大数据分析技术在网络领域中的研究综述

冯贵兰[1] 李正楠[2] 周文刚[3]
(中国民航飞行学院现代教育技术中心 四川 广汉 618307)[1]
(中国民航飞行学院航空工程学院 四川 广汉 618307)[2]
(中国民航飞行学院飞行技术学院 四川 广汉 618307)[3]

摘 要 随着移动互联网、物联网、5G通信网等新兴技术的迅猛发展,数以亿计的网络接入点、联网设备以及网络应用产生的海量数据,给网络故障排查、网络安全保障等带来了极大的挑战,同时也为人们深度挖掘和充分利用网络大数据的巨大价值带来了机遇。大数据分析可以处理海量数据,并从中抽取有价值的潜在知识,帮助决策者发现隐藏的关系和模式,近年来引起了学术界和工业界的广泛关注。文中围绕大数据分析技术应用于网络领域的最新研究成果,首先阐述了网络大数据的概念、分类和数据分析方法;然后从无线网络、SDN网络、光纤网络和网络安全4个层面着重介绍了大数据分析技术在故障检测、流量监控、网络优化、流量预测、APT攻击检测、网络异常检测等网络领域中的解决方案,重点分析和归纳了这些解决方案中大数据分析技术的思路;接着回顾了大数据分析技术在工业界中应用的情况;在此基础上,给出了基于大数据分析的网络设计周期;最后总结了大数据分析技术在网络领域中面临的机遇和挑战,并指出下一步需要关注的研究方向。

关键词 大数据分析,网络优化,频谱管控,流量预测,网络安全
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.06.001

Research on Application of Big Data Analytics in Network

FENG Gui-lan[1] LI Zheng-nan[2] ZHOU Wen-gang[3]
(Modern Education Technology Center, Civil Aviation Flight University of China, Guanghan, Sichuan 618307, China)[1]
(Institute of Aviation Engineering, Civil Aviation Flight University of China, Guanghan, Sichuan 618307, China)[2]
(Institute of Flight Technology, Civil Aviation Flight University of China, Guanghan, Sichuan 618307, China)[3]

Abstract With the rapid development of new technologies like mobile internet, Internet of Things and 5G communication network, more and more infrastructures, devices and data are generated, such as hundreds of millions of network access points, networked devices, applications as well as massive data. Thus, great difficulties and challenges are brought to fault tolerance, cyberspace security, leading to some traditional solutions become inefficient to such large scale and complex security problems. Meanwhile, the increase of network big data presents unprecedented opportunities on deeply mining and taking full advantage of the big value of network big data. Big data analytics can extract hidden, valuable patterns, and useful information from big data. Therefore, both academia and industry have been attracted again by network field based on big data analytics, and have made certain research achievement. Researches on network field mainly involve four research directions, namely wireless network, SDN network, optical network and cyberspace security. First, the survey starts with the introduction of the big data basic concepts, data model and data analytics. Second, there is a detailed review of the current academic and industrial efforts toward network design using big data analytics. Third, the main network design cycle is illustrated by employing big data analytics. This cycle represents the umbrella concept that unifies the surveyed topics. Forth, the challenges confronted by the utilization of big data analytics in network design are identified. Finally, several future research directions are highlighted.

Keywords Big data analytics, Network optimization, Spectrum management, Traffic prediction, Cyberspace security

网络以快速、大规模和多样化的方式产生数据,每天约产生2.5艾字节的数据[1]。Facebook拥有16.5亿用户,每月有10亿活跃用户;微信拥有10多亿用户,其中有7亿活跃用户[2]。随着网络中海量数据的不断积累、大数据分析的迅速发展,隐藏在数据背后的巨大价值也逐渐显现出来。网络运营商充分利用大数据这种宝贵的资源,可以优化网络性能,最

到稿日期:2018-11-18 返修日期:2019-02-17 本文受民航飞行数据分析研究项目(XM2852)资助。
冯贵兰(1988—),女,硕士生,工程师,主要研究领域为大数据与网络安全,E-mail:fengguilan1016@sina.com(通信作者);**李正楠**(1985—),男,主要研究领域为网络与信息安全;**周文刚**(1981—),男,博士生,讲师,主要研究领域为网络管理、人工智能等。

大限度地提高网络收益[3]。面对大型网络数据时,传统数据分析技术存在以下不足。(1)传统数据分析技术主要处理结构化数据,但大量基于应用程序的数据通常是非结构化的。(2)数据分析的实现通常局限在一个部门或业务单位,最终的分析结论是基于非常有限的局部角度给出的,而不是全局角度。(3)传统分析技术主要针对交易数据,对运营数据关注较少,无法实时决策。大数据分析能够提取比传统数据分析更具洞察力的信息,可以整合分散于网络的各种零星信息,并利用数据挖掘技术深入研究网络中各要素之间的关系,在网络规划与优化、流量预测与控制、网络安全保障等方面发挥重要作用。例如,与用户相关的完整数据通常分散在不同的业务部门中,大数据分析能够收集分散的数据,从多个角度了解用户行为和偏好,从而描绘出完整的用户画像。大数据分析的另一重要特征是实时处理,通过大数据分析,运营商可以实时监控基础架构,并做出自主和动态的决策。随着互联网的发展,各应用(如社交网络、物联网、智能电网等)对现有网络提出了更多要求,如更灵活、更快捷、更安全、更智能。为满足这些需求,可以结合网络中收集的大量数据和分布式高性能计算平台构建基于大数据分析的网络平台,将网络从无视数据管理转换为富有洞察力的上下文感知网络。

目前,已有学者对大数据在网络领域中的部分应用进行了梳理和总结。Qian 等[4]从应用层、网络层、数据传输层、数据层 4 个层面总结了大数据在无线网络中的研究工作。徐全盛等[5]综述了大数据分析在无线通信技术中的研究工作,主要包括无线频谱管控、网络规划与优化、无线资源管理。付钰等[6]则是介绍了基于大数据分析的 APT(Advanced Persistent Threat)攻击检测工作。陈兴蜀等[7]总结了大数据在网络安全与情报分析中的作用。总体而言,国内针对大数据分析在网络领域的应用集中在某一子领域,仅有国外的少量学者对其进行了整体研究[1,4]。

对大数据在网络领域中的研究进展进行综述,对促进国内在该方向的研究有着十分重要的意义。本文研究了大数据分析在网络平台中发挥的作用,从可靠的底层传输到灵活控制的网络架构,再到保证网络自身安全以及网络上所传输信息的安全,旨在为各种网络应用提供可靠性、可用性、安全性的基础平台。通过对近年来大数据分析在网络领域中的应用成果进行归类和梳理,形成如图 1 所示的研究体系。

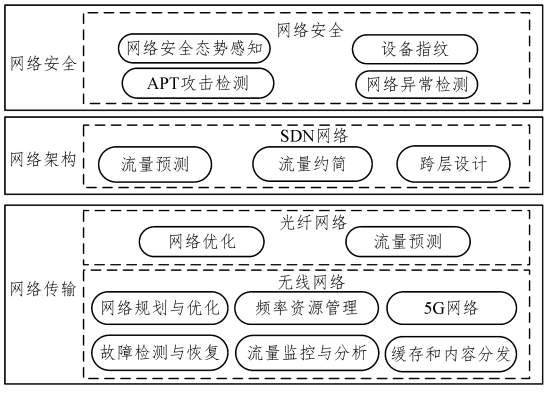

图 1 大数据分析在网络领域研究中的典型应用

Fig. 1 Typical applications of big data analytics in network

将内容从下到上分成 3 层:网络传输、网络架构和网络安全。鉴于篇幅,本文未能涵盖所有重要主题,仅从无线网络、软件定义网络(Software Defined Network,SDN)、光纤网络和网络安全 4 个研究领域对大数据分析在这些领域的典型应用进行分析和讨论,并着重阐述研究成果的技术思路。然后,总结基于大数据分析的网络设计周期,以便于研究人员全面地理解大数据分析在网络领域中的一般应用流程。最后,对存在的挑战和未来的发展趋势进行总结和分析。

本文第 1 节概述网络大数据的相关概念和分类;第 2—5 节分别对大数据分析在无线网络、SDN 网络、光纤网络和网络安全中的研究进行梳理和总结;第 6 节介绍产业界对大数据分析的应用情况;第 7 节总结基于大数据分析驱动的网络设计周期,并分析所面临的挑战;第 8 节指出未来的研究方向;最后总结全文。

1 网络领域的大数据分析

1.1 概念和分类

目前,网络大数据的应用得到了人们的广泛研究,但对于网络大数据,目前还没有一个公认的概念。文献[4]认为网络大数据通常是指无法在有限时间内用现有通信和网络系统进行传输、访问、处理和服务的数据集合。文献[8]则认为网络大数据是指“人、机、物”三元世界在网络空间中彼此交互与融合所产生并在互联网上可获得的大数据。

通过对已有网络大数据定义的研究,本文归纳总结出网络大数据的 5V 特点,即 Volume(体量浩大)、Velocity(生成快速)、Variety(模态繁多)、Value(价值巨大但密度很低)和 Volatility(波动性),如图 2 所示。(1)网络中数据的体量不断扩大,尤其是蜂窝网的高速发展带来了大量数据流量,数据规模已经从 TB、PB 扩大到了 EB。(2)网络中以极高的速率不断产生数据,具有很强的时效性。(3)网络大数据类型繁多,包括结构化数据、半结构化数据和非结构化数据。(4)网络大数据常常蕴含着潜在的价值,例如无线信令数据可以帮助改善网络部署和服务质量,呼叫详细记录(Call Detail Records,CDR)可以揭示用户的社交网络和用户行为。(5)由于网络具有突发性,移动设备具有移动性等特点,网络大数据是非平稳的,具有波动性。

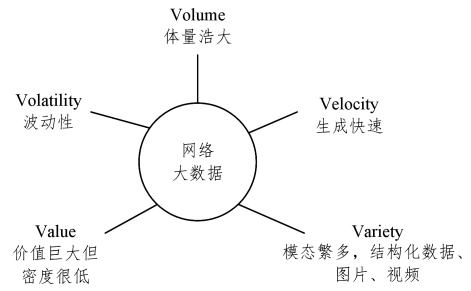

图 2 网络大数据的特点

Fig. 2 Characteristics of network big data

针对科技企业、研究学者和数据分析师们由于各自的关注点不同而对网络大数据有着不同的分类的问题,表 1 进行相关总结,以帮助研究人员更好地理解网络大数据的深刻内涵。

表 1　网络大数据的分类

Table 1　Categories of network big data

文献	分类角度	分类名称
[9]	电信运营商	IT 系统中的数据 接入网和核心网中的数据 运营商互联网应用中的数据
[10]	潜在应用	流记录数据 网络性能数据 移动终端数据 附加数据
[4]	网络社会生态	原始无线大数据 衍生无线大数据 发展中无线大数据
	应用领域	蜂窝网络、Wi-Fi 热点 智能手机 D2D、智能电网 无线传感器网络、物联网

Chen 等[9]从电信运营商的角度将数据分为 3 类：1）IT 系统中的数据：用户属性、业务消费信息、终端信息等。这些数据来自于客户关系管理系统、计费系统和终端注册平台，可以用于描述基本用户的画像和特征。2）接入网和核心网中的数据：移动信令、网络数据包、M2M（Machine-to-Machine）数据等。当客户使用语音、SMS（Short Messaging Service）服务或网络服务时，这些数据被收集在有线/无线网络中。数据的基础结构很复杂，因此需要针对不同类型的数据进行有针对性的分析和处理，以实现基于用户位置和偏好的场景描述。3）运营商互联网应用中的数据：网上营业厅数据、掌上业务数据、翼支付数据等。所有的数据，包括用户访问方式、地址、时间、商业偏好和消费习惯，都完全保存在应用程序的后台，容易获取。

Zhang 等[10]从潜在的应用角度出发，将移动蜂窝网络中的数据分为流记录数据、网络性能数据、移动终端数据和附加数据。首先，蜂窝网络中的流记录数据是描述无线用户行为最重要的数据，包括数据记录和信令记录等。其次，网络性能数据旨在为无线用户提供网络性能和服务质量的评估，主要包括关键绩效指标（Key Performance Indicator，KPI）和测量报告（Measurement Report，MR）。最后，移动终端数据可以通过移动应用程序收集，包含设备信息和无线参数等。

Qian 等[4]从网络社会生态的角度将无线大数据分为 3 类：原始无线大数据，衍生无线大数据，发展中无线大数据。首先，原始无线大数据表示由大量无线用户生成的数据集，包括无线接入行为、无线应用需求等。其次，衍生无线大数据是为向无线用户提供有效通信服务而产生的频谱、传输、接入和网络数据，包括频谱利用的分布、超密集部署小区的空间统计和传输信号的资源分配。最后，发展中无线大数据是指在对未知频谱、新型传输技术、创新接入和革命性网络结构的性能进行测试和评估的过程中产生的数据集。同时，他们还指出无线大数据也可以根据其特定领域进行分类，包括蜂窝网络、Wi-Fi 热点、智能手机 D2D（Device-to-Device）、智能电网、无线传感器网络和物联网等。

1.2　数据收集

网络大数据可以从内部或外部来源进行收集。外部数据来自国家/地方统计局、市场研究机构、客户投诉部门等。内部数据来源于运营系统、业务系统等支撑系统。数据收集方法可以分为两类：通过数据源收集和通过辅助工具收集[11]。移动设备本身就是数据收集工具。例如，用户可以通过麦克风收集音频信息，通过摄像头收集图片、视频等多媒体信息，通过 GPS、WiFi 或蓝牙等收集地理位置信息。网络数据可以通过包捕获技术或专业软件获得，如 ComView 和 SmartSniff 等。此外，专业人员可以对网络接口（如空中接口、A/Gn 接口）进行某些探测以收集信令数据。

1.3　数据模型

随机矩阵理论模型可以用来表示从多个源收集的不同数量的数据。文献[12]研究了基于随机矩阵理论和机器学习的移动蜂窝网络大数据分析统一数据模型。为阐明基于随机矩阵理论的大数据分析方法的性能，文章提出 5 个数据类型的例子，包括大信号数据、大流量数据、大位置数据、大无线电波形数据和大异构数据，利用时空数据集的高维度表征了大数据与蜂窝网络间的相互关系和独特特征。文献[13]引入大规模随机矩阵作为构建块，对大规模多输入多输出系统（Multiple Input Multiple Output，MIMO）收集的海量大数据建模，并将其转发到基站进行处理和存储。该模型适用于分布式频谱感知和网络监控。软件定义的无线电平台，配备通用软件无线电外设，用于模拟基站中的天线并演示 CPU 中的数据处理。

针对网络大数据来源众多的特征，文献[14]引入了统一张量模型来表示来自多个源的数据，并基于张量扩展算子将不同的数据类型表示为子张量的形式；同时，使用上述模型描述了一种降低大数据维度的高阶奇异值增量分解方法，并以智能交通为例验证了数据表示模型和增量降维方法的性能，可以看出该模型能作为数据表示的大数据系统模型来实现。

1.4　数据分析

面对时空维度中的海量数据集，需要更强大的分析理论和方法才能获得新的见解。本节将讨论几种常用的数据分析技术，包括时间序列分析、机器学习和博弈论框架。

网络大数据具有时空维度，适合采用时间序列分析技术。文献[15]提出了一种无线网络时序数据简化的方法，该方法使用时间序列分析来分解常规分量和随机分量，并使用时间序列来预测基于常规分量的流量模式，具有较高的可预测性。

