文章编号：1005-8451（2016）09-0014-04

铁路客运大数据平台架构及技术应用研究

单杏花，王富章，朱建生，张军锋
（中国铁道科学研究院　电子计算技术研究所，北京　100081）

摘　要：文章结合我国铁路当前运输和信息化形势，分析大数据来源和相关技术，包括大数据采集、大数据预处理、大数据存储及管理、大数据分析及挖掘、大数据展现及应用等技术，在此基础上设计了铁路客运大数据平台，并对铁路客运大数据应用场景和应用技术进行分析，提出铁路客运大数据建设和应用的推进思路。

关键词：铁路客运；大数据技术；用户画像；产品画像；智慧营销
中图分类号：U293：TP39　**文献标识码：**A

Architecture and technology application of big data platform for railway passenger transport

SHAN Xinghua, WANG Fuzhang, ZHU Jiansheng, ZHANG Junfeng
(Institute of Computing Technologies, China Academy of Railway Sciences, Beijing　100081, China)

Abstract: Combined with the current situation of China railway transport and informatization, this article analyzed the big data sources and related technologies, including big data collection, big data preprocessing, big data storage and management, big data analysis and mining, big data show and application, etc., designed big data platform for railway passenger transport, analyzed the big data application scenarios and application technology, put forward the idea to promote the construction and application of big data for railway passenger transport.

Key words: railway passenger transport; big data technology; user portrait; product portrait; wisdom marketing

随着互联网及信息处理技术的快速发展，大数据已经无处不在，而且比以往任何时候都重要，大数据已经从技术研究阶段进入实用阶段。阿里、腾讯、百度、高德等互联网运营公司通过大数据采集、处理和洞察，提升了用户体验，实现了产品的互联网精准投放。银行、保险、证券等金融行业，也通过大数据平台的搭建、数据的采集和分析，不断优化自身产品设计，提高产品的市场满意度。

我国铁路作为传统的交通运输行业，随着高速铁路的建设和运营，在运输产品设计、运输指挥方式、运输组织形式等方面，相比传统方式都发生了巨大的变化，但如何制订更合理的列车开行方案，如何进行更科学的售票组织，如何做好站车服务尤其是不可抗力发生时的应急处置，是铁路行业旅客运输组织面临的重大课题。解决这一问题，大数据技术是重要且有效的技术手段。

当前，铁路客运相关信息系统初步建成，为铁路旅客运输服务质量的提升奠定了很好的技术基础，同时也为铁路客运大数据平台提供了大量的原始生产数据。铁路客运相关信息系统如下：

（1）铁路客票发售和预订系统。
（2）铁路 12306 互联网售票系统。
（3）铁路旅客服务系统。
（4）铁路客运管理信息系统。
（5）铁路客运延伸服务系统。
（6）清算、营销、动车组检修等系统。

1　大数据来源及技术分析

大数据来源于每个人日常工作和生活的方方面面，包括企业生产系统、互联网服务系统以及银行、通信系统等，物理网、云计算、移动互联网、车联网、手机、PC 机以及遍布地球各个角落的各式各样的传感器，无一不是大数据的来源。数据形式可以是结构化数据，也可以是非结构化数据，如图片、视频、

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（2016-X004-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：单杏花，研究员；王富章，研究员。

Word、PDF、PPT 等。

大数据关键技术一般包括：大数据采集、大数据预处理、大数据存储及管理、大数据分析及挖掘、大数据展现和应用。

1.1 大数据采集技术

针对不同的数据对象，可以采用不同的数据采集技术：

（1）对于企业生产经营数据或学科研究数据等保密性要求较高的数据，可以通过与企业或研究机构合作，自行开发特定的系统接口等相关方式采集数据。

（2）对于系统日志，很多互联网企业都有自己的海量数据采集工具，用于系统日志采集。如 Hadoop 的 Chukwa，Cloudera 的 Flume，Facebook 的 Scribe 等，这些工具均采用分布式架构，能满足每秒数百 MB 的日志数据采集和传输需求。

（3）对于非结构化的网络数据采集，可以通过网络爬虫或网站公开 API 等方式从网站上获取数据信息。该方法可以将非结构化数据从网页中抽取出来，将其存储为统一的本地数据文件，并以结构化的方式存储。支持图片、音频、视频等文件或附件的采集，附件与正文可以自动关联。除了网络中包含的内容之外，对于网络流量的采集可以使用深度包检测（DPI）或深度 / 动态流检测（DFI）等技术进行处理。

