文章编号：2095-2163(2020)12-0164-06 中图分类号：U231+.4；U293.1 文献标志码：A

基于 ATS 与 AFC 数据的地铁乘客出站走行时间估计方法

张凌波，郝妍熙，胡 华
（上海工程技术大学 城市轨道交通学院，上海 201620）

摘 要：乘客出站走行时间的估计研究对于帮助检验站内拥挤度、为线路行车计划提供一定依据等具有重要的意义。AFC 数据中包含了大量的乘客出站数据，列车的 ATS 到站时间数据可以得到列车到达每一地铁站的时间，通过建立二者之间的关系，得到乘客出站走行时间的估计值，并用正态分布函数拟合出乘客的走行时间函数。同时，通过实地调查，得到乘客出站走行时间的真实值，并拟合出实地调查的走行时间函数。最后，通过比较模型计算值与真实值之间的差异，验证了此方法的可行性。

关键词：地铁；走行时间；AFC 数据；正态分布

A method for estimating walking time of subway passengers based on ATS and AFC data

ZHANG Lingbo, HAO Yanxi, HU Hua
(College Of Urban Railway Transportation, Shanghai University Of Engineering Science, Shanghai 201620, China)

【Abstract】 Estimating the walking time of passengers leaving the subway station is of great significance to help check the congestion in the station and provide evidence for the route plan. The AFC data contains a large amount of the time data of passengers walking out of the subway station. The train´s ATS arrival time data can obtain the time that the train arrives at each subway station. By establishing the relationship between the two, the estimated value of the passenger´s walking time can be obtained and used. The normal distribution function fits the passenger´s walking time function. At the same time, through the field survey, the real value of the walking time of the passengers walking out of the station is obtained, and the walking time function of the field survey is fitted. Finally, the difference between the fitted value and the true value is compared to verify the feasibility of this method.

【Key words】 Subway; Walking time; AFC data; Normal distribution

0 引 言

随着轨道交通线网的不断扩大，地铁由于速度快、安全可靠、准点、舒适、环保的特点，在乘客日常出行中所乘坐交通工具的比例越来越高。尤其是在高峰时期，地铁列车承载的客运量越来越大，在列车到站后，下车人数过多时就会造成站内乘客的拥挤。对乘客从地铁列车下车到走出出站闸机这一时间段内的走行时间进行估计，能够帮助检验站内拥挤度、规划乘客走行路径，帮助管理人员进行客流引导，改善车站客流组织，为线路行车计划提供一定依据、对提高车站运营组织管理水平和服务质量等都具有十分重要的意义。

地铁站乘客的走行时间分布规律的文献中，杜鹏设计了抽样调查的方法，通过调查通道换乘时乘客的走行时间，得出换乘走行时间近似服从对数正态分布，其均值和方差与换乘走行距离相关，拥挤加剧会增大均值、减小方差[1]；童燚杰通过人工调查得到重庆地铁 1 号线和 3 号线的换乘站内的乘客的走行时间，分别用正态分布、对数正态分布函数进行数据拟合，采用极大似然估计法对里面参数进行估计[2]；吕慎在基于换乘乘客到站时间服从正态分布的前提下，通过人工实地调查高峰时段、平峰时段换乘乘客到达换乘公交站点的走行时间，用回归分析分别建立换乘乘客到站时间均值的模型[3]；郝勇针对城市轨道交通车站的进出站闸机等延滞性步行设施对乘客走行造成的影响，通过客观实测数据，运用曲线估计的方法，分别建立乘客走行时间的 BPR（路阻函数）函数模型[4]；王志刚采用 BPR 模型来确定乘客的走行特征及客流的仿真过程，通过实际调查在自由流条件下分别得到换乘通道、楼梯上行和下行流量以及相应的走行速度，进而得到相应的走行时间函数，最后得出结论：换乘通道的走行速

基金项目：上海市科委地方院校能力建设项目(19030501400)。

作者简介：张凌波(1996-)，女，硕士研究生，主要研究方向：交通运输规划与管理；郝妍熙(1990-)，女，博士，讲师，主要研究方向：交通大数据挖掘与智慧运营；胡 华(1979-)，女，博士，教授，主要研究方向：轨道交通智慧运营与安全管理。

