研究与开发

基于流量大数据采集分析的互联网流量疏导策略研究与实践

马蕴颖[1]，王晟寰[2]
（1. 中国电信集团政企信息服务事业群，北京 100010；
2. 美国加州大学戴维斯分校，美国 加利福尼亚州 95616）

摘　要：随着我国数字经济的发展，互联网流量呈现稳步增长态势。针对互联网流量分布不均，将影响网络长期稳定发展的问题，论述了互联网流量获取、识别的方法，构建了实现流量采集的网络架构，以及完成流量分析的软件架构，提出了调节疏导互联网流量的可行性市场化策略。该策略经过运营商的实践，已经取得了初步效果，为引导互联网流量合理分布提供了解决思路和参考依据。

关键词：互联网流量；大数据分析；疏导策略
中图分类号：TN929.1
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2024177

Research and practice of Internet traffic regulating strategy based on big data acquisition and analysis of flow

MA Yunying[1], WANG Shenghuan[2]
1. Government and Enterprise Information Service Business Group of China Telecom Co., Ltd., Beijing 100010, China
2. University of California, Davis, California 95616, USA

Abstract: With the development of China’s digital economy, Internet traffic has shown a steady growth trend. In view of the problem that the uneven distribution of Internet traffic will affect the long-term stable development of the network, the methods of Internet traffic acquisition and identification were discussed. The network architecture for traffic collection, and the software architecture for traffic analysis were constructed. A feasible market-oriented strategy for regulating and channeling Internet traffic was proposed. The propoced strategy has achieved preliminary results through the practice of operators, and it provides a solution and reference basis to guide the distribution of Internet traffic in a proper way.

Key words: Internet traffic, big data analysis, regulating strategy

收稿日期：2024−02−27；修回日期：2024−07−01

0　引言

近年来，随着我国数字基础设施建设的进一步加快，互联网用户规模和各类互联网应用持续增长，我国数字经济获得稳步发展，有力助推了国民经济的提升。

在电信运营商端，移动互联网与千兆光纤网络建设步伐加快。中国互联网络信息中心（China Internet Network Information Center，CNNIC）2023 年 8 月 28 日在京发布的第 52 次《中国互联网络发展状况统计报告》显示，2023 年上半年，我国 IPv6 地址数量为 68 055 块/32，相比 2022 年 12 月增长 1.0%。全国光缆线路总长度达到 6 196 万千米，比 2022 年年末净增 238.1 万千米。累计建成开通 5G 基站 293.7 万个。在用户端，用户规模持续增长。截至 2023 年 6 月，我国网民规模达 10.79 亿人，比 2022 年 12 月增长 1 109 万人，互联网普及率达 76.4%。5G 移动电话用户达 6.76 亿户，蜂窝物联网终端用户超过 21.2 亿户。同时，互联网相关应用持续发展。截至 2023 年 6 月，我国网络视频（含短视频）用户规模为 10.44 亿人，比 2022 年 12 月增长 1 380 万人，网民使用率达到 96.8%。视频应用继续保持高位增长态势，几乎已经发展成为全民化的应用。

从以上数据可以看出，互联网流量也在同步增长。流量的流动属性，决定了其在一定条件下，可以选择不就近服务而不影响使用，这种舍近求远的做法增加了网络的复杂度和资源的浪费。例如，互联网用户位于广东省，广东省的网络条件好且就近部署时延短，但是内容源厂商出于性价比考虑，选择将内容源部署在广东周边省份，这使得广东省的用户不能就近访问，需要跨省访问。这就导致了网络流量的不合理分布，浪费了宝贵的网络资源。运营商曾经采取路由策略进行调节，但实践证明，效果不佳。为了解决以上问题，让有限的网络资源保障互联网流量的增长，而不是任由流量无序流动导致失衡且低质，就需要在了解互联网流量增长趋势及分布规律的基础上，采取合理可行的措施调节，使其均衡分布。

本文从实际出发，突破网络思维，探索了一条获取互联网流量、分析并调控互联网流量分布的市场化方法。该方法已在实践中验证并落地实用，为引导互联网流量从无序到有序，从无控到可控提供了积极有效的解决路径。

1　互联网流量大数据采集模型

1.1　互联网流量总体趋势

根据工业和信息化部发布的移动互联网数据，近年来移动互联网的流量呈现高速增长的趋势。自 2016 至 2023 年 8 年以来，移动互联网流量增长超过 30 倍，达到 3 015 亿 GB。移动互联网接入流量及增长率如图 1 所示。

