基于大数据分析的青岛地铁客流画像分析*

罗情平[1]　左旭涛[1]　张蓓蓓[1]　杜可亮[2]

（1. 青岛地铁集团有限公司，266045，青岛；2. 北京北大千方科技有限公司，100193，北京//第一作者，正高级工程师）

摘　要　青岛地铁线网运营管理与指挥中心采用大数据分析方法实现了城市轨道交通客流分析及画像功能。采用AFC（自动售检票）数据结合ATS（列车自动监控）信息的方法实现更精确的出行路径匹配，克服了传统客流分析算法的准确性缺陷。从客流角度实现了乘客、车站、列车、区间画像功能，结合ISCS（综合监控系统）数据实现电扶梯画像功能，为更精确的客流预测及设备维修维护提供了数据支撑。

关键词　地铁；客流画像；大数据分析

中图分类号　U293.1⁺3；F530.7

DOI：10.16037/j.1007-869x.2020.10.028

Research on Passenger Flow Portrait Analysis of Qingdao Metro Based on Big Data Analysis

LUO Qingping, ZUO Xutao, ZHANG Beibei, DU Keliang

Abstract　Qingdao MMCC (metro management and control center) uses the big data analysis method to realize passenger flow analysis and portrait function for urban rail transit. By combining AFC (automatic fare collection) data with ATS (automatic train supervisory) data, Qingdao MMCC has achieved more accurate travel path matching, which overcomes the accuracy defects of traditional passenger flow analysis algorithm. From the perspective of passenger flow, MMCC realizes the portrait functions of passenger, station, train and section, as well as the function of escalator portrait combined with ISCS system data to provide a data support for more accurate passenger flow prediction and equipment maintenance.

Key words　metro; passenger flow portrait; big data analysis

First-author's address　Qingdao Metro Group Co., Ltd., 266045, Qingdao, China

轨道交通运营公司为合理调配运能运力，需要预测日常客运量情况[1]。客流量预测的方法较多，如时间序列法、回归分析法、人工神经网络模型等[2]。基于大数据分析的客流分析和乘客画像能够更清晰地掌握乘客出行特点，如日常通勤人数及居所、工作位置等，提供了客流量与工作日、淡旺季的变动指数，为准确预测客流量提供了新的方法，同时乘客画像亦为精准营销提供了数据支撑。本文利用青岛地铁线网运营管理与指挥中心（以下简称“MMCC”）大数据平台，采用改进的客流分析方法对青岛市地铁客流特征进行分析，得到了青岛地铁客流量分析指标及常旅客画像，为青岛地铁资源配置优化提供了数据基础，同时基于客流量的电扶梯画像为电扶梯设备维修维护提供了支撑。

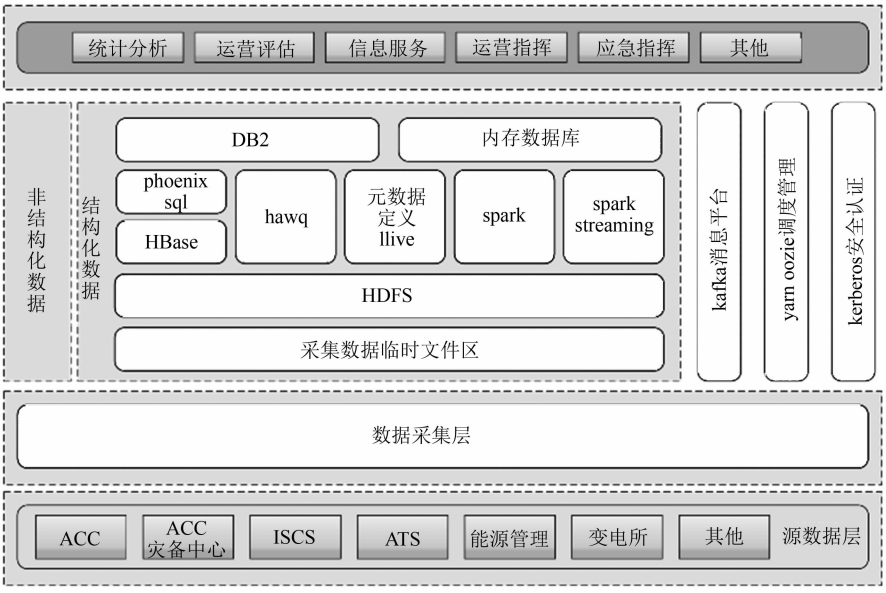
目前国内已经建设轨道交通指挥中心的城市，如北京、深圳等，均采用Teradata（天睿公司的关系数据库管理系统）作为大数据平台。但由于其成本太高，一般城市难以承受。MMCC是国内第一个采用开源的Hadoop/Spark大数据平台软件，主要用于城市轨道交通客流分析和画像研究，不仅节约了投资成本，而且提高了系统平台的可扩展性，填补了国内该领域的空白。