近年来，机器学习[16]，尤其是深度学习[17]，在许多领域里显著提高了建模和预测的性能。深度学习[18]试图通过使用多层神经元和多个非线性变换来建模高级数据表示，以进行大数据分析[19]。已有学者尝试利用机器学习和深度学习来解决网络问题。Buczak 等[20]综述了基于数据挖掘和机器学习算法的入侵检测方法，从方法的复杂性、准确性、可理解性以及使用训练过的模型对未知实例进行分类的时间等方面对比了不同的入侵检测方法。张蕾等[21]从系统安全、网络安全和应用安全 3 个层面阐述了机器学习在网络安全空间中的应用。Alsheikh 等[22]提出了一个基于 Spark 的可扩展学习框架，该框架可支持分布式深度学习。在包含数百万条记录的真实数据集上，该框架的加速效率得到展示。Ma 等[23]将重点放在电信运营商非常关心的电话变化预测问题上，并在

3 种情况下验证了 4 种预测模型的性能：Logistic 回归、随机森林、支持向量机(Support Vector Machine，SVM)和 E-BP(Enhanced Back Propagation)神经网络。

在考虑网络管理和控制时，无论是网络节点还是终端节点，博弈论都是分析多个对象之间交互的强大工具。Yang 等[24]提出了一种基于多认知代理的分治网络管理和控制架构，并提出了马尔可夫博弈论建模框架。此外，他们还重点研究了状态空间的构建、状态转换的计算以及并行 Q-learning 技术的收敛。该技术为无线大数据网络提供了一种合适、有效的建模工具。

1.5 数据分析的平台和工具

Hadoop[25]和 Spark[26]是典型的可用于大数据分析的平台。

Hadoop 是一个用于存储和处理大数据的开源框架，其核心是 HDFS 和 MapReduce。HDFS 为海量数据提供存储，MapReduce 为海量数据提供计算。Hadoop 的架构如图 3 所示。

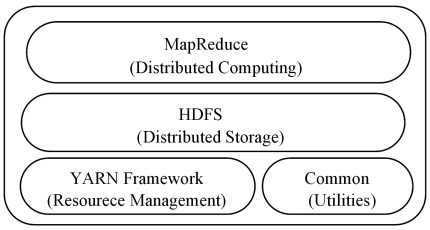

图 3 Hadoop 架构图

Fig. 3 Architecture of Hadoop

Spark 是运行在分布式计算集群上的大规模数据处理的快速和通用引擎，是一个新兴的大数据处理引擎。Spark 实现了内存计算机制，省去了大量的磁盘 I/O 操作，非常适合用于迭代机器学习任务。应用程序通过 Spark 的资源管理器申请所需 CPU 和内存等资源，从节点启动相应的进程等待主节点分配任务，集群上的节点协同完成任务并把结果返回给应用程序，如图 4 所示。

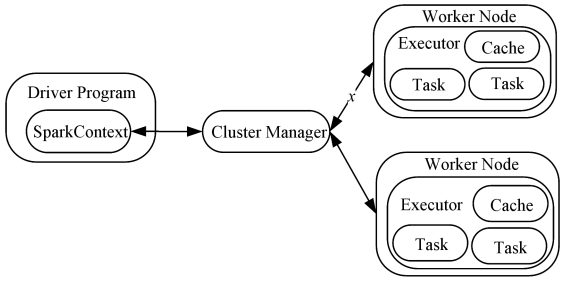

图 4 Spark 集群的部署

Fig. 4 Deployment overview of Spark cluster

Hadoop 适用于批处理操作，具有应对大量持久网络数据的优势，如 CDR 和 GPS 数据、网络日志等，这些数据都需要频繁从存储系统中提取出来进行分析。在网络领域，Hadoop 被用于流量分析、日志分析等历史数据分析任务。Spark 适用于流式数据处理和交互式数据查询，具有实时性和数据快速交互式查询的优势，常用于网络安全与情报分析。用户可根据具体应用场景下不同阶段的数据计算要求选择相应的平台。

2 大数据分析技术在无线网络中的应用

本节介绍大数据分析在无线网络中的相关研究。与其他网络子领域相比，无线网络这一领域已有大量研究成果发表。这些研究成果大致分为 6 类：故障检测、流量监控、缓存相关、网络优化、频谱管控和 5G 通信网。

2.1 故障检测

基于大数据分析的故障检测，主要是通过分析切换成功率、用户通话记录、带宽趋势、测量报告等数据，来解决弱覆盖、用户异常、设备故障、休眠小区等网络故障。

最令用户沮丧的遭遇之一，是手机在通话过程中忽然中断。用户从一个小区移动到另一个小区，如从 3G 基站移动到 2G 基站时，会发生掉话、无法接通等现象。此类问题的常用解决方案是路测、网络仿真和基于 KPI 方法，但路测的开销大且费时，网络仿真的结果部分不可靠，基于 KPI 的方法不够精确。因此，Celebi 等[27]利用大数据提出了 3G 网络覆盖分析方法，其通过挖掘运营商存储的大量网络测试数据来优化现有网络的覆盖性能。该方法基于 Hadoop 平台分析基站系统应用在基站子系统和移动交换中心之间的信息交换。实例表明，3G 网络覆盖分析方法具有较高的精度，优于基于 KPI 的方法。网络运营商借助运营支撑系统(Operation Support System，OSS)大量收集网络性能测量数据，并通过大数据技术对 OSS 数据进行处理，实现自动、高效的网络优化。Gao 等[28]提出一种基于大数据分析的下行覆盖自优化算法，通过对数据的获取和分析，记录现有无线网络的性能状况，定位具有异常覆盖性能的小区；在此基础上，通过调整天线参数提高下行覆盖性能。分析结果表明，覆盖优化算法对于电信运营商来说是高效、低成本的。

Karatepe 等[29]提出基于 CDR 的异常检测方案，用于检测用户移动的异常行为。首先从网络节点中收集 CDR 数据到中介部门；然后把收集到的 CDR 分配到相关部门，如数据仓库、计费部门和收费部门；随后使用 Hadoop 平台检测异常，将发现的异常反馈给中介部门，以便其采取适当的行动。实验表明，该方案可以检测出基于位置的异常，并提高蜂窝网的性能。Parwez 等[30]提出一种基于聚类的移动网络大数据异常检测方法，该方法基于 CDR 中包含的时空信息分析用户在不同时间和地点的活动，使用 k 均值聚类和分层聚类检测用户异常。与文献[29]不同的是，Parwez 采取了进一步的措施来提取异常背后的含义和原因，以便相关部门采取适当的措施，如主动分配资源或避免电池中断。Yang 等[31]提出深度网络分析器(Deep Network Analyzer，DNA)——一个用于移动蜂窝网络的异常检测和根本原因分析的大数据分析平台。首先，使用关联规则挖掘方法从历史数据集中学习 KQI(Key Quality Indicators)和 KPI 之间的关联模式；然后，使用这些模式构建指纹知识库，并使用传入的数据流量定期更新；最后，检测 KQI 和 KPI 数据集中的异常行为，并将检测到的异常模式与数据库中的指纹进行比较，以确定根本原因。基于 Spark 平台实现 DNA，并使用实际生产数据进行测试。测试表明，DNA 是异常检测和根本原因分析的高效平台，适用

于为数千万移动用户提供服务的大规模蜂窝系统。

Sahni 等[32]通过在小区级上分析带宽趋势来预测设备和基础设施的故障。由于收集到的数据具有大量性和多样性特征，因此使用大数据分析技术来处理数据是非常必要的。对此，可以在特定的时间段内(月或年)获取用户的接收带宽，然后对来自不同源的数据进行集成与分析，从而预测带宽趋势。

蜂窝网络的自愈功能可以自动、快速、准确地检测和定位影响网络性能的故障，并自动恢复，以确保用户连续、高质量地通信。具备自愈功能的小区能独立或联合地调整无线参数及相关算法，将无线系统的性能损失降到最低，同时还大大降低了维护成本和人员投入。为保证良好的服务质量，自愈系统必须在合理的时间内完成。网络中大量的数据与时间限制，使得需要采用大数据方法来解决自愈问题。Khatib 等[33]介绍了移动网络中 3 个使用大数据分析进行自愈的示例：数据归约、休眠小区检测和基于 KPI 的关联检测。Imran 等[34]通过收集并分析用户设备(User Equipment，UE)定期报告的测量值，解决了 5G 自组织网中休眠小区检测的问题，实验表明该方法具有较高的检测准确率。Jiang 等[35]通过大数据分析研究了基站在移动蜂窝网络中的行为和活动及其可预测性，利用大数据平台对蜂窝网络大数据进行分析，用于分析的具体数据集包括 1000 多座蜂窝基站、100 多万条 CDRs 线路、数百万用户和持续 100 多天的呼叫详细记录 CDR。文章首先详细介绍了呼叫数据的预处理和清理方法，以获得有价值的大数据集；其次提出特征提取和呼叫可预测性两种方法，以捕获基站的行为并分析其可预测性；接着进行详细的活动模式分析，包括呼叫分布、互相关特征、呼叫行为模式和日常活动；然后提出详细的分析方法来挖掘基站行为；最后使用一个研究案例来验证基站行为和活动的可预测性。研究表明，大数据技术可以有效地捕捉网络行为，预测网络活动，有助于实施高效的网络管理。

2.2　流量监控

大型蜂窝网络具有较高速率，能满足人们对移动多媒体业务的需求。通常，这些网络使用高性能、大容量的服务器来执行流量监视和分析。但随着人们对数据速率、数据量和详细分析需求的不断增加，此方法存在很大的局限性。因此，Liu 等[36]提出了基于 Hadoop 的移动互联网流量监控和分析系统，它部署在一个大型蜂窝网络的核心端，每天可以有效地处理来自 123Gbit/s 链路的 4.2 TB 的流量数据，具有成本低、性能高的优点。

Ocampo 等[37]提出了一个基于 Spark 的企业网络流量监控框架，其允许流处理和批处理。流处理可以分析活动会话、使用端口和每个会话的带宽使用等参数，能在不同的时间跨度上监视用户的网络活动。该框架能批量处理存储的用户流量，旨在进一步比较当前用户活动与历史数据。流式和批量分析都将形成异常检测的基础，能识别异常、恶意行为和错误配置的服务。实验表明，该框架在不同数量的用户和工作负载下能很好地扩展。

Qiao 等[38]为运营商和数据分析师提供了移动大数据框架(Framework for Mobile Big Data，FMBD)，它提供数据收集、存储、处理、分析和管理的功能，以监控和分析海量数据流量。FMBD 已在真实移动大数据上运行了 5 年多，并且可以推广到具有大量数据流量或大数据的其他环境。

2.3　缓存相关

大数据分析在缓存中的应用分为 5G 网主动缓存、优化缓存节点部署、缓存和跟踪热点内容、优化带宽分配 4 个研究点。

在 5G 网主动缓存方面，Bastug 等[39-40]基于大数据分析和机器学习提出了一种主动缓存机制，用于预测 5G 中内容的流行度。在收集原始数据(即用户流量)后，使用 Hadoop 提取有用信息，如位置区域代码(Location Area Code，LAC)、HTTP 请求统一资源标识符(Uniform Resource Identifier，URI)等；然后使用此信息评估之前收集的原始数据内容的流行度。基于从大城市基站收集的移动用户真实数据的实验表明，该机制能有效提高缓存性能并优化用户体验。

在认知无线网络，当次用户(Secondary Users，SU)的活动开始影响到授权用户的 QoS 水平时，SU 必须离开频谱并移动到附近的另一个自由频段。为最小化 SU 从忙信道切换到空闲信道的数据传输延迟，就需要缓存节点的存在。Omar 等[41]提出使用大数据分析技术来处理节点内随时间累积的数据，并利用这些数据来决定集群网络中缓存节点的分布，以提高认知无线电网络缓存节点的数据提取效率。

社交网络(如微信、微博)的用户数量巨大，这些网络中的多媒体通常在具有共同兴趣爱好的群体内分享。重大的新闻事件吸引了大量关注，因此很多内容在这些网络中分享。当某个视频和事件如病毒般传播时，这种分享会加重网络负载，因为请求内容必须从服务器沿着网络路径进行传播。为解决这个问题，Sahni 等[32]建议监控热点和社交媒体网站，分析监控数据，确定是否有对某些内容越来越大的兴趣，并根据某种类别(如年龄)缓存热点数据到某个特定基站。在这种情况下，可以使用大数据分析技术进行相应分析，使得用户更快地缓存内容，从而减少时延并减轻网络负载。

移动互联网中通常包含大量用户，随着基于互联网应用的增加，分配满足用户期望的带宽和高服务质量变得至关重要。蜂窝网络可以为用户随时提供互联网连接，但视频(尤其高质量的)内容仍然缓慢而且相对昂贵。从基站的角度来看，在同一个基站上将相同的视频内容转发给多个用户的影响是巨大的。因此，Fan 等[42]提出了一种动态带宽分配算法，希望下载相同内容的用户共享基站的无线信道。该算法通过分析从用户设备收集的用户和网络数据，将用户划分至不同集群并共享带宽，以提高网络的资源利用率，加快内容分发速度。

2.4　网络优化

在网络大融合时代，各制式通信手段/网络之间的互操作越来越频繁，其协调部署与优化越来越复杂，对网络管理系统处理各项数据的能力要求不断提高；同时，入网设备的急剧增长、业务类型的多样化、数据空-时域分布的不均匀以及高能耗等未来网络特征，也给网络规划与优化带来了极大的挑战。影响网络规划与网络质量的因素非常多，众多因素和网络性

能之间的关系极其复杂。传统的因果关系建模非常困难或不现实，而大数据分析擅长挖掘数据之间潜在的、有价值的相关关系，为网络分析带来了更有利的条件。鉴于大数据分析技术运用于网络规划与优化方面的技术优势，国内外开展了大量研究，主要分为事前优化、事中/后优化和结构优化，如表 2 所列。

表 2 大数据分析在网络优化中的应用

Table 2 Big data analysis in network optimization

优化类型	主要思想	文献
事前优化	基于关联位置数据、服务使用率和上下文数据来预测消耗趋势，从而选择节点的最优位置	[32]
	以 GSM 系统的海量运维数据为依托，使用 BP 网络模型实现对每个小区级无线网络质量趋势的预测	[43]
事中/后优化	提出 NCL 自配置算法来实现全自动化、自优化切换	[44]
	提取 XDR 关键字来定义网络性能指标，优化性能指标不理想的网络设备	[45]
	挖掘性能数据、投诉数据、测试数据等，优化 LTE 网络规划	[46]
	在 5G 网优化框架中展示了异构网资源管理、CDN 部署缓存服务器、基于 QoE 的网络优化 3 个案例	[47]
结构优化	基于 MapReduce 的虚拟网络的拓扑优化机制	[48]
	基于 MapReduce 的资源优化分配机制	[49]
	大数据共享平台的架构设计，整合分散的网络数据，制定统一的共享接口	[50]

事前优化是指通过分析历史数据(如资源消耗、小区网络质量)来预测未来的趋势，从而提前做好资源准备。消耗预测可实现网络的灵活部署，消耗分析涉及用户位置和服务类型两个因素。为达到预测的准确性，Sahni 等[32]假设用户数据(如 GPS 定位和服务使用)可以与其他数据(如新闻、社交网络、事件和天气状况)相关联。通过大数据分析技术来分析这些关联，网络运营商可以在不影响用户满意度的情况下决定何时何地部署节点。王磊等[43]通过 BP(Back-ProPagation)网络模型的自动学习和训练功能实现了对每个小区级无线网络质量趋势的预测，并进一步给出了资源配置建议和需要网络优化重点关注的小区列表，从而实现了网络优化工作从事后处理向预先评估与预警模式的转变。

