



基于大数据的网页浏览质差分析方法研究

程卫华, 何肖嵘

(中国电信股份有限公司上海分公司, 上海 200120)

摘要: 传统的网络运维往往聚焦网络和设备的运行状态, 难以解决用户主观感知问题, 迫切需要建立一套端到端互联网应用感知管理体系。围绕网页浏览的质量分析环节, 基于大数据技术研究用户感知数据的质量分析框架和方法, 设计提出了一套基于大数据技术的质量分析系统框架, 用于实现互联网业务质量分析。以某国际网站为例, 提出了“二望五切”诊断法用于定界溯因, 提高人工分析效率, 验证了所提方法的有效性。

关键词: 大数据; QoE; 质量分析; KQI; KPI

中图分类号: TN929.5

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2020126

Quality analysis method of Web browsing based on big data

CHENG Weihua, HE Xiaorong

Shanghai Branch of China Telecom Co., Ltd., Shanghai 200120, China

Abstract: The traditional network operation and maintenance often focuses on the operation status of the network and equipment, which is difficult to solve the user's subjective perception. It is urgent to establish an end-to-end Internet application awareness management system. Focusing on the quality analysis of web browsing, the quality analysis framework and method of user perceived data based on big data technology was studied, a quality analysis system framework based on big data technology was designed and proposed, which was used to realize the quality analysis of Internet business. Taking an international website as an example, a diagnosis method of "two looks and five cuts" was proposed to delimit and trace the cause, which improves the efficiency of manual analysis and verifies the effectiveness of the proposed method.

Key words: big data, QoE, quality analysis, KQI, KPI

1 引言

随着城市光网发展, 用户带宽提速, 网络持续扩容, 传统的网络运维往往聚焦网络和设备的运行状态, 难以满足用户需求。因此, 迫切需要建立一套端到端互联网应用感知管理体系, 形成

完整的采集评测、质差分析、优化改进和效果验证用户体验质量 (QoE) 运营机制。

1.1 传统的 QoS 运维

传统的业务质量 (QoS) 运维一般采用网络 KPI (关键性能指标), 从两个角度来关注网络质量。

(1) 网元级 KPI (设备性能)

通过网管监测路由器、交换机、服务器等网元设备 CPU 利用率、内存占用率、磁盘占用率, 以及端口状态、循环冗余校验码 (CRC error)、光衰值等, 分析网元设备是否良好支持现网业务, 以此推进设备扩容和优化。

(2) 网络级 KPI (网络性能)

通过网管监测时延、分组丢失率、流量等, 以此来推进网络扩容和优化。

传统网管视角只立足于运营商和设备商, 存在较大局限性, 并未从端到端 (用户到网站) 角度, 全面了解互联网所承载业务的感知。因此, 往往会存在自查好、用户感知却不佳的棘手问题。

1.2 演进的 QoE 运营

作为用户对服务满意程度的一种评价方式, ITU-T 定义了 QoE, 由用户主观感知决定, 通过量化方式建立 KPI、KQI (关键质量指标) 和 QoE 映射关联, 从已知或容易测量的影响因素来预测难以直接测量的主观感知。基于 QoE, 应用性能管理 (application performance monitor, APM) 领域近几年也逐步兴起^[1]。不仅需要保障网络质量, 还需要关注上层应用性能和用户感知, 这样才能更好地适应瞬息变化的宽带市场。QoE 指标体系如图 1 所示。

QoE 运营方法, 即通过指标采集、感知评测、质差分析、优化改进、效果验证 5 个步骤, 建立以 QoE 为导向的客户体验管理体系, 从而改进网络和业务的不足环节, 提升核心竞争力。其中, 质量分析是至关重要的一环, 针对感知劣化问题, 通过精准分析定界溯因, 辅助优化改进。QoE 运营方法如图 2 所示。

