划分为 3 个以上时,基于路况相似性的预测结果会比较好.另外当 k 取值大于 3 的时候预测结果的 APE 值相差不大,因此,为了提高计算效率,k-means

算法 k 值可以选择 3.

基于 SORC 的到站时间预测算法选择候选行程采用的 KNN 算法,使用不同的 K 值其预测精度会有变化.如图 8 所示,最相似行程选择 m 值取值为 4.当 K 值取 5 时 APE 值最低为 5.4%左右;当 K 的取值大于 5 时,APE 值随着 K 值增大逐渐提高,因此 KNN 算法 K 值可以选择 5.

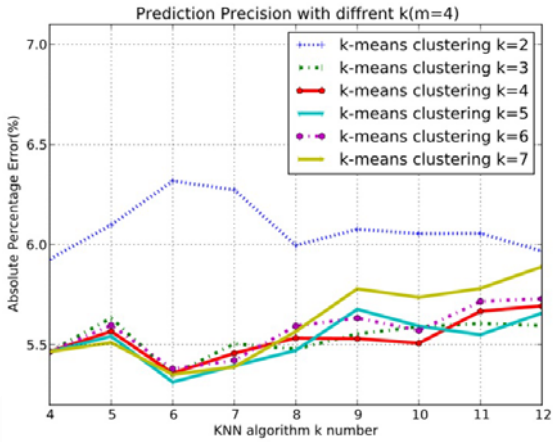

Fig.8 The prediction accuracy under different k-mean and KNN parameter

图 8 k-means 算法和 KNN 算法取值的预测精度

到站时间预测算法在 K 个候选行程中选择 m 个路况最相似的行程作为预测参考,因此在 K 值确定的情况下采用不同的 m 值其预测精度会有变化,如图 9 所示.在图 9 中 KNN 算法 K 取值为 6,m 取值范围从 1~5,可以看出 m 取值为 4 的时候预测算法的 APE 值最低.

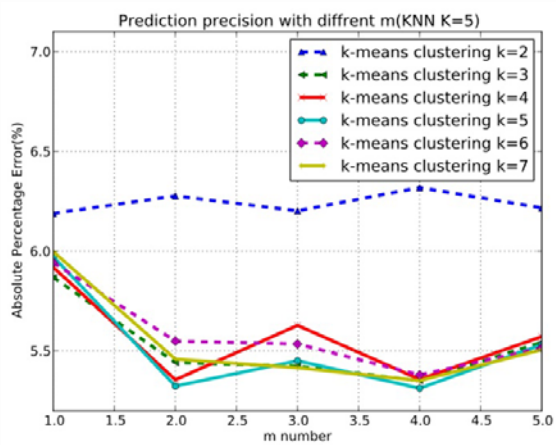

Fig.9 The prediction accuracy under different k-mean parameter

图 9 预测算法 m 取值的预测精度对比

4.2 算法性能分析

为了验证提出的预测算法性能,本文选择相关工作中 Ran[3]在博士论文中提出的基于神经网络的机器学习预测算法(ANN),和 Coffey[7]提出的基于典型范例的历史趋势预测方法(history)作对比.因为这两种算法的应用背景与本文相同,都是在应用在城市公交车系统.机器学习预测模型采用 3 层 18×37×15 的神经网络,利用 MATLAB 的神经网络工具箱进行网络训练,训练函数采用利文贝格-马夸特最优化算法,学习函数采用感知器(perceptron weight and bias).基于典型范例的历史趋势预测算法其范例的时间长度为 5 分钟.3 种预测方法都是

从公交车离开始发站的时刻开始预测公交车辆到达终点站的时间以及停靠中间站点的时间.

图 10 是机器学习预测方法、传统的到站时间预测算法和基于 SORC 的到站时间预测算法方法分别在 3 条公交线路上下行路线预测结果与实际运行时间的绝对百分误差 APE.图 10 中基于 SORC 的到站时间预测算法 APE 值都低于 6%,在大部分情况下都低于另外两种预测算法.