由图 1 可知，移动互联网接入流量在 2017—2019 年大幅增长。其中，2017 年和 2018 年的增长率超过 100%，2019 年的增长率也超过 70%，这主要得益于视频业务的高速发展。自 2020 年以来，移动互联网接入流量的增长率下降到 40% 以下，主要是因为缺乏大流量新应用的广泛采用，虽然也出现了 AR/VR 等互联网应用，但是迄今尚未实现普及，因此总体流量增长主要来自终端数量及终端的在线时长。根据估算，终端数量的年增长率不超过 5%，在线时长的年增长率不超过 8%。从用户使用习惯推测，时长增长主要贡献给视频应用，包括短视频、长视频和视频会议等，估算年度互联网流量总体增长约在 20% 以内。从 2023 年的移动互联网增长率数据 15.18% 可以推算，今后几年的趋势也基本符合此测算。

1.2　互联网流量采集模型

为了了解互联网流量的具体情况，首先需要采集流量。由于互联网流量数据非常庞大，终端分布在全国各个地市，因此在什么位置采集、采

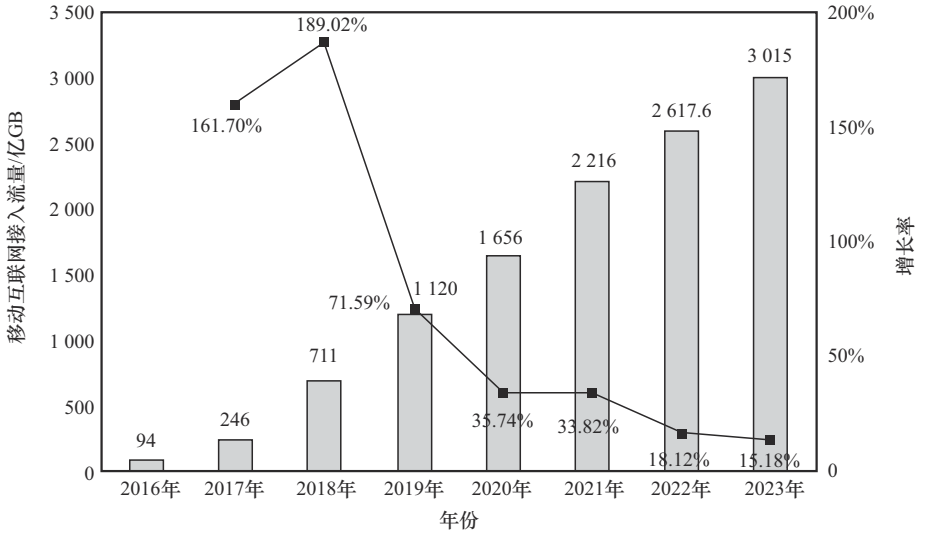

图1　移动互联网接入流量及增长率（数据来源：工业和信息化部）

集的数量多少非常重要，要在可行性和准确性间取得平衡。

互联网流量采集模型的采集点通常有两种：一种是靠近源端，即互联网接入层采集方式；另一种是靠近核心，即互联网骨干层采集方式。其中，互联网接入层采集方式采集的设备量相对较大，一般单区域设备量达千数量级，电路为10万数量级，多区域则再增加一个数量级，这种采集方式的优点是全量采集无遗漏、准确性高，缺点是采集量大，需要多套网管系统，投资和实现难度较大。互联网骨干层采集方式的采集量较小，设备量为百数量级，电路为万数量级，该采集方式的优点是可通过一套网管系统部署实现，实现较容易，不足是无法掌握不经过骨干层的流量，难以满足较精细的分析需求。

为了确保互联网流量分析的精度，且采集量和复杂度在可控的范围内，本文构建了互联网流量采集模型，如图2所示。该模型采用了在区域核心周边进行采集的方案，即把图2中深灰色的区域周边作为流量采集点。该方案综合考虑了数据准确性和采集可行性，采集量为单区域设备数百台、电路数万条，多个区域综合起来设备数千台、电路数十万条，通过新建一套网管系统实现，适用于区域间、区域内部业务间流量分布计算，是一种性价比较高的方案。

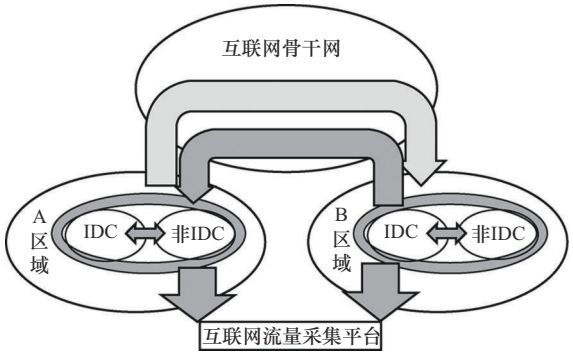

图2　互联网流量采集模型示意图

采集的数据需要包含Netflow数据，即数据包头中的“五元组”信息，包括边界网关协议（border gateway protocol，BGP）路由表信息（即自治系统（autonomous system，AS）路径信息）、设备物理端口的简单网络管理协议（simple network management protocol，SNMP）信息（即采集端口流量）、IP地址信息，以及其他的相关信息等。这些数据将经过传输电路被收集到互联网流量采集平台。

