DOI: 10.11992/tis.201612005
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180409.1137.004.html

城市轨道交通线网数据中心与评估决策平台

张铭
（中国铁道科学研究院 电子计算技术研究所，北京 100081）

摘　要：在分析网络化运营条件下大规模数据特征的基础上，根据业务系统的数据融合需求，提出城市轨道交通数据中心平台的分层框架和功能定位。探讨了线网管理的数据结构体系、数据仓库的递阶逻辑建模、面向运营业务决策的应用集市等构建方法，并以线网客流特征识别的业务应用为对象，提出了数据集市的关联规则挖掘原理、预测立方体在贯通多类运营评估应用的计算方法。结合某城市轨道交通数据中心建设案例，描述了基于数据仓库的搭建过程及相关业务的调用逻辑，表明了线网数据管理对跨业务系统融合数据的意义，有效地提高了运营管理效率。
关键词：城市轨道交通；数据中心；网络化运营；数据仓库；运营评估；数据挖掘；决策；指标
中图分类号：TP319　　**文献标志码：**A　　**文章编号：**1673−4785(2018)03−0458−11

中文引用格式：张铭. 城市轨道交通线网数据中心与评估决策平台[J]. 智能系统学报, 2018, 13(3): 458–468.
英文引用格式：ZHANG Ming. A platform for a data center and decision making in urban rail transit[J]. CAAI transactions on intelligent systems, 2018, 13(3): 458–468.

A platform for a data center and decision making in urban rail transit

ZHANG Ming
(China Institute of Computing Technologies, China Academy of Railway Sciences, Beijing 100081, China)

Abstract: Based on large datasets for network operations in urban rail transit (URT), an approach on the multilayered framework and functions of an urban rail transit data center is presented. Critical network data management technologies are also discussed, including united data structures, hierarchical logical modeling of data warehouses, decision making, and passenger behavior recognition. Then, an algorithm is proposed based on data association rules and mining principles of forecast cube for evaluation purposes. Using a URT data center as an example, it describes data warehousing and related operations and points to the value of network data management in business-systems integration and in operational efficiency.
Keywords: urban rail transit; data center; network operation; data warehouse; operational evaluation; data mining; decision making; index

随着近年各特大城市轨道交通快速形成网络，其他城市规划建设也正在向网络化迈进。从运营管理角度，对日益庞大的线网进行全面的信息掌握与综合监察，作为辅助运营决策的手段，是必不可少的基础保障。由于传统的关系型数据库等存储和管理手段，已无法承受几何级数增长的数据量和适应快速获取分析结果的需求，对大存储、高效检索、即时分析、数据挖掘提出了更高要求。因此，搭建线网数据中心平台，通过采集各线路的运营信息，进行统一存储、处理、规划、共享，供日常运营监控、应急管理和运营组织优化等业务应用。此外，不同于积累多年的单线运营管理方式，线网条件下的运营指标核算、服务水平评估、线路间及枢纽的换乘接驳、网络客流的动态分析等频繁衍生出的新问题，在大数据应用的时代，提出了新的诉求。

数据中心及数据挖掘方面近年在各行业已有前瞻性探索[1-3]，王德文等[4]提出了基于云计算的新一代电力数据中心的基础架构，为智能电网的业务系统、数据挖掘与辅助决策等提供海量数据的存储、

收稿日期：2016−12−05.　**网络出版日期：**2017−03−17.
基金项目：国家自然科学基金项目 (U1334210)；北京市重点科技支撑计划项目 (Z151100001315002).
通信作者：张铭. E-mail: zm_zhangming@hotmail.com.

管理与计算环境; 汪祖云等[5]提出了交通行业的数据中心局域网和共享交换平台的架构设计理念; 罗亮等[6]从能耗业务角度提出了面向云计算数据中心的设计; 张彧锋等[7]从城市轨道交通运营安全保障角度提出了基于数据中心的应用管理系统; 梁艳平等[8]分析了轨道交通部分基础数据库元数据的内容, 基于各类设备故障数据进行诊断和挖掘分析[9-10]。本文从城市轨道交通网络化运营角度, 面向数据资源整合和挖潜, 提出线网数据中心的构建方案和线网运行监控状态、故障报警、近线和离线业务数据的数据中心资源池的机制, 以及为运营评估及业务提供决策平台。

1　网络化运营对数据融合的需求

1) 快速增长的数据规模

城市轨道交通各类系统覆盖机电专业监控系统、业务系统和办公系统等, 根据收集的数据, 列车运行和设备监控系统产生的报警数据日达 5 GB; 客流量因线网规模差异, 北京、上海地铁工作日均客运量 1 000 万人次以上, 广州地铁日均客运量 700 万人次以上, 深圳地铁日均客运量 300 万人次以上, 南京、武汉、成都、西安等城市地铁日均客运量 100 万人次以上, 进出站、断面、换乘客流及统计等各类数据量十分庞大。非结构化数据, 如一条线路 (按 30 站计) 产生的视频监控数据量 (按 15 日循环周期) 达 500 GB。按 5 条线路规模计算, 线网级系统的累计结构化数据量可达 3 TB/年, 非结构化数据因业务量差异数据量更大。随着线路开通里程的增长, 存储数据量很快达到 1 PB 及以上。数据结构、格式、类型混杂, 缺乏与业务的关联性, 存在基础数据不全而无效数据大量存储的现象, 为了提高数据质量, 有必要通过容纳大数据量级的数据仓库和标准化建模, 使数据资源效益得以发挥。

2) 多源异构的数据共享

各类数据资源包括来自互联网的现场报送信息、来自办公网的信息、来自生产内网的专业监控和行车信号信息。针对跨网、复杂业务数据的接口, 需要保障信息安全的同时, 采用高频数据采集、多通道队列、通信服务协议等多种通信方式实现采集, 不同类型数据的获取方式与业务系统特点及数据内容融合紧密相关。

3) 网络化运营统计分析与评估需求

线网条件下, 对行车类、客流类、能耗类、服务类等考核运营效果的各项指标计算, 不是简单地由各条分线路指标的叠加, 而是对网络化运营效益的综合考量, 需要对线网实际运行的数据深入分析。计算方法和评估指标体系等有待论证和检验, 这就需要历史数据资源的收集和对比校验。

4) 线网数据资源的挖潜

数据中心平台, 对累积的数据进行特征分析、建模和高效运算, 通过仿真、数据挖掘等方法, 为制订有效的节能方案、运营组织优化方案、指导新线规划和设备选型等提供决策依据。