事中/后优化是指通过分析网络的运行数据(如切换成功率、性能指标)，及时调整策略并反馈到网络配置中。移动通信网中移动的关键点是切换成功率，它确保了用户从一个小区移动到另一个小区时的通话连续性。网络运营商为每个小区手动配置相邻小区列表，但当网络突然变化而需要快速响应时，这些小区无法适应的可能性很高。Lee 等[44]提出了基于大数据分析的切换自优化方法，包括初始化相邻小区列表的自配置和自优化，基于性能测量进一步细化相邻小区列表，以提高切换成功率，从而提升服务质量。Chih-Lin 等[45]介绍了一种基于信令的智能网络优化方案，提取 XDR (call/transaction Detail Records)关键字来定义网络的性能指标，通过分析性能指标不理想的网络设备发现问题，并提供相应的解决方案，为用户提供最佳体验。刘毅等[46]针对以往网络规划与优化数据来源单一、无法辨别价值热点等问题，通过价值评估、干扰评估、覆盖评估等多维度的评估体系，利用大数据挖掘技术处理工参数据、性能数据、经分口数据、MC 口数据、投诉数据、测试数据这六大项数据，并从中提取有用信息进行分析，优化 LTE (Long Term Evolution)网络规划，提升网络建设质量。Zheng 等[47]提出了基于大数据驱动的 5G 网优化框架，通过异构网络资源管理、移动内容分发网络部署缓存服务器、基于 QoE(Quality of Experience)的网络优化 3 个案例验证了所提框架对提升网络性能的作用。

结构优化是指对网络拓扑结构、资源分配方式、接口等进行优化。Xu 等[48]基于 MapReduce 框架提出了一种虚拟网络的拓扑优化机制，并通过仿真实验证明所提机制可以极大地提升网络的整体性能。Kiran 等[49]基于 MapReduce 提出了一种资源优化分配机制，以从网络记录文件、配置文件、数据库条目、监控警报等大量信息中挖掘有用的信息，从而优化每个用户的资源分配，解决现有无线网资源分配存在的诸多问题。许汝鹏等[50]针对现有数据存储分散、利用率低和使用不便等问题，提出了大数据共享平台的架构设计，整合分散的网络数据，制定统一的共享接口，以提高数据利用效率、网络优化效率和质量。

2.5 频谱管控

随着无线通信技术的快速发展，频谱资源变得越来越紧张。一方面，有限的频谱资源难以满足日益增长的频谱使用需求；另一方面，相当数量的频谱资源利用率非常低。根据美国联邦通信委员会提供的数据，已分配的频谱利用率只有 15%～85%。大数据技术可以有效采集频谱监测数据，充分挖掘海量检测数据中隐藏的有用信息，预测频谱需求，评估用频效率，解决频谱短缺等难题。

在采集频谱监测数据方面，Li[51]提出了基于底层的频谱数据库架构，使用 SpectrumMap 设计频谱数据可视化，为 SpectrumMap 开发基于 Hadoop 的 Web 系统，以改善大数据处理效果并提供频谱数据可视化功能，在用户友好的访问界面和参与模型中，鼓励用户加入 SpectrumMap，以便启用基于底层的数据收集。Wu 等[52]引入了频谱设备互联网(Internet of Spectrum Devices，IoSDs)的新概念，并在未来无线网络上为 IoSDs 开发了基于云的架构，目的是建立一个连接各种频谱监测设备(Spectrum Monitoring Devices，SMD)和大规模频谱利用设备(Spectrum Utilization Devices，SUD)的桥接网络，从而在未来无线网络中实现高效的频谱共享和管理。

在分析频谱资源方面，Gvk 等[53]基于 ELK(Elasticsearch Logstash and Kibana)堆栈提出了一种架构，用于分析支持动态频谱接入(Dynamic Spectrum Access，DSA)的 LTE-A (Long Term Evolution Advanced)网络中的大频谱数据；同时，他们通过实验验证了所提架构的性能，包括生成支持 DSA 的 LTE-A 网络的数据集，设置用于 LTE-A 日志数据频谱分析的 ELK 堆栈，以及可视化光谱数据分析样本。为提高用频效率，Zhu 等[54]提出了一种新颖的拍卖博弈方案，用于解决发射功率和无线频谱等无线资源分配的问题。规划无线资源分配问题为一个拍卖过程，每个移动用户与其他虚拟用户相互竞争，投标物理无线网络有限的资源。

在预测频谱需求方面，为应对海量频谱数据给频谱管理

带来的挑战，Li[55]基于贝叶斯网络设计了一种频谱占用预测算法，其通过学习历史频谱数据来预测未来频谱占用的情况。Baltiiski 等[56]在认知 C-RAN 网络中提出了基于大数据分析和机器学习的未来频谱占用预测机制，从长期频谱检测数据中获得频谱的未来占用情况。

2.6　5G 通信网

相比 4G，5G 具有更高的速率和更大的容量，其预期的连接能力至少支持 1000 亿设备和每个用户 10gbps 的高速传输速率。利用大数据促进新兴通信技术的发展，是一个重要的机遇。5G 网中大数据的分析方法包括描述性分析、诊断分析、预测分析和规范性分析等[57]。

描述性分析基于历史和当前网络数据进行分析，可识别过去发生的事件以及测量报告中的各种决定因素和相关性等，常用于网络管理。Imran 等[34]提出了一个 5G 网基于大数据的自组织网络（Big data empowered SON，BSON）框架，其核心思想是开发端到端的网络可见性。Imran 认为 BSON 与自组织网络（Self-Organizing Network，SON）的区别是：对当前网络状态有完整的情报；具有预测用户行为的能力；具有连接网络响应和网络参数的能力。

BSON 框架如图 5 所示。1）从网络中的所有信息源（如用户、小区）收集数据，构成聚合数据集。2）数据转换，即将大数据转换为正确的数据。首先，根据关键运营和业务目标（Operational and Business Objectives，OBO）对数据分类；其次，通过统一多个绩效指标（Performance Indicators，PIs）得到更显著的 KPI，并根据每个 OBO 的 KPI 影响对 KPI 排序；然后，过滤影响 OBO 小于预定义阈值的 KPI；接着，关联每个 KPI，找到影响 KPI 的网络参数（Network Parameter，NP）；最后，根据每个 KPI 的关联强度排序相关 NP。3）建模，即从步骤 2）获得的正确数据中学习，建立网络行为模型。4）运行 SON 引擎，确定新的 NPs 和 KPIs。5）验证。若一种新的 NP 能通过专家知识或运维人员的经验进行评估，则继续更改；否则为新 NPs 进行新的网络模拟行为，如果模拟行为符合 KPIs，则继续新的 NPs。6）重新学习/改进。若步骤 5）的验证不成功，则反馈到概念漂移块，更新行为模型。为保持模型的准确性，即使验证步骤中有积极结果，也要周期性地触发概念漂移。

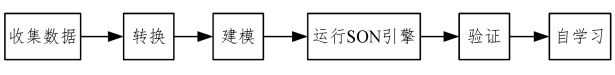

图 5　BSON 框架

Fig. 5　BSON framework

诊断分析是基于历史数据进行分析，提供有关过去某些结果的根本原因见解。网络运营商可根据诊断分析的结果，对网络运维采取更好的措施，从而避免错误和过去的负面结果，常用于网络优化和故障诊断。Chiu 等[58]通过构建干扰和干扰小区站点的测量建模，应用基于回归模型的机器学习算法识别干扰对并提供相应的优化决策，同时根据优化建议对一些被列为高风险的小区站点进行天线倾斜变化和参考信号传输功率的修改。利用大数据平台上的数据分析方法，能够处理大量的小区站点数据，寻找匹配上述干扰模式的小区对，干扰源和被干扰源小区站点都可以应用不同的优化方案，从而实现网络的整体改进。

预测分析是关于预测和提供未来结果概率的估计，定义未来的机会或风险。预测分析的典型应用是预测用户行为。Yan 等[59]构建了一个将基于大数据分析的预测与主动推送和缓存技术相结合的“人在环”系统。首先将用户需求分为个人兴趣驱动需求（Demands driven by Personalized Interests，DPI）和社交网络驱动需求（Demands driven by Social Networking，DSN）；然后基于大数据分析预测用户需求；接着提出“人在环”系统的基本框架，根据预测结果确定推送和缓存策略，并使用缓存命中率（Cache Hit Ratio，CHR）测试系统的性能。仿真结果表明，基于数据分析预测的“人在环”系统的性能远远优于没有数据分析和需求预测的情况。“人在环”系统中的关键核心是使用了基于大数据分析的预测模块。一方面，大数据技术能应对用户数据量巨大的情况。以微信为例，用户数接近 $9 * 10^8$，每天都有大量用户在“时刻”转发和请求数以百万计的内容，即使只输入用户一天内的踪迹到预测模块，请求总数也可能达到数十亿。另一方面，应用大数据分析技术可以减少预测延迟，预测结果越准确，下次推送越能获得更高的缓存命中率。大数据分析的基本过程是：1）网络划分，即将用户群集分为多个组。现实的社交网络倾向于具有清晰的集群结构，并且用户受同一组中用户的影响更大，因此首先根据集群结果将网络划分为小的网络。2）内容分割，即根据主题将内容集合划分为若干子集。内容相关图中同一主题上两个文件之间边缘的权重通常较大，因此基于该主题的划分将保留原始图的更多信息。模块计数仅分析每个子社交网络中最常访问的几个主题。3）计算。将具有少量主题的每个小型网络中的 DPI 和 DSN 预测任务分配给各种计算服务器。4）归约。从分布式服务器收集结果，然后提供一个包含用户预测结果的基站。

规范性分析可以预测未来行动的影响，并回答“可能发生的事情”作为组织行动的结果，以帮助用户采取最佳决策。规范性分析的典型应用是网络资源分配。Raza 等[60]介绍了一种基于 BDA（Big Data Analytics）的切片准入策略，该策略可以确定资源分配的最佳解决方案，显著提高基础设施提供商（Infrastructure Providers，InPs）的利润。软件定义网络和网络功能虚拟化（Network Function Virtualization，NFV）实现了传输、无线电和云资源的联合协调，在 5G 中发挥着核心作用。这使得 InPs 能够在其物理基础设施（即网络切片）的基础上创建虚拟网络，从而在具有各种需求的不同租户之间共享资源。InPs 可以依靠大数据分析来预测切片资源需求的变化。基于 BDA 的切片许可策略是：对于每个输入切片请求 s，协调器检索 s 和当前正在运行的每个切片（即在集合 S 中）的预测资源需求；然后，协调器检查可用基础结构资源的状态，如果它意识到接受 s 将导致 s 或 S 中的任一切片退化，则协调器拒绝 s，否则接受 s 并分配当前所需的资源。实验结果表明，BDA 可以增加 49% 以上的利润。

此外，Zhang 等[61]概述了大数据和 5G 无线网络的协同

和互补特性。前者利用 5G 的通信、缓存和计算等异构资源来支持大数据应用和服务。后者利用大数据技术收集无线大数据挖掘出数据中的知识或洞察力,以改善网络的规划和运营。Chih-Lin 等[45]研究了 5G 中大数据分析技术的应用,阐述了大数据分析如何显著地促进本地内容供应、动态网络和功能部署、用户行为感知、微调网络运营以及异构网络节能。

2.7 小结

本节回顾了大数据分析技术在无线网络中的研究,如表 3 所列。总体上,已有部分研究人员利用大数据的优势特点对无线网络进行了大量尝试与研究,但还存在以下问题及未来可能的研究方向:1)在故障检测层面,需进一步提高网络故障的检测精度,并增强网络的自愈能力;2)在网络流量监控层面,提高实时监测效率以及流量监控的可视化有待进一步深入研究;3)在缓存相关层面,需要收集大量用户流量及其感兴趣的内容,因此在未来的研究中要加强用户的隐私保护;4)在网络优化层面,影响网络规划与网络质量的因素成百上千,需进一步挖掘影响网络质量的关键因素,构建性能更高的基于大数据分析技术的网络规划工具;5)在频谱资源管理层面,当前的研究主要通过历史频谱数据预测未来频谱的占用情况,还应兼顾不同终端所支持的频段信息、不同用户的未来业务需求、不同用户的用频偏好等对频谱空洞预测的影响;6)在 5G 网层面,使用大数据分析技术可以提升用户体验,但还需进一步研究基于大数据分析技术的大规模数据交换和海量设备连接。

表 3 大数据分析技术在无线网络中的应用

Table 3 Applications of big data analytics in wireless network

研究分类	问题描述	研究内容	相关技术	参考文献
故障检测	覆盖问题	通过收集和分析无线网络的网络性能测量数据,优化小区覆盖性能	Hadoop	[27-28]
	用户异常及原因分析	收集 CDR 和 KQI 等数据,使用聚类等技术检测用户异常,并提取异常背后的原因	Hadoop、k-means 聚类、分层聚类、Spark	[29-31]
	设备故障	通过在小区级上分析带宽趋势来预测设备和基础设施故障	MapReduce,HBase	[32]
	休眠小区	蜂窝网络中大数据分析技术进行自愈的示例,分别是数据归约、休眠小区检测和基于 KPI 的关联检测	MapReduce、知识库系统 KBSs	[33]
流量监控	流量监控与分析	对网络流量进行监控与分析,监视用户的网络活动,识别异常、恶意行为和错误配置的服务	Hadoop,Spark	[36-38]
缓存相关	主动缓存	收集用户流量后,提取 URI 等有用信息,预测数据内容的流行度,从而主动缓存	Hadoop	[39-40]
	优化缓存节点部署	利用大数据分析技术来决定集群网络中的缓存节点分布,提高数据提取的效率	MapReduce	[41]
	缓存和跟踪热点内容	监控热点和社交媒体网站,分析数据,确定是否有对某些内容越来越大的兴趣,并根据某种类别缓存热点数据到某个特定基站	MapReduce,HBase	[32]
	优化带宽分配	分析从用户设备收集的数据,将用户划分至不同集群并共享带宽,提高网络资源利用率	D2D、基于数据驱动的集群算法、带宽分配算法	[42]
网络优化	事前优化	通过分析历史数据(如资源消耗、小区网络质量)预测未来趋势,提前做好资源准备	多层感知器预测模型	[32,43]
	事中/后优化	分析网络运行数据(如切换成功率、性能指标),及时调整策略并反馈到网络配置中	用户行为聚合模型、聚类、神经网络模型	[44-47]
	结构优化	对网络拓扑、资源分配方式、接口等进行优化	MapReduce,OpenStack Neutron	[48-49]
频谱管控	采集频谱监测数据	利用大数据技术采集频谱监测数据,充分挖掘海量检测数据中隐藏的有用信息	Hadoop	[51-52]
	分析频谱资源	分析频谱资源,提高用频效率	博弈论	[53-54]
	预测频谱需求	通过大数据分析或机器学习技术学习历史频谱数据,以此预测未来频谱的占用情况	贝叶斯网络	[55-56]
5G 通信网	描述性分析	基于大数据的自组织网络 BSON 框架,其核心思想是开发端到端的网络可见性	高斯过程回归、最佳线性预测	[34]
	诊断分析	构建干扰和干扰小区站点的测量建模,应用基于回归模型的机器学习算法识别干扰对并提供相应的优化决策	回归模型的机器学习算法	[58]
	预测分析	一个将基于大数据分析的预测与主动推送和缓存技术相结合的"人在环"系统	线性阈值模型 LTM	[59]
	规范性分析	基于 BDA 的切片准入策略,可以确定资源分配的最佳解决方案	SDN、NFV、网络切片	[60]