通讯作者：郝妍熙 Email：haoyanxi@ sues.edu.cn

收稿日期：2020-09-15

度都要比楼梯快[5]。

乘客走行时间的研究，对于城市轨道交通线网规划、列车运行图编制的研究等都提供重要的依据。邵远忠将 BPR 函数移植于地铁站内的行人交通流并对其进行线性改进，得到调研地铁站内各个出入口的 AFC 设备处行人通行时间与流量的关系式，为地铁站内 AFC 设备的通行效率评估提供依据[6]；何韬等在基于乘客换乘走行时间路径固定且近似服从同一对数正态分布的假设前提下，分析了换乘等待时间的影响因素，构建了不同列车到达间隔情况下的换乘等待时间优化模型，该模型优化了乘客换乘时间[7]；李智等通过分析列车延迟时间和乘客换乘走行时间的概率分布，并据此计算乘客换乘等待时间，提出基于换乘最优的城市圈城际铁路周期运行图编制模型，模型优化后的列车运行图可以有效地减少乘客的换乘等待时间[8]；李玉书等对轨道交通车站单位时间内换乘通道客流量与走行时间的数据进行分析，利用 BPR 函数来表征车站换乘通道的通行能力及客流压力情况[9]；还有的研究在基于乘客的换乘走行时间已知且固定的假设前提下，对城市轨道交通末班车时刻表问题进行建模[10]。

目前国内外对于城市轨道交通乘客走行时间的研究大多研究的是换乘站的走行时间，且很多关于走行时间的研究都是在基于乘客走行时间路径固定、走行时间固定的假设前提下，与实际有一定的偏差。本文所研究的是乘客出站走行时间，对于城市轨道交通线网规划、列车运行图编制的研究等都提供重要的依据。通过挖掘 AFC 刷卡数据以及列车的 ATS 到站时间数据，建立相应的模型，得到乘客的走行时间，拟合出乘客的走行时间概率密度函数。同时，通过实地调查，得到乘客出站走行时间的真实值，并拟合出实地调查的走行时间函数。最后通过比较二者差异，验证了此模型的可行性。

1　乘客走行时间模型概述

现有研究中常用正态分布、对数正态分布函数来描述乘客走行时间的概率密度分布情况，以及使用 BPR 路阻函数来描述乘客走行时间与客流量之间的关系。

1.1　正态分布

现有交通研究中，常常通过人工跟随调查得到乘客走行时间的分布情况，再使用正态分布函数拟合出乘客走行时间的概率分布密度函数。正态分布是一种服从均值为具有两个参数 μ 和 σ^2 的连续型随机变量的分布，第一参数 μ 是遵从正态分布的随机变量的均值，第二个参数 σ^2 是此随机变量的方差。正态分布是一种概率分布，其概率密度函数(1)为：

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{1}$$

1.2　对数正态分布

除了正态分布，现有交通研究中还常常使用对数正态分布对走行时间概率密度函数进行拟合，找到更适合描述走行时间规律的函数。对数正态分布是指对数为正态分布的任意随机变量的概率分布，即假设 $Y=\ln x$ 服从正态分布，则 x 服从对数正态分布。其概率密度函数(2)为：

$$f(x)=\begin{cases}\dfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, & x>0;\\ 0, & x\geqslant 0.\end{cases} \tag{2}$$

1.3　BPR 路阻函数

BPR 路阻函数用来表示车辆行驶时间与交通流量的关系，其函数表达式(3)为：

$$t=t_0[1+\alpha(v/c)^{\beta}]. \tag{3}$$

其中，t 为两交叉路口之间的路段行驶时间，min；t_0 为交通量为零时的路段行驶时间，min；v 为路段机动车交通量，辆/h；c 为路段通行能力，辆/h；α、β 为系数。

由于地铁通道处等客流状态与道路交通流具有较高的相似性，因此 BPR 路阻函数现也用来描述乘客走行时间与流量(如进出站闸机等步行设施流量，换乘通道的断面客流等)之间的关系。在此表达式中，t_0 是自由流条件下的乘客走行时间，s；t 表示在流量 v 下乘客的走行时间，s；v 表示单位宽度上的流量，人次/min；c 是步行设施单位宽度上的容量或通行能力，人次/min；α、β 为延滞系数。

目前乘客走行时间模型的数据获取方法都是通过人工跟随实地调查得到的。但是，由于每个地铁站都不相同，通过人工调查得到乘客走行时间的方法比较繁琐，人工调查得到的数据仅能适用于所调查的地铁站。本文通过基于 ATS 列车到站时间数据和 AFC 刷卡数据，分别得到高峰、平峰时期乘客出站走行时间数据，利用走行时间数据，建立一种乘客出站走行时间估计模型，通过此模型拟合出出站走行时间曲线，得到乘客出站走行时间的均值、标准差。