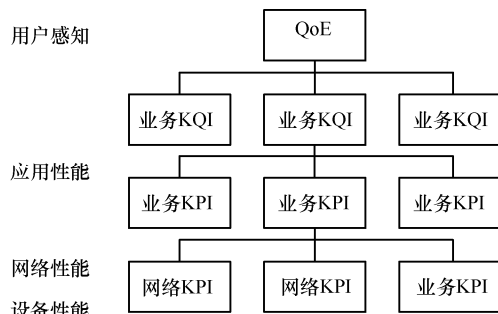


图 1 QoE 指标体系

业界内有利用各类终端和软件工具采集感知数据的相关案例, 也形成了各种各样的质差分析经验, 存在的难点和挑战如下。

(1) 数据分散, 难于管理

大数据时代, 企业只有将运营支撑系统 (operation support system, OSS)、业务支撑系统 (business support system, BSS) 和管理支撑系统 (management support system, MSS) 等数据资源有效集中管理起来, 才能更好地挖掘数据的价值和支撑精准分析。

(2) QoE 质差分析方法缺乏专业指导

工业和信息化部先后发布过《YD/T 2400-2012 宽带速率测试方法固定宽带接入》《YD/T 2691-2014 宽带速率测试方法客户上网体验》《YD/T2016-1918 宽带互联网业务体验网络评分计算方法网页/视频/测速业务》等行业标准, 主要涉及 KQI 的定义和算法。

但是, 这仅对 KQI 采集环节有指导意义。运维人员在得到 KQI 后, 对于如何将各种网管采集的 KPI 数据结合起来以及进行关联分析挖掘, 却是一筹莫展、杂乱无章, 或凭个人经验、顾此失彼。因此, 端到端感知质量分析环节往往成为耗

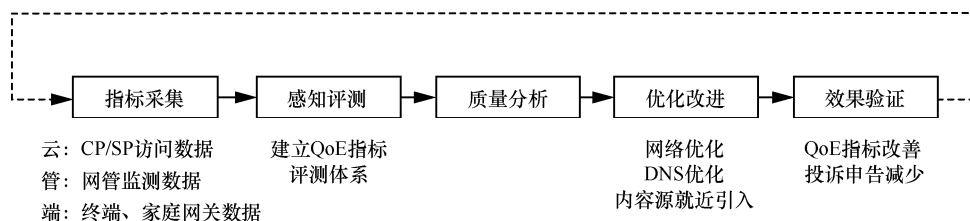


图 2 QoE 运营方法



费人力、难以攻克的一个难关。

(3) 人工分析和抽样排查容易以偏概全

由于缺乏大数据技术的有效支撑,很难真正用数据说话,发现质差因子,主要依靠人工分析局部数据,或进行抽样排查,难免有失偏颇。

2 网页浏览质量分析方法研究

2.1 网页浏览交互过程

Web 页面由一个基础文档和多个内嵌资源组成,基础文档一般采用超文本标记语言(hypertext markup language, HTML)编写,代表网页框架;其中又引用了其他内嵌资源,即图片、CSS 文件、JS 脚本等元素。网页浏览交互过程如下:

(1) DNS 解析;

(2) TCP 建立;

(3) HTTP 报头交互至首分组到达;

(4) HTML 文本页面下载;