1.2 大数据预处理技术

数据预处理的目的在于整合数据，把一些与数据分析、挖掘无关的项清除，保证数据的正确性和有效性，通过对数据格式和内容的调整，使数据更符合挖掘的需要，给后续数据挖掘提供更高质量的数据。具体包括数据清理（解决空缺值、错误数据、孤立点、噪声问题）、数据集成、数据变换（平滑、聚集、数据概化、规范化）、数据归约（维归约、数据压缩、数值归约、概念分层）[1]。

1.3 大数据存储及管理技术

大数据存储及管理技术主要还是数据库存储技术，根据数据处理阶段不同，涉及的大数据存储和管理技术除关系型数据库外，还有以下 3 类 [2]：

（1）分布式存储与计算。是指大量普通 PC 机服务器通过网络互联，对外作为一个整体提供存储服务，具有可扩展、低成本、高性能和易用等特点。目前大数据领域应用最为广泛的就是 Hadoop。

（2）NoSQL 数据管理系统。是没有固定数据模式且可以水平扩展的系统的统称。NoSQL 指的是“Not Only SQL”，是对关系型 SQL 数据系统的补充。采用的技术有简单数据模型、元数据和应用数据分离、弱一致性、高吞吐量、高水平扩展能力和低端硬件集群等。如 Hbase、Cassandra、MongoDB 等。

（3）NewSQL 数据管理系统。是指解决了传统关系型数据库有关通信、日志、锁、闩以及缓冲区管理等问题，通过使用冗余机器来实现复制和故障恢复，可扩展、高性能的 SQL 数据库，如 RethinkDB、VoltDB 等。

1.4 大数据分析及挖掘技术

（1）常用的数据分析技术包括：聚类分析、因子分析、相关分析、对应分析、回归分析、方差分析等。

（2）常用的数据挖掘技术包括：A/B test、top N 排行榜、地域占比、文本情感分析和关联规则分析等。

1.5 大数据展现及应用技术

（1）展现技术包括大数据检索、图表展现、可视化展现、地理信息系统展现等。

（2）大数据典型应用包括构建 360° 立体用户画像、构建产品标签模型、进行市场预测、执行个性化的智慧营销、实现高效准确的风险控制、挖掘数据价值、实现历史数据归档查询、分析用户舆论评价等。

2 铁路客运大数据平台架构

大数据平台或大数据系统的建设需要各种大数据技术的支撑，大数据时代下的平台或系统需求可以概括为以下 3 个方面：（1）高并发读写需求，能够高并发、实时动态获取和更新数据；（2）海量数据的高效率存储和访问的需求，类似 SNS（Social Network Sofware）网站，海量用户信息的高效率实时存储和查询；（3）高可扩展性和高可用性的需求，需要拥有快速横向扩展能力、提供 7×24 h 不间断服务。

按照上述需求，结合铁路客运大数据相关系统间关系，铁路客运大数据平台架构设计如图 1 所示。

（1）外部系统层，是指能提供客运大数据平台相

图1 铁路客运大数据平台架构

关数据的外部系统。

（2）数据层，是指从外部系统获取客运大数据平台需要的相关数据，以及经过整理的内部数据。

（3）存储层，是指存储客运大数据平台数据计算资源、存储资源、网络资源和系统管理资源。

（4）分析层，是指支持对客运大数据平台中的数据进行分类、统计、聚合等分析操作的分析服务。

（5）展示访问层，是指支持对客运大数据平台中的数据进行展示服务，包括图、表以及地理信息系统服务等。

（6）应用层，是指结合应用场景，利用分析层和展示访问层提供的服务，为用户提供的具体应用服务。

3 铁路客运大数据技术应用

大数据的应用，无论是在工业，还是在农业，抑或是在金融和互联网运输业，最终都是通过大数据技术获知事情发展的真相，利用这个“真相”更加合理地配置资源。也就是说，大数据的核心价值就是“优化资源配置”。

相关研究表明，要实现大数据的核心价值，在技术上主要包含以下 3 个步骤：

（1）通过“众包”的形式收集海量数据；

（2）通过大数据的技术途径进行“全量数据挖掘”获知“真相”；

（3）利用分析结果进行“资源优化配置”。

我国铁路客运大数据技术应用则主要包含构建用户和客运产品 360° 立体画像、客流及市场预测、客运产品设计、智慧营销、风险发现、应急指挥等 6 个方面。

3.1 构建用户和客运产品360° 立体画像

构建用户 360° 立体画像，包含用户基本属性信息、用户事件信息、用户关系信息、用户沟通信息、用户财务信息、用户风险信息等，并根据需要支持业务的拓展，不断予以调整和丰富。