1　Hadoop大数据平台的构建

Hadoop大数据平台最核心的框架设计是HDFS（分布式存储系统）和MapReduce（分布式计算系统）。其中HDFS为海量数据提供了存储，而MapReduce则为海量数据提供了计算。与Teradata的数据仓库相比，MapReduce的运行速度较慢，执行较大数据量分析时，常常需数分钟，无法满足指挥中心的需求。为克服该问题，采用Spark/Spark Streaming计算引擎，使得大数据平台的运算速度达到了Teradata的水平。Hadoop大数据平台采用HIVE作为数据仓库；HBase作为海量数据存储；DB2作为集市层数据存储媒介，为上层应用提供数据服务；Redis作为实时数据库/内存数据库，为上层应用系统提供实时数据服务；HAWQ和Phoenix

* 南京工程学院科研创新基金面上项目（CKJB201311）

SQL 作为分析引擎。Hadoop 大数据平台采用 Kafka 作为数据中心实时、离线、近线消息平台；采用 Oozie 作为数据中心任务调度，对分析层数据的加工、分析、汇总等工作流及任务提供任务调度；采用 Kerberos 认证对平台数据进行保护。Hadoop 大数据平台技术架构如图 1 所示。



注：ACC代表自动售检票清分中心；ATS代表列车自动监控；ISCS代表综合监控系统；DB2代表关系型数据库管理系统

图 1　Hadoop 大数据平台技术架构

2　改进的客流分析方法

城市轨道交通客流分析的主要作用[3]包括：为票务清分提供数据基础、统计分析客流规律[4]、调配运力资源和车站组织、评价轨道交通运营状况、辅助突发大客流预警[5]及为规划建设提供数据支持等。传统的客流分析算法主要依靠 AFC（自动售检票）系统/ACC 提供的客流数据，但这种方法在数据采集的完备性、准确性和时效性上存在缺陷[3]。

为避免上述问题，MMCC 大数据平台采用了 AFC 系统/ACC 数据结合 ATS 信息的算法[6]，实现更精确的出行路径匹配，在一定程度上克服了准确性缺陷。为减少 AFC 系统至 ACC 的传输延时，大数据平台直接从线路 AFC 系统采集数据，并且只使用乘客进出站刷卡数据和 AFC 系统降级运营信息作为数据源，避免了各线路 AFC 系统厂商统计算法不一致的问题。

目前，MMCC 大数据平台已接入青岛地铁 2 号线、3 号线、11 号线等 3 条线路，实现进站量、出站量、进出量、OD（起终点）客流、周转量、换乘量、去路/乘距、负荷强度、客流不均衡度等 9 大类、70 多小类客流指标的统计。与传统客流统计相比，大数据分析算法能够更灵活方便地统计分析各种客流指标，减少了对 ACC 的依赖。对乘客进出站行为特点，特别是闸机-出入口的对应关系进行了人工抽样统计，得到了传统客流分析算法无法实现的车站出入口客流指标。自 2019 年 7 月该平台进入试运行以来运行稳定，为上层系统提供了有力的数据支撑。青岛地铁 2 号线芝泉路站站口 5 min 客流进站量统计如表 1 所示。

表 1　青岛地铁 2 号线芝泉路站站口 5 min 客流进站量统计表

名称	不同时刻站口客流进站量/人次								
	08:00	08:05	08:10	08:15	08:20	08:25	08:30	08:35	08:40
A 出入口	53	39	47	9	43	36	48	35	34
B 出入口	26	18	29	23	26	17	21	17	27
C 出入口	10	4	7	7	7	10	8	3	3