3 大数据分析技术在光纤网络中的应用

光纤是网络的重要资源,它的合理分配是网络流量需求得到满足的基础和前提。近年来,使用大数据分析技术解决光纤网络优化和流量预测等问题,均取得了一定的研究成果。

3.1 网络优化

3.1.1 解决路由与波长分配问题

路由与波长分配算法(Routing and Wavelength Assign-

ment，RWA)[62] 在光纤网络中发挥着重要的作用。当前，RWA 算法的解决方案主要有基于线性规划模型、基于分层图模型、基于快速有效启发式算法等[63]，应用大数据技术来解决 RWA 的研究较少。本文重点关注基于大数据平台来解决 RWA 问题。Shen 等[64] 开发了一个由 10 台低端桌面计算机组成的 Hadoop 云计算系统，每台计算机上独立运行 RWA 算法来处理一定数量的需求序列，以便在短时间内评估足够数量的需求序列。他们比较了所有评估的需求序列的结果，并选择最佳的一个作为 RWA 问题的最终解决方案，详细流程如图 6 所示。

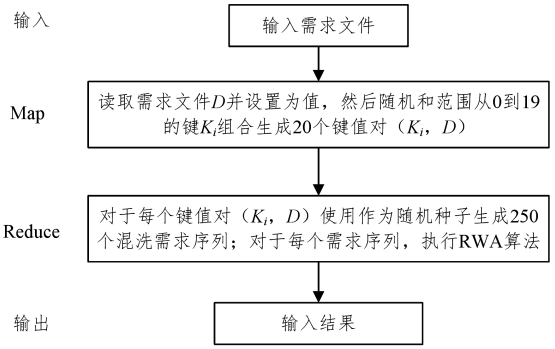

图 6　并行执行 RWA 算法的详细流程

Fig. 6　Map/Reduce process for RWA algorithm

首先，将包含光路需求请求的文件输入到 HDFS。其次，Map 函数以〈key，value〉键值对的形式作为输入，map 函数读取需求文件 D 并将其设置为 value，与不同的键 K_i 组合，其值范围为 0～19。这些键在稍后的 Reduce 函数中被用作随机种子。Map 函数总共生成 20 个键值对（K_i，D）。然后，将键值对转发到 20 个 Reduce 函数以进行并行计算。对每个键值对（K_i，D），使用 K_i 作为随机种子对光路需求列表 D 混洗 250 次。针对每个混洗的需求序列，执行 RWA 算法，以找到所需波长的数量。通过比较 250 个混洗需求序列的结果，在每个 Reduce 函数中找到最佳的一个（K_i，$D_{i,j}$）。对于整个 Hadoop 系统，比较 20 个 Reduce 函数的结果，以找到全局最优。因此，在整个计算过程中总共有 5 000（250×20）个混洗需求序列，以获得最终的最佳结果。

仿真研究表明，评估多个混洗需求序列的方法可以达到与最优值相同或非常接近的性能。对于具有 500 个节点、1 000条链路和 4 000 个请求以及 5 000 个并行混洗光路需求序列的大型网络，实验记录证明了 Hadoop 系统能够在 3 h 内为所有序列运行相同的 RWA 算法，比普通桌面计算机快 30 倍。

3.1.2　使用 Hadoop 解决多步优化问题

Li 等[65] 试图解决光纤网络中的 3 类优化问题：1）能耗最小化问题，目标是最小化网络的整体功耗；2）共享备份通路保护（Shared Backup Path Protection，SBPP）网络规划问题，目标是最小化网络中频率槽（Frequency Slots，FSs）的最大数量；3）自适应前向纠错分配问题，目标是最大化用于用户数据传输的 FSs 总数。上述问题都属于装箱问题。在满足网络流量需求时，应考虑需求大小和路由等问题。由于计算复杂度高、服务需求的顺序不同，试图解决这些问题的启发式算法的性能无法得到保证。通过评估多个混洗需求序列并选择具有最佳性能的序列，可以有效地解决光纤网络中的装箱优化问题。然而，这面临着计算复杂度高的困难。

为缩短计算时间，降低计算复杂度，Li 等构建了一个 Hadoop 系统来并行评估多个混洗需求序列。Hadoop MapReduce 平台可以并行评估多个混洗需求序列。启发式算法为每个混洗的需求序列提供服务，并且每次都产生结果，然后通过比较选择具有最佳性能的结果。在每个 Reduce 函数上重复相同的过程。最后，通过比较所有 Reduce 函数的结果，找到最终的全局最优值。经验证，该方法可有效地找到近似最优的解决方案，与单台机器相比显著缩短了计算时间。

3.2　流量预测

新服务的出现对网络提出了新要求，需要大而动态的比特率，这促使网络运营商寻求能够以动态方式应对预期流量的虚拟网络拓扑（Virtual Network Topology，VNT）架构。一种解决方案是通过网络的超供应来应对预期流量，其缺点是总成本增加。另一种解决方案是使用基于阈值的容量重配置来节省功耗[66]。与超供应方案相比，第二种解决方案的缺点是没有减少每个 IP 路由器需要安装的光转发器数量。

Morales 等[67] 提出基于大数据分析的 VNT 重构方案，通过定期分析原始目的地（Origin Destination，OD）通信流，执行相应的 VNT 重构，整个流程如图 7 所示。

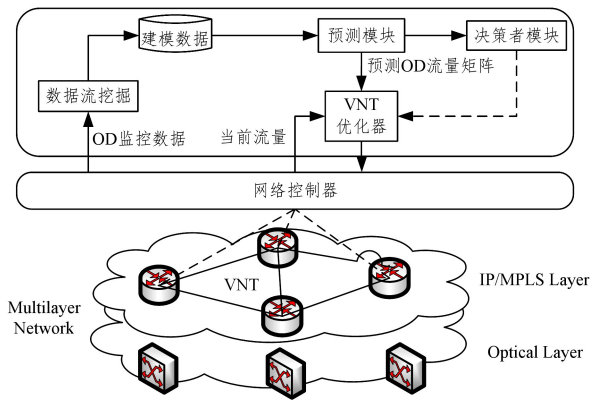

图 7　基于数据分析的 VNT 重配置

Fig. 7　VNT reconfiguration based on big data analysis

基于大数据分析的 VNT 重构方案简单概括为：在边际 IP 路由器上定期收集流量监控数据。每个边际路由器为每个其他目的路由器收集一组流量样本，这些集合存储在收集的数据库中。根据预定义的时间段，定期检索收集到的 OD 监控数据并执行数据流挖掘过程，为每个 OD 对汇总收集到的数据。该阶段的结果是建模数据库，包括每个 OD 对的最大、平均和最小比特率。预测模块利用机器学习技术（即人工神经网络），为即将到来的时段生成预测 OD 流量矩阵。决策者模块根据上述矩阵来决定是否进行 VNT 重构。如需重新配置，VNT 优化器将提供当前和预测的 OD 流量矩阵给决策者模块。解决方案返回给网络控制器，以在数据平面中实现所需的更改。

与静态和基于阈值的方法相比，基于大数据分析的 VNT

重构方案将安装的光转发器总数节省了 8%～42%；此外，该方案能够在低流量时通过停用转发器做出反应，从而实现节能；同时通过从光学层释放光路来降低成本。

3.3 小结

本节从网络优化、流量预测两个方面详细介绍了大数据分析在光纤网络中的研究工作。表 4 总结了该方向的主要研究内容、相关平台和相关文献。

表 4 大数据分析在光纤网络中的应用

Table 4 Applications of big data analytics in optical network

研究分类	研究内容	相关技术	文献
网络优化	基于大数据技术解决 RWA 问题，在 Hadoop 集中的每台计算机上独立运行 RWA 算法来处理需求序列，比较所有评估的需求序列的结果，并选择最佳的作为 RWA 问题的最终解决方案	Hadoop	[64]
	使用 Hadoop 来并行评估多个混洗需求序列，解决光纤网络中的 3 类优化问题	Hadoop	[65]
流量预测	基于大数据分析的 VNT 重构方案，通过定期分析原始目的地 OD 通信流，执行相应的 VNT 重构	ANN	[67]

光纤网络的优化主要是利用 Hadoop 的并行处理优点来减少不同优化算法的执行时间。利用大数据分析技术来优化光纤网络时，须进一步考虑多用户接入的情况，根据不同用户的需求进行资源的配置。流量预测是使用大数据分析技术对目的地通信流进行预测，从而动态重新配置网络。利用单一或少量的要素进行通信流的预测时，外部环境变量会对预测效果产生较大影响。如某个僵尸主机不断发起通信请求，造成系统误判而频繁进行网络重构，这会严重影响网络的正常通信。因此，多流量要素结合的流量预测也是未来探索的方向之一。

4 大数据分析技术在 SDN 网络中的应用

SDN 网络提供了使用集中控制器编程网络的能力，该控制器能够使用一个标准化的开放接口编程多个数据平面，从而提供灵活的体系结构支持[68]。本节主要介绍大数据分析技术在 SDN 网络中的相关研究成果，包括流量预测、流量约简、跨层设计 3 个方面。

4.1 流量预测

4.1.1 SDN 网络中的资源感知路由

目前，Internet 上的应用程序不断增多，特别是实时或接近实时的应用程序，它们具有非常大的体积和非常高的计算复杂性，不能缺少网络底层的支持。最近，SDN 作为一种新的网络样式，引起了人们极大的兴趣。SDN 的主要思想是从转发面分离出控制面，以打破垂直统一管理模式并提出网络编程能力[69]。

Cui 等[70]设计并实现了 SDN 中基于大数据分析的资源感知路由架构，包括以下几个组成部分。1）用户偏好分析。使用 Hadoop 实现预测功能，通过分析网络流量和用户应用程序信息来发现每个应用流的分布规律。对于每个数据流，发现其特定的分布规律。通过分析这一规律，针对不同的应用和领域，开发出一个初步的通用预测模型，以适应同一应用但不同领域的情况。2）SDN 控制器与数据库接口的设计。云平台负责计算和预测每个 OpenFlow 交换机的流量分布值。此外，该平台将读取链接信息并进行流量预测。数据库保存记录值，并更新最后的预测值。为确保资源分配能适应流量变化，使用 Floodlight 定期从数据库中读取最新的预测值。Floodlight 是一个基于 Java 的 SDN 控制器，可以通过加载不同的模块来适应不同的应用程序。3）基于 SDN 控制器的路由。将预测值作为选择最佳路线的首选项，并对 Dijkstra 算法进行相应的改进。通过该架构将应用感知和用户偏好预测集成到 SDN 中，以促进和增强网络资源分配，并提供更好的应用分类。大数据的作用体现在使用网络流量分析和用户行为来预测传入流量的类型和速率。

SDN 中基于大数据分析的资源感知路由的具体流程如图 8 所示。首先，云平台读取当前网络负载（流量大小和类型），使用基于大数据驱动的预测算法来预测总体流量并将其保存到数据库中；然后，SDN 控制器访问数据库，读取预测流量，根据改进的 Dijkstra 路由算法创建资源分配方案并将其发送到相关交换机；最后，SDN 交换机根据此方案转发数据。

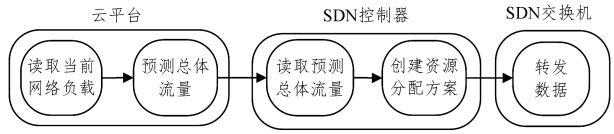

图 8 资源感知路由的流程图

Fig. 8 Flowchart of cognitive routing resources

大数据分析技术可以基于用户的需求为网络提供动态的资源分配和应用分类，从而为网络提供更好的负载均衡。通过对相关交换机动态发布流量表，验证了该方案对流量变化的自适应能力，提高了资源利用率，实现了整体的负载均衡。

4.1.2 预测数据中心网络的数据通信量

处理大数据应用程序的网络可能会受到数据大小和速度的影响。例如，网络的总体响应时间会受到 MapReduce 的重量级通信阶段的影响，如果通信模式受到严重的倾斜影响，这个问题就会加剧。Neves 等[71]提出了 Pythia 系统，其利用 Hadoop 实时通信预测来优化运行中的数据中心网络；将端到端的流映射到底层网络；利用 SDN 提供的可编程性来实现数据传输的高效和及时的网络资源分配。实验证明，该系统能有效减少任务完成的时间。

4.2 流量约简

不管是 Google 和 Facebook 这样的大型网络，还是中小型企业网络，都面临着流量过多的问题。流量过多的问题主要是由于在批处理或实时应用程序中处理大量数据造成的。常见的解决方案是增加企业中的可用带宽。但 Costa 提出了另一种提高网络性能的方法[72]，将数据聚合从边缘推向网络，从而减少流量。该方法利用 CamCube[73]平台的属性实现了高性能，提供了数据流聚合的完整路径。通过运行 CamCube 平台上类似 MapReduce 的服务 Camdoop，构建了聚合树，其根位于服务器上，子项与中间数据源相似，从而最终减少流量。在 CamCube 上对 Camdoop 小型原型进行了测试，结果显示，Camdoop 能有效减少流量，比运行在交换机上的

Camdoop，Hadoop 和 Dryad/DryadLINQ[74-75]有更高的性能。

4.3　跨层设计

传统网络被划分成不同的层，一组协议用于相邻层之间的通信，不相邻的层之间不能直接通信。然而，跨层设计的最新研究表明，非相邻层可以在运行时共享信息，并显著改善网络性能。虽然不同层之间的信息共享可以改善性能，但模块化的原则已经被打破，网络变得复杂，传统的网络设计和优化方法不再适用。大数据可以使 SDN 的跨层设计受益，SDN 的逻辑集中控制器具有全局性的网络视图，能够从任意粒度的所有层次获取大数据，如物理层的信道状态信息、数据链路/网络层的分组信息和应用层的应用信息。将大数据分析应用于网络控制和管理，可以极大地改善网络的控制和管理过程。

Cui 等[76]提出了一种将大数据和 SDN 相结合的体系结构，该体系结构可以在大数据的帮助下，方便 SDN 的跨层设计。该体系结构包括基础结构层、数据处理和控制层以及应用层，如图 9 所示。