构建客运产品 360° 立体画像，包含客运产品的基本属性信息、运营特征信息、营销活动信息、成本信息、与其他客运产品的关系信息等，也可根据需要进行优化和调整。

3.2 客流及市场预测

在用户画像基础上，分析旅客出行行为，引入时间序列、神经网络、模型树等客流预测模型，并予以优化和调整，进行客流及市场预测，包括长期客流预测、短期客流预测以及车票预售过程中的动态预测，指导客运产品设计和营销策略的制订。

3.3 客运产品设计

在客运产品画像基础上，设计客运产品吸引模型，分析客运产品的市场接受度和行业竞争性，辅助客运产品优化设计。

3.4 智慧营销

综合用户画像和客运产品画像，可以设计用户流失模型，分析旅客出行行为，判断旅客忠诚度，对流失用户进行挽留；可以设计产品推荐引擎，利用客户画像判断客户喜好，进行产品推荐，如新开行的列车产品、广告的投放、酒店产品、餐饮等。

（下转 P30）

2.2 分析结果应用

2.2.1 产品结构调整

北京出发至京广线沿线车站的动车组换乘旅客，在武广深间一次乘车抵达目的地的需求没有完全满足，特别是至岳阳、长沙，综合京广线运行图铺画特征，可考虑调整北京至广州深圳方向列车在武汉至广州深圳间的经停车站。

2.2.2 售票组织调整

就绝对换乘人数分析，北京经过武汉，换乘至长沙这一高峰换乘区间，换乘数量不足以支撑新增开行列车，但通过大数据分析和换乘人数预测，票额组织调整可以实现让北京至长沙间的旅客通过北京至长沙的列车完成运输需求，让换乘流通过票额组织成直通流。

2.2.3 旅客服务调整

在换乘成为武汉车站的一个必要业务时，可以通过大数据分析，获得换乘群体集中换乘时间和换乘车次，对应武汉站可以通过构建车站的武广换乘专用通道和旅客引导专员，实现换乘客流的顺畅运输，提升旅客的铁路出行体验。

综上所述，基于客运大数据的用户群体特征分析，在当前建立以旅客为中心的客运营销体系下，其提取的数据价值显而易见，分析结果应用于指导生产，产生的经济效益与社会效益明显。但大数据分析结果在庞大的铁路客运体系中的及时流转与控制流转是应用的一个局限和难点，也是我们下一步研究的重点。

3 结束语

铁路客运大数据分析是铁路信息化和铁路营运管理部门当前的研究热点和重点。重拾客运数据价值，检查铁路客运运营管理有效性和合理性，指明客运管理与组织的优化调整方向，实现大数据技术指导下的铁路客运高效生产是铁路客运大数据平台的重要意义所在。本文从铁路用户群体分析应用角度出发，对目前基于客运大数据平台的群体分析应用流程、架构进行阐述，分析结果用于指导铁路客运生产，证明了将大数据蕴含价值转化为服务于旅客的信息是优化铁路客运管理与组织的一个重要方法。

参考文献：

[1] Philip Kotler,Gary Armstrong. 市场营销原理 [M]. 赵　平，译 . 北京 : 清华大学出版社，1999，10 : 156-179.

[2] 缔元信 . 用户分群画像 : 抽样“猜想”让位于大数据“观察”[EB/OL].http://www.thebigdata.cn/YingYongAnLi /12697.html, 2014-12-05.

责任编辑　杨琍明

（上接 P16）

3.5 风险发现

基于用户画像的风险标签，可以设计铁路售票或客运组织过程的风险控制模型，识别风险用户，防范 12306 互联网售票过程中抢票、倒票和囤票行为，防范客运组织过程中的闯闸、动车组列车上吸烟、异常退签获利等行为。

3.6 应急指挥

综合运用客运大数据平台数据模型，可以设计客运应急指挥模型，提高应急事件识别能力，建立应急指挥流程和步骤，提高应急指挥的处置能力，提升突发事件发生时铁路旅客的出行体验。

4 结束语

客运是铁路行业的核心支柱产业，客运大数据技术的运用将成为未来客运增运增收的信息化支撑手段。铁路客运大数据平台的建设不可能一蹴而就，从基础计算资源和存储资源的投入，到画像系统的设计和调整；从路内外数据标签的获取，到各应用场景的技术实现，都将是一个不断持续迭代的过程。在此过程中，可以根据应用的紧迫性，边设计边建设边应用，逐步取得相应的效果。

参考文献：

[1] 邵明豪 . 数据预处理技术的具体实现形式研究 [J]. 网络安全技术与应用，2009（6）: 52-53.

[2] 陆嘉恒 . 大数据挑战与 NoSQL 数据库技术 [M]. 北京 : 电子工业出版社，2013.

责任编辑　王　浩