基于大数据的审计技术研究

徐 超[1],陈 勇[1],葛红美[1],何炎祥[2]
(1. 南京审计大学信息工程学院,江苏南京 211815;2. 武汉大学计算机学院,湖北武汉 430072)

摘 要: 审计是党和国家监督体系的重要组成部分,在维护国家财政经济秩序、提高财政资金使用效益、促进廉政建设、保障经济社会健康发展等方面发挥了重要作用.大数据时代的来临引领了审计技术方法的革新,应用大数据技术是实现审计全覆盖目标的必由之路,大数据审计建设是影响审计事业未来发展的核心技术工程.本文首先介绍了我国现阶段审计信息化建设的意义及发展历程,概括了大数据审计概念、特征及研究现状,探讨了大数据审计面临的新机遇、新挑战,以及对大数据审计在采集、存储、分析和可视化工作中的相关研究等进行了总结、比较和分析,并以蓝天保卫计划、精准扶贫等典型审计类型进行具体应用,最后展望了大数据审计未来发趋势.本文阐述的科学问题研究源于多学科领域交叉,具有鲜明的学科交叉特征,旨在通过交叉研究促进多学科知识融通发展,以期望对相关的理论研究有着重要的借鉴与参考意义.

关键词: 大数据;审计;电子数据;信息化
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2020)05-1003-15
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.05.023

Audit Technology Research Based on Big Data

XU Chao[1], CHEN Yong[1], GE Hong-mei[1], HE Yan-xiang[2]
(1. *School of Information Engineering, Nanjing Audit University, Nanjing, Jiangsu* 211815, *China*;
2. *School of Computer Science, Wuhan University, Wuhan, Hubei* 430072, *China*)

Abstract: Audit is an important part of the supervision system of the Party and the state. It plays an important role in maintaining the national financial and economic order, improving the efficiency of the use of financial funds, promoting the construction of a clean government, and ensuring the healthy development of the economy and society. The advent of the era of big data has led to the innovation of audit technology and methods. The application of big data technology is the only way to achieve the goal of full coverage of audit. The construction of big data audit is the core technology engineering that affects the future development of audit. Firstly, this paper introduces the significance and development course of auditing informationization construction at present in China, summarizes the concept, characteristics and research status of big data auditing, summarizes, compares and analyses the related research of big data auditing technology in collection, storage, analysis and visual display, and probes into the problems faced by big data auditing. New opportunities, new challenges, and typical audit scenarios such as Blue Sky Security Plan, Precision Poverty Alleviation. Finally, the future trend of big data audit is prospected. The research on scientific issues in this paper originates from the interdisciplinary field and has distinct interdisciplinary characteristics. It aims to promote the development of interdisciplinary knowledge integration through interdisciplinary research, with a view to providing important reference and reference for relevant theoretical research.

Key words: big data; audit; electronic data; informatization

1 引言

审计是国家治理的基石和重要保障,是依法用权力监督制约权力的制度安排.其本质是国家治理这个大系统中的一个内生的具有预防、揭示和抵御功能的“免疫系统”,核心是推动民主法治,实现国家良好治理,促进国家经济社会健康运行和科学发展,从而更好地保障人民的根本利益,是国家治理的重要组成部分.

收稿日期:2019-10-31;修回日期:2020-01-02;责任编辑:马兰英
基金项目:国家自然科学基金面上项目(No. 71972102, No. 61640220);教育部人文社会科学研究规划基金(No. 19YJAZH100)

新时代审计要求对公共资金、国有资产、国有资源和领导干部履行经济责任情况实行审计全覆盖,做到审计监督无死角.这是首次从国家治理高度完善审计制度,是我国审计史上重要的里程碑,也是新常态下对审计事业提出更高的要求.审计与国家治理之间有着天然的联系,在维护国家财政经济秩序、提高财政资金使用效益、促进廉政建设、保障经济社会健康发展等方面发挥了重要作用.特别是党的十八大以来,为促进党中央令行禁止、维护国家经济安全、推动全面深化改革、促进依法治国、推进廉政建设等作出了重要贡献.