2　走行时间估计方法

2.1　出站走行时间数据获取方法

地铁 AFC 刷卡数据详细记录了进出站的每一

位乘客的刷卡信息,记录的信息有乘客的身份证号、刷卡时间、进站站名、出站站名、刷卡类型、闸机编号、票卡类型等,见表 1。其中,刷卡类型分为进站、出站。根据刷卡数据,选取一个固定的地铁站,并在刷卡类型中选择出站类型,就可以得到该地铁站每一位乘客的出站时间,记为 $T_{出}$。

表 1 AFC 刷卡数据表

Tab. 1 AFCcredit card data

身份证号	刷卡时间	进站站名	出站站名	刷卡类型	闸机编号	票卡类型
0152 * * * * 75	64203	外环路	莘庄	进站	111027081	普通成人卡 1
3273 * * * * 60	64203	沈杜公路	徐家汇	出站	118021039	全终端手机交通卡
…	…	…	…	…	…	…

上海轨道交通 ATS 列车到站和发车时间数据包括了列车的车次号、速度等级、运行线、不同车次的列车在每一个地铁站的实际到站时间、实际发车时间等,见表 2。通过列车 ATS 的到站时间数据,可以得到不同车次的列车在同一地铁站的实际到站时间,记为 $T_{到}$。

表 2 ATS 列车到站和发车时间表

Tab. 2 ATS train arrival and departure schedule

车次号	运行线	速度等级	运行方向	车站	实际到点	实际发点
918CL	308	节能	上行	漕河泾开发区	05:50:14	05:51:04
923SR	206	正常	下行	九亭	05:48:09	05:51:14
…	…	…	…	…	…	…

乘客下车后的出站走行时间记为 $T_{走}$,即为乘客的出站时间 $T_{出}$减去乘客的下车时间 $T_{下}$,即 $T_{走} = T_{出} - T_{下}$。由于列车的每扇门的前面几位下车的乘客基本都是在列车屏蔽门开启后就下车,因此这部分乘客的下车时间即为列车 ATS 到站时间 $T_{到}$+列车屏蔽门开启时间记为 $T_{屏}$,即式(4):

$$T_{下} = T_{到} + T_{屏}. \tag{4}$$

由于并不是每一位乘客都是在列车屏蔽门开启后就立即下车,因此需要通过实地调查,探究每位乘客在列车屏蔽门开启后的下车时间。

2.2 下车时间

以上海轨道交通九号线漕河泾开发区地铁站为例,在高峰时段 7:30-9:30、平峰时段 10:00—11:00、13:00-14:30 选取站台中间的一扇屏蔽门,在每一列列车到站时,统计这扇车门的每一位乘客从列车屏蔽门开启后到下车的时间,高峰时段、平峰时段都分别统计 50 列列车到站后的乘客下车时间数据。

将高峰时期 50 组乘客的下车时间数据,按照相同下车乘客人数进行分组,人数分布在 2~17 人之间。再分别求出第一位乘客、第二位乘客至最后一位乘客的平均下车时间,绘制出图 1。图 1 中横坐标表示第 x 位乘客(x 为 1 到 17 之间的自然数),纵坐标表示第 x 位乘客的平均下车时间。拟合出的函数(5)为:

$$y = 0.863\ 2x + 0.807\ 4. \tag{5}$$

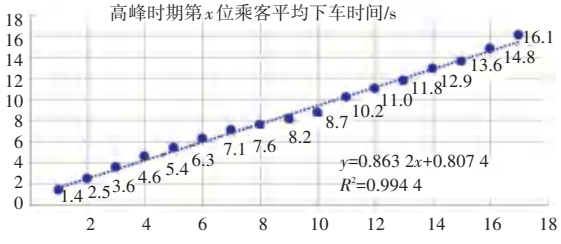

图 1 高峰时期乘客平均下车时间

Fig. 1 Average time for passengers to get off during peakperiod

同理,将平峰时期 50 组乘客的下车时间数据,按照相同下车乘客人数进行分组,人数分布在 2-7 人之间。再分别统计第一位乘客、第二位乘客至最后一位乘客的平均下车时间,绘制出图 2。图 2 中横坐标表示第 x 位乘客(x 为 1~7 之间的自然数),纵坐标表示第 x 位乘客的平均下车时间。拟合出的函数(6)为:

$$y = 0.8372x + 0.448. \tag{6}$$

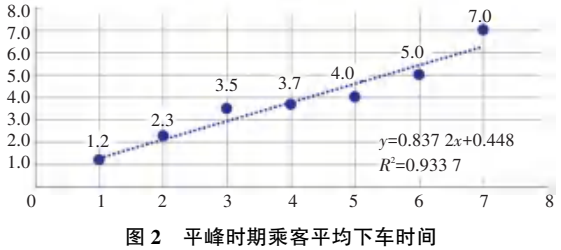

图 2 平峰时期乘客平均下车时间

Fig. 2 Average time for passengers to get off during flat period

2.3 模型假设

从高峰时期乘客平均下车时间图 1 可以看到,高峰期从列车屏蔽门开启后前后下车的乘客下车时

间间隔大致维持在一人/秒。由于高峰时期客流量较大，假设同一列列车上后面下车的乘客的下车时间误差在楼扶梯的排队过程当中被抵消掉，即忽略掉高峰时期乘客前后下车的乘客的下车时间差异。

由于平峰时期从列车屏蔽门开启后前后的乘客的下车时间前后相差不大，假设忽略掉平峰时期前后下车的乘客的下车时间差异。

少部分乘客由于各种原因下车后并未直接出站，而是在站内停留一段时间后再出站，这部分乘客可能会和下一班列车到达的乘客一起出站。这部分乘客的走行时间无法准确得出，由于这部分所占比例较小，因此忽略掉这部分人对统计数据造成的差异。

2.4　模型定义

在模型假设的前提下，忽略高峰时期、平峰时期前后下车的乘客的下车时间差异，根据地铁 AFC 刷卡数据和列车 ATS 到站时间数据建立乘客的出站走行时间模型。乘客的走行时间即为乘客出站时间 $T_{出}$ 减去列车的到站时间 $T_{到}$ 再减去屏蔽门开启时间 $T_{屏}$，设屏蔽门开启时间为 5 s，即式(7)：

$$T_{走} = T_{出} - T_{到} - T_{屏}. \tag{7}$$

3　实例分析

3.1　AFC 刷卡数据分析

选取上海轨道交通九号线 2018 年 7 月 16 日的刷卡数据，进站站点选取全部地铁站，出站站点选取漕河泾开发区站，以秒为时间单位绘制出漕河泾开发区站出站客流量的频数分布图，如图 3 所示。图中横坐标为时间，纵坐标为客流量的频数。从图 3 可以看出，刷卡数据呈现的是一簇一簇的类似正态分布的波形图，每一簇数据表示的是每一列列车到站后乘客的出站刷卡数据，每一簇数据的波谷表示没有乘客出站，波峰表示乘客出站人数比较密集。将每一簇数据最前面处于波谷处的数据与 ATS 列车到站时间数据进行匹配，从而确定每一组乘客的走行时间。

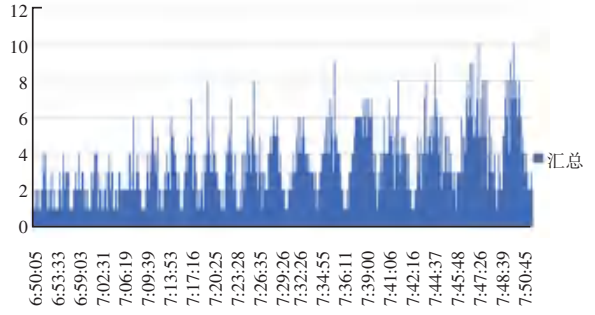

图 3　AFC 刷卡频数图

Fig. 3　AFC swipe frequency

当同站上行和下行的列车到站时间间隔较小时，会造成两列列车上的乘客出站时间紧凑，甚至会造成不同方向到站的列车上的乘客同时出站的情况。因此需要将列车到站时间数据与 AFC 刷卡数据进行对比分析，筛除掉 ATS 列车到站时间紧密的数据以及 AFC 刷卡数据中乘客出站时间前后间隔较长的数据组，过滤掉不准确的数据。