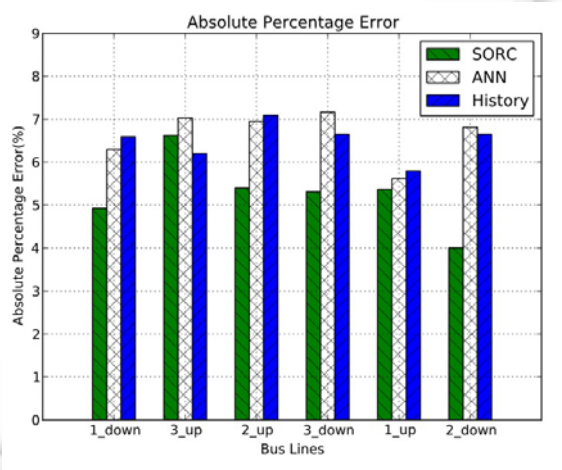

Fig.10 Absolute percentage error of prediction under different mechanism

图 10 不同预测算法 APE 比较

图 11 是 3 种预测方法与实际运行时间的平均绝对百分误差 MAPE.图中基于 SORC 的到站时间预测算法 MAPE 值都稳定在 10%左右,低于其他两种预测方法的 MAPE 值.这说明基于 SORC 的到站时间预测算法预测公交线路中每一站的到站时间都比较精确,平均误差在 10%左右.

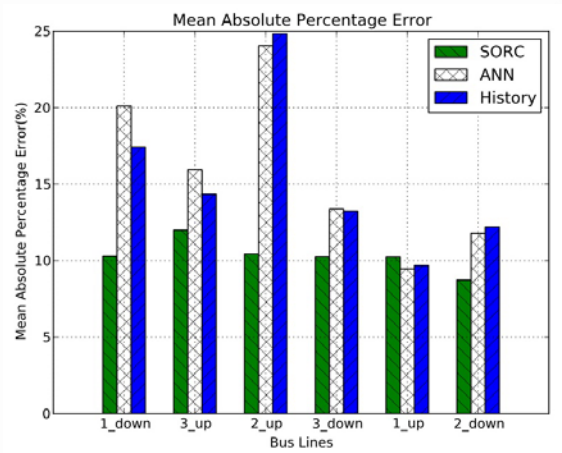

Fig.11 Mean absolute percentage error of prediction under different mechanism

图 11 不同预测算法 MAPE 比较

基于 SORC 的到站时间预测算法需要存储大量的历史行程记录,存储 3 条公交线路一个月的行程记录大概需要 500M 存储空间.基于 SORC 的到站时间预测执行时间主要包括了 KNN 算法查询时间和基于 Tanimoto 相似度系数的计算和排序时间.由于基于 Tanimoto 相似度系数的计算和排序是在 K 个历史路况元组上进行,因此两者的执行时间都取决于 KNN 算法的 K 值.

图 12 是在一台处理器为 2.93G 双核的 PC 机上运行基于 SORC 的到站时间预测算法执行的结果分析,数据库存储了存储 3 条公交线路一个月的行程记录.可以看出预测算法的查询执行时间随着采用的 KNN 算法 *K* 值增大呈现出线性增长的趋势.其他两种算法的预测查询的执行时间都在毫秒级,而基于 SORC 的到站时间预测算法每次预测的执行时间相对比较长,但每小时能为 3 万多辆次公交车提供到站时间预测,基本能够满足应用需求.另外基于 SORC 的到站时间预测算法可以在多台服务器存储不同的公交线路历史行程数据库,通过分布式查询处理提高预测计算速度.

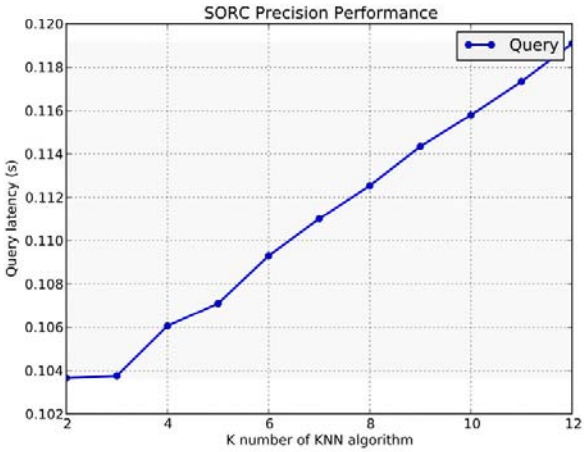

Fig.12 The execution time of bus arrival time prediction system based on SORC

图 12 基于 SORC 的预测算法执行时间

5 总 结

本文提出一种基于路况相似性的城市公交车到站时间预测机制,主要贡献是提出了一种基于 *k*-means 聚类算法的路段交通状况计算方法,对公交车辆历史运行数据分析并提出评估标准,在运行阶评估公交线路各个路段的实时交通状况;提出了一种基于 Tanimoto 系数的路况元组相似度计算方法,并利用该路况相似度,在最相似时间、距离的候选历史行程记录中,挑选路况最相似的行程记录,最后预测公交车辆到站时间.

