文章编号：1005-8451（2016）09-0025-03

基于大数据技术的铁路客流预测系统架构研究

张军锋，贾新茹，李　永，张利明
（中国铁道科学研究院　电子计算技术研究所，北京　100081）

摘　要：分析现有铁路客流预测理论研究和系统应用，针对预测过程中缺乏对不确定因素考虑的情况，提出预测系统中事件的概念，基于大数据技术构建事件处理平台和预测系统架构，该架构为铁路客流预测实时性和准确性的提升奠定基础。

关键词：大数据；客流预测；体系架构
中图分类号：U293：TP39　**文献标识码：**A

Architecture of Railway Passenger Flow Prediction System based on big data technology

ZHANG Junfeng, JIA Xinru, LI Yong, ZHANG Liming
(Institute of Computing Technologies, China Academy of Railway Sciences, Beijing　100081, China)

Abstract: Based on the analysis of the existing theory and system application of railway passenger flow prediction, aiming at the lack of consideration to uncertainty factors in the prediction process, this article proposed the concept of uncertain event processing, built a data processing platform and prediction system architecture based on big data technique. This architecture is the basis for the improvement of real-time and accuracy of railway passenger flow prediction.

Key words: big data; passenger flow prediction; system architecture

在铁路客流预测领域，研究人员提出了很多客流预测模型及改进方法，如黄召杰、陈伟[1]采用Box—Jenkins模型和灰色预测模型相组合，Clark, S[2]提出了采用多元非参数回归的交通流预测模型，王卓、王艳辉、贾利民等[3]对BP神经网络模型进行改进，潘亮[4]在其硕士学位论文中应用改进经验模态分解（EEMD）模型等。这些研究表明，通过模型的改进和组合应用可以提高预测精度。但是这些模型和算法大多停留在研究阶段，因为需要的因素难以获取或量化，计算所需资源和时间消耗较大，无法满足实际需要，所以鲜有投入实际应用的案例，更缺乏成型的系统软件。铁路客票发售和预订系统研发团队通过多年研究，结合铁路客运业务，开发了铁路客流预测系统。

铁路客流预测系统建设在客票系统、客运营销辅助决策系统的基础上，适应铁路快速发展的需要，针对客运人员的业务需求，利用时间序列、神经网络等客流预测模型算法，对列车、区域、线路的总量或者始发站—终点站（OD）客流进行预测，为票额智能预分、列车开行方案设计、列车收益管理等客运业务提供科学决策的依据。自投入使用以来，在辅助业务人员完成产品设计与开发，实现精细化管理和科学决策方面起到了很大作用。但是在使用过程中也发现了一些问题和不足，最突出的问题是以历史客流规律为预测基础的时间序列模型，对预售过程中因特殊情况导致的客流需求突变等情况没有考虑，在预测精度上还有待提高。因此，本文在引入事件这一概念的基础上，提出了预测事件处理大数据平台和基于大数据技术的铁路客流预测系统架构方案，以期弥补现有系统的不足，实现提高客流预测精度的目标。

1　铁路客流预测中的事件处理

铁路客流是由一个个独立的旅客个体随机组成，是一个高度复杂的非线性动态系统，其变化规律既有一定的趋势性，同时又受到其他诸多因素的影响。

收稿日期：2016-06-15
基金项目：国家自然科学基金（U1334207）；中国铁路总公司科技研究开发计划课题（2016X005-B）；中国铁路总公司科研计划课题（2016X004-G）；中国铁道科学研究院科研专项课题（研发中心）（J2016X009）。
作者简介：张军锋，副研究员；贾新茹，副研究员。

铁路客流历史数据是典型的时间序列，而时间序列会经常受到突发情况的影响，诸如天气变化、大型比赛、重大社会活动等。客流预测模型一般建立在客流波动较小的情况下，当这些突发情况发生时，预测模型不能取得很好的效果，为了定量分析评估这些突发情况对客流造成的影响，可以将导致预测指标发生随机变化的不确定因素定义为“事件”，并在系统中建立事件处理模块对预测结果进行调整干预。增加事件处理模块后预测系统的功能结构如图 1 所示。