3.2　实例分析

根据 AFC 刷卡数据及列车 ATS 到站时间数据，以漕河泾开发区地铁站为例，高峰期和平峰期分别选取 5 组合适的数据，通过走行时间模型 $T_{走} = T_{出} - T_{到} - T_{屏}$ 分别得出乘客的走行时间，发现其符合正态分布，拟合出正态分布概率密度函数图。同时，在高峰时期及平峰时期通过人工跟随调查乘客的出站走行真实值时间值，同样拟合出正态分布概率密度函数图，比较差异。分别对高峰及平峰时期走行时间的模型值与真实值的方差进行独立样本均值检验，查看走行时间模型的合理性。

高峰时期分别选取了 5 列列车乘客的 AFC 刷卡数据及 ATS 列车到站时间数据，按照走行时间估计模型计算出每位乘客的走行时间，并分别计算出每一组数据的平均值及客流量，见表 3。

表 3　高峰时期 5 列列车乘客走行时间表

Tab. 3　Passengers walking time of 5 trains during peak period

列车号	均值/s	客流量
1	95	237
2	96	243
3	98	324
4	90	323
5	99	383

绘制出通过走行时间模型，得到的 5 列列车的乘客在高峰时期的走行时间概率密度图以及高峰时期通过人工实地调查得到的乘客走行时间概率密度图，如图 4 所示。均值、标准差见表 4。

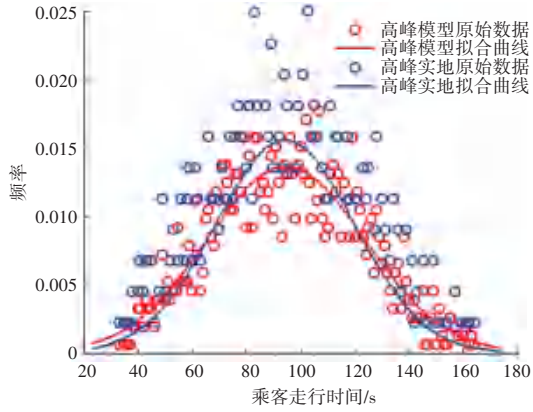

图 4　高峰期乘客走行时间概率密度图

Fig. 4　Probability density of passengers walking time during peak period

表 4 高峰期乘客走行时间概率密度函数参数

Tab. 4 Probability density function parameters of passengers walking time during peak period

	高峰模型	高峰实地
正态函数	$f(x)=\frac{1}{\sqrt{2\pi}*29.29}e^{-\frac{(x-94.6)^2}{2*29.29^2}}$	$f(x)=\frac{1}{\sqrt{2\pi}*25.64}e^{-\frac{(x-94.28)^2}{2*25.64^2}}$
均值	94.6	94.28
标准差	29.29	25.64

同理,绘制出平峰时期的图表,见表 5、表 6 和如图 5 所示。

表 5 平峰时期 5 列列车乘客走行时间表

Tab. 5 Passengers walking time of 5 trains during flat period

列车号	均值	客流量
1	77	70
2	83	68
3	80	82
4	78	25
5	79	80

表 6 平峰期乘客走行时间概率密度函数参数

Tab. 6 Probability density function parameters of passengers walking time during flat period

	平峰模型	平峰实地
正态函数	$f(x)=\frac{1}{\sqrt{2\pi}*18.92}e^{-\frac{(x-81.48)^2}{2*18.92^2}}$	$f(x)=\frac{1}{\sqrt{2\pi}*18.06}e^{-\frac{(x-78.66)^2}{2*18.06^2}}$
均值	81.48	78.66
标准差	18.92	18.06

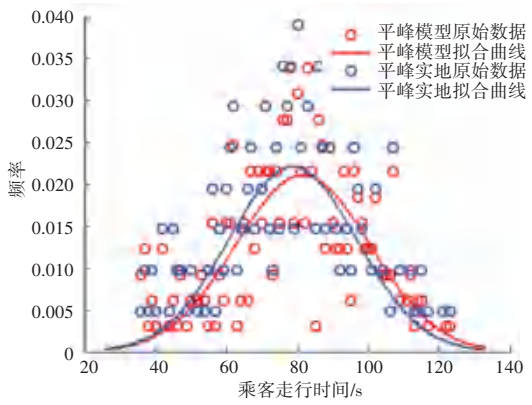

图 5 平峰期乘客走行时间概率密度图

Fig. 5 Probability density of passengers walking time during flat period

3.3 方差分析

对乘客在高峰时期、平峰时期通过模型得到的走行时间与实地调查的走行时间真实值的方差是否相等进行检验,结果见表 7、8。高峰期、平峰期方差检验的 F 统计量值的显著性概率分别为 0.256 和 0.432,均大于 0.1;再对总体均值的比较采用 t 检验法,得到显著性概率为分别 0.103 和 0.215,均大于 0.1。即在显著性水平为 0.1 的条件下,认为高峰期和平峰期的模型走行时间估计方法值和实地调查的走行时间真实值的方差相等。为此,可认为该乘客出站走行时间模型是合理的。