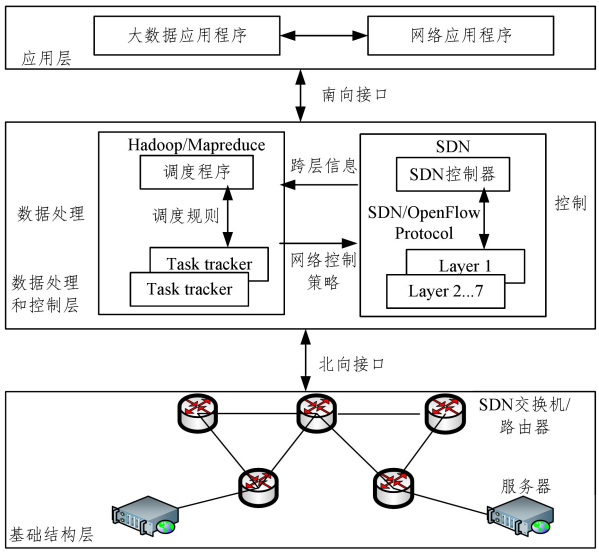

图 9　SDN 与大数据的跨层设计

Fig. 9　Cross-layer design with SDN and big data

基础结构层由交换机/路由器、服务器和数据中心设备组成。交换机或路由设备根据存储在本地存储器中的转发规则将数据包转发到下一跳。数据中心的服务器存储大数据并运行任务。在数据处理和控制层，SDN 控制器和 Hadoop 将密切配合大数据的处理和决策。SDN 控制器向 Hadoop 提供来自所有层的跨层信息，Hadoop 为 SDN 控制器提供网络控制策略（物理层参数自适应、资源分配、拓扑结构、路由机制、拥塞控制等）来操作和优化 SDN 的性能。大数据应用程序和网络应用程序都运行在数据处理和控制层的顶部。此外，Cui[76]还指出了大数据与 SDN 存在密切关系，SDN 能为大数据提供基础，而大数据能在流量工程、跨层设计、克服网络攻击、数据中心网络等方面提升 SDN 网络的性能。

4.4　小结

本节主要介绍了大数据分析技术在 SDN 网络中的研究现状，如表 5 所列。

表 5　大数据分析技术在 SDN 网络中的应用

Table 5　Applications of big data analytics in SDN network

研究分类	研究内容	相关平台	文献
流量预测	使用 Hadoop 预测用户需求，为网络提供动态的资源分配和应用分类，从而为网络提供更好的负载均衡	Hadoop	[70-71]
流量约简	利用 CamCube 平台实现了高性能，提供了数据流聚合的完整路径，将数据聚合从边缘推向网络，从而减少流量	CamCube 平台、Camdoop	[73]
跨层设计	将大数据分析技术应用于 SDN 网络控制和管理，可以极大地改善网络的控制和管理过程	Hadoop	[76]

在流量预测层面，大数据分析可以从大流量数据中获取洞察力，进而指导 SDN 控制器制定流量工程策略。根据这些流量策略，改变 SDN 设备的切换行为，打开/关闭设备和链路，以最小化功耗和链路拥塞现象，但快速、频繁的表更新请求以及大量的数据传输和处理会导致其性能下降。因此，需进一步设计能容纳更大流量表和大数据的网络控制器。在流量约简层面，除了运用聚合树的方法压缩流量，还可以考虑利用聚类、分类等方法对流量进行聚合，以减少数据量。在跨层设计层面，大数据分析为 SDN 控制提供了网络控制策略，交换机根据这些策略来进行转发，但由于这些网络交换机没有任何智能，因此只是向网络控制器发送原始数据包。这样在交换机中不做任何预处理，会导致控制器的负载过重。因此，需要为 SDN 转发平面赋予智能。

5　大数据分析技术在网络安全中的应用

网络安全是网络的重要支柱，为互联网的可靠运行提供了基础，各项网络安全检测措施为互联网各项活动的开展提供了安全的通信保障。本节主要介绍大数据分析技术在网络安全方面的相关研究成果，包括 APT 攻击检测、网络安全检测、网络安全态势感知和设备指纹 4 个方面。

5.1　APT 攻击检测

随着全球网络信息化的高速发展，具有隐蔽性、渗透性和针对性的 APT 攻击日益增多，使国家、企业的网络信息系统和数据安全面临着严峻挑战。例如，2016 年，黑暗力量不仅入侵了乌克兰的电力系统，还攻击了其矿业和铁路系统。APT 攻击检测的一大挑战是在检测异常时须筛选大量数据，这使得检测任务看起来像大海捞针一般。由于网络数据体量巨大、来源多样、增长速度快，传统的网络外围防御系统在检测目标攻击方面可能变得无效。大数据分析技术是一种适合 APT 检测的方法，可以实现海量网络安全数据的深度关联分析，在检测 APT 攻击方面具有明显优势。

2012 年，Giura 等[77]提出了一种可成功检测 APT 攻击的方法。该方法基于攻击树的概念，建立了一个概念攻击模型——攻击金字塔。攻击金字塔在顶层包含可能的攻击目标（如敏感数据、数据服务器），采用横向平面表示与攻击相关联的事件环境（如用户平面、网络平面）。检测框架先将所有记录在组织中的可能与安全相关的事件分组为多个场景，然后在每个场景及跨场景中使用 MapReduce 进行并行处理，再应用不同的算法检测可能的恶意活动。

美国 RSA 实验室提出的 Beehive 系统[78]利用大数据分析技术在短时间内处理大量的日志信息，检测组织机构中的资源使用模式，发现以往会被忽视的策略违背与恶意软件感染，并根据 APT 攻击多个阶段的行为与正常通信存在的细小行为差异，关联检测到的看似孤立的事件，发现攻击者 APT 入侵的证据。

Marchetti 等[79]提出了一种能有效分析大量网络流量的方法，以揭示与数据泄露和其他可疑 APT 活动相关的微弱信号。最终是对最可疑的内部主机进行排名，这种排名允许安全专家将他们的分析集中在数千台典型的大型组织机器中的一小部分主机上。实验评估表明了该方法的可行性和有效性。

张小松等[80]提出了一种基于树型结构的 APT 攻击预测方法，如图 10 所示。

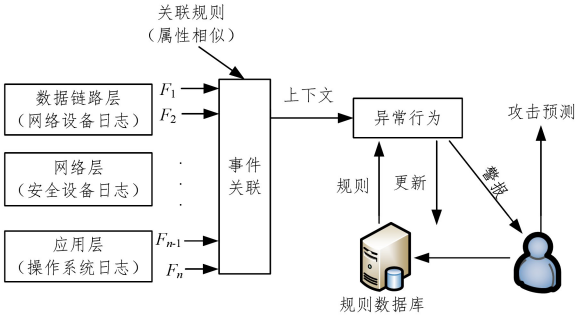

图 10 APT 预测框架

Fig. 10 APT framework

该方法使用大数据分析技术对海量的网络设备日志、安全设备日志和操作系统日志进行关联分析，形成攻击上下文，然后通过引入可信度和 DS 证据组合规则确定攻击事件，计算所有可能的攻击路径。实验表明，该预测模型能有效地对攻击目标进行预警，具有较好的扩展性和实用性。

5.2 网络安全检测

网络安全检测一直是网络安全领域内最为活跃的研究分支之一，主要指对网络的安全状态或者面临的风险进行分析和检测，对接入网络的不同行为进行分析和控制，以发现潜在的威胁或正在进行的攻击。鉴于篇幅，本节只对每种类型介绍两项工作。

5.2.1 DDoS 检测

DoS(Denial of Service)是通过强制被绑架的计算机启动或消耗其资源(如 CPU、带宽)来实现的。分布式拒绝服务(Distributed Denial of Service，DDoS)是指 DoS 攻击由各种分布式计算机生成。随着网络流量的不断增长，以往的检测方法无法从如此大规模的网络流量中检测到网络攻击行为，而大数据技术有利于 DDoS 检测技术和水平的不断提高[81-82]。

Hameed 等[83]提出的基于 Hadoop 的实时 DDoS 检测框架 HADEC，可以在经济实惠的时间内分析 DDoS 攻击。HADEC 捕获实时网络流量，处理它们后形成简短的日志相关信息，并使用 MapReduce 和 HDFS 运行 DDoS 泛洪攻击检测算法。评估结果显示，对于约 15.83GB 的实时网络流量，HADEC 从捕获到检测生成 1GB 的日志文件只需不到 5min 的时间，从而表明了 HADEC 能够近乎实时地处理和检测 DDoS 攻击。

Hoon 等[84]研究了 DDoS 攻击检测中应用大数据分析技术的机器学习模型，其使用数据挖掘工具 WEKA 和 H2O 实现监督学习和非监督学习模型，在 NSL-KDD 入侵数据集上对模型进行训练和测试，并对模型的效率和准确性进行评估。总体而言，监督学习算法比非监督学习算法表现更好。综合考虑准确性和训练时间，NaïveBayes，Gradient Boosting Machine 和 Distributed Random Forest 是最适合 DDoS 检测的模型。

5.2.2 入侵检测

网络入侵检测是根据网络流量或主机数据来判断系统的正常行为或异常行为。大数据分析技术可以将多个信息源关联成一个连贯的视图，识别异常和可疑活动，最终实现高效的入侵检测。

Mylavarapu 等[85]利用 Apache Storm 开发了一种实时混合入侵检测系统，其中 Apache Storm 充当分布式、容错、实时的大数据流处理器。该系统由两个神经网络组成：CC4 瞬时神经网络作为未知攻击的异常检测，多层感知器神经网络作为已知攻击的误用检测。基于这两个神经网络的输出，输入的数据将被分类为“攻击”或“正常”。该混合检测系统的平均准确率为 89%，误报率为 4.32%。此模型适用于实时检测。

Rathore 等[86]提出了一种使用 Hadoop 实现的超高速大数据环境的实时入侵检测系统，如图 11 所示。该系统由捕获层、过滤和负载平衡层、Hadoop 处理层和决策层组成，使用了 5 种主要的机器学习方法进行评估，包括 J48、REPTree、随机森林树、联合规则、SVM 和 NaïveBayes 分类器。实验结果表明，在所有的分类器中，REPTree 和 J48 在准确性和效率方面表现最佳。

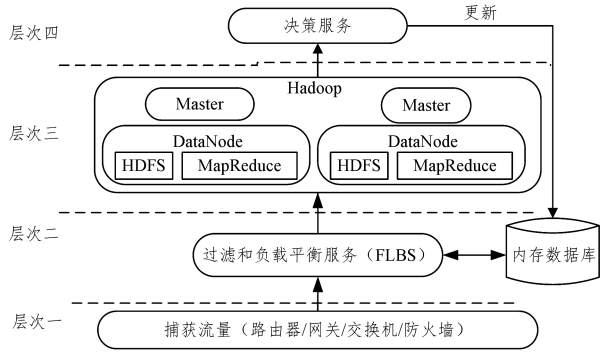

图 11 基于 Hadoop 的实时入侵检测系统

Fig. 11 Hadoop-based IDS architecture

Buczak[20]从模型的构建及部署角度介绍了入侵检测和机器学习结合的研究工作。Wang 等[87]对用于网络入侵检测的大数据分析技术进行了概述。

5.2.3 僵尸网络检测

互联网上的很多安全问题(如垃圾邮件、网络钓鱼)都是由僵尸网络引起的。僵尸网络是指由攻击者控制的受恶意软件感染的机器组成的网络[88]。僵尸网络攻击能够在一次攻击中利用 90000 个 IP[89]，这在国际上是一项安全性挑战，特别是它们可能造成的财务损失。

为检测和消除此类攻击，安全研究人员和网络分析师认为数据包捕获和网络跟踪是最受欢迎的手段之一。但是，分析这些大型数据集对于当今的计算机来说并不容易。为克服

这一挑战，Singh 等[90]提出了一个基于大数据分析框架的 P2P(Peer-to-Peer)僵尸网络检测系统。该系统包含流量嗅探、特征抽取模块和机器学习模块 3 个组件，如图 12 所示。

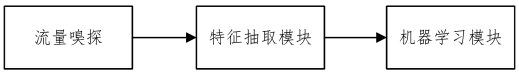

图 12　僵尸网络检测的组件

Fig. 12　Components of botnet detection

首先，利用 Dumpcap[91]进行抓包，用 Tshark 提取域后提交到 HDFS 进行存储；然后，使用 Apache Hive[92]实现动态网络特征的提取；最后，利用 Mahout 的并行处理能力建立基于随机森林的决策树模型，并将其应用于准实时 P2P 僵尸网络的检测中。

Terzi 等[93]提出了一种基于 Spark 的无监督网络异常检测方法，其使用 NetFlow 协议收集可以用于检测网络异常的流量信息，并使用主成分分析方法(Principal Component Analysis，PCA)来减少维度，结果以 3D 可视化的方法呈现。该方法可以很容易地检测到可疑或恶意流量、异常、可疑设备和越权资源访问。

5.2.4　网络异常检测

网络异常检测是探寻表征目标对象属性、状态与变化的特征，然后构建检测模型，对违背策略或偏离正常行为模式的行为进行判定。

Yao 等[94]提出了一种基于大数据分析技术的网络流量异常检测算法，其可以避免网络流量分布调整带来的影响，降低漏报率，提高检测精度。Bachupally 等[95]介绍了一种利用大数据技术分析网络流量的方法。该方法在 Hadoop 分布式文件系统 HDFS 环境中，通过处理和加载流量数据到 Hive 数据库，检测正在进行的异常活动和通过网络传输的恶意数据，并使用 Hive 查询分析数据，以可视化的方式展示了用该方法检测样本数据集攻击的结果。

5.3　网络安全态势感知

面对不断增加的多层面网络安全威胁和安全风险，企业和组织需要及时发现网络中的异常事件，实时掌握网络安全状态，由过去的“亡羊补牢”转向事前自动评估，以降低网络安全风险，提高网络安全防护能力。网络态势感知能够对引起网络态势变化的安全要素进行获取、理解和显示，从而对未来的安全趋势进行预测。大数据分析技术可以将看似毫无联系、混乱无序的各类安全数据转化成直观的可视化信息，为大规模网络安全态势感知带来了更好的契机。网络态势感知已经成为当前网络安全领域的研究热点。

Puri 等[96]介绍了一种新的实时操作和态势感知实现，将大数据架构、图分析、流分析和交互式可视化技术应用到安全用例中。管磊等[97]提出了一种集安全数据采集、处理、分析和安全风险发现、监测、报警、预判于一体的安全态势感知平台。该平台整合安全区域内的用户终端、网络链路、应用系统、数据流量等各类感知数据源，经统一汇聚存储后，利用机器分析，结合数据处理、安全规则模型、攻击推理模型等分析算法，将安全日志、报警数据转化成直观的可视化安全事件信息，并从海量数据中挖掘威胁情报，从而实现风险发现、安全预警和态势感知，进而提升安全监测的攻击发现和安全态势感知的能力。赵梦等[98]借助大数据的处理和分析能力对成千上万的网络事件等信息进行自动分析与深度挖掘，基于大数据的分析和挖掘结果可以对网络的安全状态进行分析评估，感知网络的异常事件与整体安全态势，并对未来的安全态势进行预测。

通过对已有的网络安全态势感知方案进行分析，总结出基于大数据分析技术的网络安全态势感知架构，如图 13 所示。

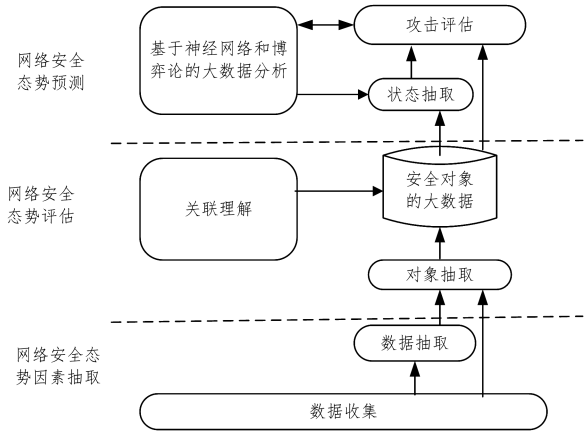

图 13　网络安全态势感知架构

Fig. 13　Security situational awareness mechanism

从图 13 中可以看出，网络安全态势感知架构分为网络安全态势因素提取、网络态势评估和网络态势预测 3 个部分。首先，从网络中提取安全状态因素，主要包括静态配置信息、动态运行信息、流量信息等。静态配置信息包含了网络拓扑、漏洞信息、状态信息等基本的环境配置。动态运行信息包括从日志中获取的威胁信息和各种保护措施的分析技术等基本运行信息。然后，评估网络态势，通过融合大量的网络安全数据信息，分析其相关性，得到网络的总体安全情况。最后，进行网络态势预测，使用数据挖掘算法和机器学习算法(如神经网络、博弈论)，对情报和事件有效地汇聚、处理与分析，从而实现安全态势预测和漏洞预警发布。