2　网络化运营数据中心框架

根据网络化运营管理和决策分析的需求, 搭建面向多用户的信息集中共享、资源高效利用、运行可靠的轨道交通线网数据服务和综合业务的数据中心平台, 实现信息的统一采集、长期存储、统计分析、业务调用的功能。根据数据源的信息特点和支撑业务分支的目标导向[11-12], 将线网数据中心系统划分为“四个业务板块”, 即数据采集、数据管理、统计分析、评估决策, 同时与轨道交通企业的各类信息系统接口, 形成稳定、长期的数据资源融合与挖掘运用。

城市轨道交通的数据中心平台具有其特殊性: 首先, 数据源来自于各分立系统, 覆盖车辆、行车、机电设备、客流、运营管理等多个专业, 数据内容具有专业的分散性; 其次, 围绕运营决策与评估考核业务, 须对应于业务主题找到各专业数据之间的关联性, 并聚合于具有高度自组织性的主题域; 再次, 数据类型和内容众多, 具有近线、离线等数据采集时效的多样性, 以及随时空变化特性、业务视角差异性和多维分析预测的复杂性。因此, 数据中心的框架、数据融合的深度及专业化的数据模型, 对于轨道交通线网级别的运营管理和决策支持具有重要意义, 也是搭建城轨数据中心平台面临的主要问题。

2.1　数据资源整合平台

1) 监控数据融合与共享

采集各线路控制中心及业务系统的信息, 包括行车、供电、设备、防灾报警、客流、视频监控等, 可归纳为 13 类运营监控系统信息, 7 种数据结构类型[13]。建立数据共享平台, 汇总各类数据, 如图 1 所示。

在数据采集的基础上, 通过统一处理对多专业的信息集成与实时监察, 可掌握线网行车、线网电力运行状态, 包括多线路共享主变电所能耗监控与联动控制; 采集线网客流的出、入站客流数据[14-15], 线路断面客流、换乘客流信息, 从实时客流监察预警和历史客流预测角度划分数据结构。划分实时数据、近线数据和离线数据, 实时信息用于线网运行状态的监察, 根据故障报警信息及时启动应急处

置；近线数据和离线数据分别载入历史库，用于各种维度的统计和评估核算。

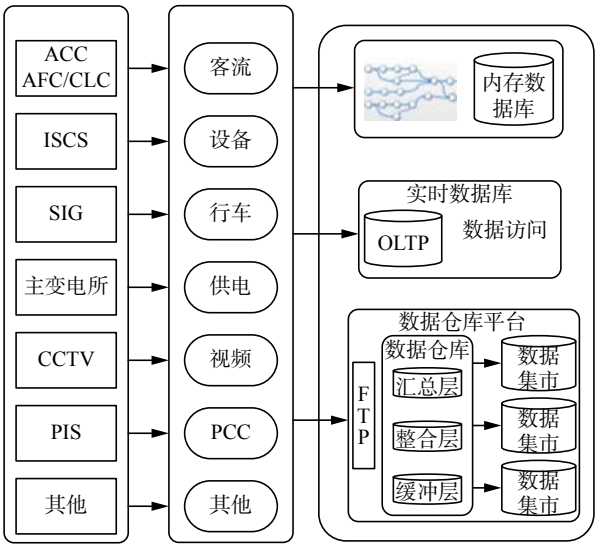

图 1　数据采集逻辑原理

Fig. 1　Principle of data collection

2) 基于数据仓库的一体化数据管理

根据业务分析需求建模导入数据仓库，将行车、设备、调度指挥、突发事件、客流等数据分类、存储、分析、挖掘，建立完整的元数据管理体系，包括元数据的定义、收集、管理和发布的流程。

3) 基于大数据与多媒体的集成应用

轨道交通企业对外发布的客流信息、运营信息、突发事件应急信息等，利用实时库的快速处理特性和应用集市的逻辑生成机制，通过内网、移动客户端、数据接口等方式，实现集通信工具、呼叫中心等方式一体化的信息发布。通过知识库及预测结果调用综合，将分析和反馈信息进一步收集，实现信息的收纳和共享。

4) 网络化运营统计分析与评估决策

针对运营考核和监管需求，构建网络化运营业务数据的统计、查询和运营评估的应用集市，形成业务调用的关联关系的统一视图，并进一步结合远期规划，建立评估决策模型，为多维、分段的历史数据分析挖掘和预测提供基础。

2.2　数据中心的分层框架

根据不同的业务对象，建立分层架构，即数据接口层、数据模型层、应用集市层、业务访问层，上层面向用户访问，应用框架如图 2 所示。

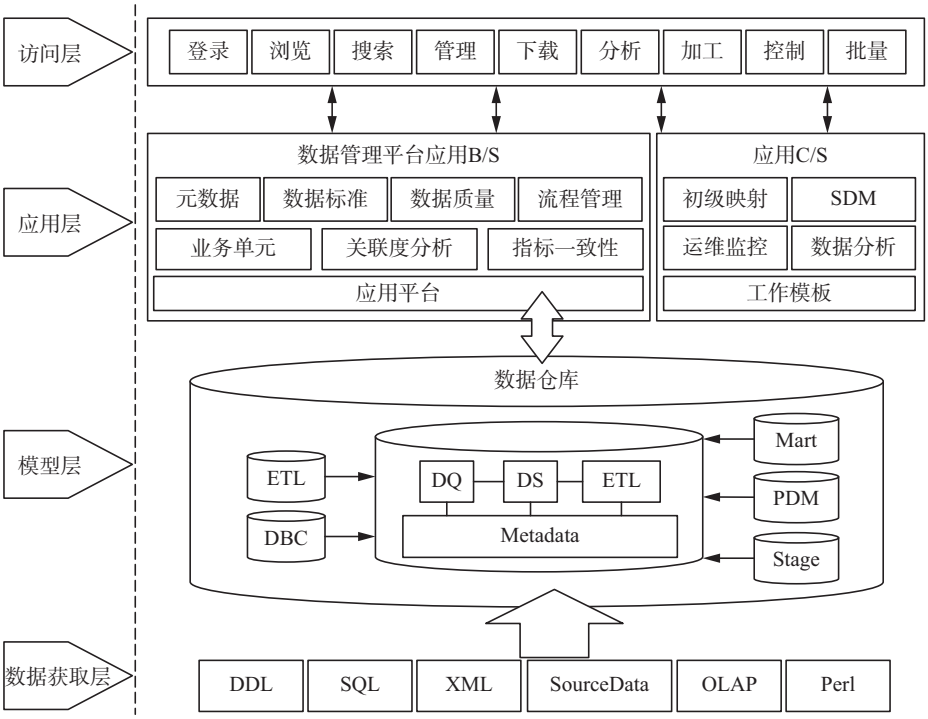

图 2　数据中心平台的分层框架

Fig. 2　Schematic of the data center platform

1) 数据接口层

主要承担数据的采集，作为系统接口通道，根据接口数据的实效性、数据量、数据内容等不同条件，设置接口通信协议转换实现数据的获取，即可设计接口模型，将接口类型标准化、规则化。获取的数据通过抽取、清理、转换、加载过程转入数据建模，根据业务规则建立统一视图后，为数据结构标准化做准备[16-17]。按照不同分类形式划分为：

