计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(8):217−221 [doi: 10.15888/j.cnki.csa.006990]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于历史数据的高速多义路径概率识别方法①

路　珊[1], 徐　刚[2], 赵卓峰[1], 丁维龙[1]

[1](北方工业大学 大规模流数据集成与分析技术北京市重点实验室, 北京 100043)
[2](兖州煤业股份有限公司, 邹城 273500)
通讯作者: 路　珊, E-mail: 49760775@qq.com

摘　要: 高速公路多义路径问题是指如何在具有多条可选路径的高速公路网中确定车辆的一条驶经路径. 目前普遍采用的基于识别点的多义路径识别方法在某些情况下 (如设备故障、环境亮度或透明度不够等) 存在识别率低的问题, 导致一些时段存在车辆多义路径难以识别. 针对以上情况, 本文提出一种基于历史数据的多义路径概率识别方法, 通过基于路段的聚类方法计算各路段概率值, 然后结合贪心算法找出车辆的驶经路径, 用来在识别设备故障时辅助识别多义路径. 该方法可以有效的在识别设备故障时识别多义路径, 提高了该方法的准确度.

关键词: 数据缺失; 多义路径识别; 概率; 历史数据

引用格式: 路珊,徐刚,赵卓峰,丁维龙.基于历史数据的高速多义路径概率识别方法.计算机系统应用,2019,28(8):217–221. http://www.c-s-a.org.cn/1003-3254/6990.html

Probabilistic Recognition Method of High Speed Polysemy Based on Historical Data

LU Shan[1], XU Gang[2], ZHAO Zhuo-Feng[1], DING Wei-Long[1]

[1](Beijing Key Laboratory on Integration and Analysis of Large-scale Stream Data, North China University of Technology, Beijing 100043, China)
[2](Yanzhou Coal Mining Company Limited, Zoucheng 273500, China)

Abstract: The highway polysyllabic path problem refers to how to determine a driving path of a vehicle in a highway network with multiple optional paths. At present, the identification point-based polysemy path identification method commonly used in some cases (such as equipment failure, ambient brightness or insufficient transparency) has a low recognition rate, which makes it difficult to identify the vehicle polysemy path in some time periods. Aiming at the above situation, this study proposes a multi-sense path probability identification method based on historical data. The road segment-based clustering method is used to calculate the probability values of each road segment, and then the greedy algorithm is used to find the vehicle's driving path, which is used to identify equipment faults. It assists in identifying polysemy paths. The method can effectively identify the ambiguous path when identifying the equipment failure, and improves the accuracy of the method.

Key words: missing data; ambiguous path identification; probability; historical data

1　引言

随着高速公路的迅速建设, 路网结构越来越复杂, 在车辆进出两个收费站点之间具有多条可能行驶的路径, 从而产生了多义路径问题. 当前, 高速路网中关键位置设置了大量配备车牌拍照系统或 RFID 装备的识别点, 用来帮助准确判定车辆驶经路径, 但这些识别设备受环境亮度或透明度不够、硬件设备故障的影响造成车辆识别不清, 使得车辆在进出收费站点间的监测

① 收稿时间: 2019-01-03; 修改时间: 2019-01-24; 采用时间: 2019-02-26; csa 在线出版时间: 2019-08-08

数据缺失, 车辆实际行驶路径难以准确识别. 因此, 高速公路运营中亟待解决的一个重要问题是如何在多义路径中确定一条车辆的驶经路径, 进而实现通行费用的正确收取和合理拆分, 维护道路使用者和业主的合法利益.

目前, 在高速路中多义路径识别采用基于识别点的路径识别方法, 该方法通过车牌拍照系统或 RFID 装备实现路径识别, 但是这种方法在有不利因素存在的时候, 准确度低. 近几年, 为了更好的监测车辆的行驶路径, 在高速路的重要路段上设置了大量的识别点, 识别点的前端设备会实时上传车辆的监测数据, 同时, 高速路收费站会上传收费数据, 这两类数据上传至数据中心进行存储, 积累了大量的车辆历史通行数据. 能否充分利用这些历史数据在数据缺失时辅助进行多义路径识别成为提高路径识别精度的一个新解决思路.