(5) 其他元素下载直至整个首页加载完成。

网页浏览交互过程如图 3 所示。

2.2 网页浏览指标采纳和采集

针对网页浏览应用规范定义了 KQI、KPI 等分级指标用于 QoE 评测。网页浏览应用监测体系如图 4 所示。

课题研究中,主要采纳 2 个 KQI 用于感知监测、4 个业务 KPI、3 个网络 KPI 和 1 个统计指标辅助质差分析。网页浏览业务 KQI 见表 1。

主要的业界数据采集方式如下。

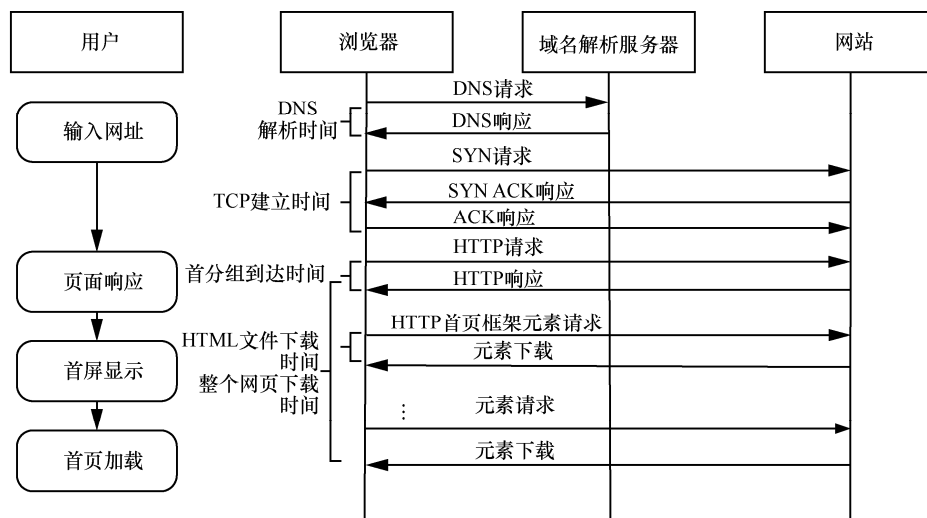


图 3 网页浏览交互过程

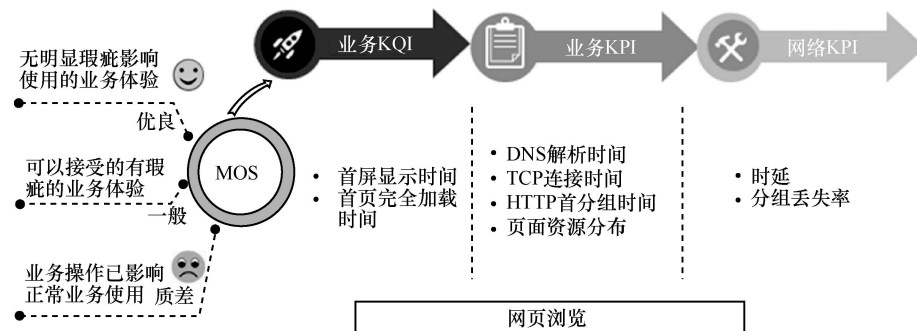


图 4 网页浏览应用监测体系

表 1 网页浏览业务 KQI

对比项	指标	指标含义
KQI	首屏显示时间	浏览器中 1 024×768 区域呈现网页内容的时间
	首页完全加载时间	浏览器中首页页面全部元素加载完成的时间
业务 KPI	DNS 解析时间	客户端发起访问请求, DNS 解析完成的时间
	TCP 连接时间	客户端发起访问请求, TCP3 次握手建立连接的时间
	首分组时延	从客户端发起 HTTP 请求开始, 到收到第一个 HTTP 响应报文的时间
	页面资源分布	页面元素的大小、类型、所在服务器 IP 地址等特征
网络 KPI	时延	ping 测试目标地址的双向时延
	分组丢失率	ping 测试目标地址的分组丢失率
	路径	tracer 测试路径
统计指标	首页访问成功率	访问某网页的成功率

(1) 主动拨测方式

根据 W3C 提出的 Navigation Timing API 规范, 主流浏览器一般都提供标准的 Performance Timing 性能接口^[2], 可以直接获得浏览器访问的相关质量指标。通过各类主动拨测探针, 模拟用户采集网络、应用层的 KPI 和 KQI。优点是实现方便, 缺点是仿真行为不能绝对代表用户感知。

(2) 被动监听方式

通过植入代码、深度分组检测 (deep packet inspection, DPI) 等技术被动记录数据或监听报文。优点是精确获取用户感知, 缺点是部署难度较大, 使用范围有限。例如植入代码只适用于内容提供商、自营网站、IDC 等场景。