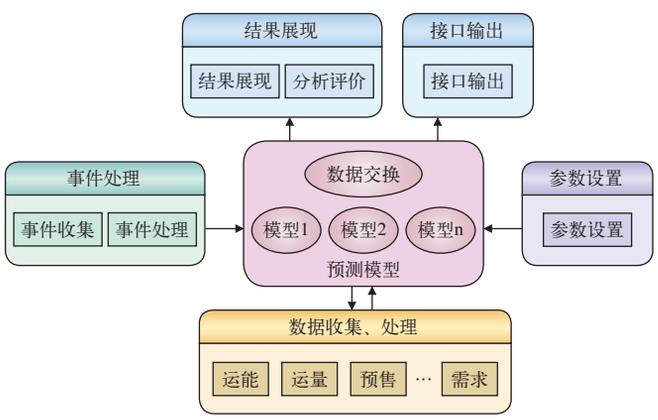

图1 铁路客流预测系统功能结构

预测系统中的事件处理模块是以 12306 网站和手机客户端交易记录以及日志信息为数据源，对旅客需求的变化进行感知和挖掘，在统计分析后对预测值提出修订的方向和幅度。

2 铁路客流预测事件的大数据处理平台

目前的客流预测系统是以客票系统、客运营销辅助决策系统中的数据仓库系统为支撑建立的。现有的数据仓库系统作为客运营销辅助决策系统的核心要素，为铁路企业的决策层、管理层和业务层提供了重要的分析手段，为挖潜提效、改善服务质量做出了重要贡献。

然而，随着移动互联网时代的到来，12306 网站、手机客户端售票渠道已经上线运行，产生了大量的交易、日志信息，事件处理模块需要对这些数据进行存储和处理，从而获取有用的信息，但现有数据仓库的性能已经无法应付。主要原因是：(1) 移动互联网需要对大量半结构化、非结构化的信息进行有效分析，现有数据仓库系统的结构化存储和处理架构难以有效应对；(2) 事件处理需要对市场需求和突发情况及时掌握、响应，也对数据处理速度提出了更高的实时性要求，移动互联网产生的流数据需要能够实时进行处理和分析，现有的以批量加载为主的数据仓库系统难以有效支撑。因此，需要采用大数据技术处理这些宝贵的、大规模的数据，以应对复杂的、即时性要求较高的事件处理需求，从而使预测系统能提供更好的预测精度。

铁路客流预测事件的大数据处理流程与传统数据处理流程基本相似，整个处理流程可以概括为数据采集、数据导入和预处理、数据存储、数据处理和存储等。由于事件处理涉及到大量的、多种形式的数据，其中还有一部分强实时性的数据，如及时反映旅客需求情况的日志记录等，所以采用在线实时流数据处理和离线批量处理相结合的计算框架。采用分布式文件存储系统 HDFS、分布式数据库 Hbase/GreenPlum 和现有数据仓库 SybaseIQ 存储数据。平台搭建如图 2 所示。

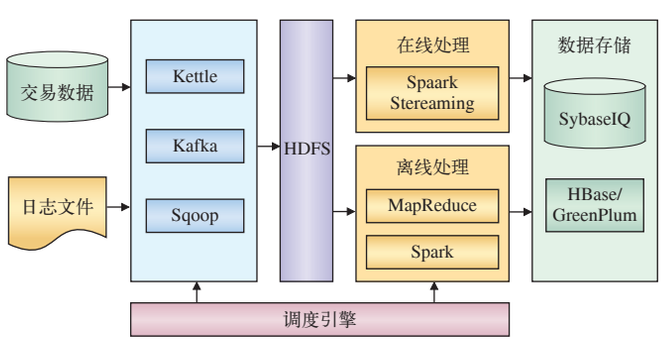

图2 铁路客流预测事件处理大数据平台

平台使用构建在 Hadoop 生态系统上的分布式日志处理系统收集 12306 互联网用户访问和订票日志，用数据同步传输工具 kettle 和 Sqoop 实现各种数据源与 Hadoop 分布式文件系统 HDFS 间的数据传输。因存在强实时性的处理需求，平台在并行批处理 MapReduce 和 Spark 的基础上，引入了 Spark Streaming 的实时流数据处理计算框架。数据处理后形成多个主题，综合考虑预测过程中的访问频率和其他业务应用需求，分别放置于 SybaseIQ、Hbase 和 GreenPlum 数据库。