车辆的历史通行数据主要有收费数据和识别点监测数据两类. 其中, 监测数据是由高速路上的识别点上传, 某省高速路上收费站有数百个, 识别点有近千个, 高峰时平均每分钟就会上传上万条收费监测数据, 随着时间的变化, 积累了大量的车辆历史通行数据, 历史通行数据数据量庞大. 为此, 需要设计实现一个可以快速处理大批量历史数据的计算模型, 通过该模型实现车辆行驶路径的准确高效识别.

本文设计了一种基于历史数据的高速多义路径概率识别方法, 对车辆的历史收费数据和监测数据使用基于路段聚类的方法进行路段概率计算并与贪心算法相结合进行路径识别. 文章的具体组织如下: 首先, 提出本文需要解决的问题并从研究方法和所用技术两个方面介绍路径识别的相关工作, 然后介绍基于路段聚类的路段概率值计算流程和贪心算法与概率矩阵相结合实现路径识别的方法; 最后给出总结.

2 相关工作

目前, 针对高速路上多义路径识别方法以及所用技术, 在国内外已经有了许多研究成果[1,2]. 其中多义路径识别方法主要分为两类, 分别是概率识别和精确识别方法, 技术方面又分为传统技术和大数据技术.

研究方法方面, 路径识别方法主要分为概率识别方法和精确识别方法. 概率识别方法是指依据交通均衡与非均衡理论通过数理统计方法去计算路径, 精确识别方法是指依靠高速路中前端设备采集信息的功能去记录路径. 关于概率识别, 文献[3]在分析了影响路径识别方法选择的各种相关因素基础上对布瑞尔交通分配模型进行了改进, 重新定义和标定其参数. 该方法依赖于公式中的大量参数, 而参数的标定受很多因素的影响, 降低了方法的精确度. 文献[4]利用每个车辆在公路网上的行驶时间来估计每个 OD 对间的车辆行驶时间, 然后对比实际和估计的 OD 间行驶时间, 提出修正遗传算法以获得车辆的路径流, 该方法需要统计大量的数据, 消耗了大量的人力物力. 关于精确识别, 文献[5]研究了车牌拍照的路径识别方法, 详细分析了其中的关键原理与技术, 其中车牌识别过程受环境影响较大, 不能满足路径识别的高精确度要求. 文献[6]对 RFID 射频识别技术的工作原理、特点进行分析, 并结合高速公路运输特点, 设计了基于 RFID 射频技术的高速路收费系统. 该方法中使用的无源射频卡的发射距离有限, 有时会导致车辆难以被识别到.

处理技术方面, 处理交通数据使用的技术主要有传统技术和大数据技术. 传统技术是指用关系型数据库处理数据, 大数据技术是指用 Hadoop 等大数据框架处理数据. 关于传统技术, 文献[7]研究了动态交通信息处理技术, 其中使用关系型数据库存储交通信息, 但由于关系型数据库存储量有限, 而涉及到的交通数据量超过了该数据库的存储范围, 所以采用定期删除数据的方法来接收新的数据. 这种方法使得历史数据不完整, 影响后续数据的使用. 关于大数据技术, 文献[8]提出了一个适合于城市交通网两节点间计算最短路径的算法, 并将任意两节点间最短路径计算过程移植至 MapReduce 框架上, 得出最短路径的路径矩阵和权值矩阵. MapReduce 是大数据平台 Hadoop 上的分布式计算框架, 可快速处理大批量的历史数据. 为此, 本文利用车辆历史的收费数据和监测数据, 通过基于路段的聚类方法进行概率数据计算, 并结合贪心算法进行路径识别. 其中针对数据量庞大的历史数据, 使用 Map Reduce 分布式计算框架进行处理, 提高了数据处理的效率.