该模型中，城域网出口Netflow和SNMP数据采用直采的方式，互联网数据中心（Internet Data Center，IDC）出口从现有的系统中获取。

实际中，采集设备数量为963台。由于Netflow数据量非常大，不可能全部采集，通常选择1 000∶1～10 000∶1的数据包采集比例。综合考虑有效性和可用性，实际选择采集比例为10 000∶1。

2　互联网流量大数据分析模型

2.1　互联网流量分布识别方法

为了识别互联网流量，尤其是必须准确地识别区域信息，本文建立了一套识别方法，互联网流量分布识别方法示意图如图3所示。该方法对所采集的SNMP和Netflow数据进行关联分析，即根据端口SNMP信息整理的SNMP关键字段信息（见表1），得到所有上联端口的索引信息；根据索引信息，在Netflow数据中进行数据匹配；根据Netflow数据匹配出的源地址和目的地址，核查BGP路由表关键字段信息（见表2）；根据BGP路由表关键字段信息计算出区域，匹配所得区域分布信息（见表3）。

由图3可知，其中关键字段信息为index、ip等，通过关键字段对SNMP、Netflow、BGP路由表进行综合计算得出流量的归属。

表1　SNMP关键字段信息

id	Id
common_region_id	区域ID
type	类型：IDC；城域网
device_name	设备名称
device_ip	设备IP地址
port_name	端口名称
port_desc	端口详情
circuit_properties	端口描述
circuit_type	电路类型
snmp_index	SNMP端口索引
flow_index	Netflow端口索引
time	更新时间

表2　BGP路由表关键字段信息

id	Id
prefix	地址前缀
originator	起源路由器
nexthop	下一跳地址
path	推送文件位置
begin_data_time	文件到达时间
upload_time	文件入库时间
aspath	AS路径
start_ip	起始IP地址
end_ip	结束IP地址

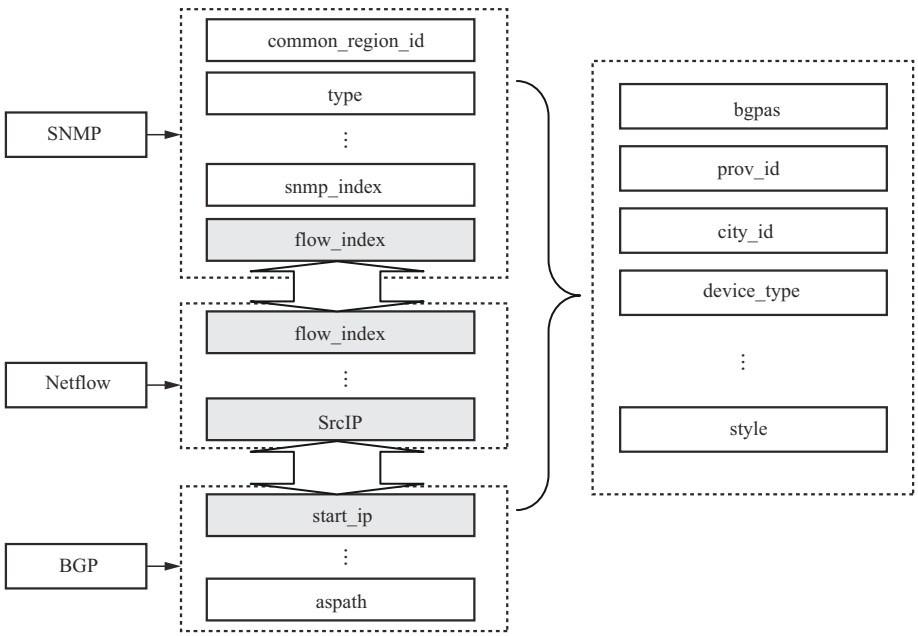

图3　互联网流量分布识别方法示意图

表3　匹配所得区域分布信息

id	Id
prov_id	省份ID
prov_name	省份名
city_id	城市ID
city_name	城市名
device_name	设备名称
device_type	设备类型
device_ip	设备IP地址
bgpas	AS号
style	网络类型
source	来源
cust_name	客户名
create_date	创建时间

2.2　互联网流量分析模型

在完成互联网流量采集并对流量数据进行整理后，本文构建了一套多维度立体的互联网流量大数据分析模型。该分析模型能清晰地展示互联网流量的分布和趋势等情况。

（1）互联网区域流量分析模型

区域定义可细化到片区、省级、城市，从而实现域间、域内的流量分析，包括3类：可以按照省级分析省内、省间的互联网流量，某省到外部各省及外部各省到某省的互联网流量；可以按照城市分析省内各城市间的互联网流量；可以综合省市的互联网流量，分析某城市到外部各省的互联网流量、外部各省到某城市的互联网流量。区域流量分析示意图如图4所示。

