doi:10.16652/j.issn.1004-373x.2016.23.033

一种卡口车辆轨迹相似度算法的研究和实现

樊志英
（公安部第一研究所，北京　100048）

摘　要：依据车辆轨迹相似度在时间和空间维度上的约束，引入LCSS算法，遵循最长公共子序列的原理，抽象出轨迹中的卡口号序列，提出一种两条车辆轨迹相似度的计算方法，并结合Spark并行计算、Hive数据仓库存储等相关技术，搭建数据分析平台，实现该算法。实验表明，该算法满足实际车辆轨迹在时间和空间上的相似性，数据分析计算在性能上可以满足前台业务的检索。该算法和轨迹相似度分析业务，可作为治安卡口应用系统中关联车辆分析、团伙作案车辆分析等功能的后台支撑业务。

关键词：轨迹相似度；LCSS算法；Spark；Hive
中图分类号：TN911-34；TP311.5　　**文献标识码**：A　　**文章编号**：1004-373X(2016)23-0133-03

Research and implementation of a vehicle trajectory similarity algorithm used for security access monitoring

FAN Zhiying
(First Research Institute of the Ministry of Public Security of PRC, Beijing 100048, China)

Abstract: According to the constraints of time and space dimensions of the vehicle trajectory similarity, the LCSS (longest common subsequence) algorithm is proposed. According to the principle of longest common subsequence, the access monitoring sequences in the trajectory are abstracted. A calculation method of two vehicle trajectories similarity is proposed. The Spark parallel calculation, Hive data warehouse storage and other correlation technologies are combined to establish the data analysis platform, and implement the algorithm. The experimental results show that the algorithm can satisfy the time and space similarity of the practical vehicle trajectory, and the data analysis and calculation can meet the search performance of foreground business. The algorithm and trajectory similarity analysis business can be used as the background support service of the vehicle relevance analysis and gang crime vehicle analysis in the security access monitoring application system.

Keywords: trajectory similarity; LCSS algorithm; Spark; Hive

0　引　言

随着城市经济的快速发展，各地机动车保有量迅速增加，与车辆相关的刑事和治安案件也在逐年上升，除了传统的违法涉案车辆的缉查管控外，基于重点车辆的行驶轨迹和出行规律分析等业务也将为侦查破案提供有力的依据。

随着治安卡口、电子警察等应用系统的建设和使用，各地已积累了大量的车辆通行记录和违法记录，这些记录中涵盖了车牌号码、经过时间、车辆颜色、车辆类型、行驶方向、行驶状态等车辆信息，为开展车辆出行规律分析等业务提供了强大的数据支撑。

本文使用某地区已有的大量车辆通行记录，结合大数据相关技术，对车辆轨迹和轨迹相似度进行分析和实现，该方案可作为治安卡口应用系统的车辆数据分析的实现思路，为其提供业务支撑。

1　车辆轨迹相似度计算

车辆轨迹相似度分析业务指的是计算指定车辆和其他车辆的行驶轨迹，分析出与指定车辆具有相似轨迹的多个车辆的通行记录，进而为治安卡口应用系统的关联车辆、团伙作案车辆等功能提供后台业务支撑。

车辆轨迹相似度分析分别在时间和空间维度上进行了限制，首先，其他车辆与指定车辆经过同一个卡口的时间要在一定范围内，如2 min以内；其次，其他车辆

收稿日期：2015-12-27
基金项目：十二五国家科技支撑计划“基于视频及动态信息的智能研判技术研究及应用示范”：基于重点目标自动跟踪采集技术的智能视频监控系统研发(2013BAK02B02)

与指定车辆经过多个卡口的顺序要一致，一致性越高，相似度越高。

1.1 两条轨迹的相似度计算公式

根据车辆轨迹相似度分析在时间和空间维度上的约束，在计算两条轨迹的相似度时，需要考虑以下两点：

（1）时间相似度的保证

计算相似度前，对轨迹数据按照时间进行筛选，保证时间跨度上的接近。

计算相似度时，将经过每个卡口的时间偏差均值纳入考虑范围。

（2）空间相似度的保证

轨迹中的空间维度对应的就是卡口号的序列。通过时间排序后的卡口号序列，看作是两条对象序列，每个对象元素就是一个卡口号字符串。在空间相似度上，引入LCSS（Longest Common Subsquence）算法[1-2]，此处类似于字符串的LCSS算法，比较的最小单元是每个卡口号字符串，而不是单个字符。通过对这两条对象序列做LCSS算法，求出最长公共子序列，即满足指定时间差内经过同一个卡口条件的卡口数量，作为后续计算相似度的主体轨迹序列。

依据以上约束性，定义两条轨迹的相似度计算公式如下：

$$\begin{cases} E(\Delta T)=\dfrac{\sum_{i=1}^{p}|\Delta t_i|}{p} \\ \mathrm{sim}=\dfrac{p}{p_0}\cdot\dfrac{1}{1+E(\Delta T)/tt}\cdot Mf \end{cases}$$