3 基于路段聚类的历史数据处理

高速路路网结构复杂, 路网中的路径可以看成是由多个相互连接的路段所组成, 可将收费站和识别点看做是路网中的点, 它们之间由路段相连接, 因此车辆行驶某一路径的概率其实是该路径所包含的各路段通行概率的一个组合[9]. 针对这种情况, 在路径识别前, 应

预先计算出每个路段的通行概率作为基础数据, 以便路径识别时, 根据路径中各路段的概率去辨别车辆究竟走哪一条路径. 现实中, 影响一个司机路径选择的因素很多, 例如道路路况、路长、拥挤程度等因素. 而历史通行记录是综合各种因素后道路使用者的最终道路选择情况, 所以基于车辆历史通行数据进行路段概率计算是更为精确的方法. 高速路历史通行数据包括收费数据和监测数据, 数据量庞大, 采用传统的数据处理方式效率低、时效性差. 所以这里采用 Hadoop 中的 MapReduce 计算框架来进行数据处理, MapReduce 框架是分布式计算框架, 适合处理海量数据, 处理数据速度快.

步骤 1. 车辆的历史通行数据, 包括高速路收费数据和识别点监测数据, 这两类数据都属于单点数据, 不利于计算路段车流量. 所以首先, 需要根据收费数据和监测数据构建车辆历史路径, 再在路径数据中去判断每个路段的通行概率, 这样结果会更为精确. 构建车辆路径的 MapReduce 处理流程如图 1 所示.

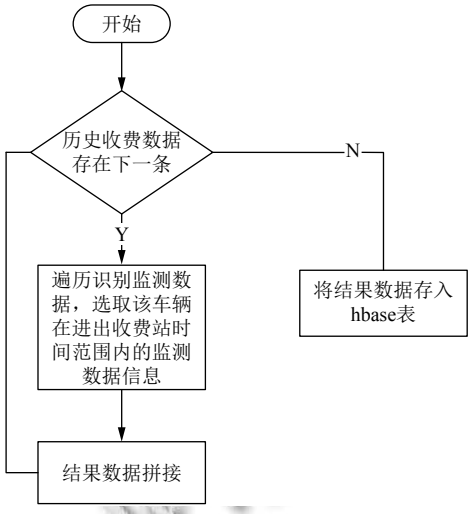

图 1 构建路径流程

该 MapReduce 作业中, 输入是历史收费数据、监测数据, 输出是拼接的车辆历史路径数据. 在 map 阶段, 扫描每条车辆收费记录, 提取其中的车牌号、进出站点时间, 然后遍历历史监测数据, 查询该车在进出收费站点时间范围内的监测数据, 并按照时间先后顺序将车辆经过的站点数据拼接. 中间键值对被发送到 reduce 阶段进行存储, 其中 key 为车牌号, value 为拼接的车辆路径数据.

步骤 2. 上述步骤构建出了车辆的历史路径, 接下来要将路径拆分为不同的路段, 并统计每个路段的通行次数, 这里使用基于路段的聚类方法去统计路段通行次数, 可以把每条车辆历史路径都看成是多个彼此相连接的路段的组合, 遍历车辆历史路径, 遍历每一条路径数据时, 判断该条路径中包含哪几个路段, 然后将对应的路段的通行次数各加 1, MapReduce 处理流程如图 2 所示.

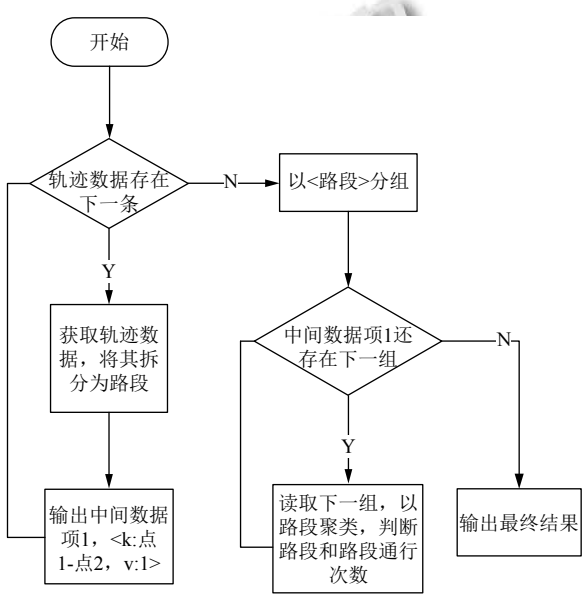

图 2 统计通行次数流程

该 MapReduce 作业中, 输入是历史路径数据, 输出是各路段的通行次数. 在 map 阶段, 扫描每条车辆路径数据, 将其拆分为多个路段. 中间键值对被发送到 reduce 阶段进行处理, 其中 key 为路段的两个端点 (形式为‘路段端点 1+路段端点 2’), value 为 1, 在 reduce 阶段, 中间键值对将根据 key 值被聚集和计数, 最终统计出各路段的历史通行次数.