①按业务类型，划分为客流数据、列车运行数据、设备运营数据、票务数据、清算数据、应用系统的融合数据、统计数据等。

②按数据类型，划分为结构化数据和非结构化

数据。结构化数据包括可建立数据表统一存储在数据库中的数据, 如基础设施、业务类数据; 非结构化数据主要包括站点监视视频信息和规范与图纸, 以及预案、数据接口等文件类信息。

③按时效性可划分为实时数据、非实时数据。

实时数据: 在数据变化时立即由控制端控制器传给采集端, 即发送端、接收端同步, 包括行车运行信息、设备运行信息和故障报警信息等。非实时数据: 这类数据在数据变化时经一定间隔时间后传给采集端, 包括各线路的运营数据、阶段统计数据等。

不同分类间互有交叉, 例如: 列车运行类数据包括列车运行的具体位置、时间等实时信息, 及列车运行图等非实时信息, 因此可对数据多级划分:

①阶段性信息: 按照设定的采集周期自动接收各线路上传的运营数据, 如车站一段时间内的温湿度统计、线路的用电量统计、各站的客流数据统计等, 为运营人员分析整体情况进行决策积累数据。

②实时采集: 用于满足数据中心实时、非实时业务需求, 通过特定通信协议, 监控源系统实时上传所需数据, 上传时间可通过参数化设置。

③定时采集: 用于满足数据中心离线业务需求, 系统通过特定通信协议, 在预定的时段内 (通常为非运营时段) 向生产系统采集所需数据。各生产系统在预定的时段前, 须以预定的格式存档。

此外, 数据仓库形成统一的数据资源池, 为上层业务的调用封装出接口供访问数据。

2) 数据模型层

根据大规模数据和线网综合业务的处理需求, 采用数据仓库作为线网数据中心平台的基础数据库。由于数据源系统很多[18], 从分散而异构的源数据到最终的层次分明的展示数据, 需要设置多层级过滤, 对数据仓库进行分层设计。

业务建模划分为 5 个层面: 调度管理、客运管理、车辆管理、设备管理、安全监察。

①调度管理模型: 行车、设备、消防环控调度、指挥与运营调度、突发事件应急处置、事故处理及调查、夜间施工管理。

②客运管理模型: 运输计划及运行图、运营与应急协调、质量分析与控制考核、质量管理、客运组织与服务、站务与乘务。

③车辆管理模型: 检修计划、故障分析、采购、车辆调度运力优化、技术改造、机务管理。

④设备管理模型: 维修计划、固定资产管理、故障排查、新线及试运行管理、多专业协同检修。

⑤安全监察模型: 安全巡查、应急预案管理、事故统计、安全考核评估、案例知识库。

将逻辑建模作为重要环节, 使其直观映射业务部门的需求, 如设定对外预警与预防准备和运营组织调整方案的逻辑关联模型等。依据业务规则转译为模型内的关系, 清晰地反映业务操作模式。设计的逻辑模型满足第三范式 (3NF), 减少数据冗余, 提高访问效率[19]。建模的过程中, 对各种原始数据、衍生数据和元数据进行标准化处理, 形成有序的标准数据并进行统一管理和维护, 保证存储数据的安全, 具备保护机制。

3) 语义应用层

面向轨道交通日常业务进行应用集市的设计, 包括统计分析集市、运营评估集市、决策分析集市、客流查询集市。采取在数据仓库中划分空间, 建立逻辑集市, 单独划定逻辑区域用于存放前端应用访问的实体表或视图, 不放置处理的中间数据, 并严格遵循命名规则, 同时多个应用集市之间数据重复利用。以客流管理的应用集市为例, 逻辑分区设置为“乘客分群、客流特征分析、路网不均衡性分析、客流预测、重大活动与节假日分析、车站限流分析、突发事件应急响应、换乘枢纽接驳、客流预测”。

因为应用集市依赖于业务需求和数据仓库的整体建设规划, 所以对数据仓库的总体设计的高度稳定性提出极高要求。为各数据集市分配独立的数据库区域, 空间大小可根据实际使用大小灵活调整。通过负载管理来分配资源, 实现提升数据集市的服务能力。根据“不同的业务策略”在“不同时段”为“不同类型的对象”提供“不同的资源权限”, 从而为不同类型用户提供差异化服务, 资源权限的切换由数据仓库平台自动完成, 资源权限由系统自动分配或执行变更。

4) 安全管理体系

由于生产运营调度系统通常位于企业生产内网, 属于信息安全等级保护三级, 而日常业务系统位于办公网, 其中部分系统对外发布信息, 如时刻表、乘客查询信息等, 则与互联网相连。因此, 对应不同级别网络, 建立信息安全管理体系, 各系统数据进入数据仓库融合。将线网数据中心平台的系统划分多个区, 包括应用区、数据区、接口区等, 设置安全管理中心, 通过配置硬件安全设备, 如网闸、防火墙、堡垒机、入侵检测、入侵防御、审计系统等, 配置防病毒软件、用户认证、数据安全等安全过滤和控制, 保障信息安全。

3　基于数据仓库的建模

3.1　线网数据结构预定义

线网数据中心需建立统一的数据结构体系, 在构建数据仓库前首先执行 ETL(extract-transform-

load) 过程，即数据从不同的数据库或异构数据源中，流向统一的目标数据库，去映射源数据，载入业务模型的数据仓库或数据集市。ETL 连接着数据仓库和汇集数据的业务系统，确保新的业务数据持续流入数据仓库，同时保证生成的结果反映最新的业务动态。

1) 数据抽取

数据抽取包括增量、全量及自定义抽取方式，具备异步和同步抽取，灵活设定抽取频率。对行车、设备监控、故障报警、时刻表文件等大批量数据以日为单位增加抽取，对客流类数据以文件存储的，以单个文件传输的周期为单位，作为数据抽取频率可全量抽取。

2) 数据转换

从数据采集系统获取源数据时进行数据转换，包括数据的定义、数据结构和错误数据的转换处理等，如时刻表文件的解析分为工作日、非工作日、节假日，各自成表。转换的内容包括格式和类型转换、数据的翻译、匹配、聚合等。

3) 数据加载

将常规格式的数据以批量模式加载到数据仓库，并对部分业务类数据分别处理入库，如以 5 min 为单位积累的客流文件。也可并行加载，如 BAS 和 PSCADA 数据表，采用自动加载模式，但对于线路控制中心 OCC 上报的运营日报、月报等需手工加载，如直接追加、全部覆盖、更新追加。