本文研究主要采用主动拨测方法, 利用各类软硬件探针仿真用户浏览, 采集感知数据。

2.3 质量分析系统

2.3.1 质量分析框架

本文设计提出了一套基于大数据技术的质量分析系统框架, 用于实现互联网业务质量分析^[3-4], 具体如图 5 所示。

(1) 数据采集层

统一收集 OSS、BSS、MSS 等数据。

(2) 接口层

通过高吞吐量的 Kafka 分布式消息队列和传统 FTP/SFTP 等接口, 结合数据总线, 使原始数据通过抽取 (extract)、转换 (transform) 和加载 (load) 过程, 将数据汇聚到大数据平台。

(3) 计算存储层

通过 MapReduce/Spark 计算组件, 实现数据多维度的关联计算和处理, 并存储在数据仓库中。通过 elastic earch 组件, 采用 SPL (search processing language) 进行高效搜索分析。

(4) 分析挖掘层

基于大数据平台实现网页浏览感知问题的各项分析挖掘, 提高运维人员分析工作效率, 快速定位关键问题。

2.3.2 质量分析方法

将探针采集数据作为训练集, 结合用户体验反馈, 采用回归和决策树分类算法, 将首屏显示时间和首页完全加载时间两项 KQI 作为评估网页浏览应用的主要业务指标, 主观感知采用 MOS 值和优、良、中、次、劣 5 档体验等级表示^[5-7]。