步骤 3. 经过上述步骤, 已经计算出了各个路段的通行次数, 接下来需要求出路段的通行概率, 应先计算出总的路段通行次数, 然后用各个路段的通行次数值与其相除求得各个路段的概率值.

将统计出的各个路段通行次数求和, 记为 sum, 概率值计算的 MapReduce 处理流程如图 3 所示.

该 MapReduce 作业中, 输入是路段通行次数数据, 输出是各路段的通行概率. 在 map 阶段, 扫描每条路段通行次数数据, 将其值与 sum 相除得到概率值. 中间键值对被发送到 reduce 阶段进行数据存储, 其中 key 仍

为路段的两个端点 (形式为‘路段端点 1+路段端点 2’), value 为路段的通行概率值.

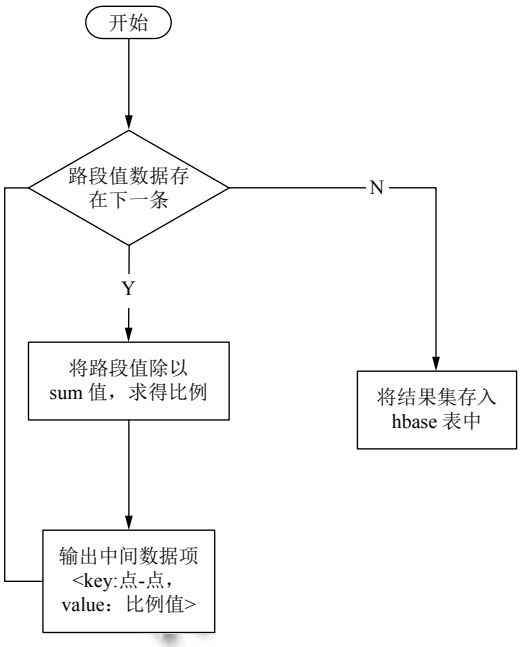

图 3 计算概率值流程

最终结果路段概率值存储的逻辑结构是邻接矩阵, 邻接矩阵的上下标是路段的两个端点, 值是该路段的通行概率值. 由于路网中路段众多, 路段概率数据量庞大, 所以物理存储采用适合存储海量数据的 Hbase 数据库, 存储路段概率数据的 Hbase 表结构设计如表 1 所示, 其中行键设置为路段的两个端点, 用横线相连接, 如‘路段端点 1-路段端点 2’的形式, 这种形式可以清晰的表示出路段的结构, 然后将列中存储的值设置为路段的通行概率. 这种存储结构有助于后续使用贪心算法进行路段的选择, 以构成相应的路径.

表 1 表结构设计

RowKey	Section	Timestamp
	probability	
点 1-点 2	路段概率值	时间戳 T1

4 贪心算法和概率矩阵的结合

高速路路网结构复杂, 在进出收费站点间可能经过一个或多个识别点. 因此在数据缺失的条件下判断车辆的通行路径时, 应以路段为单位, 不仅需要考虑单个识别点数据缺失, 还需要考虑多个识别点数据缺失的情况.

在前一节中基于历史通行数据计算出的路段概率值是道路路况, 时长、拥挤程度等因素的综合体现, 路网中并行的路段相比较, 道路使用者会偏向于选择概率更大的路段去行驶. 贪心算法的原则是每次选取最有利的选择作为当前的选择, 这与道路使用者的路径选择规律相符. 所以本文采用贪心算法和概率矩阵相结合的路径识别方法, 利用贪心算法搜索出从收费站入口到出口的路径, 从入口点开始, 寻找路网中可通行的下一点, 若有一个, 则直接选取这个路段作为路径一部分. 若有多个, 则选取并行的路段中概率值最大的那条路段作为路径一部分, 然后继续寻找下一个点, 做相同的判断和处理, 直到到达出口点为止. 这样一直选取概率最大的路段去构建路径, 最终会得到一条从收费站入口到出口的车辆通行路径. 利用贪心算法寻找车辆行驶路径的流程如图 4 所示.