习近平总书记在中央审计委员会第一次会议上指出,"审计是党和国家监督体系的重要组成部分,改革审计管理体制,组建中央审计委员会,是加强党对审计工作领导的重大举措.要深化审计制度改革,解放思想、与时俱进,创新审计理念,及时揭示和反映经济社会各领域的新情况、新问题、新趋势.要坚持科技强审,加强审计信息化建设."国家审计署胡泽君审计长指出:"必须坚持科技强审,革新传统审计方法,加强信息化基础建设,更好运用互联网技术和信息化手段开展审计,向信息化要资源、向大数据要效率,通过信息化、数字化、网络化,提高审计监督、过程控制、决策支撑能力,积极推进审计全覆盖."由此可见,新时代信息技术下审计技术与方法的研究已经成为了当下以及今后一段时期亟需研究的重大问题.长期以来,受人力、被审计单位信息化水平以及审计自身的信息化手段的限制,审计比较依赖于抽样分析.大数据时代的来临,引领了审计技术方法的革新.审计利用大数据及其相关技术,有利于从庞大的数据源中迅速挖掘出对决策有用的审计信息,并从多层面探索有效的审计新思路、新方法,在不断演进的过程中发挥审计的"免疫系统"功能.大数据时代给出了"样本=总体"全数据模式,使全覆盖审计成为可能.审计可以依照法律权限采集各公共管理部门数据及社会公开的海量数据,利用跨领域、跨层级、跨行业、跨系统的全维度数据,开展智能化的数据挖掘与分析,进行综合审计判断,形成审计结论.这样的大数据审计,不仅是技术方法层面的创新,更是审计理念、审计制度、审计方式的变革.

本文基于我国现阶段审计信息化建设现状,梳理了电子数据审计发展脉络、总结了大数据技术在审计电子数据的采集、存储、分析和可视化工作中的相关研究,探讨了大数据环境下审计面临的新机遇、新挑战,对未来的大数据审计研究方向进行了探讨和展望,为进一步研究作参考.

2 审计信息化发展历程

信息化背景下的审计技术总体概括起来可以分为三个阶段:"计算机辅助审计→联网审计→大数据审计".第一阶段,计算机辅助审计主要是将计算机作为一种工具为审计服务,针对各种类型的财务数据报表进行处理(尤其处理在 EXCEL 报表方面),进行计算、分类、汇总等;根据审计流程,以人工审计为主,凭借审计人员的"经验",计算验证为辅的审计方式,起到一种辅助计算作用.目前,在县市一级的基层审计局依然以这种模式为主.第二阶段,为了探索适合我国国情的联网审计实施方案以及一些数据采集与处理方法,国家审计署已经成功开展了"金审工程"一期和二期的建设工作.具有典型代表为:"现场审计实施系统(Auditor Office)"、"基于平台生长的审计信息系统"等,主要作用对被审计单位的数据信息进行采集与清洗,以便实时审计,在线审计,审计预警.使得审计方式由"事后审计"转变为"事前审计"、"事中审计".目前,联网审计在省市一级的住房公积金及医疗保险基金领域应用较好.第三个阶段,面对日益增长的"海量"大数据,世界各国开始建立一些大数据平台和项目用来提高政府管理能力.国务院印发《关于加强审计工作的意见》,第 19 条明确提出:探索在审计实践中运用大数据技术的途径,加大数据综合利用力度,提高运用信息化技术查核问题、评价判断、宏观分析的能力.这是国家首次在文件中将大数据审计列入审计信息化工作重点.

2.1 计算机辅助审计

Alali[1] 等人最早提出 CAATTs(Computer Assisted Audit Tools and Techniques,计算机辅助审计工具与技术)这一术语.Robert[2] 认为计算机辅助审计技术是指在帮助完成审计的过程中使用任何技术.中国审计署将计算机辅助审计技术定义为"审计机关、审计人员将计算机作为辅助审计的工具,对被审计单位财政、财务收支及其计算机应用系统实施的审计.帮助审计人员收集审计证据、提高审计效率和降低审计风险".具体流程是根据审计任务的需要,利用审计软件采集电子数据,然后对这些电子数据进行预处理并完成数据分析得到审计证据.审计软件主要包括通用数据分析软件及专业审计软件,这些软件一般具有数据采集和分析功能.通过数据采集将被审计单位的电子数据导入到审计软件的数据库中,并利用数据抽样、统计概化、数据查询、异常检测等方式发现审计线索,最终提交审计部门取证形成审计结论[3].相比于手工审计,计算机辅助审计可有效扩大审计面、提高审计效率.但也存在一定局限性.如对于显式违规活动有效;对于更为复杂与隐蔽的电子数据分析比较低效甚至无效;对于审计中存在的信息孤岛无从下手,缺乏对于各自独立数据的关联考虑;进行电子数据采集耗时耗力,无法跨地区、跨行业审计;比较依赖小样本经验地毯式排查,效率低.