通过常态化跟踪监测工作, 收集到越来越多的海量质量数据进行校验, 使得感知评估模型不断迭代和完善。网页浏览业务感知评估模型见表 2。

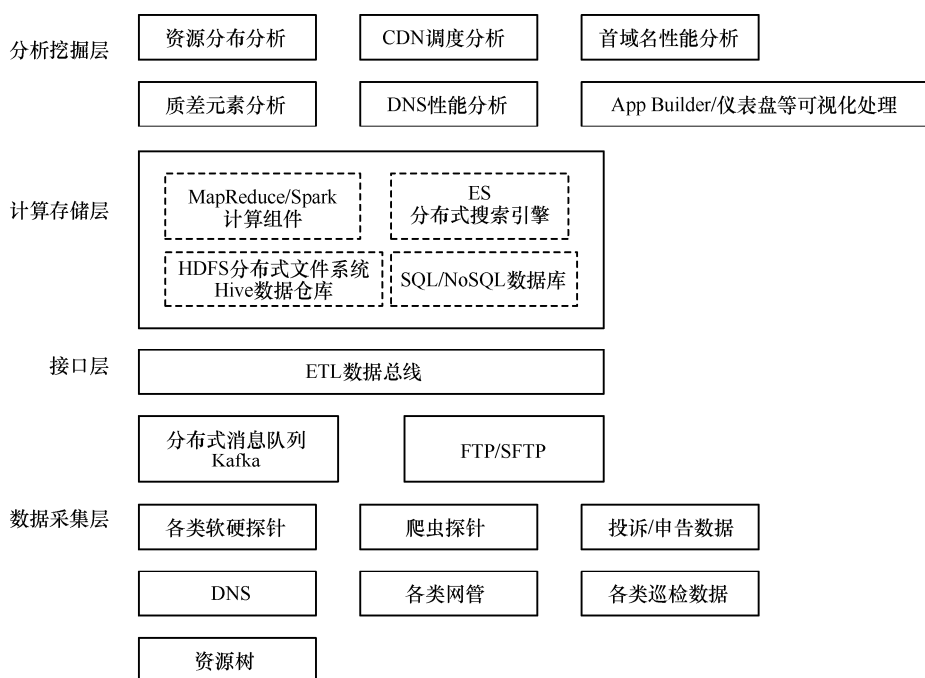


图5 质量分析框架

表2 网页浏览业务感知评估模型

体验等级	体验指标	用户感知
优, MOS 值: $[y_1, y_2]$	首屏显示时间: $[a_1, a_2]$	网页浏览“零”等待, 页面主体内容“秒”显
	首页完全加载时间: $[b_1, b_2]$	
良, MOS 值: $[y_3, y_4]$	首屏显示时间: $[a_3, a_4]$	网页浏览稍有等待感, 页面主体内容显示快速
	首页完全加载时间: $[b_3, b_4]$	
中, MOS 值: $[y_5, y_6]$	首屏显示时间: $[a_5, a_6]$	网页浏览有明显等待感, 可感觉到页面主体内容的逐级显示过程
	首页完全加载时间: $[b_5, b_6]$	
次, MOS 值: $[y_7, y_8]$	首屏显示时间: $[a_7, a_8]$	网页浏览等待时间较长, 感觉到页面主体内容空白区域的时间较长
	首页完全加载时间: $[b_7, b_8]$	
劣, MOS 值: $[y_9, y_{10}]$	首屏显示时间: $[a_9, a_{10}]$	网页浏览白屏时间过长, 加载超出用户忍耐程度, 选择放弃浏览
	首页完全加载时间: $[a_9, a_{10}]$	

本文结合运维分析经验, 提出“二望五切”诊断法。具体如下。

- 首先通过网络对标、基线跟踪等手段采集国内外知名网站的 QoE 指标。
- “二望”: 通过 KQI 和 KPI 宏观发现问题。
- “五切”: 进一步从以下 5 个角度依次排查诊断质差问题。

(1) 资源分布分析

结合 DNS 日志和网络爬虫技术, 分析国内外

热点网站资源占比和网页资源分布构成。前者是为了全局掌握资源引入情况, 支撑市场决策; 后者是为了从网站设计和 CDN 调度角度评估对感知造成的影响。

(2) CDN 调度策略分析

通过软硬探针数据, 分析 CDN 主机调度优先顺序, 这对内容调度策略的优化具有较大参考价值。

(3) 首域名性能分析

针对首域名, 从网络性能和应用性能两个角

度进行排查分析。

(4) 质差元素分析

基于大数据挖掘出每个网站的质差元素集,再针对每个质差元素进行 KPI 分解分析和质差元素性能分析。KPI 分解可以细分为 DNS 解析时延、TCP 连接时延、首分组时延等情况,逐一检查指标是否劣化以判断质差可能涉及的因素;质差元素性能分析从网络性能和应用性能两个角度进行排查分析。

(5) DNS 性能分析

通过软硬探针数据,分析 DNS 服务器的域名解析时间,这可以帮助运维人员发现 DNS 性能劣化的问题,指导扩容优化。网页浏览质差分析流程如图 6 所示。

2.3.3 质差问题分类

质差问题主要可以划分为:用户侧问题、网络侧问题、资源平台侧问题、资源引入类问题和调度策略类问题,如图 7 所示。定位质差原因后,在运维可控范围内,可以通过网络优化、DNS 优化、资源合作引入、调度策略协同优化等方式改善用户感知。