表 7 高峰时期乘客走行时间方差分析表

Tab. 7 Analysis of variance of passengers walking time during peak period

		方差方程 Levene 检验		均值方程 t 检验				差分的 95% 置信区间	
		F	Sig.	t	df	Sig.(双侧)	均值差值	标准误差值	下限	上限
高峰期走行时间	假设方差相等	1.292	0.256	1.631	1950	0.103	2.396	1.469	−0.485	5.277
	假设方差不相等			1.595	696.246	0.111	2.396	1.502	−0.553	5.345

表 8 平峰时期乘客走行时间方差分析表

Tab. 8 Analysis of variance of passengers walking time during flat period

		方差方程 Levene 检验		均值方程 t 检验				差分的 95% 置信区间	
		F	Sig.	t	df	Sig.(双侧)	均值差值	标准误差值	下限	上限
平峰期走行时间	假设方差相等	0.619	0.432	1.536	525	0.125	2.563	1.668	−0.714	5.840
	假设方差不相等			1.520	411.68	0.129	2.563	1.686	−0.751	5.877

4 结束语

本文根据 AFC 刷卡数据和 ATS 列车到站时间数据,建立一个乘客出站走行时间估计模型,通过实例分析:实地调查高峰时段与平峰时段的乘客出站走行时间,分别拟合出模型与实地调查的乘客走行时间概率密度函数图,并分析模型数据与实地调查数据的方差,得出结论:此乘客出站走行时间模型是合理的。通过实例分析中的漕河泾开发区站高峰时期与平峰时期的数据对比分析,可以得出,高峰时期的乘客出站平均走行时间比平峰时期的乘客出站平均走行时间长约 16 s。以往关于走行时间的研究都是在基于乘客走行时间路径固定、走行时间固定的假设前提下,与实际有一定的偏差。此模型解决了以往走行时间模型只适用于一个地铁站以及人工调查数据量过大的问题,对于城市轨道交通线网规划、列车运行图编制的研究等都提供重要的依据。

参考文献

[1] 杜鹏,刘超,刘智丽. 地铁通道换乘乘客走行时间规律研究[J]. 交通运输系统工程与信息,2009,9(4):103-109.

[2] 童燚杰,王利鑫,潘明轩,等. 基于数理统计的地铁车站换乘走行时间估计研究[J]. 交通科技与经济,2016,18(1):7-11.

[3] 吕慎,田锋. 轨道交通枢纽换乘乘客到站时间均值特征研究[J]. 交通标准化,2013 (21):21-25.

[4] 郝勇. 地铁车站延滞性步行设施影响乘客走行时间的研究[J]. 铁道运输与经济,2009,31(2):70-72.

[5] 王志刚,石嵘,高伟君. 上海轨道交通车站乘客走行时间函数的分析[J]. 城市轨道交通研究,2010,13(12):57-60.

[6] 邵远忠,邵巍跃,张宁,等. 利用行人路阻函数评估地铁站内 AFC 设备运营状况[J]. 都市快轨交通,2013,26(2):49-52,118.

[7] 何韬,毛保华,杨远舟,等. 地铁换乘站线路间列车到站间隔优化问题研究[J]. 物流技术,2011,30(11):118-121.

[8] 李智,张琦,袁志明. 基于换乘最优的城市圈城际铁路运行图研究[J]. 交通运输系统工程与信息,2015,15(3):114-119,139.

[9]李玉书,孙越,万衡,等. 城市轨道交通车站换乘通道客流压力的评估方法[J]. 城市轨道交通研究,2020,23(1):106-109,144.

[10]ZHOU Yu, WANG Yun, YANG Hai, et al. Last train scheduling for maximizing passenger destination reachability in urban rail transit networks[J].Transportation Research Part B,2019,129:79-95.