4) 数据检查与异常控制

由于各数据源的数据质量不可控，因此进行数据检查，包括接口数据的及时性、完整性和正确性，设置各种类型的数据质量检查规则、检查规则的上下阈值，在第一时间根据规则提醒相关人员处理数据质量故障，并对各类异常数据进行必要的处理。经过处理的数据划分为以下 3 种类型。

①基础数据：基础数据层面定义为全局概念，以便对一些基础或通用类信息保持一致的认识，如管理者、设备。

②公共代码：对多个源系统不一致的数据定义进行整合，供其他系统引用，以保证可识别的一致性，如基础设施、专业。

③统计指标：设置以业务为导向的公式化计算引擎，提供可分解的全局性统计指标，并使计算调取的数据遵循这些指标的数据标准。

系统中元数据的业务流程逻辑关系如图 3 所示。通过建立完整的元数据管理体系，包括元数据的发布、浏览、查询、关联分析及追溯等，业务人员从而及时准确地了解数据仓库的数据内容。以此为基础，以便快速进行数据查询、数据资源管理、数据模型管理、业务信息以及变更管理等。

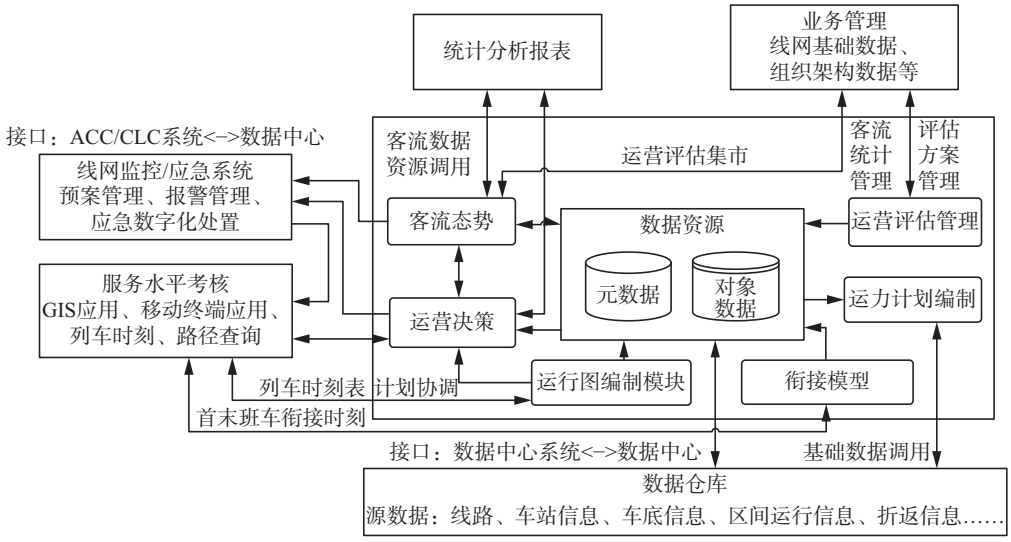

图 3　系统元数据的业务流程

Fig. 3　Business flow of system metadata

3.2　综合业务承载的主题域设计

根据数据仓库面向主题的特性，按照数据模型分主题组织和存放数据，对所有数据分类，根据各自业务划分不同的主题，由主题域来建模。主题域是对某个主题进行分析后确定主题的边界。根据线网数据中心的业务，将数据仓库的数据模型设计为 10 个主题域，分别为当事人 (party)、线网 (subway network)、位置 (location)、设备 (equipment)、行车 (trip)、OD(origination and destination)、客流 (passenger flow)、票务 (ticket)、清算 (clearing)、事件 (event)，其构件关系如图 4 所示。

以行车信息的主题域为例，数据视图如图 5 所示。

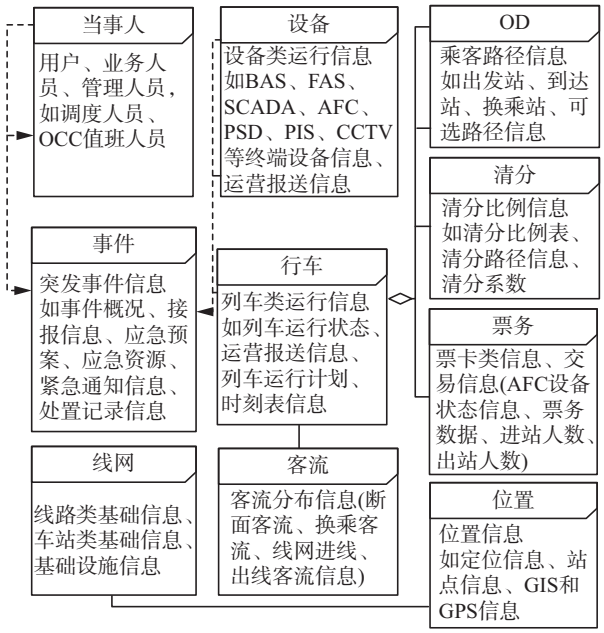

图 4　数据仓库专业主题域的构件关系

Fig. 4　Component relation of commerial data warehouse

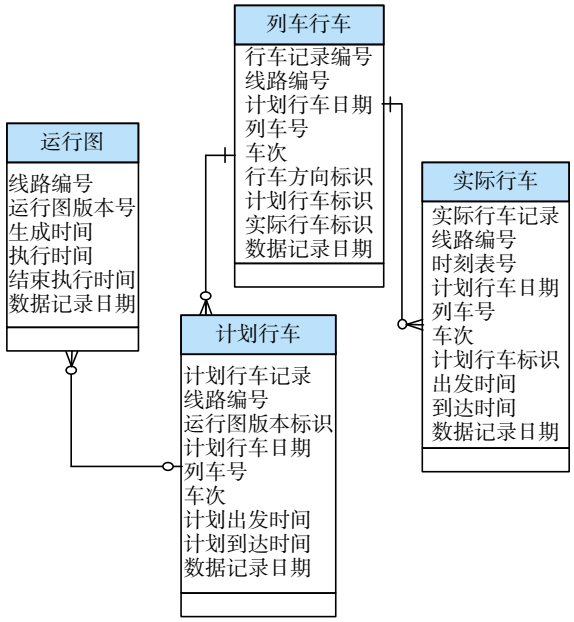

图 5　行车信息主题域的数据视图

Fig. 5　Data flowchart of a train-based domain

划分不同类型主题域，由于基础数据是业务系统或各渠道采集进入数据仓库的，通过唯一定义，归纳资源、行程、当事人等业务数据；而公用数据是在基础数据基础上按照一定的业务规则汇总的数据，被多业务主题所共享；专用数据则是从部门视角或独立业务主题出发，经过特定业务智能产生的数据，如预测评估、故障监视等专用数据等。