2.3.4 质量分析案例

针对国际网站 A,采用“二望五切”诊断法进行分析。

(1) 质量评测

通过质量跟踪,发现该网站感知指标较差,首屏时间 6.18 s,首页完全加载时间 11.21 s,用户

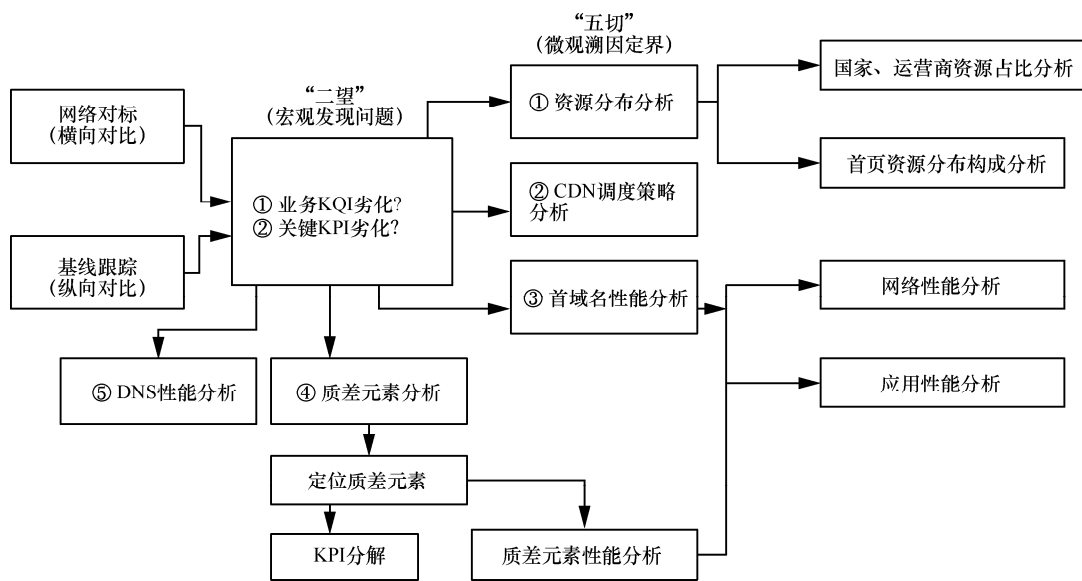


图6 网页浏览质差分析流程

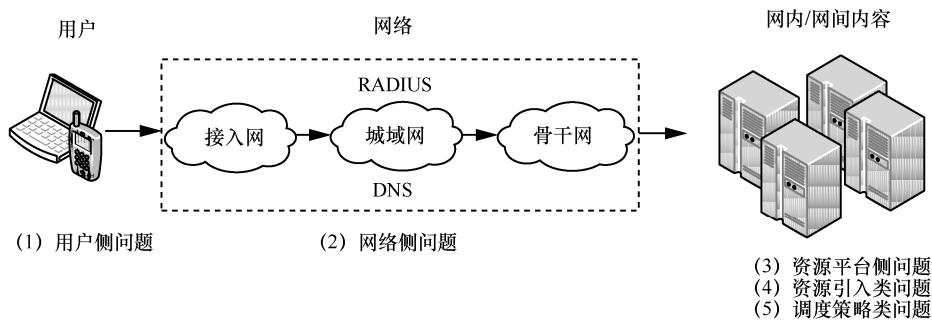


图7 质差问题分类



主观感知较差。

(2) 资源分布分析

利用爬虫技术爬取网站两层网页进行资源汇总占比分析,资源主要集中在欧美、日本,占比分别为 63%和 17%。再更进一步从首页资源构成来看,首页元素主要分布在所属欧美、日本的 3 个域名下,其网页元素占比高达 61.5%。

(3) CDN 调度策略分析

通过 CDN 主机访问量分析调度优先级,发现该网站虽然有中国的 CDN 节点,但是却优先调度到了欧美和日本的 CDN 主机。

(4) 首域名性能排查

汇总首域名 IP 地址集合,根据时延、分组丢失率定位出网络性能质差 IP 地址集有 5 个。

针对每个质差 IP 地址收集 path ping 遍历到的网络路径集,时延最长一条路径为 123 ms,路径共有 12 跳。逐跳排查,发现劣化点处于第 7 跳,时延增大为 104 ms,存在分组丢失现象。进一步关联网管数据,诊断为该段落链路流量在忙时存在一定程度拥塞。

通过分析,得到以下结论:

- 网络段落忙时拥塞对感知有影响。
- 资源引入和 CDN 调度策略也是不容忽视的重要因素。

根据分析结论,提供市场和运维决策参考,在可控范围内,考虑通过扩容或疏忙、资源合作探讨等优化举措进行改进。

跟踪国际网站 B,发现存在质差元素,导致网络好但感知却未必好的棘手问题。网页渲染过程中,用户感知差,白屏等待直到 68 s 才打开网页。

(1) 质差元素分析

结合大数据分析和元素瀑布图,发现首页中存在 4 个质差元素,访问连接失败。排除网络因素,由于内容中关键元素访问超时导致体验较差。

(2) 分析结论

该网站已经排除网络因素,但不排除资源平台侧等不可控因素。

此类问题短期内改善可能性较小,但通过质量分析定界溯因,已能很好地诠释了用户感知的复杂性,不仅受网络影响,还受 CP/SP 内容源、CDN 分布,甚至更多不可控的其他因素影响。