（2）互联网网络流量分析模型

依据网络类型定义出的IDC、城域网和移动通信网等相应网络，分析各网络内、网络间的互联网流量，包括IDC到IDC的互联网流量、IDC到城域网的互联网流量和城域网到IDC的互联网流量等。

（3）互联网区域网络综合流量分析模型

依据区域和网络属性，综合分析某区域网络内、某区域网络到其他区域网络间的流量，包括某省IDC到某省城域网的互联网流量、某省IDC到外部各省IDC的互联网流量、某省IDC到外部各省城域网的互联网流量、某省城域网到本省IDC的互联网流量、某省城域网到外部各省城域网的互联网流量、某省城域网到外部各省IDC的互联网流量、某省某地市IDC到外省某地市城域网的互联网流量等。

3　互联网流量大数据分析结果

3.1　互联网流量大数据分析区域分布情况

（1）工业和信息化部移动互联网接入流量大数据区域分布

从如图5所示的工业和信息化部统计的移动互联网接入流量大数据区域分布可知，近3年来，东部地区的移动互联网接入流量占比超过40%，呈增长的趋势；中部地区的移动互联网接入流量占比稳定在22%左右；西部地区的移动互联网接入流量占比接近30%，呈下降的趋势；东北地区的移动互联网接入流量占比约5%，呈下降的趋势。

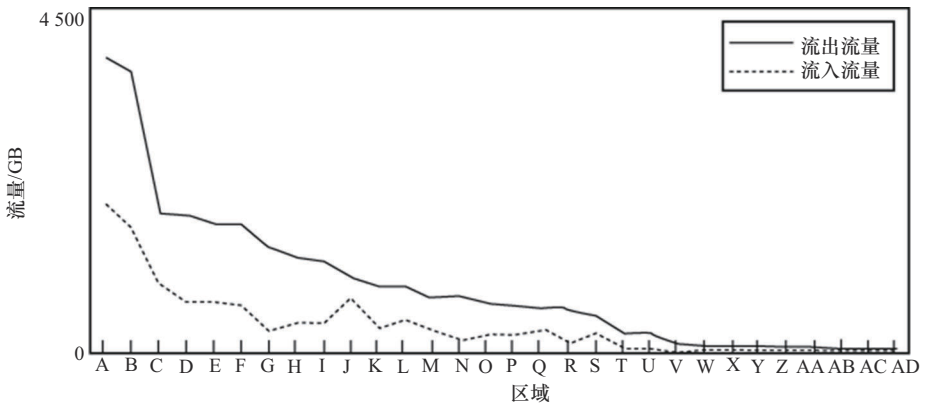

图4　区域流量分析示意图

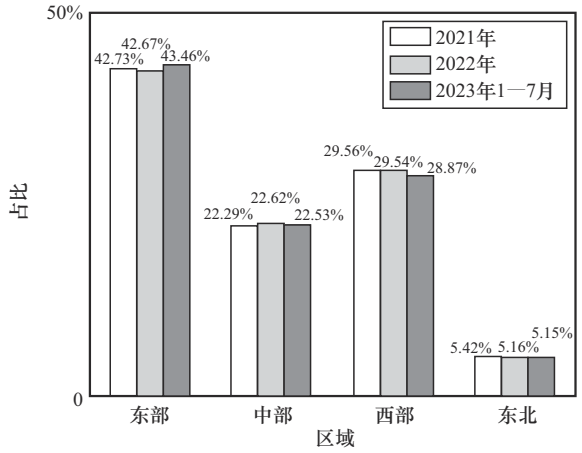

图5　移动互联网接入流量大数据区域分布

（2）运营商骨干网互联网流量大数据区域分布

由如图6所示的某大型运营商近3年骨干网互联网流量大数据区域分布可知，东部、中部、西部和东北的整体分布与工业和信息化部统计的分布基本一致。其中，东北地区的互联网流量占比差异较大，如某运营商的互联网流量占比为3.52%，而工业和信息化部统计的东北地区的互联网流量占比约5%，主要原因在于某运营商在该区域不是最大的运营商，因此相应占比偏低。

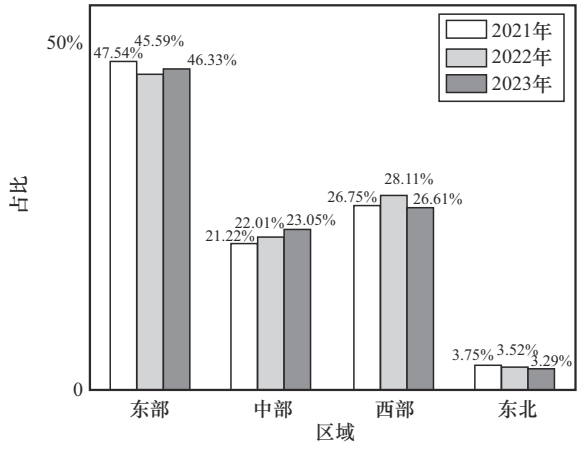

图6　某大型运营商骨干网互联网流量大数据区域分布

（3）运营商IDC互联网流量大数据区域分布

由如图7所示的某运营商IDC互联网流量大数据区域分布可知，东部区域占比最大，与工业和信息化部及运营商骨干网的数据一致。与前面2个分析的不同之处在于，2023年东部地区的互联网流量占比超过60%，比运营商骨干网东部地区的互联网流量占比高出15.27%，中部、西部和东北地区的互联网流量占比分别比骨干网相应地区的互联网流量占比低5.86%、7.37%、1.74%。这说明提供互联网流量的信息源更集中于东部地区。从各个互联网大厂的布局也可以看出，此数据与实际情况契合。