5.4　设备指纹

Xu 等[99]对无线网络中的设备指纹方法进行了研究。设备指纹是指收集设备信息以生成设备特定签名的过程，通过分析协议栈上的信息来完成，用于对抗无线网容易受到内部攻击和节点伪造的漏洞。设备指纹的采集主要分为 3 个步骤，如图 14 所示。首先，识别协议栈中所有层的相关特性，提取设备特征。然后，根据提取的特征进行建模，生成签名。由于无线信道的动态性，其特征具有随机性，因此模型也具有随机性。最后，通过监督和非监督的机器学习算法进行设备识别。

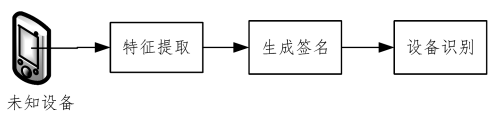

图 14　设备指纹的采集步骤

Fig. 14　Process of device fingerprinting

Xu 对比了监督学习和无监督学习两种指纹算法。无监督学习方法的计算复杂度高，仅被用于检测可能存在的攻击者，不能准确定位恶意设备。但与监督学习方法相比，无监督学习方法更实用，因为它无须预先注册和人工干预。

5.5 小结

本节主要介绍了大数据分析技术在网络安全领域的研究现状。在 APT 攻击检测中，主要利用大数据分析技术对海量的网络设备日志、安全设备日志和操作系统日志进行关联分析，并关联检测到的看似孤立的事件，以发现攻击者 APT 入侵的证据。但是，APT 攻击是经过精心策划的高级隐蔽性攻击，且由多个阶段的攻击组成，具有较高的检测难度，因此设计出能够区分隐蔽性、持续性网络通信行为和正常行为的大数据分析方法，在海量网络数据中有效识别隐蔽性、持续性网络通信行为，是当前需要解决的问题。网络安全检测的研究如表 6 所列。一方面，大多数的网络安全检测方法是基于监督技术的，不能有效地确认非预期的异常行为和及时检测出新的异常现象，因此还须研究更多的基于无监督的异常检测算法，以识别更多未知攻击。另一方面，DDoS、僵尸网络以及入侵检测具有攻击流量大、形式多样化的特点，对该类攻击的检测要求能够做出实时的响应。因此，构建基于大数据分析的网络安全检测模型，提高检测率和实时性，均需进行深入研究。目前，网络安全态势感知的方法多种多样，数据融合领域内几乎所有的理论方法都被应用到网络安全态势评估中，因此需要考虑如何根据态势感知评估的具体对象选择合适的评估方法，以提高评估的准确率和效率。在利用大数据分析技术进行设备指纹的采集时可能都会涉及到用户隐私，因此在未来的研究中要加强用户隐私的保护。

表 6 大数据分析在网络安全检测中的应用

Table 6 Applications of big data analytics in network anomaly detection

文献	研究内容	异常类型	数据集	相关技术
[81]	Spark 集群实现的 DDoS 检测方法	DDoS	2000 DARPA LLDOS 1.0 和产生的正常流量	ANN，Spark
[82]	基于多元降维分析的实时 DDoS 攻击检测机制	DDoS	KDD Cup 1999	PCA、多变量相关分析、MATLAB
[83]	基于 Hadoop 的实时 DDoS 检测框架 HADEC	DDoS	Mausezahn 产生的攻击	MapReduce，HDFS
[84]	DDoS 检测中应用大数据分析技术的机器学习模型	DDoS	NSL-KDD	NaïveBayes，GBM
[85]	基于 Apache Storm 的实时混合入侵检测系统	DDoS	ISCX 2012	Storm、CC4 神经网络、多层感知机
[86]	使用 Hadoop 的超高速大数据环境的实时 IDS	DoS，U2R，R2L，Probing	DARPA，KDD 99，NSL-KDD	J48、随机森林，e、SVM、Hadoop
[90]	使用随机森林的准实时 P2P 僵尸网络检测	Bot attacks	CAIDA 和校园网络流量	Hive、Tshark、Mahout、随机森林
[93]	基于 Spark 平台的无监督网络异常检测方法	Botnet	CTU-13 dataset	Spark，netflow
[94]	基于大数据的模型，可以避免网络流量分布调整带来的影响，减小误检率并增加检测准确率	Dos，U2R，R2L，Probe	KDD CUP99	K-means、KNN、决策树、随机森林
[95]	使用大数据技术来分析网络流量的方法	SYN Flood，Scan，XMAS Scan，SYN/FIN Attack	NCCDC	HDFS，Hive

6 大数据分析技术在产业界的应用

网络时代，很多企业纷纷提出基于大数据分析技术的产业解决方案，如表 7 所列。由于商业产品的特性，这些产品背后的算法或方法无法在公开的文献中获得，因此在每种解决方案中添加了相对应的学术文献。NetReflex IP 和 NetReflex MPLS 利用大数据分析技术提供异常分析和流量分析等服务，如文献[45，47，100]。诺基亚提供了很多面向无线领域的解决方案，如 tradica 是实时流量监控工具，用于分析用户行为以获得网络洞察力，如文献[36，101]。Wireless Network Guardian 用于检测移动网络中的用户异常，文献[102]讨论了类似的话题。Preventive Complaint Analysis 利用大数据分析技术预测可能出现客户投诉的位置，并相应地优先考虑网络优化。文献[103-104]提供了类似的方法。惠普提出了 Vertica——一种利用 CDRs 进行网络规划、优化和故障预测的解决方案，文献[29，105]研究了类似的方法。Amdoc 的 Deep Network Analyzer 为蜂窝网络提供预测性维护和前瞻性的网络部署，文献[106]实现了类似的方法。日志分析可用于各种各样的目的。Aprevi 的 ARLAS (Apervi's Real-time Log Analytics Solution)解决方案实现了网络日志的实时收集和存储。日志分析的相关学术研究有文献[107-109]。

此外，北塔软件的 BTSO(Betasoft Smart Operation & Maintenance)产品为网络运维中的大数据采集分析提供了有力的支持。BTSO 采用专有技术提供高效、海量的数据采集能力，能以 5min 为间隔对数十万的数据样点进行不少于 1 年的数据记录，并提供数据的初步分析，为后续的各类数据分析、挖掘工具提供数据接口，可满足业务分析管理需要。华三公司的 H3CData 无线大数据产品可提供精确到个人终端的位置感知、人员室外运动轨迹描绘，同时提供曾经、现在、未来轨迹的查询、展现及预测服务。

通过研究上述解决方案，可以发现大多数解决方案都是在无线领域，这与学术界的研究方向是一致的。通过对所提供的解决方案进行抽样，可以看出人们对异常预测和网络节点的部署越来越感兴趣。运营商的目标为：为用户尽可能提供接近最优的服务，同时最小化网络扩展支出。

表 7　大数据分析驱动的产业解决方案

Table 7　Big data analytics-powered industrial solutions

公司名称	解决方案	文献	主要功能
Juniper	NetReflex IP	[45,47,100]	消除网络错误；监控 QoS/QoE；容量规划，流量路由，缓存和其他优化
	NetReflex MPLS		预测 MPLS VPN 使用率以便拥塞规划；确定流量利用率和趋势以优化运营成本；能够根据 VPN、服务成本(CoS)等分割网络性能，从而实现更高效的计划
Nokia	Traffica	[36,101]	实时问题检测和网络故障排除；获得有关流量、网络、设备和用户的实时信息
	Wireless Network Guardian	[102]	通过实时用户级信息改进端到端网络分析和报告；检测异常并报告通话时间、信令和带宽资源消耗；主动检测问题，包括自动检测用户异常和低 QoE 分数警报
	Preventive Complaint Analysis	[103]	检测网络元素的行为异常；预测可能出现客户投诉的位置，并相应地优先考虑网络优化
	Predictive Care	[102,104]	通过分析以往网络运行日志，对网络故障进行预测；简化警报的准确率约为 98%，从而减少了运营工作量
HP	Vertica	[29,105]	为通信服务提供商提供 CDR 分析；检查掉话记录和其他维护数据，以确定在基础设施上投资的位置；故障预测和主动维护
Amdocs	Deep Network Analytics	[106]	将 RAN 信息与 BSS 和客户数据相结合，以主动部署网络；预测性地进行网络维护
Apervi	Apervi's Real-time Log Analytics Solution (ARLAS)	[107-109]	实时收集、聚合和存储日志数据

7　大数据分析驱动的设计周期和挑战

7.1　大数据分析设计的循环周期

大数据时代的研究者依靠大数据分析技术提供的功能来改变网络的设计方式。这包括使用大数据分析技术来预测和最大化带宽利用率、预测和准备潜在故障，并预测精确的能耗需求。因此，使用大数据分析技术能创建一个停机更少、用户满意度更高、性能更好的网络。

基于前文所述文献，概括出网络领域中大数据分析技术的循环周期，如图 15 所示。该循环周期包括 3 个部分：首先，从网络中收集大数据；然后，在大数据集群中存储和处理网络大数据，抽取数据中的有用信息，如趋势、模式和关联等，把推断知识传递给决策平台；最后，决策平台使用推断知识评估网络的最佳配置，把新的设计决策作为反馈配置参数发送到网络，实现重新配置。值得注意的是，上述周期的持续时间因网络应用类型的不同而不同。例如，企业网络可以在短时间内生成大量数据，利用企业网络生成的数据，能快速消除企业网络中的配置错误。另外，医疗网络随着时间的推移通常产生较少的监控数据，在有足够多的可用数据之前不应被重新配置，因为频繁的重新配置可能导致失败，从而严重影响人们的健康。

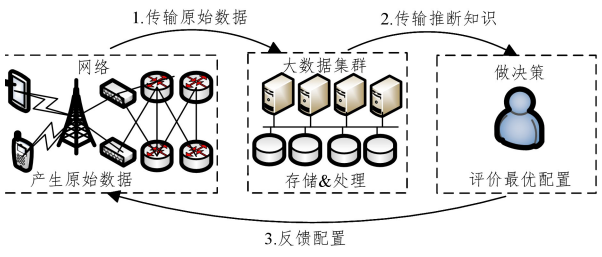

图 15　大数据驱动的网络设计循环周期

Fig. 15　Big-data-powered network design cycle

网络领域中基于大数据分析的应用架构如图 16 所示，包括数据产生、数据收集、数据存储、数据分析和数据应用 5 个模块。底层运行的网络会源源不断地产生各种数据，数据收集器根据不同的应用目的收集不同数据，并将收集到的数据进行预处理后存储到数据库、分布式文件系统；然后应用 MapReduce、神经网络等方法对数据进行分析，并将分析结果应用到用户行为分析、网络规划与优化等应用中。

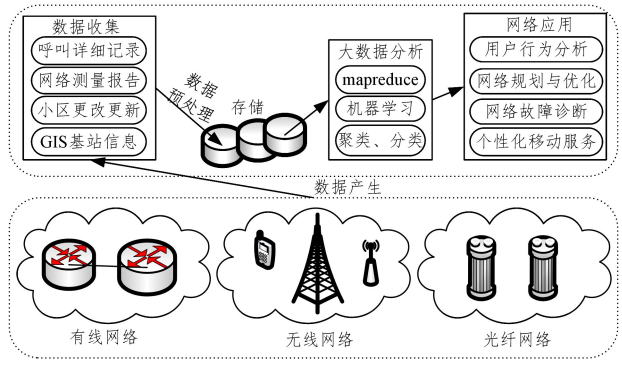

图 16　基于大数据分析的应用架构

Fig. 16　Application architecture based on big data analysis in network

7.2　在网络设计中大数据分析面临的挑战

目前，尽管大数据分析技术在无线网络、SDN 网络、光纤网络和网络安全领域中已有不少解决方案和方法，但网络领域的大数据分析依然面临着各种问题和挑战。

(1)如何保护用户隐私。出于优化网络、监控网络等目的，大数据分析系统会从各个角度充分挖掘用户的偏好信息，这将很大程度上威胁到用户的隐私。例如，移动手机中与银行交易相关的数据是高度私密的信息，在移动蜂窝网络的大数据分析中应谨慎处理。此外，政府和企业关于隐私和数据保护的法规在保护网络中大数据的敏感性方面发挥着基础和必要作用。

(2)如何过滤无用数据。网络以惊人的速度不断产生数据，其中很多都不值得关注。如何过滤原始数据并按数量级进行压缩是今后的研究重点，尤其是怎样定义数据过滤器，使得过滤后不会丢弃有用的信息。

(3)如何自动生成正确的元数据，以描述记录了哪些数据以及如何记录和测量数据。这种元数据对后续分析至关重

要。通常收集的信息不会以可供分析的格式出现,在对数据进行分析之前必须处理错误数据,如某些关于无线电强度的质量报告中的帧错误率和包错误率是不准确的。

数据分析比简单定位、识别、理解和引用数据更具挑战性。为有效地进行大规模分析,数据分析的过程应以自动化的方式进行。目前,数据分析师被从数据库导出数据、执行非SQL过程并将数据存回数据库的冗长过程所阻碍。如果用户无法理解大数据的分析,那么拥有大数据分析能力的价值是有限的。因此,一个决策者如果提供了分析结果,那么必须对这些结果进行解释。

7.3 大数据分析在网络应用中的分类与趋势

网络中产生的海量数据,无论是网络中直接产生的还是由网络服务提供商收集的,都为网络应用提供了前所未有的机遇。本文从应用领域的角度,将大数据分析技术在网络中的应用分为5类,如表8所列。

表8 大数据分析技术在网络中的应用

Table 8 Application of big data analytics in network application

应用领域	相关应用	相关文献
社交网络	内容推荐系统、定向广告推送、个性化移动服务、市场营销、舆情分析	[110-112]
智慧城市	城市规划、智能交通管理、交通流量可视化	[113-116]
智能电网	电力数据可视化、电力系统运营与规划、节能系统、可再生能源集成	[117-118]
医疗健康	安全智能医疗监控报警系统、以病人为中心的蜂窝网络	[119-121]
物联网	物流和供应链管理、智慧数字化城市、智能医疗、车载内容中心网络、广播风暴处理等	[122-124]

社交网络分析、个性化移动服务、智慧城市等基于大数据分析技术的应用已得到广泛关注。物联网的日益普及,以及终端设备的便携性和移动性,使得数据采集更加灵活有效。基于大数据分析的应用已经变得更加直接,并正在向实时方向发展。移动支付等数字化交易彻底改变了人类的行为和资本流动方式,人与机器的互动产生了丰富的大数据,从人与机器互动产生的大数据中挖掘知识和洞察力,从而设计出更适合人类生活的模式。此外,政府拥有最多的数据(除了媒体和社交媒体应用),数据涵盖资源、金融、交通、安全、医疗、环境、食品等。政府的开放数据政策对整个数据产业的发展至关重要。随着5G技术、物联网、人工智能等技术的发展,基于大数据的解决方案将会逐步应用于各个行业。