2.2　持续审计与联网审计

网络信息技术的快速发展使得审计向持续、动态、实时方向发展,持续审计(Continuous Auditing,简称CA)成为审计的一个重要发展方向[4,5]. Alexander 认为,持续审计是指使用在线计算机系统将审计部门和被审计部门连接起来,使得相关事件发生的同时或之后很短的时间内就能产生审计结果的一种审计类型[6]. 而联网审计则成为实现持续审计的主要方式.

我国政府"金审工程"始于 2002 年,主要通过"预算跟踪+联网核查"审计模式,逐步实现"三个转变"即从单一事后审计转变为事后审计与事中审计相结合,从单一静态审计转变为静态审计与动态审计相结合,从单一现场审计转变为现场审计与联网审计相结合[7]. 相比于计算机辅助审计模式,联网审计的主要优点如下:(1)当被审计单位的业务复杂、数据量急剧扩大情况下,可有效提高审计效率;(2)减少了审计人员差旅和驻地的时间与经费成本;(3)通过组成一个单位的内部信息网,可以将被审计单位的各个部门联系在一起,将分散的信息集中归拢,提高审计数据采集和分析效率;(4)能够在动态的监督中关注资金与项目的效益,能够及时、准确地为决策部门提供决策信息,从而提高审计质量,充分发挥审计在经济监督中的作用.

2.3　大数据审计

大数据审计定义　通过大数据技术手段采集审计证据,对被审计单位的经营、财务、管理等各类数据的真实性、可靠性、有效性和安全性进行综合审查与评价活动[8]. 审计领域的大数据包括结构化、半结构化和非结构化数据,具备海量、多样等基本特征. 但大数据审计不仅仅是汇聚集中大量数据、运用先进技术方法,它更是审计工作在新形势新环境新要求下的全新体现,始终和最高审计机关全面履职尽责密切相关,体现出更为鲜明的具体特征.

大数据审计特征　可以总结为 6M,即多对象(Multi-agency)、多目标(Multi-objective)、多关系(Multi-relationship)、多时点(Multi-timepoint)、多工具(Multi-tool)、多模式(Multi-model). 正是大数据审计所具备的这些特点,促进审计工作从样本向总体转变,从局部向整体转变,从微观向宏观转变,从事后向事中、事前审计转变.

大数据审计研究现状　大数据研究也同样给学术界带来了巨大挑战和机遇,并已成为学术研究热点.《Nature》和《Science》等刊物相继出版专刊探讨大数据研究[9~11]. 紧随着全球大数据研究热潮,国内外实务界和学术界也开始关注大数据在审计中的应用. Earley[12]分析了大数据技术给审计工作带来的机遇和挑战. Yoon[13]认为大数据因其充分性、可靠性和关联性等特点,将成为传统审计取证方式的有力补充. Juan Zhang[14]等分析了大数据与当前持续审计数据分析的能力在数据一致性(consistency)、完整性(integrity)、聚合性(aggregation)、识别性(identification)和机密性(confidentiality)等方面存在的鸿沟. Appelbaum 等认为[15]现代审计管理需将大数据与复杂商务分析方法相融合以产生更具预测性的决策. 国内方面,大数据审计的相关研究始于 2013 年. 刘碧湘[16]首先讨论了大数据对计算机审计的挑战,并展望了如何利用大数据推进计算机审计的发展. 随后,若干文献[17,18]探讨了大数据环境下审计思维模式、审计技术方法、审计人才培养与管理模式等方面的发展建议. 郑伟等[19]分析了大数据环境给数据审计模式带来的影响和改进可行性,并从逻辑流程、网络架构和应用架构等角度对数据审计模式进行完善性设计及应用指标设计. 可以看出,国内外学术界对大数据在审计行业中应用的研究日渐丰富和细化.