(上接第 163 页)

4.1 Hadoop 底层 Linux 优化

(1)关闭 swap 分区。在 Linux 操作系统中,swap 分区是为了解决若一个进程所需的内存空间不足而设立的,通常会在磁盘存储器中分配一块数据临时存储区,需要其相关数据时,再动态将该存储数据移到内存中。这一操作的弊端在于任务的执行效率或多或少都会受到影响,更严重的会导致远程访问登录系统会被挂起。除此外,服务器一旦发生交换,尤其是 HBase 与 ZooKeeper 交换信息时,会话可能无法正常进行,后者会认定该会话已经超时并失效(即租约失效)。这种情况产生的直接后果是,之前部署的 Region 会被重新部署到其他服务器中。若集群在遇上额外压力后,也有可能会发生类似问题。

(2)更好的利用磁盘空间。可使用 tune2fs 命令为一个大型集群分配可用存储空间。默认磁盘保留空间占磁盘百分比总量的 5%,通常在操作系统盘中会默认该设置,这样有利于将数据存储盘安全性提升到最高。

(3)Data Node 处理线程数。参数 xcievers 主要用于 HDFS 的 Data Node 设置服务时收到处理的文件数量限制。在数据节点的日志中,发现参数 xcievers 使用量超过限额后,通常会抛出客户端,以块丢失异常处理。因此,在 Hadoop 的 hdfs-site.xml 文件中,有必要将 xcievers 参数值至少设定不低于4 096。

4.2 针对数据倾斜的优化

数据热点主要是由数据分布的不均衡、数据大规模集中造成。数据热点带来的最大弊端就是无法充分体现分布式系统的优势。一般有两种情况会导致数据倾斜:

(1)reduce 端的数据倾斜。在 mapper 端重写 Combiner,在 mapper 端先做一次 reduce 合并,减少 reducer 端的计算压力。如果两个数据量差异较大的表做 join 时,发生数据倾斜的常见解决方法,是将小表广播到每个节点去,这样就可以实现 map 端 join,从而省掉 shuffle,避免了大量数据在个别节点上的汇聚,执行效率也大大提升。

(2)数据分区的严重不均衡。重写 Partitioner,人为根据数据设置合理分区。如果数据量较大,或数据经常变动,可以采用 Hadoop 提供的随机抽样的方式,自动选取分区的临界点进行分区,保证数据的均衡。

5 结束语

本文提出的基于构建处理电信数据的 Hadoop 平台系统,建立仿真的 BI 前端系统。通过上述处理的数据进行套餐的分析,优化流量套餐设计,实现从访问、搜索、通话时长、短信使用量等行为构建了用户分析体系,多维度定位用户兴趣偏好,通过聚类分析得出流量套餐的适用性并能给出套餐设计的参考意见,为电信服务部门建立决策系统。该平台研究的文件备份与存放动态调整算法,可以高效、准确地挖掘电信交往圈中的频繁交往圈和个人交往圈,对企业具有较高的实用价值。

参考文献

[1] 王广钰. 基于 Hadoop 的时空大数据的分布式检索方法[D]. 北京:中国科学院大学(中国科学院国家空间科学中心), 2017.

[2]江鹤. 面向 CDN 日志业务的数据处理系统的设计与实现[D]. 北京:中国科学院大学(中国科学院工程管理与信息技术学院), 2017.

[3] 赵建喆. 大数据背景下不确定性人工智能中的知识表达、知识获取及推理[M]. 辽宁:东北大学出版社,2016.

[4] 张功水. 基于 Hadoop 技术的电信大数据分析平台的设计和实现[J]. 信息通信, 2016,(10):6-8 .

[5] 李程,柴小丽,谢彬,等. 一种 Hadoop YARN 的资源调度机制[J]. 计算机与现代化, 2017 (11):7-9 .

[6] 袁昌权,胡益群,许光,等. 基于 Hadoop 的高可用数据采集与存储方案[J]. 电子技术与软件工程, 2019 (18): 5-9.

[7] 党引,吴旻荣,李强. 基于 HBase 的海量数据分布式序列存储策略优化[J]. 自动化技术与应用, 2020,39(8): 39-43.

[8] 王伟晨. 基于非关系型数据库 HBase 存储技术的检索研究[J]. 物联网技术, 2020,10(1): 103-105.

[9] 杨大磊. 基于 HBase 的互联网用户行为日志查询系统[J]. 中国新通信, 2020,22(16): 231.