虽然本文提出预测模型取得了比较精确的预测结果,但仍然存在一些问题和改进之处.路段交通状况信息可以直接从门户地图网站或者市政信息服务网站获取,简化计算复杂度.该模型可以进一步增加一些潜在的相似性特征,例如天气情况以及路段交通信号灯的情况,但同时会增加内存空间的使用和预测算法的运行时间.另外因此可以考虑利用公交车内的摄像头和车站的摄像头预测上下车乘客数目和公交车数目,提高到站时间预测精度.

References:

[1] Ravi K, Demetsky M. Modeling schedule deviations of buses using automatic vehicle location data and artificial neural networks. Trans. Research Record, 1995. 44−52.

[2] Chen D, Zhang K, Liao T. Practical travel time prediction algorithms based on neural network and data fusion for urban expressway. In: Proc. of the 6th Inte'l Conf. on Natural Computation (ICNC). 2010. 1754−1758.

[3] Ran J. The prediction of bus arrival time using automatic vehicle location systems data. Texas A&M University, 2004.

[4] Vanajakshi L, Rilett L. Support vector machine technique for the short term prediction of travel time. IEEE Intelligent Vehicles Symp. 2007, 600–605.

[5] Chen P, Yan X, Li X. Bus travel time prediction based on relevance vector machine. In: Proc. of the IEEE Information Engineering and Computer Science (ICIECS 2009). 2009. 1–4.

[6] Wang J, Chen X, Guo S. Bus travel time prediction model with v-support vector regression. In: Proc. of the IEEE Intelligent Transportation Systems, ITSC 2009. 2009. 1−6.

[7] Coffey C, Pozdnoukhov A, Calabrese F. Time of arrival predictability horizons for public bus routes. In: Proc. of the 4th ACM SIGSPATIAL Int'l Workshop on Computational Transportation Science, CTS 2011. 2011. 1−5.

[8] Lin W, Jian Z. An experimental study on real time bus arrival time prediction with GPS data. Transportation Research Record, 1999, 101−109.

[9] Manolis K, Kwstis D. Intelligent transportation systems□□travelers information systems the case of a medium size city. In: Proc. of the IEEE Int'l Conf. on Mechatronics 2004. 2004. 200−204.

[10] Li F, Yu Y, Lin H. Public bus arrival time prediction based on traffic information management system. In: Proc. of the IEEE Int'l Conf. on Service Operations, Logistics, and Informatics, SOLI 2011. 2011. 336−341.

[11] Tetreault P, Eigeneidy A. Estimating bus run times for new limited stop service using archived avl and APC data. Transportation Research Part A: Policy and Practice, 2009,44(6):390−402.

[12] Ishak S, Al-deek H. Performance evaluation of short-term time-series traffic prediction model. Journal of Transportation Engineering, 2002,128(6):490−498.

[13] Zhu T, Ma F, Ma T. The prediction of bus arrival time using global positioning system data and dynamic traffic information. In: Proc. of the 4th Joint IFIP Conf. on Wireless and Mobile Networking Conf. (WMNC). 2011. 1−5.

[14] Kim D, Kim Y, Estrin D. Sensloc: Sensing everyday places and paths using less energy. In: Proc. of the ACM Conf. on Embedded Networked Sensor Systems. 2010. 43−56.

[15] Macqueen J. Some methods for classification and analysis of multivariate observations. In: Proc. of the 5th Berkeley Symp. on Mathematical Statistics and Probability. 2009. 281−297.

[16] Beckmann N, Kriegel H, Schneider R. The R*-tree: An efficient and robust access method for points and rectangles. In: Proc. of the 1990 ACM SIGMOD Int'l Conf. on Management of Data. 1990. 322−331.

[17] Balan R, Nguyen X, Jiang L. Real-Time trip information service for a large taxi fleet. In: Proc. of the 9th Int'l Conf. on Mobile Systems, Applications, and Services. 2011. 99−112.

孙玉砚(1982－),男,甘肃永昌人,助理研究员,主要研究领域为无线自组网,车载网络,物联网安全.

刘燕(1971－),女,博士,副教授,主要研究领域为计算机网络,软件工程.

周新运(1979－),男,博士,高级工程师,主要研究领域为无线传感器网络,物联网安全.

孙利民(1966－),男,博士,研究员,博士生导师,主要研究领域为无线传感器网络,无线自组网.