此外,世界审计组织数据工作组显示:中国审计署开展了跨行业、跨领域、跨部门的数据比对和关联分析,提出了"集中分析、发现疑点、分散核实、系统研究"的大数据审计工作模式,以及"中央到省市的纵向关联,一级、二级预算单位的横向关联,财政、金融、企业的数据关联,财政与其他多部门、多行业的数据关联,财政数据与业务数据、宏观经济数据的关联"五个分析要求,并在企业审计、金融审计、资源环境审计中多次运用了云计算、智能挖掘、社交网络、自然语言理解、可视化、词云分析和地理信息技术等大数据分析技术. 美国审计署采用了多种新技术进行非结构化数据的分析以及网页数据挖掘. 在具体审计实践中,通过将社会死亡人员名单与领取联邦补贴人员名单相关联,发现了潜在的欺诈行为. 奥地利审计法院在进行大数据分析时,主要采用 R 工具,开展了"对 2100 个奥地利社区多年来债务情况进行监测"、"通过相关性排序,确定要审计的社区名单"、"社区财务关键数据分析"、"医疗保险资金流审计"等项目. 英国审计署的大数据分析侧重于增加价值、降低成本,主要采用统计分析、机器学习、文本挖掘和可视化等技术,通过创新自动化的审计模式来降低流程成本、减少时间、提高效率. 芬兰审计署在"中央政府的预算和其他财务数据审计"中尝试使用可视化技术,以及正在利用大数据技术寻找试点审计主题. 还有印度尼西亚、厄瓜多尔、巴西、印度、泰国等国家最高审计机关也尝试了一些大数据审计技术方面的实践. 如印度尼西亚审计委员会使用 crisp-dm 跨行业数据挖掘标准流程,将审计业务分为商业理解、数据理解、数据准备、建模、评估、部署六个步骤,并尝试使用了预测分析、仿真技术、文本和多媒体分析等技术;厄瓜多尔审计署

开发了诸如“家庭地址查询”、“负债证明”等应用程序;巴西联邦审计法院开发了“财政风险分析模型”;印度审计署使用可视化技术构建了审计计划的标准模型;泰国审计署开展了 IT 审计等.

3 基于大数据的审计技术

大数据审计是随着大数据技术的发展而产生的一种新的审计方法,其内容包括大数据环境下的电子数据审计(如何利用大数据技术审计电子数据、如何审计大数据环境下的电子数据)和对大数据环境下的信息系统进行审计两个方面的内容,其中大数据环境下的电子数据审计是研究的热点[20]. 基于大数据的审计技术总体框图如图 1 所示.

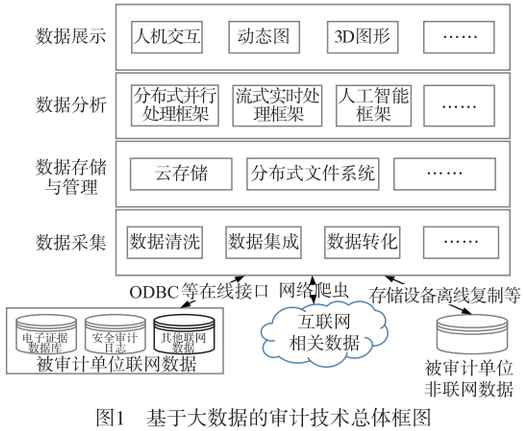

图1 基于大数据的审计技术总体框图

大数据审计是基于电子数据的审计,因此首先是通过数据采集获得尽可能全而真实的审计数据. 在该阶段通常需要根据审计目标,从被审计单位的联网系统、离线存储数据以及互联网中,利用 ODBC 数据库接口、网络爬虫等技术,获得尽可能全相关审计电子数据,然后利用数据清洗、数据集成、数据转换等数据预处理方法,保证电子数据的质量. 其次为充分利用大数据的优势,采集的审计数据通常体积较大,因此,利用云存储、分布存储等技术进行数据存储和管理就成为需要考虑的重要方面. 还有对于存储的大批量数据,为提取出对审计目标有价值的信息,只有利用分布式并行处理、流式实时处理以及人工智能等大数据分析手段,才能及时准备的完成数据分析,获得可靠的审计结果. 最终通过数据可视化,将分析的结果以友好的方式展示给审计员,使审计员能够根据分析结果获得直观的认识,便于审计结果的快速确认,提升审计效率.

本文将首先对比大数据审计方法与现有电子数据审计方法的区别,以理清大数据审计的特点,然后将从审计电子数据的采集、存储、分析和可视化四个方面,从技术层面对现有的大数据审计进行分析总结,以为后续相关研究奠定基础.

3.1 大数据审计方法与电子数据审计方法比较

电子数据审计的数据分析技术,主要依据数据分析模型实现审计疑点挖掘和审计线索发现. 一般来说,数据分析过程主要包括采集、清洗、整理、挖掘和可视化. 传统环境下,常用方法包括帐表分析、数据查询、统计分析、审计抽样和数值分析等[4]. 在这类数据分析方法中,Excel、Oracle、AO、ACL、IDEA 等作为主流审计软件被广泛使用. 例如,文献[21]以卫生计生系统中的医院业务数据、财务数据为背景,利用 SQL 数据库和 Excel 函数实现系统对账审计、重号查找审计、断号查找审计、班福(Benford)定律审计这四种审计方法. 文献[22]针对查找潜在错报问题,利用 Excel 的回归分析功能预测错报数据与真实数据的偏差,以提高审计效率.