8 未来的研究方向

近年兴起的一些技术也为网络大数据分析的进一步应用提供了可能的方向。

(1)数据安全。当用户在微信、微博等社交网络上分享信息时,大量个人信息不仅存储在终端里,也分散在互联网中。大数据的汇集加大了用户隐私泄露的风险,因此开展大数据安全研究,强化数据加密、加固智能终端、保护个人敏感信息,对网络大数据的推广应用具有重要意义。He等[2]提出了一种数据清理策略,其既保留了社交网络数据的优点,又保护了潜在的敏感信息。

(2)最小化处理。大数据处理的第一步是数据收集和预处理。数据清理是数据预处理的一种形式。预处理的特定示例是使用计算射频识别(Computational Radio Frequency Identification,CRFID)传感器,无线传感器可以在接近移动的收集器对象(如车辆)时使用诸如磁力共振[125]的技术来无线供电,使得某些预处理任务能够朝CRFID传感器侧移动,从而收集已经清理并减少的数据量。这样可以提高效率,缩短分析时间,提高存储利用率,并促进实时分析。

(3)人口密集和贫穷地区的卫星互联网连接。SpaceX[126]这类项目已经出现,其中包含4000多颗卫星,成本超过1亿美元,埃隆·马斯克宣称该项目旨在向全世界提供高速的互联网卫星。通过利用卫星通信网络的大数据分析,可以集中更多的功率,降低信号接收要求。大数据分析技术可用于关联地面数据(如地理信息、天气状况)和经济相关数据,以帮助识别这些区域。

(4)物联网节点的部署。据惠普预测,2030年物联网传感器的数量将达到1万亿,这将使得物联网数据成为大数据中最重要的部分[127]。但是,有效收集数据需要将物联网传感器部署在可以收集尽可能多的数据的位置。许多传感器由于被部署在错误的位置而浪费(这个位置对提供有价值的数据/类型没有帮助)。为实现最佳的物联网设计,大数据分析技术可以关联多个参数(如流量模式、社交事件、网络参数等),以确定传感器的最佳部署位置。

(5)低能耗绿色网络。用户在享受社交网络、搜索引擎、在线视频等应用时,无论在网络端还是用户终端均需消耗很多计算、存储、能量等资源。在无线接入网络端,超密集化的网络站点部署会显著加剧高能耗问题。在用户终端,多样化的应用也会增加能耗。因此,可以利用大数据分析用户的移动性(如位置信息、移动速度、移动方向等),并结合分析结果实现超密集异构网络中不同层级基站的动态激活与休眠、资源共享、协作传输等,从而降低网络的整体能耗,提高传输效率。同时,用户终端针对具有相似特征的业务需求或地理位置临近的终端组,可以通过终端间直接通信(D2D)的方式完成信息的交互与推送,从而节省网络资源,降低发射功率,实现低能耗通信。

结束语 在网络领域,可以利用大数据分析技术的应用很多。网络中存在着大量的网络流量、日志信息、系统信号等数据,收集网络数据并将其与用户趋势、服务需求相关联,可以设计以用户为中心的自适应网络。大数据分析作为数据科学最热门的研究方向之一,在商业、金融、医疗等领域取得了一系列令人瞩目的研究成果,吸引了越来越多网络领域研究人员的关注,并取得了一系列的重要研究成果。本文对这些成果进行了系统的总结和分析,以大数据分析为手段,着重总结了大数据分析技术如何应用于网络领域,并对大数据分析技术在无线网络、SDN网络、光纤网络和网络安全方面的应用进行了深入分析,最后探讨了大数据分析技术在网络领域的发展趋势及挑战。

参 考 文 献

[1] HADI M,LAWEY A Q,EL-GORASHI T,et al. Big Data Analytics for Wireless and Wired Network Design:A Survey[J]. Computer Networks,2018,132:180-199.

[2] HE Z,CAI Z,YU J. Latent-data Privacy Preserving with Customized Data Utility for Social Network Data[J]. IEEE Tran-

sactions on Vehicular Technology，2017，67(1)：665-673.

[3] BI S，ZHANG R，DING Z，et al. Wireless communications in the era of big data [J]. IEEE Communications Magazine，2015，53(10)：190-199.

[4] QIAN L，ZHU J，ZHANG S. Survey of wireless big data[J]. Journal of Communications & Information Networks，2017，2(1)：1-18.

[5] XU Q S，GE L Q，ZOU Q Y. Wireless Communication Technology based on Big-Data Analysis[J]. Communications Technology，2016，49(12)：1635-1641. (in Chinese)
徐全盛，葛林强，邹勤宜. 基于大数据分析的无线通信技术研究[J]. 通信技术，2016，49(12)：1635-1641.

[6] FU Y，LI H C，WU X P，et al. Detecting APT attacks：a survey from the perspective of big data analysis[J]. Journal on Communications，2015，36(11)：1-14. (in Chinese)
付钰，李洪成，吴晓平，等. 基于大数据分析的 APT 攻击检测研究综述[J]. 通信学报，2015，36(11)：1-14.

[7] CHEN X S，ZENG X M，WANG W X，et al. Big Data Analytics for Network Security and Intelligence[J]. Advanced Engineering Sciences，2017，49(3)：1-12. (in Chinese)
陈兴蜀，曾雪梅，王文贤，等. 基于大数据的网络安全与情报分析[J]. 工程科学与技术，2017，49(3)：1-12.

[8] WANG Y Z，JIN X L，CHENG X Q. Network Big Data：Present and Future[J]. Chinese Journal of Computers，2013，36(6)：1125-1138. (in Chinese)
王元卓，靳小龙，程学旗. 网络大数据：现状与展望[J]. 计算机学报，2013，36(6)：1125-1138.

[9] CHEN K F，ZHOU H C，AMP J P. Research on Realization Mode of Telecom Operators' Big Data Resource and its Strategy [J]. Mobile Communications，2016，40(1)：63-67.

[10] ZHANG X，YI Z，YAN Z，et al. Social Computing for Mobile Big Data[J]. Computer，2016，49(9)：86-90.

[11] CHEN M，MAO S，LIU Y. Big Data：A Survey[J]. Mobile Networks & Applications，2014，19(2)：171-209.

[12] HE Y，YU F R，ZHAO N，et al. Big Data Analytics in Mobile Cellular Networks[J]. IEEE Access，2016，4：1985-1996.

[13] ZHANG C，QIU R C. Massive MIMO as a Big Data System：Random Matrix Models and Testbed[J]. IEEE Access，2015，3：837-851.

[14] KUANG L，HAO F，YANG L T，et al. A Tensor-Based Approach for Big Data Representation and Dimensionality Reduction[J]. IEEE Transactions on Emerging Topics in Computing，2017，2(3)：280-291.

[15] XU F，LIN Y，HUANG J，et al. Big data driven mobile traffic understanding and forecasting：a time series approach[J]. IEEE Transactions on Services Computing，2016，9(5)：796-805.

[16] MURPHY K. Machine learning：a probabilistic perspective [M]. Cambridge：MIT Press，2012.

[17] GOODFELLOW I，BENGIO Y，COURVILLE A. Deep Learning [M]. Cambridge：MIT Press，2016.

[18] DONAHUE J，HENDRICKS L，GUADARRAMA S，et al. Longterm recurrent convolutional networks for visual recognition and description [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2015：2625-2634.

[19] ZHANG Q，YANG L T，CHEN Z，et al. A survey on deep learning for big data[J]. Information Fusion，2018，42：146-157.

[20] BUCZAK A L，GUVEN E. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection [J]. IEEE Communications Surveys & Tutorials，2017，18(2)：1153-1176.

[21] ZHANG L，CUI Y，LIU J，et al. Application of Machine Learning in Cyberspace Security Research[J]. Chinese Journal of Computers，2018，41：1-35. (in Chinese)
张蕾，崔勇，刘静，等. 机器学习在网络空间安全研究中的应用[J]. 计算机学报，2018，41：1-35.

[22] ALSHEIKH M A，NIYATO D，LIN S，et al. Mobile big data analytics using deep learning and apache spark [J]. IEEE Network，2016，30(3)：22-29.

[23] MA Q，ZHANG S，ZHOU W，et al. When Will You Have a New Mobile Phone? An Empirical Answer From Big Data[J]. IEEE Access，2016，4：10147-10157.

[24] Yang C. Learning methodologies for wireless big data networks：a Markovian game-theoretic perspective [J]. Neurocomputing，2016，174：431-438.

[25] LANDSET S，KHOSHGOFTAAR T M，RICHTER A N，et al. A survey of open source tools for machine learning with big data in the Hadoop ecosystem[J]. Journal of Big Data，2015，2 (1)：24.

[26] Apache Spark™-lightning-fast cluster computing [EB/OL]. [2017-03-20]. http://spark. apache. org/.

[27] CELEBI O F，ZEYDAN E，KURT O F，et al. On use of big data for enhancing network coverage analysis [C]// International Conference on Telecommunications. IEEE，2013：646-655.

[28] GAO J，CHENG X，XU L，et al. A downlink coverage self-optimizing algorithm for LTE cellular networks based on big data analytics[C]// Proceedings of the 3rd International Conference on Signal and Information Processing，Networking and Computers. Springer，2017：373-380.

[29] KARATEPE I A，ZEYDAN E. Anomaly Detection In Cellular Network Data Using Big Data Analytics[C]// Proceedings of VDE. 2014：1-5.

[30] PARWEZ M S，RAWAT D B，GARUBA M. Big Data Analytics for User Activity Analysis and User Anomaly Detection in Mobile Wireless Network[J]. IEEE Transactions on Industrial Informatics，2017，PP(99)：1-1.

[31] YANG K，LIU R，SUN Y，et al. Deep Network Analyzer (DNA)：A Big Data Analytics Platform for Cellular Networks [J]. IEEE Internet of Things Journal，2017，4(6)：2019-2027.

[32] SAHNI A，MARWAH D，CHADHA R. Real time monitoring and analysis of available bandwidth in cellular network-using big data analytics[C]// International Conference on Computing for Sustainable Global Development. IEEE，2015：1743-1747.

[33] KHATIB E J，BARCO R，MUNOZ P，et al. Self-healing in mobile networks with big data [J]. Communications Magazine IEEE，2016，54(1)：114-120.

[34] IMRAN A，ZOHA A. Challenges in 5G：how to empower SON with big data for enabling 5G[J]. Network IEEE，2014，28(6)：27-33.

[35] JIANG D，HUO L，SONG H. Rethinking behaviors and activities of base stations in mobile cellular networks based on big data analysis[J/OL]. IEEE Transactions on Network Science and Engineering，https://ieeexplore. ieee. org/abstract/document/8423194.

[36] LIU J,LIU F,ANSARI N. Monitoring and analyzing big traffic data of a large-scale cellular network with Hadoop[J]. Network IEEE,2014,28(4):32-39.

[37] PALACIO A F,WAUTERS T,VOLCKAERT B,et al. Scalable distributed traffic monitoring for enterprise networks with Spark Streaming[C]// European Conference on Cyber Warfare and Security. 2018:563-570.

[38] QIAO Y,XING Z,FADLULLAH Z M,et al. Characterizing Flow,Application,and User Behavior in Mobile Networks:A Framework for Mobile Big Data[J]. IEEE Wireless Communications,2018,25(1):40-49.

[39] BAŞTUŠ E,BENNIS M,ZEYDAN E,et al. Big data meets telcos:A proactive caching perspective[J]. Journal of Communications and Networks,2016,17(6):549-557.

[40] ZEYDAN E,BASTUG E,BENNIS M,et al. Big data caching for networking:moving from cloud to edge[J]. IEEE Communications Magazine,2016,54(9):36-42.

[41] OMAR A. Improving Data Extraction Efficiency of Cache Nodes in Cognitive Radio Networks Using Big Data Analysis[C]// International Conference on Next Generation Mobile Applications,Services and Technologies. IEEE,2016:305-310.

[42] FAN B,LENG S,YANG K. A dynamic bandwidth allocation algorithm in mobile networks with big data of users and networks [J]. IEEE Network,2016,30(1):6-10.

[43] WANG L,WANG X D,CHENG N. Intelligent System of Wireless Network Optimization Based on Big Data Technology[J]. Telecommunications Science,2015,31(12):159-163. (in Chinese)
王磊,王西点,程楠. 基于大数据技术的智能化无线网络优化体系[J]. 电信科学,2015,31(12):159-163.

[44] LEE C L,SU W S,TANG K A,et al. Design of handover self-optimization using big data analytics[C]// Network Operations and Management Symposium. IEEE,2014:1-5.

[45] CHIH-LIN I,LIU Y,HAN S,et al. On Big Data Analytics for Greener and Softer RAN[J]. IEEE Access,2016,3(94):3068-3075.

[46] LIU Y,LIU K,KONG J K. TD-LTE Network Planning based on Big Data Mining[J]. Communications Technology,2015,48(2):194-198. (in Chinese)
刘毅,刘珂,孔建坤. 基于大数据挖掘的 LTE 网络规划研究[J]. 通信技术,2015,48(2):194-198.

[47] ZHENG K,YANG Z,ZHANG K,et al. Big data-driven optimization for mobile networks toward 5G[J]. IEEE Network,2016,30(1):44-51.

[48] XU C,YANG J,YU H,et al. Optimizing the Topologies of Virtual Networks for Cloud-Based Big Data Processing[C]// High PERFORMANCE Computing and Communications,2014 IEEE,Intl Symp on Cyberspace Safety and Security,2014 IEEE,Intl Conf on Embedded Software and Syst. IEEE,2014:189-196.

[49] KIRAN P,JIBUKUMAR M G,PREMKUMAR C V. Resource allocation optimization in LTE-A/5G networks using big data analytics[C]// International Conference on Information Networking. IEEE Computer Society,2016:254-259.

[50] XU R P,CUI J Y,GUAN Z L,et al. Design and application of Wireless Network optimal data sharing architecute[J]. Telecommunications Science,2016,32(4):152-158. (in Chinese)
许汝鹏,崔晶也,关则洛,等. 网优大数据共享架构设计及应用实践[J]. 电信科学,2016,32(4):152-158.

[51] LI Y. Grass-root based SpectrumMap database for self-organized cognitive radio and heterogeneous networks:Spectrum measurement,data visualization,and user participating model[C]// Wireless Communications and Networking Conference (WCNC),2015 IEEE. IEEE,2015:117-122.

[52] WU Q,DING G,DU Z,et al. A Cloud-Based Architecture for the Internet of Spectrum Devices Over Future Wireless Networks[J]. IEEE Access,2017,4:2854-2862.

[53] GVK S,DASARI S R. Big Spectrum Data Analysis in DSA Enabled LTE-A Networks:A System Architecture[C]// IEEE,International Conference on Advanced Computing. IEEE,2016:655-660.

[54] ZHU Q,ZHANG X. Effective-capacity based gaming for optimal power and spectrum allocations over big-data virtual wireless networks [C]// The IEEE Global Communications Conference (GLOBECOM). IEEE,2015:1-6.

[55] LI D. Application of Spectrum Data Management based on Bayesian Network[J]. International Journal of Future Generation Communication and Networking,2015,8(7):13-24.