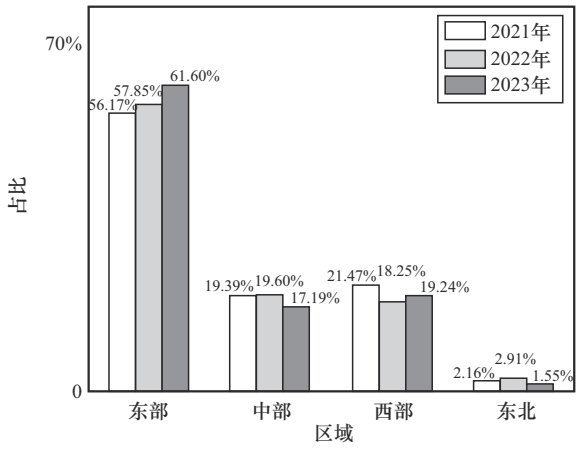

图7　某运营商IDC互联网流量大数据区域分布

3.2　流量趋势与区域分布的关系

移动互联网接入流量反映的是互联网流量使用方的增长情况，运营商骨干网互联网流量反映的是互联网流量提供方及使用方综合的增长情况。某运营商骨干网流量及增长率如图8所示，对比图1和图8可以看出，二者流量均为正增长且趋势基本一致，说明互联网流量提供方所提供的互联网流量与用户使用的互联网流量相匹配。

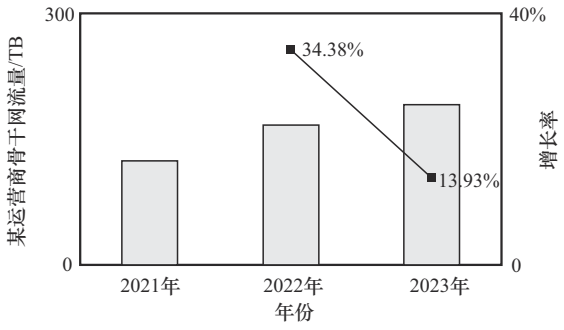

图8　某运营商骨干网流量及增长率

运营商IDC互联网流量反映的是互联网流量提供方的情况，其趋势的下降反映了运营商为互联网流量提供方所提供的互联网流量在下降，但这并不等于互联网流量真正在下降，因为运营商骨干网互联网流量在增长，其下降说明有一部分互联网流量被其他流量所替代。某运营商IDC流量及增长率如图9所示，由图9可知，近2年IDC流量持续下降。

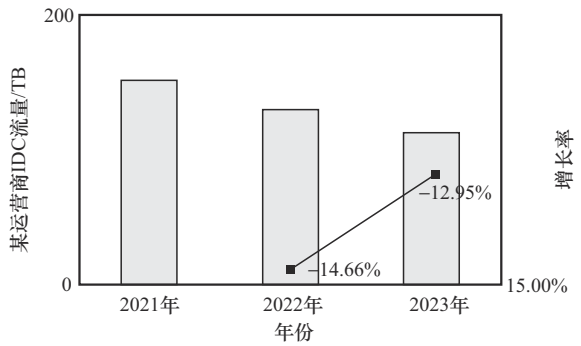

图9　某运营商IDC流量及增长率

运营商IDC互联网流量的区域分布，呈现出明显的东高西低的态势，是因为互联网流量提供方主要位于东部地区，这与中国经济发展的整体分布相一致。

由此就产生了需要解决的问题，即运营商IDC互联网流量与整体互联网流量的发展趋势存在背离，不符合整体互联网流量的发展趋势，不能满足互联网业务长期稳定运行发展的需求。在实际生产中，运营商为了改变背离问题，首先采取的是网络手段，如流量劫持、丢包等，对流量进行限制和干扰。由于这些方法对网络质量造成了影响，常常遭到用户的投诉，因此难以得到有效实施，导致流量越来越偏离预期，使得运营商陷入了进退两难的境地。为了有效解决流量不均衡的问题，必须改变思路，为此，本文提出了一个解决此问题的系统方案。