大数据时代的审计工作往往涉及国民经济运行中的各种数据,这些数据跨行业、跨领域、跨部门,天然的具有大规模、多样化、高价值、低密度等大数据特征. 从数据类型上看,既包括结构化数据、半结构化数据,也包括如文档、图像、视频等非结构化数据. 从数据来源看,既包括被审计单位的财务业务数据、相关支撑资料、也包括来自互联网的公开数据. 从应用目标上看,审计目标逐步转向发现线索、评估风险、关注效益等方面. 审计工作不仅需要发现违法违规问题,更需要揭示管理制度方面存在的问题,评估内控风险,通过对经济社会相关大数据的获取和分析,洞察行业整体走向,探索发展规律,对国家、行业、部门的制度出台与发展策略做出前瞻性思考与战略性分析[17]. 因此,大数据审计的上述特征使其所采用的数据采集、存储、管理、分析挖掘和可视化等方法与传统方法有所不同,本文总结了大数据环境下的审计方法与现有电子审计方法的差异,如表 1 所示.

表 1 大数据审计方法与现有数据审计方法对比

对比条目	大数据审计方法	现有数据审计方法
数据采集	采集的数据源不仅包括被审计单位的结构化、非结构化数据,还包括其他辅助数据,如互联网公开数据、关联部门/单位的数据.	采集的数据源主要为被审计单位的部分数据,以结构化数据为主.
数据存储	大数据环境下表格、文本、图像等各类数据共存,数据类型复杂,往往需要 NoSQL 数据库(如 HBase)实现数据存储.	数据类型以表单为主,一般采用 SQL 数据库,如 Oracle、DB2,或审计软件自带的数据库即可满足存储需求.

续表 1

对比条目	大数据审计方法	现有数据审计方法
数据管理	数据规模大、共享程度高、还有安全性和保密性需求,需要采用专门用于大数据管理的技术框架,如 Hadoop、云存储等;此外,在存储设施、网络架构、访问机制等方面也有相应的要求,以确保数据实时访问、安全可控.	一般情况下,数据量较小,数据类型单一,采用常规服务器或联网数据存储系统管理数据库.
数据分析	以发现审计疑点为目标,传统数据分析方法、大数据挖掘、自然语言处理、模式识别等技术方法综合运用;此外,数据处理的实时性需求也应考虑.	一般采用帐表分析、查询统计、抽样分析等,部分审计任务需考虑聚类、回归预测等数据挖掘方法.
数据可视化	除传统图表方式外,针对数据规模大、分析结果信息量较大的问题,需要更精细的可视化工具,如人机交互图、动态图、3D 图形等.	一般采用审计分析软件,如 Excel 或数据库自带可视化软件,如柱状图、折线图等展示分析结果.

3.2　大数据审计采集技术

电子数据采集决定审计工作能否高效准确,获取真实和完整的电子数据,是开展审计数据分析的第一步,学者们在采集模式和理论方法上作了大量的研究,主要包括两个方面:(1)数据采集和转换技术;(2)对采集的数据的完整性和有效性进行验证.

在数据采集和转换技术上,目前的研究主要是针对特定领域、特定来源的数据构建针对性采集和处理方法. 陈伟[23]分析了目前大数据审计数据采集过程中存在的问题,对比分析了现有网络爬虫的优缺点,提出了面向审计数据采集的网络爬虫技术;陈琦[24]提出一种基于 C# 的审计数据采集方法,分析了如何采用 Visual Studio 2008 实现文本格式数据、Excel 数据和 Access 数据 3 种数据格式的采集方法. 赵华[25]设计一种 Oracle 审计数据采集的方法,并开发出实用的转换工具,该工具能辅助审计人员快速掌握数据含义,有效辅助审计分析. 董海韬[26]为解决互联网上使用安全套接层/传输层安全协议保密的数据难以审计的问题,提出了一种基于中间人原理的安全套接层/传输层安全保密网络数据的明文采集方法. 赖春林[27]以医疗保险基金审计为例,如何有效采集审计电子数据. 王志之[28]从审计获取的原始数据存在的质量问题入手,如何改善审计所采集的原始数据质量,提高数据可分析性.