式中：p 是按时间顺序排序的两条轨迹的卡口序列的最长公共子串长度，即在给定时间差内（如2 min以内）两辆车经过相同卡口的数量；p_0 是指定车辆的轨迹中经过的卡口数量；$E(\Delta T)$ 是两条轨迹中经过所有相同卡口的时间差（单位：s）的绝对值的期望；Mf 是相似度计算结果的放大因子，如果 $Mf=1$ 则计算的相似度在[0，1]之间；tt 是时间偏差的容忍度（单位：s），不能为零，即两辆车经过同一个卡口时可接收的最大时间偏差（如2 min）。

1.2 计算过程分析

以指定车辆的轨迹 X 和其他车辆的轨迹 Y 为例，说明轨迹相似度的计算过程，如图1所示。

（1）将轨迹 X 和待比较的轨迹 Y 中的轨迹点数据按照时间顺序升序排列。在指定时间段内，轨迹 X 为：卡口1（t_1），卡口2（t_2），卡口3（t_3），卡口4（t_4），卡口5（t_5），卡口6（t_6），量化为6。此时，$p_0=6$；轨迹 Y 为：卡口2（$t_2-\Delta t_{2Y}$），卡口3（$t_3+\Delta t_{3Y}$），卡口4（$t_4+\Delta t_{4Y}$），卡口5（$t_5-\Delta t_{5Y}$）。

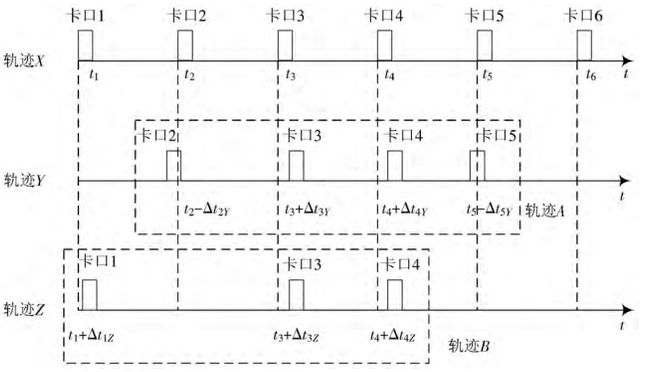

图1 多条轨迹示意图

（2）比较轨迹 X 和轨迹 Y 中经过相同卡口的时间跨度和时间偏差容忍度 tt，从轨迹 Y 中截取符合时间跨度要求的轨迹 A。

（3）对目标轨迹 X 和截取后的轨迹 A 做LCSS（X，A）算法，求出它们的最长公共子序列。此时最长公共子序列 p 为：卡口2，卡口3，卡口4，卡口5，量化为4。此时，$p=4$。

（4）依据两条轨迹的相似度计算公式，求出相似度。

假设 $tt=120$ s，$\Delta t_{2Y}=20$ s，$\Delta t_{3Y}=20$ s，$\Delta t_{4Y}=40$ s，$\Delta t_{5Y}=50$ s，$p_{XY}=4$，$p_0=6$。$Mf=1$。$\Delta t_{1Z}=30$ s，$\Delta t_{3Z}=40$ s，$\Delta t_{4Z}=35$ s，$p_{XZ}=3$。

计算轨迹 X 和轨迹 Y 的相似度：$E(\Delta T)_{XY}=32.5$，$\mathrm{Sim}_{XY}=0.525$。

同理，计算轨迹 X 和轨迹 Z 的相似度：$E(\Delta T)_{XZ}=35$，$\mathrm{Sim}_{XZ}=0.387$。

从以上计算结果可以得出，该计算公式符合实际情况，可作为轨迹相似度分析的基础算法。随着车辆通行数据量的累积和对大量车辆轨迹分析计算的结果，可以对其中的参数进行调优，进而满足轨迹相似度业务的需求。

1.3 使用大数据技术的轨迹相似度计算的实现

随着车辆通行记录的数量累积，需要考虑海量数据存储和高效快速的数据分析，采用大数据的相关技术对轨迹相似度分析算法进行实现。

1.3.1 数据分析平台的结构

本文将数据抽取、存储和并行计算的逻辑模块称为大数据分析平台，它采用开源的分布式文件系统和并行计算框架，并结合传统数据库的检索便利，实现了轨迹相似度的分析业务。

原始车辆通行记录存储于Oracle数据库中。使用Sqoop工具将存储在Oracle数据库中的卡口车辆通行记录定期抽取到Hive数据仓库中，启用Spark并行计算车辆行驶轨迹和与指定车辆的轨迹相似度，将计算结果保