3　地铁客流画像分析

MMCC 大数据平台利用客流分析的结果,建立乘客以及各线路所有车站、列车、区间、电扶梯的画像。

3.1　乘客画像

乘客画像是对轨道交通乘客及乘客出行进行不同层面、不同角度的描述,乘客画像分析功能输出的数据包括乘客的基本属性和乘客的出行特征。乘客的基本属性包括乘客 ID(标志)、年龄段、是否通勤等。乘客的出行特征包括居所、工作位置、通常单程里程、通常单程时间、累计乘车天数、累计乘车里程、累计乘车时间、累计地铁耗时、累计乘车次数、工作日历、上班时间、下班时间、居所同行人、工作同行人、宵夜活动场所、节假日活动场所、工作日异常 OD、工作日最新异常 OD、节假日异常 OD、节假日最新异常 OD、进站客流量最大的 5 个入口、出站客流量最大的 5 个出口等。

对比 2019 年 9 月和 2019 年 10 月的乘客画像数据(见表 2)可知,青岛地铁购买非单程票的常旅客人数约为 120 万人,10 月比 9 月常旅客增长为 14.04%,其中儿童人数增长率最快为 27.57%。常旅客中成年人占比最高,通勤的人数仅约占 5%。

表 2　青岛地铁不同类型的常旅客人数

月份	不同类型常旅客人数及占比								总人数/人
	儿童		成年人		老年人		通勤		
	人数/人	占比/%	人数/人	占比/%	人数/人	占比/%	人数/人	占比/%	
9 月	60 275	5.30	847 827	74.53	172 586	15.17	56 880	5.00	1 137 568
10 月	76 892	5.95	955 098	73.96	200 385	15.52	58 962	4.57	1 291 337

2019 年 9 月和 2019 年 10 月常旅客通常单程乘车时间 t 对比,如表 3 所示。由表 3 可知,t 为 10~30 min 的人数最多。各乘车时间段 10 月比 9 月的常旅客增长率分别如下:$t \leqslant 10$ min 时为 2%,10 min$<t\leqslant$20 min 时为 6%,20 min$<t\leqslant$30 min 时为 12%,30 min$<t\leqslant$40 min 时为 20%,40 min$<t\leqslant$50 min 时为 26%,50 min$<t\leqslant$60 min 时为 28%,$t>$60 min 时为 48%。可见乘车时间越长,常旅客增长率越高。这说明青岛市民远距离出行偏爱选择乘坐地铁。

表 3　青岛地铁不同单程时间的常旅客人数

月份	不同单程时间的常旅客人数及占比													
	$t\leqslant$10 min		10 min$<t\leqslant$20 min		20 min$<t\leqslant$30 min		30 min$<t\leqslant$40 min		40 min$<t\leqslant$50 min		50 min$<t\leqslant$60 min		$t>$60 min	
	常旅客人数/人	占比/%	常旅客人数/人	占比/%	常旅客人数/人	占比/%	常旅客人数/人	占比/%	常旅客人数/人	占比/%	常旅客人数/人	占比/%	常旅客人数/人	占比/%
9 月	69 855	6	322 634	30	341 885	32	188 188	17	84 132	8	38 929	4	35 065	3
10 月	71 495	6	343 302	28	383 950	31	225 791	18	105 954	9	49 952	4	51 930	4

青岛地铁常旅客月乘车天数对比如表 4 所示。表 4 中,每月不同乘车天数时 10 月比 9 月的常旅客增长率分别如下:每月只乘车 1 d 时为 14%,每月只乘车 2 d 时为 17%,每月只乘车 3 d 时为 18%,每月乘车 4~10 d 时为 14%,每月乘车 11~20 d 时为 6%,每月乘车超过 20 d 时为 10%。由表 4 可知,每月只乘车 1 d 的客流最多,且大多数常旅客每月只乘车不到 3 d。这也印证了表 1 中通勤人数仅占 5%的结论,说明常旅客的乘坐黏性不高,有很大的提升空间。

表 4　青岛地铁不同乘车天数的常旅客人数

月份	不同乘车天数的常旅客人数及占比											
	1 d		2 d		3 d		4~10 d		11~20 d		>20 d	
	常旅客人数/人	占比/%	常旅客人数/人	占比/%	常旅客人数/人	占比/%	常旅客人数/人	占比/%	常旅客人数/人	占比/%	常旅客人数/人	占比/%
9 月	423 002	39	195 941	18	109 402	10	222 243	21	101 977	9	28 123	3
10 月	482 727	39	228 904	19	128 588	10	253 557	21	107 802	9	30 797	2