3 结束语

本文聚焦网页浏览应用的质量分析环节,综合探针、网管、资源树等各类数据,基于大数据挖掘技术研究用户感知的质量分析框架和方法,提出了“二望五切”诊断法用于定界溯因。这套方法通过不断实践,将主观的运维经验升华为较为客观的分析理论,定量和定性分析相结合,提高了运维人员网络优化工作的效率和准确性。

然而,本文主要涉及探针质量数据和网管基础数据,随着企业 OSS、BSS、MSS 数据的迅猛增长,需要基于大数据技术不断深入挖掘,形成更多切实有效的分析方法,补充到现有质量分析体系中。

目前,针对用户感知只做到初级的质差定界溯因,希望后续能通过机器学习和预测等手段,力求变事后分析为事前预警。

参考文献:

- [1] GUITART J, TORRES J, AYGUADE E. A survey on performance management for internet applications. concurrency and computation[J]. Practice and Experience, 2010, 22(1): 68-106.
- [2] 王懿. Web 页面加载性能监测平台的设计与实现[D]. 武汉: 华中科技大学, 2015.
WANG Y. Design and implementation of web page loading performance monitoring platform[D]. Wuhan: Huazhong University of Science and Technology, 2015.
- [3] 孙雪松. 基于大数据分析的电信网络质量评估系统的实现和应用[D]. 北京: 北京邮电大学, 2017.

- SUN X S. Implementation and application of telecom network quality evaluation system based on big data analysis[D]. Beijing: Beijing University of Posts and Telecommunications, 2017.
- [4] 蔡森, 沈俊良. 面向 5G 的移动网络用户感知大数据分析体系研究[J]. 中国新通信, 2019(8): 34.
- CAI S, SHEN J L. Research on big data analysis system of mobile network user perception for 5G[J]. China New Communications, 2019(8): 34.
- [5] 康波, 刘胜强. 基于大数据分析的互联网业务用户体验管理[J]. 电信科学, 2013, 29(3): 32-35.
- KANG B, LIU S Q. Management of the user experience based on the big data analysis[J]. Telecommunications Science, 2013, 29(3): 32-35.
- [6] 党永亮. 大数据分析在移动通信网络优化中的应用研究[D]. 武汉: 华中师范大学, 2015.
- DANG Y L. Application research of big data analysis in mobile communication network optimization[D]. Wuhan: Central China Normal University, 2015.
- [7] 邹洁, 郑淑琴, 肖慧. 基于 DPI 大数据分析的移动互联网端到端业务感知 KQI 质量定段预定位方法[J]. 广东通信技术, 2017(4): 2-6, 23.
- ZOU J, ZHENG S Q, XIAO H. The KQI quality segment pre-positioning method of the mobile internet end-to-end business perception based on DPI data analysis[J]. Guangdong Communication Technology, 2017(4): 2-6, 23.

[作者简介]



程卫华 (1979—), 女, 中国电信股份有限公司上海分公司高级工程师, 主要研究方向为互联网业务感知。

何肖嵘 (1976—), 男, 中国电信股份有限公司上海分公司工程师, 主要从事云网融合、核心网、光宽端到端的研究工作。

《电信科学》第五届编辑委员会

主任委员:

韦乐平 中国电信集团有限公司

副主任委员:

朱 峰 中国通信学会

刘华鲁 人民邮电出版社有限公司

彭木根 北京邮电大学

编委:(按姓氏笔画排序)