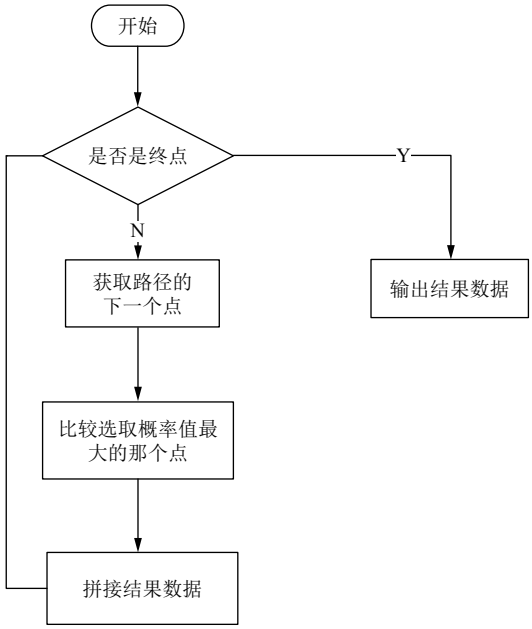

图 4 贪心算法流程

首先判断下一个点是否是路径终点, 即收费站出口点, 如果是终点就输出结果数据, 否则继续获取路径的下一个点, 比较当前点与下一个点相连的并行路段的概率值, 选取概率最大的那个路段加入路径中. 这里贪心算法的规则是一直选择与当前点相连接的路段中概率值最大的路段, 直到收费站出口为止.

5 结语

当前高速公路路网形态复杂, 车辆在进出收费站点之间形成了多条可选择的行驶路径. 同时高速路上

设置了大量的识别点, 识别点上配备的车牌拍照系统或 RFID 装备受环境亮度或透明度不够、硬件设备故障的影响, 使车辆监测数据缺失, 车辆实际行驶路径无法准确识别. 针对以上情况, 本文设计了基于历史数据的多义路径概率识别方法, 首先利用车辆历史通行数据, 使用基于路段的聚类方法计算各路段概率值, 然后将贪心算法与概率矩阵相结合进行多义路径识别. 该方法可以在监测数据缺失情况下有效辅助多义路径识别, 给车主通行费收取、路公司通行费拆分提供了合理的依据.

在下一步的研究工作中, 将对车辆通行数据出现错误的情况下进行多义路径识别计算. 本文仅考虑了数据缺失的情况, 而实际中车辆通行数据不仅仅会出现数据缺失的问题, 还可能会出现数据错误的情况, 日后可研究方法进行通行数据的修正, 以保证数据的完整性, 进而实现数据错误情况下的高速路多义路径识别.

参考文献

1 刘涛. 高速公路联网收费系统中的多义性路径识别. 科技情报开发与经济, 2010, 20(12): 105–106. [doi: 10.3969/j.issn.1005-6033.2010.12.047]

2 孙文波. 高速公路联网收费多路径通行费拆分研究[硕士学位论文]. 天津: 天津大学, 2014.

3 陈洪星, 孙洋. 改进的布瑞尔交通分配模型在高速公路路径识别问题中的应用. 交通与运输, 2008, (2): 37–40.

4 Li SG, Zhou QH. Multiple path toll distribution problems with weight toll on highway networks. Proceedings of 2010 International Conference on Optoelectronics and Image Processing (ICOIP). Haikou, China. 2010. 24–26.

5 杨佳莉. 基于车牌识别的高速公路网多路径精确识别研究[硕士学位论文]. 西安: 长安大学, 2014.

6 雍新琳. 基于 RFID 技术高速公路不停车收费系统设计及实现[硕士学位论文]. 西安: 长安大学, 2017.

7 沈涛, 李娟, 邵春福, 等. 动态交通信息处理技术研究. 道路交通与安全, 2009, 9(4): 34–37.

8 万一红. 基于大数据分析的城市交通网最短路径算法设计[硕士学位论文]. 南昌: 江西财经大学, 2018.

9 温程. 并行聚类算法在 MapReduce 上的实现[硕士学位论文]. 杭州: 浙江大学, 2011.