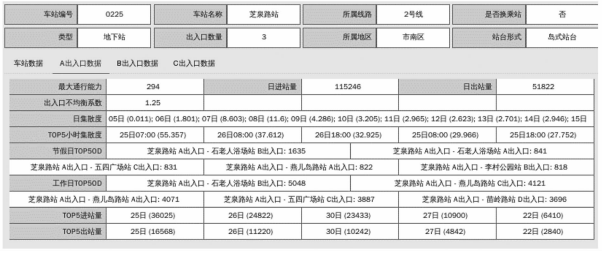
3.2　车站画像

车站画像包括车站的基本属性、车站出入口基本属性、车站客流、车站出入口客流等。车站基本属性包括车站名称和编号、所属线路、是否为换乘

站、车站类型(地面站或地下站)、出入口数量、所属地区、站台型式等。出入口基本属性包括电扶梯、售票机、检票机、连接站厅名称、最大通行能力。车站客流统计指标包括最大/最小/平均进站量、最大/最小/平均出站量、最大/最小/平均换乘量、最大/最小/平均车站集散量、出入口不均衡系数、5 个客流量最大的 OD、潮汐特征等。车站画像如图 2 所示。

3.3 列车画像

列车的基本属性包括列车识别号、列车型号、列车规格、列车定员等。列车客流统计指标包括区间列车满载率、列车满载率里程分布、列车满载率区间分布、列车客运强度等。列车画像的满载率分布如图 3 所示。



车站编号	0225	车站名称	芝泉路站	所属线路	2号线	是否换乘站	否
类型	地下站	出入口数量	3	所属地区	市南区	站台形式	岛式站台

车站数据　A出入口数据　B出入口数据　C出入口数据

最大通行能力	294	日进站量	115246	日出站量	51822
出入口不均衡系数	1.25				
日集散度	05日 (0.011); 06日 (1.801); 07日 (8.603); 08日 (11.6); 09日 (4.286); 10日 (3.205); 11日 (2.965); 12日 (2.623); 13日 (2.701); 14日 (2.946); 15日				
TOP5小时集散度	25日07:00 (55.357)	26日08:00 (37.612)	26日18:00 (32.925)	25日08:00 (29.966)	25日18:00 (27.752)
节假日TOP50D	芝泉路站 A出入口 - 石老人浴场站 B出入口: 1635		芝泉路站 A出入口 - 石老人浴场站 A出入口: 841		
芝泉路站 A出入口 - 五四广场站 C出入口: 831	芝泉路站 A出入口 - 燕儿岛路站 A出入口: 822	芝泉路站 A出入口 - 李村公园站 B出入口: 818			
工作日TOP50D	芝泉路站 A出入口 - 石老人浴场站 B出入口: 5048		芝泉路站 A出入口 - 燕儿岛路站 C出入口: 4121		
芝泉路站 A出入口 - 燕儿岛路站 A出入口: 4071	芝泉路站 A出入口 - 五四广场站 C出入口: 3887		芝泉路站 A出入口 - 苗岭路站 D出入口: 3696		
TOP5进站量	25日 (36025)	26日 (24822)	30日 (23433)	27日 (10900)	22日 (6410)
TOP5出站量	25日 (16568)	26日 (11220)	30日 (10242)	27日 (4842)	22日 (2840)

图 2　车站画像截屏图

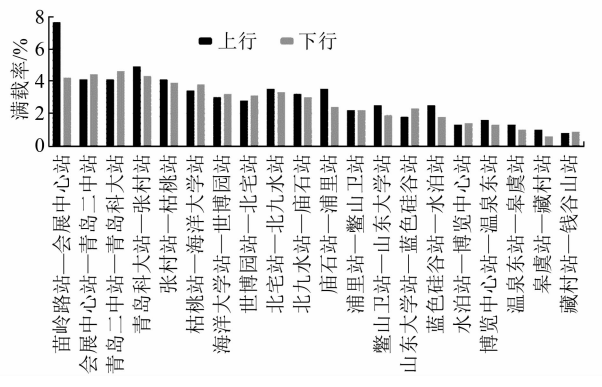


图 3　列车画像的满载率分布图

3.4 区间画像

区间画像包括区间的基本属性、区间能力、客运属性等。区间的基本属性包括区间所属线路、区间两端车站、区间长度等。区间的能力体现为断面运输能力,可分为工作日图运力和节假日图运力。区间的客运属性包括区间断面流量、区间断面满载率。区间画像如图 4 所示。