山世光 中国科学院计算技术研究所

王 宇 中国科学院电子学研究所

王继业 国家电网有限公司信息通信部

方景龙 杭州电子科技大学

龙 腾 北京理工大学

毕 军 清华大学

卢光跃 西安邮电大学

曲 桦 西安交通大学

吕文俊 南京邮电大学

刘永祥 国防科技大学

刘建明 工业和信息化部产业发展促进中心

闫 江 北方工业大学

江 涛 华中科技大学

孙晓颖 吉林大学

阳小龙 北京科技大学

纪越峰 北京邮电大学

杜百川 国家广播电视总局科学技术委员会

李 丹 清华大学

李长海 中国工信出版传媒集团

李玉峰 上海大学

杨小康 上海交通大学

吴 巍 中国电子科技集团公司第五十四研究所

余少华 中国信息通信科技集团有限公司

沈连丰 东南大学

宋令阳 北京大学

张云勇 中国联合网络通信集团有限公司

张成良 中国电信集团有限公司

张同须 中国移动通信有限公司研究院

张宏莉 哈尔滨工业大学

陈 巍 清华大学

陈山枝 中国信息通信科技集团有限公司

陈芳炯 华南理工大学

陈铁明 浙江工业大学

陈章渊 北京大学

易东山 北京信通传媒有限责任公司

金 石 东南大学

周 涛 中国电信股份有限公司上海西区电信局

周一青 中国科学院计算技术研究所

周华春 北京交通大学

孟维晓 哈尔滨工业大学

赵军辉 华东交通大学

赵慧玲 工业和信息化部通信科学技术委员会

段晓东 中国移动通信有限公司研究院

高 鹏 中国移动通信集团设计院有限公司

高新波 重庆邮电大学

唐雄燕 中国联合网络通信有限公司研究院

曹卫平 桂林电子科技大学

曹蓟光 中国信息通信研究院

章坚武 杭州电子科技大学

梁海滨 北京信通传媒有限责任公司

董振江 高新兴科技集团股份有限公司

蒋 力 中国电信股份有限公司研究院

蒋林涛 中国信息通信研究院

程 方 重庆邮电大学

奚 笠 中国铁塔股份有限公司通信技术研究院

蔡 康 中国电信股份有限公司研究院

收录声明

本刊已被中国核心期刊遴选数据库(万方)、中国学术期刊数据库(知网)全文收录,被本刊录用的文稿,将一律由编辑部统一纳入中国核心期刊遴选数据库(万方)、中国学术期刊数据库(知网)。若作者不同意被收录,将在来稿时向本刊说明,本刊将做适当处理。若无特别声明,视同将录用文稿的电子网络版、汇编或合订本的使用权授予本期刊社,本刊支付的稿费中已包含上述费用。

The articles of *Telecommunications Science* have been embodied by WANFANG DATA and CNKI. The accepted articles will be unified into WANFANG DATA and CNKI. If the author does not agree to his or her article to be included, the author must illustrate when the article is submitted. The editorial department will properly handle. If the authors do not give a special announcement, the editorial department believes that the authors have granted the right of their articles to *Telecommunications Science*. The expenses have been included in the fee paid by editorial department.

道德声明

本刊发表的文稿是作者独立取得的原创性研究成果,无一稿多投;文稿内容不涉及国家机密;未曾以任何形式用任何文种在国内外公开发表过;文稿内容不侵犯他人著作权和其他权利。若发生一稿多投、侵权、泄密等问题,文稿作者将承担全部责任。

The authors of *Telecommunications Science* guarantee that their submitted articles are original and contain nothing confidential. The said article is only submitted to *Telecommunications Science*. The said article has not been published before and has not been submitted elsewhere for print or electronic publication consideration. The said article is in no way whatever a violation or an infringement of any existing copyright or license from the third party. Otherwise, the authors of the said article shall take the blame for the violation or infringement of the related copyright and the leakage of secrets.

电信科学

Telecommunications Science



ISSN 1000-0801



9 771000 080200

邮发代号：2-397 国外代号：M841 定价：68.00元