[56] BALTIISKI P,ILIEV I,KEHAIOV B,et al. Long-Term Spectrum Monitoring with Big Data Analysis and Machine Learning for Cloud-Based Radio Access Networks[J]. Wireless Personal Communications,2016,87(3):815-835.

[57] VASSAKIS K,PETRAKIS E,KOPANAKIS I. Big Data Analytics:Applications,Prospects and Challenges[M]//Mobile Big Data. Cham:Springer,2018.

[58] CHIU P,REUNANEN J,LUOSTARI R,et al. Big Data Analytics for 4.9G and 5G Mobile Network Optimization[C]// Vehicular Technology Conference. IEEE,2017:1-4.

[59] YAN Q,CHEN W,POOR H V. Big Data Driven Wireless Communications:A Human-in-the-Loop Pushing Technique for 5G Systems[J]. IEEE Wireless Communications,2018,25(1):64-69.

[60] RAZA M R,NATALINO C,VIDAL A. Demonstration of Resource Orchestration Using Big Data Analytics for Dynamic Slicing in 5G Networks[C]//2018 European Conference on Optical Communication. 2018.

[61] ZHANG N,YANG P,REN J,et al. Synergy of Big Data and 5G Wireless Networks:Opportunities,Approaches,and Challenges [J]. IEEE Wireless Communications,2018,25(1):12-18.

[62] RAMASWAMI R,SIVARAJAN K N. Routing and wavelength assignment in all-optical networks[J]. IEEE/ACM Transactions on Networking (TON),1995,3(5):489-500.

[63] YAN S F. Heuristic Algorithm for Routing and Wavelength Assignment Problem [D]. Wuhan:Huazhong University of Science and Technology,2016. (in Chinese)
燕圣峰. 基于启发式算法求解路由与波长分配问题[D]. 武汉:华中科技大学,2016.

[64] SHEN G,LI Y,PENG L,et al. Almost-optimal design for optical networks with hadoop cloud computing:Ten ordinary desktops solve 500-node,1000-link,and 4000-request RWA problem within three hours[C]//2013 15th International Conference on Transparent Optical Networks (ICTON). IEEE,2013:1-4.

[65] LI Y,SHEN G,CHEN B,et al. Applying Hadoop Cloud Computing Technique to Optimal Design of Optical Networks[C]// A-

sia Communications & Photonics Conference. Optical Society of America，2015：Asu3H. 3.

[66] AUTENRIETH A，AGUADO A，MAYORAL A，et al. Dynamic Virtual Network Reconfiguration Over SDN Orchestrated Multitechnology Optical Transport Domains[J]. Journal of Lightwave Technology，2016，34(8)：1933-1938.

[67] MORALES F，RUIZ M，GIFRE L，et al. Virtual network topology adaptability based on data analytics for traffic prediction [J]. IEEE/OSA Journal of Optical Communications & Networking，2017，9(1)：A35-A45.

[68] QADIR J，AHAD N，MUSHTAQ E，et al. SDNs，Clouds，and Big Data：New Opportunities[C]// International Conference on Frontiers of Information Technology. IEEE Computer Society，2014：28-33.

[69] MCKEOWN N，ANDERSON T，BALAKRISHNAN H，et al. OpenFlow：enabling innovation in campus networks[J]. ACM SIGCOMM Computer Communication Review，2008，38(2)：69-74.

[70] CUI H，ZHANG Y，MA C，et al. Design and Realization of Cognitive Routing Resources Using Big Data Analysis in SDN[C]// IEEE International Congress on Big Data. IEEE Computer Society，2015：424-429.

[71] NEVES M V，KATRINIS K，FRANKE H. Pythia：Faster Big Data in Motion through Predictive Software-Defined Network Optimization at Runtime[C]// Parallel and Distributed Processing Symposium，2014 IEEE，International. IEEE，2014：82-90.

[72] COSTA P，DONNELLY A，ROWSTRON A，et al. Camdoop：Exploiting In-network Aggregation for Big Data Applications [C]// NSDI. 2012.

[73] COSTA P，DONNELLY A，O'SHEA G，et al. CamCube：a key-based data center：Technical Report MSR TR-2010-74[R]. Microsoft Res. ，Redmond，WA，USA，2010.

[74] ISARD M，BUDIU M，YU Y，et al. Dryad：distributed data-parallel programs from sequential building blocks[C]// ACM SIGOPS Operating Systems Review. ACM，2007：59-72.

[75] YU Y，ISARD M，FETTERLY D，et al. DryadLINQ：a system for general-purpose distributed data-parallel computing using a high-level language[C]// Usenix Symposium on Operating Systems Design and Implementation(OSDI 2008). San Diego，California，USA，DBLP，2008：1-14.

[76] CUI L，YU F R，YAN Q. When big data meets software-defined networking：SDN for big data and big data for SDN[J]. IEEE Network，2016，30(1)：58-65.

[77] GIURA P，WANG W. Using large scale distributed computing to unveil advanced persistent threats[J]. Science，2013，1(3)：93-105.

[78] YEN T F，OPREA A，ONARLIOGLU K，et al. Beehive：large-scale log analysis for detecting suspicious activity in enterprise networks[C]// Computer Security Applications Conference. ACM，2013：199-208.

[79] MARCHETTI M，PIERAZZI F，COLAJANNI M，et al. Analysis of high volumes of network traffic for Advanced Persistent Threat detection[M]. Elsevier North-Holland，Inc. ，2016.

[80] ZHANG X S，NIU W N，YANG G W，et al. Method for APT Prediction Based on Tree Structure[J]. Journal of University of Electronic Science and Technology of China，2016，45(4)：582-588.(in Chinese)
张小松，牛伟纳，杨国武，等. 基于树型结构的 APT 攻击预测方法[J]. 电子科技大学学报，2016，45(4)：582-588.

[81] HSIEH C J，CHAN T Y. Detection DDoS attacks based on neural-network using Apache Spark[C]// International Conference on Applied System Innovation. IEEE，2016：1-4.

[82] JIA B，MA Y，HUANG X H，et al. A Novel Real-Time DDoS Attack Detection Mechanism Based on MDRA Algorithm in Big Data[J]. Mathematical Problems in Engineering，2016，2016：1-10.

[83] Hameed S，Ali U. HADEC：Hadoop-based live DDoS detection framework[J]. Eurasip Journal on Information Security，2018，2018(1)：11.

[84] HOON K S，YEO K C，AZAM S，et al. Critical review of machine learning approaches to apply big data analytics in DDoS forensics[C]// 2018 International Conference on Computer Communication and Informatics. IEEE，2018.

[85] MYLAVARAPU G，THOMAS J，ASHWIN K T K. Real-Time Hybrid Intrusion Detection System Using Apache Storm[C]// IEEE International Conference on High PERFORMANCE Computing and Communications. IEEE，2015：1436-1441.

[86] RATHORE M M，AHMAD A，PAUL A. Real time intrusion detection system for ultra-high-speed big data environments[J]. Journal of Supercomputing，2016，72(9)：1-22.

[87] WANG L，JONES R. Big data analytics for network intrusion detection：A survey[J]. International Journal of Networks and Communications，2017，7(1)：24-31.

[88] STONE-GROSS B，COVA M，CAVALLARO L. Your botnet is my botnet：analysis of a botnet takeover[C]// ACM Conference on Computer and Communications Security(CCS 2009). Chicago，Illinois，Usa，DBLP，2009：635-647.

[89] USCERT. WordPress Sites Targeted by Mass Brute-force Botnet Attack[EB/OL]. https：//www. us-cert. gov/.

[90] SINGH K，GUNTUKU S C，THAKUR A，et al. Big Data Analytics framework for Peer-to-Peer Botnet detection using Random Forests[J]. Information Sciences，2014，278(19)：488-497.

[91] CHRIS S. Applied Network Security Monitoring[M]// Applied Network Security Monitoring：Collection，Detection，and Analysis. Syngress Publishing，2013.

[92] LIU Y，GUO S，HU S，et al. Performance Evaluation and Optimization of Multi-dimensional Indexes in Hive[J]. IEEE Transactions on Services Computing，2016，PP(99)：1-1.

[93] TERZI D S，TERZI R，SAGIROGLU S. Big data analytics for network anomaly detection from netflow data[C]// International Conference on Computer Science and Engineering. IEEE，2017：592-597.

[94] YAO H，LIU Y，FANG C. An Abnormal Network Traffic Detection Algorithm Based on Big Data Analysis[J]. International Journal of Computers，Communications & Control，2016，11(4)：567-579.

[95] BACHUPALLY Y R，YUAN X，ROY K. Network security analysis using Big Data technology[C]// Southeastcon. IEEE，2016：1-4.

[96] PURI C，DUKATZ C. Analyzing and Predicting Security Event Anomalies：Lessons Learned from a Large Enterprise Big Data Streaming Analytics Deployment[C]// International Workshop on Database and Expert Systems Applications. IEEE Computer

Society,2015:152-158.

[97] GUAN L,HU G J,WANG Z. Research on Network Security Situational Awareness Technology Based on Big Data[J]. Netinfo Security,2016(9):45-50. (in Chinese)
管磊,胡光俊,王专. 基于大数据的网络安全态势感知技术研究[J]. 信息网络安全,2016(9):45-50.

[98] ZHAO M. Network Security Situation Awareness Based on Big Bata[J]. Netinfo Security,2016(9):90-93. (in Chinese)
赵梦. 基于大数据环境的网络安全态势感知信息[J]. 网络安全,2016(9):90-93.

[99] XU Q,ZHENG R,SAAD W,et al. Device Fingerprinting in Wireless Networks: Challenges and Opportunities[J]. IEEE Communications Surveys & Tutorials,2016,18(1):94-104.

[100] MOLINA M,PAREDES-OLIVA I,ROUTLY W,et al. Operational experiences with anomaly detection in backbone networks [J]. Computers & Security,2012,31(3):273-285.

[101] RICCIATO F. Traffic monitoring and analysis for the optimization of a 3G network[J]. IEEE Wireless Communications,2006,13(6):42-49.

[102] PARWEZ M S,RAWAT D B,GARUBA M. Big Data Analytics for User-Activity Analysis and User-Anomaly Detection in Mobile Wireless Network[J]. IEEE Transactions on Industrial Informatics,2017,13(4):2058-2065.

[103] SPIESS J,T'JOENS Y,DRAGNEA R,et al. Using Big Data to Improve Customer Experience and Business Performance[J]. Bell Labs Technical Journal,2014,18(4):3-17.

[104] JIANG Z,DAIWU S,ZHENHUA Y U. Study on Network Failure Prediction Based on Alarm Log[C]// Mec International Conference on Big Data & Smart City. IEEE,2016:1-7.

[105] SHUAN L H,FEI T Y,KING S W,et al. Network Equipment Failure Prediction with Big Data Analytics[J]. International Journal of Advances in Soft Computing & Its Applications,2016,8(3):59-69.

[106] YANG K,LIU R,SUN Y,et al. Deep Network Analyzer (DNA): A Big Data Analytics Platform for Cellular Networks [J]. IEEE Internet of Things Journal,2017,PP(99):1-1.

[107] QIAO Y,LEI Z,YANG J,et al. FLAS: Traffic analysis of emerging applications on Mobile Internet using cloud computing tools [C]// 2013 16th International Symposium on Wireless Personal Multimedia Communications (WPMC). IEEE,2013:1-6.

[108] QI G,TSAI W T,LI W,et al. A cloud-based triage log analysis and recovery framework[J]. Simulation Modelling Practice and Theory,2017,77:292-316.

[109] PARK B H,HUKERIKAR S,ADAMSON R,et al. Big Data Meets HPC Log Analytics: Scalable Approach to Understanding Systems at Extreme Scale[C]// 2017 IEEE International Conference on Cluster Computing (CLUSTER). IEEE,2017:758-765.

[110] CHANG V. A proposed social network analysis platform for big data analytics[J]. Technological Forecasting Social Change,2018,130:57-68.

[111] LEUNG C K,JIANG F,POON T W,et al. Big Data Analytics of Social Network Data: Who Cares Most About You on Facebook? [J]. Highlighting the Importance of Big Data Management and Analysis for Various Applications,2018,27:1-15.

[112] SONG P,LU D Y,ZHAO Y P,et al. New Progress of Big Data Research in Social Network[J]. Lantai World,2017(12):63-67. (in Chinese)
宋朋,陆丹玥,赵燕萍,等. 社交网络中大数据研究新进展[J]. 兰台世界,2017(12):63-67.

[113] RATHORE M M,PAUL A,HONG W H. Exploiting IoT and big data analytics: Defining Smart Digital City using real-time urban data[J]. Sustainable Cities and Society,2018,40:600-610

[114] BORJA M G,MARÍA HENAR S O,JUAN CARLOS G P,et al. Dynamic accessibility using Big Data: The role of the changing conditions of network congestion and destination attractiveness[J]. arXiv:1610.06450,2016.

[115] SENARATNE H,MUELLER M,BEHRISCH M,et al. Urban Mobility Analysis With Mobile Network Data: A Visual Analytics Approach[J]. IEEE Transactions on Intelligent Transportation Systems,2017,19(5):1-10.

[116] GOHAR M,MUZAMMAL M,RAHMAN A U. SMART TSS: Defining Transportation System Behavior using Big Data Analytics in Smart Cities[J]. Sustainable Cities & Society,2018,41:114-119.

[117] TAO H. Big Data Analytics: Making the Smart Grid Smarter [J]. IEEE Power & Energy Magazine,2018,16(3):12-16.

[118] WANG G,GUNASEKARAN A,NGAI E W T. Distribution network design with big data: Model and analysis[J]. Annals of Operations Research,2018,270(12):539-551.

[119] MANOGARAN G,VARATHARAJAN R,LOPEZ D. A new architecture of Internet of Things and big data ecosystem for secured smart healthcare monitoring and alerting system[J]. Future Generation Computer Systems,2018,82:375-387.

[120] HADI M S,LAWEY A Q,EL-GORASHI T E H,et al. Patient-Centric Cellular Networks Optimization using Big Data Analytics[J]. IEEE Access,2019,7:49279-49296.

[121] HOSSAIN M S,MUHAMMAD G. Emotion-Aware Connected Healthcare Big Data Towards 5G[J]. IEEE Internet of Things Journal,2017,PP(99):1-1.

[122] WAMBA S F,ANGAPPA G. Big data analytics in logistics and supply chain management[J]. Journal of Logistics Management,2018,29(2):478-484.

[123] FIROUZI F,RAHMANI A M,MANKODIYA K,et al. Internet-of-Things and big data for smarter healthcare: From device to architecture,applications and analytics[J]. Future Generation Computer Systems,2017,78(2018):583-586.

[124] WAHID A,SHAH M A,QURESHI F F. Big data analytics for mitigating broadcast storm in Vehicular Content Centric networks[J]. Future Generation Computer Systems,2018,86:1301-1320.

[125] IMURA T,HORI Y. Maximizing Air Gap and Efficiency of Magnetic Resonant Coupling for Wireless Power Transfer Using Equivalent Circuit and Neumann Formula[J]. IEEE Transactions on Industrial Electronics,2011,58(10):4746-4752.

[126] FINLEY K. Internet by Satellite Is a Space Race With No Winners[EB/OL]. https://www.wired.com/2015/06/elon-musk-space-x-satellite-internet/.

[127] LI P,CHEN Z,YANG L T,et al. Deep Convolutional Computation Model for Feature Learning on Big Data in Internet of Things[J]. IEEE Transactions on Industrial Informatics,2017,PP(99):1-1.