3.5 电扶梯画像

电扶梯画像包括电扶梯的基本属性、运转时间、运转里程、运转强度、运行系数等。电扶梯的基本属性包括唯一标志、所属车站、所属出入口、电扶梯运力、厂商、型号,以及是否有同空间的步行梯和节能模式等。电扶梯的运转时间包括累计开行时间、累计上行时间、累计下行时间、日开行时间、日正常速度开行时间、日缓行速度开行时间等。电扶梯的运转里程包括日累计里程、日累计上行里程、日累计上行缓行里程、日累计下行里程、日累计下行缓行里程等。电扶梯的运转强度包括累计承载量、累计开行承载量、累计关闭承载量、日承载量、日开行承载量、满载率、满载率时间分布、满载率日期分布、客运强度等。电扶梯的运行状态从 ISCS 获得。电扶梯画像如图 5 所示。

11号线 区间画像(2019年2月)

断面	区间长度	工作日图运力	节假日图运力	区间断面流量
芝泉路站-五四广场站	1.56 km	7857024	2506512	307500
五四广场站-芝泉路站	1.56 km	6969696	2233488	314509
五四广场站-浮山所站	0.93 km	6647376	2233488	1044914
浮山所站-五四广场站	0.93 km	5706960	1867560	1067090
浮山所站-燕儿岛路站	1.04 km	6325056	2089392	1044361
燕儿岛路站-浮山所站	1.04 km	6615144	2199360	1059548
燕儿岛路站-高雄路站	0.8 km	5413080	1784136	978623
高雄路站-燕儿岛路站	0.8 km	3183384	978336	1012865
高雄路站-麦岛站	1.16 km	3208032	1090200	990161
麦岛站-高雄路站	1.16 km	4618656	1562304	1029818
麦岛站-海源路站	1.16 km	1823952	638952	990991

图 4　区间画像截屏图

电扶梯画像(2019年2月)

扶梯编号	车站	出入口	累计开行时间	日开行时间	日累计里程	累计开行承载量	日开行承载量	客运强度
E07/04	浮山所站	0227A	0	0	0	6646	215	0
E07/05	浮山所站	0227A	0	0	0	6646	215	0
E07/06	浮山所站	0227B	482	16	10	6646	215	1
E07/07	浮山所站	0227B	482	16	10	6646	215	0
E07/10	浮山所站	0227D	482	16	10	18337	592	0
E07/11	浮山所站	0227D	482	16	10	18337	592	1

电扶梯编号	E07/07	所属车站	浮山所站	所属出入口	0227B	电扶梯运力	3600
是否有空间步梯	是	厂商		型号		是否有节能模式	是
累计开行时间	482	累计上行时间	482	累计下行时间	0	累计承载量	6646
累计开行承载量	6646	累计关闭承载量				客运强度	0
日开行时间	16	日正常速度开行时间	16	日缓行速度开行时间	0	日累计里程	10
日累计上行里程	10	日累计上行缓行里程	0	日累计下行里程	0	日累计下行缓行里程	0
日承载量	215	日开行承载量	215				
上行通行系数	0.95; 0.90	下行通行系数	0.95; 0.90	关闭通行系数	0.0		

图 5　电扶梯画像截屏图

4 数据存储和可视化

MMCC 采用的 Hadoop 大数据平台需要处理海量数据获得各种客流指标和画像,同时产生大量的数据输出。因此客流大数据的存储和可视化也是一个难题[7]。Hadoop 大数据平台采用 DB2 作为集市层数据存储媒介,为上层应用提供数据服务;采用 Redis 作为实时数据库/内存数据库,为上层应用系统提供实时数据服务。Hadoop 大数据平台在提供灵活查询条件的基础上,采用表格和图形结合

(下转第 123 页)

为了实现 ACLC 系统的计算、存储,以及网络资源的按需分配、统一管理和集中检测,提高整个系统资源的利用率,便于资源的快速部署和扩展,分别在主、副中心部署云计算资源池,存储资源池、网络资源池等。ACLC 应用的系统业务及互联网业务均承载在共享资源池上,并设置统一云管理平台管理本系统的各类 IT 资源,还可以在未来平滑融合扩展新线路。