存至Hive中，并向原Oracle数据库写入计算结果。用户可访问Oracle数据库检索车辆轨迹相似度分析结果。

此处，对数据分析平台（见图2）中涉及到的几个技术进行简要描述。

Hadoop：Hadoop是一个由Apache基金会开发的分布式系统基础架构，其最核心的设计就是HDFS和MapReduce。

Sqoop：它允许用户将数据从关系型数据库抽取到Hadoop中，用于进一步的处理[3]。

Hive：是一个构建在Hadoop上的数据仓库框架。

Spark：是一个通用的并行分布式计算框架，支持内存计算，能同各种迭代算法和交互式数据分析，能够提升大数据处理的实时性和准确性。

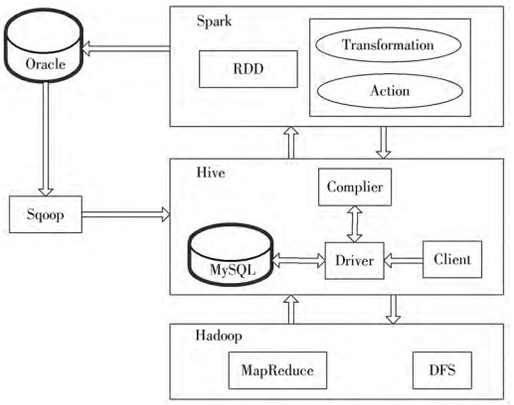

图2　数据分析平台结构

1.3.2　单个车辆轨迹的计算过程

车辆轨迹分析为定时任务，定时统计计算出前一天所有车辆的轨迹信息，并存储在数据库和数据仓库中。

基于Spark框架的并行计算车辆轨迹的过程如下：

Step1：用HiveQL从HDFS中查询出一天的过车数据信息，得到rowsRDD。

Step2：对rowsRDD做一次mapToPair操作，用车牌号码做key，卡口号和通过时间合并的字符串作为value，得到middleCombineRdd。

Step3：对middleCombineRdd做一次reduceByKey，将相同车牌的轨迹点合并成一个字符串，得到middleReduceRdd。

Step4：对middleReduceRdd做一次map操作，将合成的轨迹数据转换为数据表的数据结构，得到targetRdd。

Step5：将targetRdd保存到Hive数据仓库的cartrajectory数据表中。

Step6：将targetRdd数据保存到Oracle数据库的cartrajectory数据表中。

cartrajectory数据表为车辆轨迹表，存放着每一天每辆车的轨迹信息，包含车辆号牌和轨迹信息字符串，其中轨迹信息字符串包含车辆经过的卡口信息和经过时间。

1.3.3　轨迹相似度分析的业务流程

多车辆轨迹的相似度计算是一个后台定时任务。部署于Spark计算平台上的ServerListener监听客户端的请求，当客户端定期发出请求时，启动计算任务。在Spark平台上计算指定车辆的轨迹信息，再计算与指定车辆的轨迹大于相似度阈值的其他车辆的轨迹信息，再将计算结果返回给客户端，由客户端将结果存储于Oracle数据中，用于用户检索。轨迹相似度分析业务流程如图3所示。

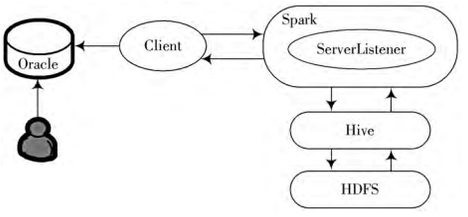

图3　轨迹相似度分析业务流程

2　实　验

依据轨迹相似度计算公式，结合大数据分析平台，对算法进行验证。

实验条件：

（1）某城市一天的过车数据量为100万条左右；

（2）搭建小规模集群：4台服务器，服务器配置：2CPU（4核），16 GB内存，1 TB高速硬盘。

实验结果：

（1）Spark并行计算效率：

计算与指定车辆在一个月内的轨迹相似度时间：5 min左右。

计算与指定车辆在一周内的轨迹相似度时间：3 min左右。

（2）用户从Oracle数据库中检索相似轨迹的时间为秒级响应（与检索时间段有关）。

以上实验说明，轨迹相似度计算公式基本符合实际情况；而Spark并行计算时间为分钟级，主要是从硬盘上获取数据文件时耗费了较多的时间。

通过对车辆历史轨迹相似度的分析，数据分析平台的并行计算也可以满足要求。但是对于当前时段（如当天）的轨迹相似度并行计算仍需要优化，后续可将当前时段的实时数据缓存于数据分析平台，当客户端触发实时检索任务时，可计算缓存中的部分数据，再结合之前计算好的历史轨迹相似度结果，统一作为检索结果再返回。

3　结　语

本文依据卡口车辆轨迹相似度在时间和空间维度

（下转第140页）

利用ADAMS验证了平顺性仿真模型。最后，以悬架刚度阻尼为设计变量，以悬架动挠度和轮胎相对动载荷为约束函数，以座椅中心加权加速度均方根值为目标函数，建立其优化模型，利用Matlab优化工具箱中的遗传算法函数进行优化。优化结果表明，利用遗传算法进行优化大大地提高了汽车的平顺性，为非线性振动系统平顺性时域仿真优化奠定了基础。

注：本文通讯作者为莫秋云。

参 考 文 献

[1] 余志生.汽车理论[M].5版.北京：机械工业出版社，2009.