在数据完整性和有效性验证上,目前的研究方法主要是基于规则的研究和数据传输可靠性保障. 基于规则的研究主要是根据具体的审计目标,结合相应的审计规范和标准,分析审计数据之间规则满足性,以此评估远程数据以及本地数据的完整性和有效性. Colombo T 等人[29]设计了一个大数据采集系统的仿真模型,在此基础上开发了仿真工具,通过比较模拟仿真结果,验证了工具的有效性. 卢学英[30]从数据采集、数据清理、数据转换、数据验证四个阶段,介绍保证电子数据真实完整准确的方法. 数据传输可靠性保障主要通过加密技术保障安全,通过数据重构技术完成缺失数据的修复. 徐超等[31]通过对 DES 和 RSA 加密技术进行分析,设计一种基于 DES 和 RSA 算法的数据加密传输系统,能够适用于不同的应用环境. Lee K M 等人[32]对远程采集和存储的流数据进行远程数据完整性检查. Fan 等人[33]提出一种在没有时间戳的条件下,使用时效约束查询数据时效的方法,以此来判定实体的最新状态信息,意在解决数据一致性问题. 杜岳峰[34]从数据一致性和时效性着手修复错误数据,提升数据质量,提出了一种基于关联数据的一致性和时效性清洗方法,并采用一种启发式的修复方法对错误进行修复,提高修复的准确性. Papenbrock 等人[35]认为,使用规则约束来验证数据一致性并对其修复是最有效的数据一致性管理方法,提出了一致性规则的发现方法. Wang 等人[36]使用 CFDs 技术解决一致性检测和修复问题. 杨东华[37]认为数据质量问题会对大数据的应用产生致命影响,提出了一种优化技术-基于任务合并的优化技术. 针对冗余计算和利用同一输入文件的简单计算进行合并,分别对实体识别模块、不一致数据修复模块和缺失值填充模块进行了优化. 冉德彤[38]提出了一种基于数据一致性的记录比较方法. 该方法利用条件函数依赖检测数据一致性信息,基于该信息计算属性相似度,并与传统方法的结果相结合,完成记录比较. 此外,最近研究者针对采集系统本身的可靠性进行了研究,Samtani S 等人[39]提出一种文本挖掘的方法来评估数据采集系统的脆弱性,该方法确定了 55000 多个数据采集系统存在漏洞.

3.3　大数据审计存储技术

在大数据审计环境中数据存储系统主要包括传统关系型数据库、新型 NoSQL 数据库和分布式文件系统等. 由于审计中经常获得的是金融机构、政府部门、企事业单位的敏感数据,审计电子数据不但要保证存储的效率,更要保证数据存储的完整性,避免其被恶意获取和篡改,因此,审计电子数据将广泛使用云存储技术. 云存储技术是通过分布式文件系统、集群等技术,将网络中大量不同类型的存储设备协同起来,共同向外提供数据存储和数据访问服务的技术. 它的高可用、低成本、高性能,是实现多源异构大数据高效存储的不二选择,其中,多副本技术和数据完整性验证[40]技术是

云存储中的两大关键技术.

(1)多副本技术

保证数据可用性,避免因磁盘故障等原因导致数据丢失是大数据存储首先需要解决的问题. 数据可用性问题可形式化表示为如下形式:设集合 D 的数据一致性、精确性、完整性、时效性和实体同一性分别为 Q_1, Q_2, Q_3, Q_4 和 Q_5,则数据可用性可以定义为:

$$usability(D)=\delta_1 Q_1+\delta_2 Q_2+\delta_3 Q_3+\delta_4 Q_4+\delta_5 Q_5$$

其中,δ_1, δ_2, δ_3, δ_4 和 δ_5 是由用户根据实际需要确定的权值,且 $\delta_1+\delta_2+\delta_3+\delta_4+\delta_5=1$.

多副本技术是提升数据可用性的关键技术,它的基本原理如图 2 所示.