ACLC 云平台遵循面向业务需求的设计思路,基于 AFC 系统的业务特点,采用云计算资源池的设计方法,实现 IT 基础架构模块与业务模块松耦合以及资源池模块化交付和横向扩展。通过 ACLC 云平台保证资源的快速交付和统一管理,支撑业务快速上线、融合运营、统一运维。

4 结语

随着城市轨道交通网络化运营和新业务应用需求的不断涌现,持续优化 AFC 系统架构从而不断解决传统架构的弊端和应对新形势下新业务的需求已成为各地不断探索的方向。本文通过项目实际应用,提出了基于双活架构的 AFC 系统体系架构,并详细分析了实现双活架构所采用的关键技术,以期为行业的发展提供借鉴。

参考文献

[1] 胡冬. 城市轨道交通 AFC 区域中心系统设计[D]. 南京:东南大学,2015.

[2] 陈青云,顾洋. 基于集中式管理体系结构的轨道交通自动售检票系统[J]. 都市快轨交通,2017(2):94.

[3] 朱嘉斌,黄问遂. 地铁清分中心灾备系统设计[J]. 都市快轨交通,2011(6):92.

[4] 黎庆,张宁,徐钟全,等. 城市轨道交通自动售检票系统区域中心总体设计[J]. 城市轨道交通研究,2015(8):71.

[5] 吴娟,徐钟全,毛建. 南京地铁 AFC 区域线路中心的规划设计[J]. 铁路通信信号工程技术,2012,9(5):63.

[6] 陈楠,李继铭. 南京地铁 AFC 系统管理方式的分析和研究[J]. 铁路通信信号工程技术,2011(6):47.

[7] 王健,张宁,黄亮,等. 南京地铁 AFC 系统网络化建设思路和再思考[J]. 都市快轨交通,2011(1):69.

[8] 李道全,赵华伟. 多线共用 AFC 系统线路中心设计探头[J]. 都市快轨交通,2012(5):71.

[9] 王浩,刘旭,杨霓霏. 北京市轨道交通自动售检票系统多线共用线路中心的设计与实现[J]. 铁路计算机系统,2013(3):60.

(收稿日期:2019-11-22)

(上接第 118 页)

的方式,保证数据使用者能够在查看大量数据的同时,可以更直观地看到数据之间的相关性。

5 结语

利用大数据平台分析客流数据已成为各城市轨道交通指挥中心的必备功能。青岛地铁 MMCC 是国内首个成功应用 Hadoop 开源大数据平台的案例。MMCC 大数据平台不仅建设成本低,而且能够更为灵活高效地实现客流分析的所有功能,包括进出站客流、OD 客流、周转量、换乘量、去路/乘距、负荷强度、客流不均衡度等。MMCC 大数据平台能够挖掘出乘客、车站、列车、区间、电扶梯等画像,为研究客流数据和乘客出行规律提供了新的视角,为设备维修维护提供了数据支撑,为其他城市的轨道交通指挥中心建设提供了借鉴。

参考文献

[1] 张啟梅,廖玉梅,任永成,等. 基于大数据下的旅客流量分析[J]. 数据挖掘,2017(1):26.

[2] 蔡昌俊,姚恩建,王梅英,等. 基于乘积 ARIMA 模型的城市轨道交通进出站客流量预测[J]. 北京交通大学学报,2014(2):135.

[3] 卢恺,韩宝明,鲁放. 城市轨道交通运营客流数据分析缺陷及应对[J]. 都市快轨交通,2014(4):25.

[4] 段卫静,陈艳艳,赖见辉. 北京地铁 4 号线客流特征分析[J]. 都市快轨交通,2013(4):43.

[5] 吴志强,黄天印,颜彦文,等. 基于大数据的城市轨道交通运营故障影响分析系统对客流影响的分析[J]. 城市轨道交通研究,2019(4):31.

[6] 刘彦君. 基于 AFC 和列车时刻表的城轨乘客时空扩展出行路径匹配[D]. 北京:北京交通大学,2016.

[7] 李伟,周峰,朱炜,等. 轨道交通网络客流大数据可视化研究[J]. 中国铁路,2015(2):94.

(收稿日期:2019-11-06)