3.3　数据仓库的递阶数据建模

将业务逻辑关系设计为运行规则，进行线网数据仓库的建模，建立递阶分层机制，递阶模型按照层次进阶关系设计为 4 阶，即临时数据、核心数据、汇总数据、专用数据。

1) 临时数据Ⅰ阶

存放从数据源采集的原始交易数据，保持与数据源系统相同的表结构，用于部分近实时性报表。为了生成业务系统的镜像区，作为核心数据层的数据来源，将保留从数据源中抽取的业务数据，数据存储的周期设计为 1 周，并定期转存至数据仓库中。

2) 核心数据Ⅱ阶

结合源系统的数据现状和业务逻辑模型，设计数据模型，形成的数据结构用于数据的管理和分析，包括基础设施公用数据、线网运行状态记录、客运、维修、资产管理等。

3) 汇总数据Ⅲ阶

对不同粒度的轻量级汇总或高度汇总，来提升专用数据阶的数据生成效率，存放的数据是专用数据阶的多个结果数据的源数据，以便重复使用。

4) 专用数据Ⅳ阶

顶级分层面向运营业务统计、评估、信息发布等应用，经过公共指标或外部计算的结果，直接供各应用功能调取。

建立递阶模型后，就可对每一个主题域进行细化、分解，直到明确模型中的业务概念后，对主题或者实体之间的关系进行建模。定义逻辑数据模型 LDM(logic data model)，适应源系统结构变化、业务规则变化或新增业务，屏蔽源系统变化对应用系统的影响，并在长时间内保持稳定。

4　运营评估决策的应用集市

利用数据仓库形成的资源池，通过分析挖掘实现数据的多维查询，为统计分析、信息服务提供服务，实现跨业务的数据整合共享，满足运营公司各部门对各项业务的需要。因此，根据业务标的，建立应用集市，为数据仓库定向提供指令集，主要包括运营评估类、统计分析类、运营业务挖掘类。

1) 运营评估体系

运营指标的创建包括数据建模、数据模型导入、业务指标创建和发布。其中基础指标定义是针对直接和数据关联的指标。运营评估类的应用集市主要用于计算线网级的运营指标，包括行车及设备类 30 项指标，客流类 39 项指标，服务水平类 18 项指标，能耗类综合 5 项指标，票卡类 14 项指标。

衍生指标和用户自定义指标通过不同的组合计算及函数定义，结合常量和其他衍生指标等计算后生成的指标，包括公式管理、指标度量、维度管理等。归纳调用的模式，包括同比分析、环比分析、趋

势预测等，实现时间维度从年到分钟的逐级钻取，时间钻取维度的最底层是 1 min。

2) 运营数据的挖掘

线网客流是数据中心的主要业务应用之一，也是占用数据仓库最大空间的数据。乘客出行特征识别与客流预测是业务挖掘应用集市的主要应用。从进出站客流、线路上下行区间断面客流、换乘客流、线网客运量等多层次、多维度时空角度分析客流运行规律，进行需求预测，能够为制定合理的列车开行方案和组织高效运输提供重要的决策依据。

客流分析的数据建模主要依托 OD 分布和清分比例，通过特征要素提取，采用基于 R 语言的关联规则算法，构建多维群组矩阵，辨识客流乘距、时段特征、客运量的分布特点。目标要素概括如表 1。

表 1　客流特征识别要素逻辑数据模型

Table 1　Logical data model of passenger features recognition

序号	客流特征识别目标	特征值	最小置信度
1	线路乘距人次分布时段	乘距分段人数	线路里程的 50%
2	线路乘距人次分布日期	乘距分段人数	线路里程的 50%
3	线路乘车站数人次分布时段	乘车站数分段人数	线路车站数的 50%
4	线路乘车站数人次分布日期	乘车站数分段人数	线路车站数的 50%
5	线路乘车时间人次分布时段	乘车时长分段人数	平峰时段全线运营累计时长的 30%
6	线路乘车时间人次分布日期	乘车时长分段人数	平峰时段全线运营累计时长的 30%
7	线路平均值时段	线路客运量	(日均客运量/累计时长)>低峰小时客运量
8	线路平均值日期	线路客运量	(日均客运量/月均客运量)>月最小客运量
9	线网乘距人数分布	乘距分段人数	线网乘距的 50%
10	线网乘距人数分布日期	乘距分段人数	线网乘距的 50%
11	线网乘车站数人数分布时段	乘车站数分段人数	线网平均乘车站数的 50%
12	线网乘车站数人数分布日期	乘车站数分段人数	线网平均乘车站数的 50%
13	线网乘车时间人数分布时段	乘车时长分段人数	平峰时段全线运营总时长的 30%
14	线网乘车时间人数分布日期	乘车时长分段人数	平峰时段全线运营总时长的 30%
15	线网旅行时间人数分布时段	旅行时长分段人数	平峰时段全线运营总时长的 30%
16	线网旅行时间人数分布日期	旅行时长分段人数	平峰时段全线运营总时长的 30%
17	线网平均值时段	路网进站量	(平峰日均进站量/累计时长)>低峰小时客运量
18	线网平均值日期	路网进站量	(平峰日均进站量/累计时长)>低峰小时客运量
19	线网 OD 客流	路网 OD 乘距	无限制

在目标导向和特征值的基础上，根据关联规则定义客流分析的应用集市，数据模型如图 6 所示。

根据特征分析结果，计算线网客流特征指标，评估客流在线网中的分布情况和服务水平，此处仅以典型指标为例。

选择线网层级的运营评估指标“线网平均运距”“线网换乘系数”的计算过程说明指标的数据模型和自定义参数的配置管理。

①线网平均运距，即统计期内线网中乘客平均一次出行全程的总乘车距离，表示为 $\varphi=\sum_{l\in L}\sigma(l_i)/\sum_{l\in L}\gamma(l_i)$，其中 φ 为线网平均运距，σ 为线路 l_i 客运周转量，γ 为第 l_i 线路的进线量。

②线网换乘系数，即统计期内，乘客在路网内完成一次出行需乘坐的平均线路条数，表示为 $\tau=\sum_{l\in L}\delta(l_i)/\sum_{l\in L}\gamma(l_i)$，其中 τ 表示线网换乘系数，δ 表示线路 l_i 的客运量。

将以上指标中线路 l_i 均从“基础数据”和“公共代码”数据识别并导入数据即可获得。而客运周转量和客运量的计算值是由客流量、正线运营里程等基础数据计算得出的中间结果，可存储于指标定义的暂存表中作为计算参数。