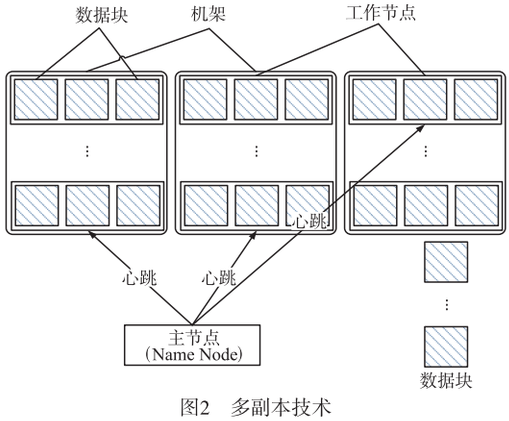

图2 多副本技术

它通常设置有一个主节点,用于维护副本间的数据一致性以及副本内数据可用性的管理. 数据存放在工作节点中,并以多副本的方式保存. 主节点和工作节点之间通过心跳包通信. 当主节点发现其中一个副本出现故障时,则自动进行数据重分布,以尽快恢复故障副本. 多副本技术由于提供了多个副本供系统使用,系统可以将同一数据的访问分散到多个副本中,在一定程度上缓解了数据访问压力,提升了系统性能,因此目前 GFS[41]、HDFS[42] 等云存储中的主流分布式文件系统,均使用了该技术. 但由于该方法是以牺牲存储容量为代价的(通常设置 3 个副本),因此存储效率较低,成本高,维护开销大.

(2)数据完整性验证技术

可证明数据持有(PDP 模型)是典型的云存储数据完整性验证方法. PDP 模型是由 Ateniese 等于 2007 年基于同态特性提出的方法[43,44],其基本原理如图 3 所示. PDP 模型包括两个阶段:预处理阶段和验证阶段. 在预处理阶段,用户根据输入文件 F 提取元数据 m 和修改后的文件 F',本地保留元数据 m,服务器端存储修改后的文件 F'. 在验证阶段,由用户生成随机挑战 R 给服务器,服务器结合存储的文件 F',计算获得文件完整性证据 P,传输给用户. 用户将获得的 P 以及预处理使用的元数据 m,计算该证据 P 的有效性,从而验证服务器端存储的文件完整性.

为实现以上两个阶段的操作,PDP 模型包含 4 个基本算法,为保证效率,这些算法都是多项式时间算法.

(a)用于生成公私密钥对的密钥生成算法,其中 l^k 为安全参数,$KeyGen(l^k)\rightarrow(pk,sk)$.

(b)用于计算 m 的元数据生成算法,$TagBlock(pk,sk,F)\rightarrow(m,F')$.

(c)用于计算证据 P 的算法,其中 θ 表示与挑战 R 相关的原文件内容及其元数据集合,$GenProof(pk,\theta,R)\rightarrow(P)$.

(d)用于验证服务器数据完整性的算法,$CheckProof(pk,sk,R,P,m)\rightarrow\{0,1\}$.

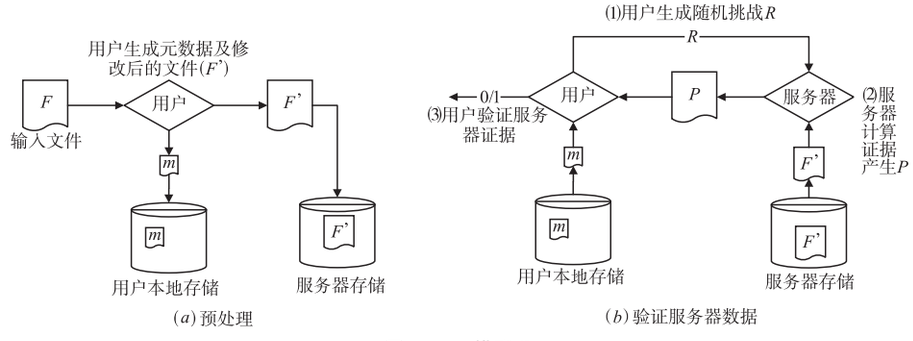

图3 PDP模型[43]

PDP 模型仅能验证数据是否完整,但无法修复错误的数据,这对于弥补损失毫无帮助,因此,研究者们又设计了可恢复证明(POR 模型[45])方案,POR 模型基于纠删码原理实现数据恢复. 随后在文献[46]中,Ateniese 等人对模型进行了改进,提出了可扩展 PDP 模型,以支持动态操作,但该方案近支持数据的更新、删除和添加,并不支持数据块插入功能. 为解决数据块插入问题,Erway 等人[47]首次探索动态 PDP 方案架构,提出了基于等级的认证跳表(RASL)概念,并在之前的可扩展 PDP 模型基础上废除了标签索引信息以实现支持包括

数据插入的完全数据更新操作 PDP 方案,但该方案不支持批量审计和数据隐私保护功能. 由于 PDP 方案仅支持验证数据的完整性,不能保证数据的可恢复性,Juels 等人[48]提出了一个专注于大文件静态存储的可恢复证明(PoR)方案,但挑战次数有限. Shacham 等人利用 BLS 签名在文献[49]中提出了一个支持公开审计的紧凑 PoR 方案,但该方案只支持静态数据且容易泄露用户隐私信息.