4　互联网流量疏导策略

4.1　现网互联网流量大数据模拟

采用互联网流量大数据采集模型对运营商现网的互联网流量进行采集，根据大数据分析模型得到了详细的区域流量分布数据。

以模拟采集的2023年7月互联网流量的数据为例，根据获取到的互联网流量均值数据（Gbit/s）计算得出各区域IDC及非IDC净流出互联网流量模拟数据（见表4）。

表4　各区域IDC及非IDC净流出互联网流量模拟数据

区域	IDC净流出互联网流量/亿GB	非IDC净流出互联网流量/亿GB
东部	7 059.82	−1 121.64
中部	−769.86	2 013.86
西部	−6 791.37	−566.81
东北	501.42	−325.40

由表4可知，东部地区的IDC为主要的数据流出源，中部地区的非IDC为主要的数据流出源。为了使流量的分布更符合实际，即东部地区的IDC与非IDC均作为数据流出的源，且增长率与骨干的增长率相一致，需要对流量进行调节。由表4数据可以看出，最需要调节的是东部和中部地区的流量。

4.2　互联网流量疏导策略研究

对互联网流量进行疏导的实践证明，聚焦流量本身的网络调节手段可行性不足，因此，需要创造性地引入市场化手段对流量进行调节，即对区域间的互联网流量进行结算，当结算价格达到一定程度时，区域间的互联网流量将会下降，从而实现对互联网流量的调节。其内在的逻辑是通过市场化的手段，引导内容源主动布局，将流量调整至合理区域。

为此，本文将IDC间的结算费用设定为a元/GB，非IDC间的费用设定为b元/GB，得到东部和中部区域模拟结算（见表5）。

为了疏导互联网流量，采用不同的价格策略将会引导互联网流量产生转移。

（1）非IDC互联网流量区域间疏导策略

假设a=0，即IDC结算价为0，说明IDC互联

表5　东部和中部地区模拟结算

区域	IDC间结算出/（元·GB^{-1}）	非IDC间结算出/（元·GB^{-1}）	总结算出/（元·GB^{-1}）
东部	7 059.82a	−1 121.64b	7 059.82a−1 121.64b
中部	−769.86a	2 013.86b	−769.86a+2 013.86b

网流量可以在东部和中部地区间无成本流动，非IDC的流量在区域间流动产生成本。按照表5计算方式可得，东部结算出负值，中部结算出正值，即中部地区向东部地区结算费用，结出金额为（2 013.86−1 121.64）b=892.22b。

如果单价b取10 000元/GB，则中部地区向东部地区结算出899.22万元。本策略将引导非IDC互联网流量从中部向东部转移。同时，单价b取值越大，将越有利于非IDC流量从中部向东部转移，最终达到平衡。

（2）IDC互联网流量区域间疏导策略

假设b=0，即非IDC结算价为0，说明非IDC互联网流量可以在东部和中部地区间无成本流动，IDC间互联网流量如果去往别的区域，就需要产生成本。从表5可以明显看出，东部结算为出，中部结算为入，金额为（7 059.82−769.86）a=6 289.96a。此策略将引导IDC流量从东部向中部转移。

（3）综合疏导策略

假设b>2.5a，根据表5计算方式得到，−769.86a+2 013.86b>7 059.82a−1 121.64b，即中部向东部地区结算的总金额将超过东部向中部结算的总金额。在这种情况下，IDC间和非IDC间的互联网流量都会产生成本。

由于a、b单价的调节，此策略不仅会引导中部互联网流量向东部互联网流量流动，还将促使非IDC互联网流量向IDC互联网流量流动。

4.3　互联网流量疏导策略实践

（1）互联网流量大数据采集网络组织方案

为了实现互联网流量疏导，首先需要采集互联网流量大数据，为此设计了一套互联网流量采集分析网络组织方案，如图10所示。总体思路为在各区域（如省级维度）部署采集设备，从各区域采集原始互联网流量数据，通过传输通道上传至互联网流量采集与分析平台。在实际应用中，根据系统流速=总峰值流数×流平均大小×8/采样比，可以计算得出总流速，从而可以得到各区域上传数据的带宽，以及处理此数据所需要的设备数量。当前，该系统每日采集的数据量达百TB。

（2）互联网流量大数据分析软件架构

为了完成对采集的互联网流量的分析，根据前文互联网流量大数据分析模型设计了一套互联网流量采集分析软件架构，如图11所示。其中，数据采集模块实现原始数据的统一采集及汇聚，主要包括Netflow数据、SNMP数据统一采集，并进行数据稽核、清洗统一格式处理；数据转发与预处理模块为数据采集底座的转发通道，实现将格式化的Netflow数据向其他功能模块的转发，同时由预处理模块补全各类关联辅助数据；数据分析模块汇聚IDC和城域网等互联网流量后，实现各个维度的互联网流量流向监控分析，提供日志监控处理和误差校验；数据呈现模块对分析数据进行最终呈现，包含流量热力分布、全域流量视图、区域流量视图和客户流量视图等多维度图形展示；系统管理模块实现用户管理、角色管理和权限管理等系统功能；基础数据管理模块实现对Netflow、SNMP基础数据的监控，并与其他系统进行数据关联分析和误差校验，确保数据源准确可靠。