3) 客流预测立方体优化

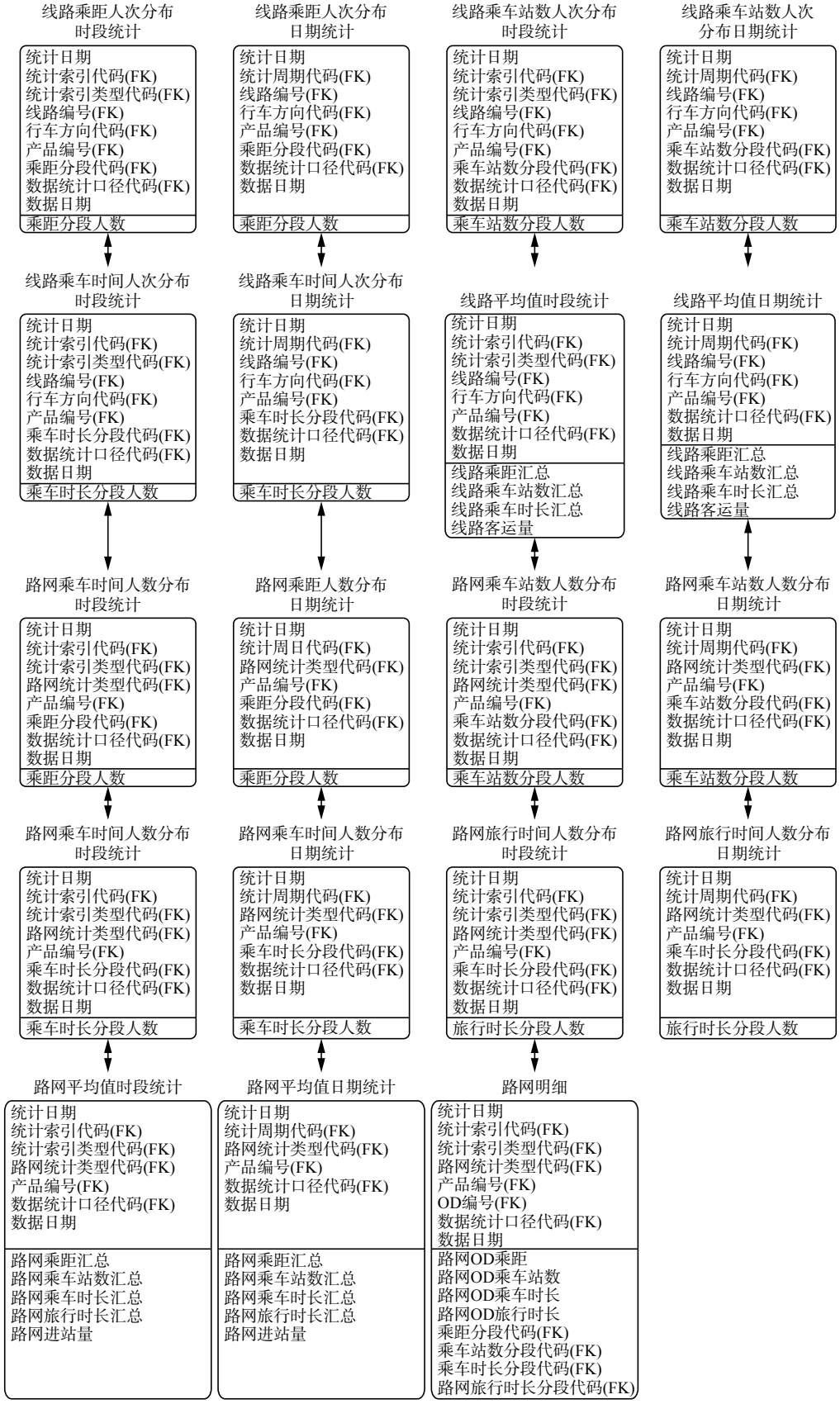

图 6　客流应用集市的数据建模

Fig. 6　Data flow model of a passenger-based application market

传统的客流预测一般通过时间序列、票价费用等要素进行需求路径分配预测客流量[20-21]。在数据中心平台中根据客流特征识别的指标要素, 对各种粒度数据的 OLAP 交互分析, 使用多维数据模型和预测立方体实现客流的多维空间预测建模。

预测立方体算法：

1) 计算聚集：在显示维度的时段 $\alpha()$、费用 $\beta()$、客流特征 $\gamma()$ 定义立方体数据空间，在其作用下的客流量聚集度量 M 用于存放立方体中的所有元组 Passenger()。首先将数组划分为块，通过访问立方体单元计算在线网上某一路径下的可能客流量。

2) 划分客流运行特征的置信区间：将满足客流特征 $\gamma(i)$ 的指定条件下置信水平为 95% 以上的客流类型记作 count()。此处指定条件包括工作日、节假日、大型活动、突发事件、早晚高峰、平峰时段。累计对应的在网车站数、时长、乘距的进出站客流量、断面客流量、换乘客流量。

3) 查询数据立方体“关注点”客流：提升小样本的置信度，如多种交通方式的枢纽集散站点、票价优惠路径、避开拥堵路径的可替代路径选择等，分配权重值，在需求客流量基础上适度扩展。需精确地度量维值与立方体值的相关性，通过语义类似值即可联机分析。

4) 计算预测客流量：调取线网的任意组合路径，使用数据立方体快速重复客流预测模型的构建，预测立方体的每个单元值等于该单元数据子集上的基础客流预测量，经加权修正计算得到预测客流量。

5) 预测值的优化：采用基于概率的组合方法，对最细粒度的单元构建模型。以断面客流量需求预测为例，给定分段路径的客流属性子集，将粒度集合$P<p_1,\cdots,p_d>$的预测立方体定义为 d 维数组，其中每个单元 (条件[O_iD_i 路径对；上行；工作日早高峰时段；>线网平均乘距]) 的值即为该单元定义的基础客流预测量估计值的预测修正量。

因此，利用线网大规模客流数据的特征分析结果，在既有客流需求预测量基础上结合各城市实际客流特点予以修正，在很大程度上改善了由单线客流预测方法直接得出线网客流预测理论计算值的单一性。

5　案例

依托某城市轨道交通的已运营构成线网，近 3 年内投入运营将达到 9 条以上线路，正在快速积累各类业务数据。随着企业信息系统衍生，形成了大量分立的小型业务系统，数据内容交叉，关联信息无法共享的问题日益显著。由于快速增长的数据量，简单整合的数据容量规模大，业务统计分析响应时间受关系型数据库的影响已无法支持实时业务，因此面向网络化运营的需求，搭建线网数据中心平台，承担数据采集和资源整合。