[2] 陈杰平，陈无畏，祝辉，等.基于Matlab/Simulink的随机路面建模与不平度仿真[J].农业机械学报，2010，41(3)：11-15.

[3] 金睿臣，宋健.路面不平度的模拟与汽车非线性随机振动的研究[J].清华大学学报(自然科学版)，1999，39(8)：77-80.

[4] 张立军，张天侠.车辆四轮相关时域随机输入通用模型的研究[J].农业机械学报，2005，36(12)：29-31.

[5] 张鄂，刘中华，邵晓春.九自由度乘坐动力学模型的人体振动特性仿真[J].交通运输工程学报，2010，10(4)：58-64.

[6] LE V Q，张建润，王园，等.基于三维动力学模型的重型卡车动态参数对平顺性的影响[J].东南大学学报(自然科学版)，2013，43(4)：763-770.

[7] 陈克，高洁，吕周泉.基于虚拟试验场技术的汽车平顺性仿真分析[J].中国工程机械学报，2010，8(2)：208-212.

[8] GOLDBERG D E. Genetic algorithms in search, optimization, and machine learning [M]. Boston: Addison Wesley, 1989.

[9] HOLLAND J H. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence [M]. Michigan: Michigan Press, 1975.

[10] 戴晓晖，李敏强，寇纪淞.遗传算法理论研究综述[J].控制与决策，2000，15(3)：263-268.

[11] 姜丽丽.基于傅里叶反变换的路面随机激励时域建模与仿真[D].长春：吉林大学，2007.

[12] 彭佳，何杰，李旭宏，等.路面不平度随机激励时域模型的仿真比较与评价[J].解放军理工大学学报(自然科学版)，2009，10(1)：77-82.

[13] 张永林.车辆道路数值模拟与仿真研究[D].武汉：华中科技大学，2010.

[14] 何宁，石成英，周保顺.路面不平度时域模型模拟方法研究[J].环境技术，2014，32(3)：50-51.

作者简介：鲍家定(1981—)，男，安徽铜陵人，硕士，助理研究员。主要研究方向为激光复合加工和机械动力学。

伍建伟(1989—)，男，湖南永州人，硕士研究生。主要研究方向为机械动力学与优化算法研究。

莫秋云(1971—)，女，广西桂林人，博士，教授，硕士研究生导师。主要研究方向为先进制造技术、机械电子工程、人机工程。

(上接第135页)

上的约束，提出了一种轨迹相似度的计算方法，并结合大数据相关技术对该算法进行验证。实验表明，该计算公式和实现方法满足后台业务分析的需求，可作为治安卡口应用系统相关功能的业务支撑。

参 考 文 献

[1] VLACHOS M, KOLLIOS G, GUNOPULOS D. Discovering similar multidimensional trajectories [C]// Proceedings of 2002 18th International Conference on Data Engineering. Riverside: IEEE, 2002: 673-684.

[2] KOLLIOS G, GUNOPULOS D, VLACHOS M. Robust similarity measures for mobile object trajectories [C]// Proceedings of 2002 International Workshop on Database & Expert Systems Applications. France: IEEE, 2002: 721-726.

[3] WHITE T.Hadoop权威指南[M].周敏奇，王晓玲，金澈清，等译，2版.北京：清华大学出版社，2011.

[4] DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters [C]// Proceedings of the 6th Conference on Symposium on Operation Systems Design Implementation. Berkeley: ACM, 2004: 107-113.

[5] ZAHARIA M, CHOWDHURY M, DAS T, et al. Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing [C]// Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation. Berkeley: ACM, 2012: 141-146.

[6] HWANG J R, KANG H Y, LI K J. Spatio-temporal similarity analysis between trajectories on road networks [C]// Proceedings of ER 2005 Workshops on AOIS. BP-UML, CoMoGIS, eCOMO, and Qols. Klagenfurt: Springer Berlin Heidelberg, 2005: 280-289.

[7] 夏俊鸾，刘旭晖，邵赛赛，等.Spark大数据处理技术[M].北京：电子工业出版社，2015.

[8] 高彦杰.Spark大数据处理：技术、应用与性能优化[M].北京：机械工业出版社，2014.

作者简介：樊志英(1982—)，内蒙古呼和浩特人，硕士研究生，工程师。主要研究方向为视频监控、信息安全、软件工程等。