（3）互联网流量疏导策略初步结果

构建上述采集模型、分析模型，某运营商已实现了对全国各省IDC、城域网的互联网流量数据的采集及分析，目前正着手开展相关的区域间

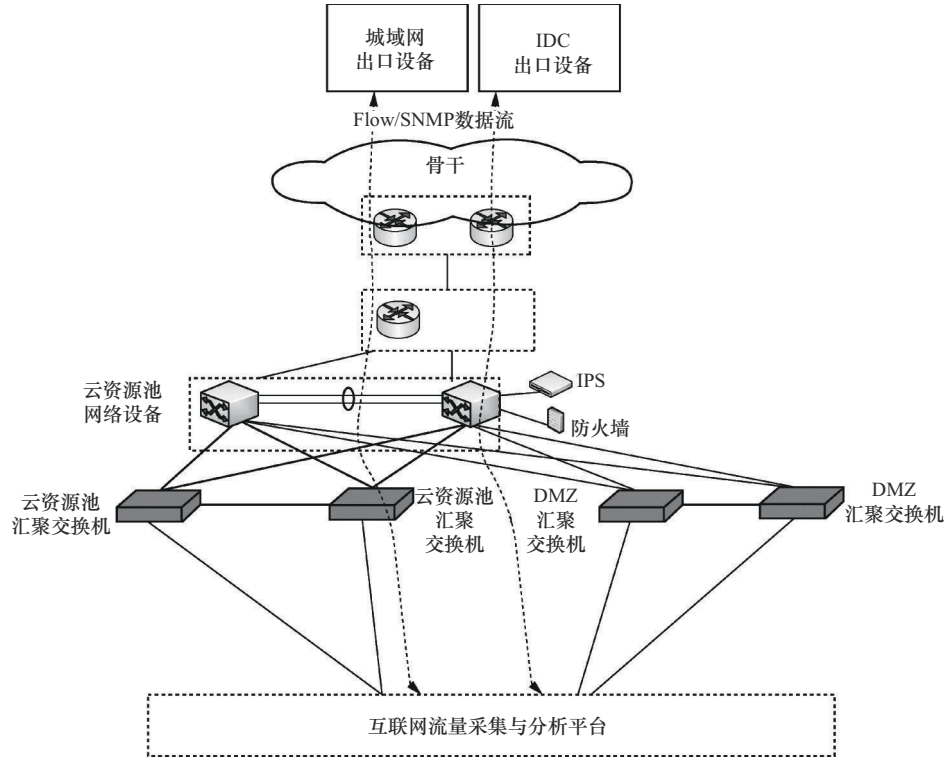

图10　互联网流量采集分析网络组织方案

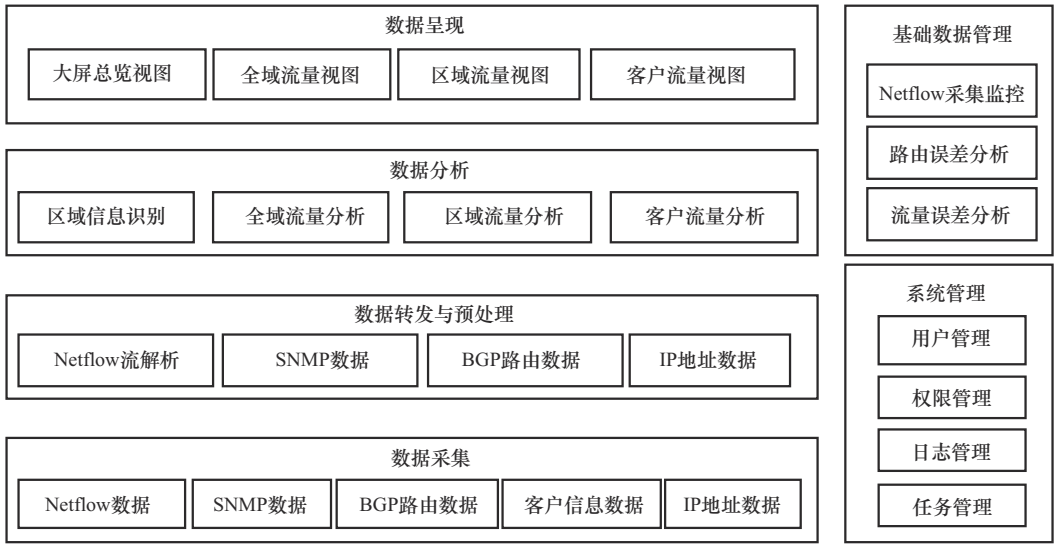

图11　互联网流量采集分析软件架构

互联网流量结算工作。某运营商已确定互联网流量结算的范围为所有省份，即如果A省的流量流向B省，A省就需要向B省结出费用，结出金额=流向B省的流量×单价，反之亦然。为了实现精细化管理，A省流向B省的流量区分了IDC和城域网流量。现网为了实现以上结算方案，除构建上述软/硬件体系外，还建立并打通了数据稽核、出账、列收的完整流程。目前该方案已基本落地，后续为了确保流量疏导的有效性，还需要持续关注流量的变化，不断优化单价取值，同时可以引入必要的参数，如免结算流量基准值等，从而对流量进行更精细化的调节。根据近3个月的