按照该轨道交通线网数据管理标准的要求，对源系统统一加工和整合，存储细粒度的历史数据区域，为各业务系统调用提供一致、规范的数据。该数据中心管理的数据包括：

1) 业务数据，包含了轨道交通内部信息系统的原始数据、衍生数据、过程数据等；

2) 线网基础数据，覆盖相关的各类文件数据、基础线网数据和基本参数数据；

3) 配置数据，主要包括用于支撑业务工具和方案的相关配置数据和业务资源数据。

建立数据仓库，设定主题域和逻辑模型，定义“维度表”，作为基础公共表，此类代码表在明确标识代码值与业务含义的基础上，还具备逐级汇总功能，细化了各个维度层级之间的上下级关系，为表的逐层汇总提供了先决条件。以公共代码表为例，说明数据仓库的基础表关联关系，设计基础代码的逻辑模型如图 7。

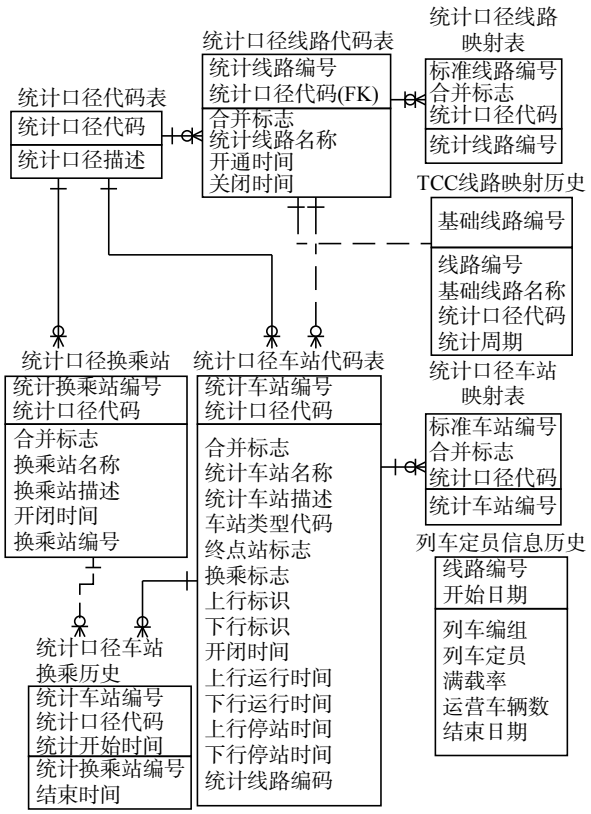

图 7　数据中心基础代码逻辑模型

Fig. 7　Logical model of a basic code for a data center

采集数据源包括行车信号系统、机电设备综合监控系统、票务清分系统等，采用 TeraData 数据仓库产品，导入数据处理，构建主题域进行数据建模，建立数据中心的系统框架。数据中心平台包括以下业务模块。

1) 数据采集系统：包括设备监控实时信息采集，文件传输，采集接口通道监控，接口数据质量管理，接口双冗余双实时采集数据配置等模块。

2) 数据管理系统：包括基础数据字典管理，数据存档备份管理，主数据管理，主题域关联视图可

视化, 数据同步管理等模块。

3) 统计评估系统: 包括行车类、设备类、客流类、服务类的基础指标, 衍生指标, 自定义指标的核算, 多维统计, 定制报表报告等模块。

4) 运营挖掘与决策系统: 包括线网行车计划智能生成, 时刻表衔接方案, 客流预测仿真等模块。

在数据仓库的基础上, 为各项业务系统接口开放应用集市的调用方法如图 8 所示, 包括: 各专业监控系统设备与资产管理系统物资编码的关联; 设备故障与运维管理系统的维修单任务派发关联; 线网供电电量计算与运营评估考核指标关联; 优化列车运力配置计划、线网列车运行计划、辅助生成列车运行图等对换乘枢纽衔接方案的关联; 提供路径查询、检索和路径可达性提示引导与售检票系统的客流量关联等。

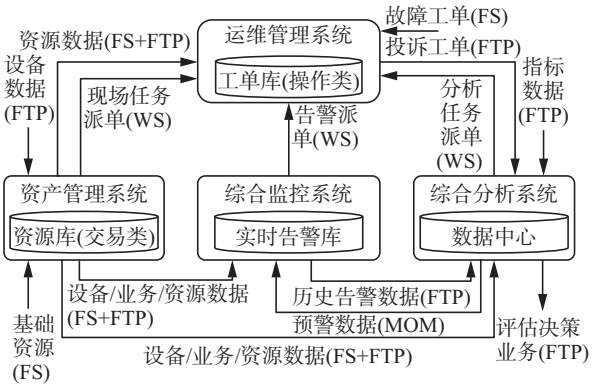

图 8　应用集市与业务系统的调用

Fig. 8　Transfer of an application market and a business system

导入 6 个月的进出站客流量、断面客流量、换乘客流量, OD 路径和清分比例表数据, 以 6 号线增量客流为模拟对象, 应用客流特征识别关联规则算法和客流需求预测模型, 将预测结果叠加到线网图上, 客流态势预测仿真结果如图 9 所示。

图 9　基于特征识别的客流预测仿真

Fig. 9　Passenger forecast simulation based on character recognition

根据业务提炼对数据仓库具有共性的数据访问、统计需求, 构建一个面向需求的、共享的访问服务的公共数据集。其数据流向是从基础共享数据层抽取数据, 再对不同数据内容详细程度、不同时间和空间维度的数据按需提取。在数据仓库的基础上, 调用运营评估应用集市接口计算相关指标, 包括线网满载率、线网能耗指标、线网设备故障率等。

6　结束语

我国多个大城市将很快面临线网级的运营管理, 构建数据中心平台, 将通用型业务系统和基础数据结构统一规划, 有利于避免随业务延展, 各种分立系统数据共享困难, 系统重复建设、功能交叉的情况。同时, 进一步分析数据融合的关联性, 从城市轨道交通业务角度, 加强数据对业务的承载内容和范围的挖掘, 为线网层面的运营评估、服务水平考核、多运营主体协调提供决策支持, 也为轨道交通网络化运营管理、辅助决策、新线规划指导等提供支撑。

参考文献:

[1] 徐俊刚, 裴莹. 数据 ETL 研究综述[J]. 计算机科学, 2011, 38(4): 15–20.
XU Jungang, PEI Ying. Overview of data extraction, transformation and loading[J]. Computer science, 2011, 38(4): 15–20.

[2] 陈慧萍, 陈岚峰, 王建东. 大型数据仓库实现技术的研究[J]. 计算机工程与设计, 2006, 27(21): 3956–3958, 3961.
CHEN Huiping, CHEN Lanfeng, WANG Jiandong. Research on issues in developing large data warehouses[J]. Computer engineering and design, 2006, 27(21): 3956–3958, 3961.