3 基于大数据技术的铁路客流预测系统架构

在引入事件处理模块后，预测系统的数据源得

到了扩充，预测所需的因素更加全面，有助于提高预测的准确性。在综合考虑了预测系统数据特性和应用需求的基础上，本文设计了基于大数据技术的铁路客流预测系统体系架构。该架构结合现有数据仓库系统，在引入大数据处理技术、构建预测事件处理大数据平台的基础上设计而成，包含 4 层结构，分别为数据源层、数据处理层、模型计算层和输出接口层，如图 3 所示。

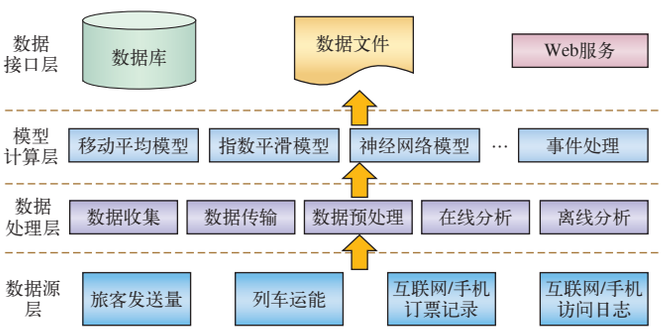

图3　基于大数据技术的铁路客流预测系统架构

（1）数据源层。客流预测系统包含旅客发送量、列车能力、客运基础字典等结构化数据，还包括用户访问日志、订票日志等半结构化数据。

（2）数据处理层。数据处理层分为 2 部分：

a. 时间序列模型所需的历史客流和列车能力数据的处理，这部分数据由数据库和应用服务器利用多线程技术进行计算处理。

b. 事件处理平台采用大数据技术对 12306 互联网和手机客户端访问日志、订票日志进行离线分析和实时在线分析，得出旅客出行需求可能的变化趋势和幅度。

（3）模型计算层。根据用户选择的模型进行预测计算，同时结合事件处理模块给出的调整策略和建议值进行修正。

（4）输出接口层。系统的预测结果可以根据需要存储到数据库、输出成文本文件，或者以 Web Service 方式提供给第三方应用，从而可以支撑多方面的应用。

4　结束语

数据是预测模型、系统运行的基础，大数据技术的应用将促进预测系统的进一步发展。本文在参阅大数据技术相关文献的基础上，结合目前铁路客流预测领域研究和开发的现状，构建了客流预测事件处理的大数据平台，设计了基于大数据技术的铁路客流预测系统架构。该架构兼顾预测过程中涉及因素多、数据量大的特点，同时兼顾了预测过程中数据处理的实时性和高效性。架构的应用，将有助于提升客流预测系统的运行效率和准确性。

后续的研究将在本文所述的平台和架构基础上，继续探索利用大数据技术对影响客流的其他因素的处理，如天气、重大活动等，研究如何获取和处理这些因素涉及的相关数据，实现数据源的完善、扩充和整合，并与现有预测系统中的模型结合，继续提升预测准确性，为开行方案设计、运力资源配置、售票组织策略调整等业务提供有力支撑。

参考文献：

[1] 黄召杰，陈　伟．组合预测方法在我国铁路客流预测中的应用 [J]. 交通科技与经济，2011，13（4）：96-98，102.

[2] Clark, S. Traffic prediction using multivariate nonparametric regression[J]. Journal of Transportation Engineering, 2003, 129(2): 161–168.

[3] 王　卓，王艳辉，贾利民，等．改进的 BP 神经网络在铁路客运量时间序列预测中的应用 [J]．中国铁道科学，2005，3（2）：127-131.

[4] 潘　亮．基于 EEMD-GSVM 的高速铁路客流短期预测 [D]. 北京：北京交通大学，2012.

责任编辑　杨琍明