流量大数据采集分析发现，区域间IDC、城域网流量已经开始下降，区域内的流量开始上升。

可以预见，本文所研究的流量大数据采集与分析方法，以及成本调节疏导策略，将能够引导互联网流量分布更趋合理。

5　结束语

互联网流量增长是大势所趋，但其不均衡分布将会影响网络的稳健性，造成网络资源的浪费和服务质量的下降。长此以往，互联网可能会陷入低质量、不均衡发展的泥沼，最终可能形成“劣币驱逐良币”的局面，导致网络整体劣化，在人工智能、工业互联网等创新领域方面，这种不均衡性可能会阻碍其发挥应有的作用。

本文介绍了一种突破传统网络思维的流量疏导策略，该策略能够在不妨碍互联网流量正常流动的前提下，有效引导互联网流量的良性流动。经过实践验证，已取得了初步效果。该策略的提出将有效解决网络低质低效的问题，确保互联网的长期稳定运行，为网络使用者提供更好的服务体验，为数字经济高质量发展提供助力。

参考文献：

[1] 中国互联网络信息中心 . 第 52 次中国互联网络发展状况统计报告[R]. 2023.
China Internet Network Information Center. The 52nd statistical report on China’s internet development[R]. 2023.

[2] 付钰, 王坤, 段雪源, 等 . 面向软件定义网络的异常流量检测研究综述[J]. 通信学报, 2024, 45(3): 208-226.
FU Y, WANG K, DUAN X Y, et al. Survey of research on abnormal traffic detection for software defined networks[J]. Journal on Communications, 2024, 45(3): 208-226.

[3] 马蕴颖, 梁汗臣 . 互联网流量分布趋势研究[J]. 通信世界, 2022(12): 36-39.
MA Y Y, LIANG H C. Research on Internet traffic distribution trend[J]. Communications World, 2022(12): 36-39.

[4] 蒲勇霖, 许小龙, 于炯, 等 . 流式大数据平台下的弹性数据迁移能效优化策略[J]. 通信学报, 2024, 45(2): 188-200.
PU Y L, XU X L, YU J, et al. Energy-efficient optimization strategy based on elastic data migration in big data streaming platform[J]. Journal on Communications, 2024, 45(2): 188-200.

[5] 马蕴颖, 王晟寰 . 互联网流量分布计算模型研究与实践[J]. 通信世界, 2023(18): 39-42.
MA Y Y, WANG S H. Research and practice on calculation model of Internet traffic distribution[J]. Communications World, 2023(18): 39-42.

[6] 郭亮, 杨裔, 秦炳峰, 等 . 基于大数据技术的甘肃智慧旅游系统[J]. 大数据, 2024, 10(1): 157-169.
GUO L, YANG Y, QIN B F, et al. Gansu smart tourism system based on big data[J]. Big Data Research, 2024, 10(1): 157-169.

[7] 马蕴颖, 王晟寰 . 互联网流量中 PCDN 业务识别模型算法研究及政策思考[J]. 广东通信技术, 2024(2): 56-60.
MA Y Y, WANG S H. Research on algorithm of PCDN service identification model in Internet traffic and policy consideration[J]. Guangdong Communication Technology, 2024(2): 56-60.

[8] 李国杰 . 大数据与计算模型[J]. 大数据, 2024, 10(1): 9-16.
LI G J. Big data and computing models[J]. Big Data Research, 2024, 10(1): 9-16.

[9] 邵杭青, 马蕴颖, 梁汉臣 . 算力调度关键要素及路径分析[J]. 江苏通信技术, 2022(6): 76-80.
SHAO H Q, MA Y Y, LIANG H C. Key elements and path analysis of computing power scheduling[J]. Jiangsu Communication Technology, 2022(6): 76-80.

[10] 孙玉娣 . 基于电信大数据的 5G 网络海量用户复访行为预测模型[J]. 电信科学, 2023, 39(2): 157-162.
SUN Y D. A prediction model of massive 5G network users’ revisit behavior based on telecom big data[J]. Telecommunications Science, 2023, 39(2): 157-162.

[作者简介]

马蕴颖（1973– ），女，中国电信集团政企信息服务事业群高级工程师，主要研究方向为互联网、IDC、网络运营。

王晟寰（2001– ），男，美国加州大学戴维斯分校在读，主要研究方向为计算机网络的发展和流量疏导解决方案。