[3] 胡侃, 夏绍玮. 基于大型数据仓库的数据采掘: 研究综述[J]. 软件学报, 1998, 9(1): 53–63.
HU Kan, XIA Shaowei. Large data warehouse-based data Mining: a survey[J]. Journal of software, 1998, 9(1): 53–63.

[4] 王德文. 基于云计算的电力数据中心基础架构及其关键技术[J]. 电力系统自动化, 2012, 36(11): 67–71, 107.
WANG Dewen. Basic framework and key technology for a new generation of data center in electric power corporation based on cloud computation[J]. Automation of electric power systems, 2012, 36(11): 67–71, 107.

[5] 汪祖云. 交通数据中心总体架构与数据共享交换平台的设计研究[J]. 交通运输系统工程与信息, 2008, 8(3): 23–28.
WANG Zuyun. Framework and data share platform of transportation data center[J]. Journal of transportation systems engineering and information technology, 2008, 8(3): 23–28.

[6] 罗亮, 吴文峻, 张飞. 面向云计算数据中心的能耗建模方法[J]. 软件学报, 2014, 25(7): 1371–1387.
LUO Liang, WU Wenjun, ZHANG Fei. Energy modeling

based on cloud data center[J]. Journal of software, 2014, 25(7): 1371–1387.

[7] 张彧锋, 韩泉叶. 城市轨道交通线网运营安全保障平台的设计与仿真实现[J]. 城市轨道交通研究, 2014, 17(12): 33–38.
ZHANG Yufeng, HAN Quanye. Design and simulation of network operation security platform in urban rail transit[J]. Urban mass transit, 2014, 17(12): 33–38.

[8] 梁艳平, 刘仍奎, 芮小平, 等. 轨道交通基础数据库元数据内容体系研究[J]. 交通运输系统工程与信息, 2005, 5(3): 61–64.
LIANG Yanping, LIU Rengkui, RUI Xiaoping, et al. A metadata content system of rail transit fundamental database[J]. Journal of transportation systems engineering and information technology, 2005, 5(3): 61–64.

[9] 赵金楼, 成俊会, 岳晓东. 基于贝叶斯网络的海洋工程装备故障诊断模型[J]. 哈尔滨工程大学学报, 2014, 35(10): 1288–1293.
ZHAO Jinlou, CHENG Junhui, YUE Xiaodong. Fault diagnosis model of marine engineering equipment based on Bayesian networks[J]. Journal of Harbin engineering university, 2014, 35(10): 1288–1293.

[10] 张天瑞, 于天彪, 赵海峰, 等. 数据挖掘技术在全断面掘进机故障诊断中的应用[J]. 东北大学学报: 自然科学版, 2015, 36(4): 527–531, 541.
ZHANG Tianrui, YU Tianbiao, ZHAO Haifeng, et al. Application of data mining technology in fault diagnosis of tunnel boring machine[J]. Journal of northeastern university: natural science, 2015, 36(4): 527–531, 541.

[11] 杨承东, 徐余明. 基于综合监控系统的线网指挥中心构建方案研究[J]. 城市轨道交通研究, 2013, 16(10): 25–29.
YANG Chengdong, XU Yuming. Structural scheme of traffic command center based on intergraded supervisor control system[J]. Urban mass transit, 2013, 16(10): 25–29.

[12] LODHA A, GUMASTE A, WANG Jianping, et al. CAVALIER architecture for metro data center networking[C]// Proceedings of the 5th International Conference on Broadband Communications, Networks and Systems. London, Britain, 2008: 169–174.

[13] 张铭, 王富章, 程超. 城市轨道交通设备故障聚类与贝叶斯网络预警[J]. 计算机工程与应用, 2016, 52(11): 259–264.
ZHANG Ming, WANG Fuzhang, CHENG Chao. Equipment fault clustering and Bayesian network pre-alarm of urban rail transit[J]. Computer engineering and applications, 2016, 52(11): 259–264.

[14] 冷彪, 赵文远. 基于客流数据的区域出行特征聚类[J]. 计算机研究与发展, 2014, 51(12): 2653–2662.
LENG Biao, ZHAO Wenyuan. Region ridership characteristic clustering using passenger flow data[J]. Journal of computer research and development, 2014, 51(12): 2653–2662.

[15] ITOH M, YOKOYAMA D, TOYODA M, et al. Visual fusion of mega-city big data: an application to traffic and tweets data analysis of Metro passengers[C]//Proceedings of 2014 IEEE International Conference on Big Data. Washington, USA, 2014: 431–440.

[16] 张文燚, 项连志, 王小芳. 支持分布式大数据应用建模的模型理论[J]. 哈尔滨工程大学学报, 2015, 36(5): 671–677.
ZHANG Wenyi, XIANG Lianzhi, WANG Xiaofang. A model theory for distributed application modeling on big data[J]. Journal of Harbin engineering university, 2015, 36(5): 671–677.

[17] 石庄彬, 陆文学, 张宁. 数据挖掘技术在轨道交通 AFC 系统中的应用[J]. 都市快轨交通, 2015, 28(1): 23–27.
SHI Zhuangbin, LU Wenxue, ZHANG Ning. Application of data mining for urban rail transit automatic fare collection[J]. Urban rapid rail transit, 2015, 28(1): 23–27.

[18] SAMADI P, WEN Ke, XU Junjie, et al. Software-defined optical network for metro-scale geographically distributed data centers[J]. Optics express, 2016, 24(11): 12310–12320.

[19] 沈斌. 复杂网络与数据挖掘: 研究范式的比较和整合[J]. 复杂系统与复杂性科学, 2014, 11(1): 48–52, 59.
SHEN Bin. Comparative study and integration of research paradigms of complex networks and data mining[J]. Complex systems and complexity science, 2014, 11(1): 48–52, 59.

[20] 李夏苗, 黄桂章, 汤杰. 基于 OD 反推模型预测客运通道客流量[J]. 铁道学报, 2008, 30(6): 7–12.
LI Xiamiao, HUANG Guizhang, TANG Jie. Passenger flow forecasting based on OD-matrix estimation model[J]. Journal of the China railway society, 2008, 30(6): 7–12.

[21] 李明, 王海霞. 轨道交通车站客流预测模型研究[J]. 铁道工程学报, 2009, 26(3): 67–72.
LI Ming, WANG Haixia. Study on the model for predicting the passenger volume of rail communication station[J]. Journal of railway engineering society, 2009, 26(3): 67–72.

作者简介:

张铭, 女, 1979 年生, 副研究员, 博士, CCF 会员, 主要研究方向为轨道交通智能系统工程、安全与应急规划。主持和参与国家自然科学基金、863 计划、国家科技支撑计划、住房与城乡建设部科技计划示范工程、北京市重大科技计划、企业联合等多项课题, 发表学术论文 30 余篇。