为解决动态审计和隐私信息泄露问题,Wang 等利用同态可验证标签技术和数据分段技术先后在文献[50,51]中引入 TPA,提出了云数据完整性公开审计方案. 他们在文献[51]中提出了在分布式情况下考虑动态数据存储且能定位错误数据的云存储安全性方案;后在文献[52]中引进 Merkle 哈希树(Merkle Hash Tree,MHT)提出了新的改进方案,其可应用在 PDP 或 PoR 方案中,支持完全数据更新操作,且 TPA 能高效进行数据完整性批量审计,并在文献[51,53]利用同态密钥随机掩码技术解决了新的隐私泄露问题. 文献[54]在文献[52]基础上通过改进 MHT 实现了支持细粒度更新数据的功能,文献[55]则将多个副本的 MHT 合并成一个 MHT 以提高效率;但这些方案都会带来 CSS 在数据不完整的情况下仍可伪造应答证据欺骗审计的安全问题. 文献[56]针对云数据完整性公开审计中隐私泄漏给第三方审计者(TPA)以及云存储服务器(CSS)发起替代攻击的问题,提出一种面向公有云的数据完整性公开审计方案,解决了现有方案隐私问题及攻击问题,且在计算开销、存储开销和通信开销方面的性能不会有数量级变化.

3.4　大数据审计分析技术

大数据分析中常用的数据挖掘和机器学习方法是实现"验证型审计"转向"发掘型审计"的重要手段. 其中,云计算架构是实施大数据审计的主要架构. 根据审计应用场景的不同,主要可分为三种架构:批处理架构、流处理架构、混合处理架构.

(1)批处理架构　批处理云计算架构通过将无依赖关系的大批量数据分为多组小批量数据,每组数据分布在不同的地方同时处理,实现数据的分布并行处理. 批处理架构具有高吞吐率,主要应用于事后审计,也是目前使用最广的一种架构,MapReduce[57]是其中最典型的一种.

(2)流处理架构　流处理云计算架构与批处理架构不同,它将数据看成像"水龙头"出水一样源源不断的到来,对于每受到的"一滴"数据,就分配相应的任务进行处理,处理的过程也是流水式的. 流处理架构如 Storm[58]主要用于实时性要求高的场景,通常可达到秒级甚至毫秒级,主要用于实时审计. 由于实时性的要求,流式处理大部分的结果均在内存中执行,并不保存到磁盘.

(3)混合处理架构　为结合批处理的高吞吐率和流处理的高实时性,基于流式处理和批处理的混合架构逐渐受到人们的重视,是未来审计应用的重要方向. 其中,Spark[59]是目前应用较为广泛的混合处理架构. Spark 将数据组织成 RDD(Resilient distributed datasets,弹性分布式数据集)的方式,所有的操作都基于 RDD 进行,其处理流程同 MapReduce 框架十分类似,但为提升效率,中间结果可以只保存在内存,不用写入磁盘. 同时,为适应流处理需求,Spark 推出了 Steaming 版本,它将输入数据流以时间片(秒级)为单位进行拆分,然后以类似批处理的方式处理每个时间片数据.

基于云计算架构,审计分析的效率可以获得极大的提升,但要获得有效的审计分析结果,还需要借助大数据挖掘技术. 大数据挖掘的目标与传统数据挖掘的目标是类似的,都是从大量复杂数据中提取对挖掘目标有价值的信息. 面向大数据的数据挖掘方法是在传统数据挖掘算法基础上发展起来的,但由于大数据挖掘的数据来源多、数据量大、数据类型复杂、数据价值密度低,面向大数据的数据挖掘方法具有自身的一些特点:(*a*)大数据挖掘算法充分利用分布式并行处理技术,通常都建立在 MapReduce、Spark、Storm 等云计算框架基础上;(*b*)大数据挖掘算法对半结构化数据、非结构化数据有较大的考虑,数据来源通常是 HBase[60]、Redis[61]、MongoDB[62]、LevelDB[63] 等 NoSQL 数据库;(*c*)大数据挖掘算法可以通过损失部分精度而追求更高的挖掘效率.

由于数据上的很多操作可以用数据库查询原语来表达,因此,为便于将传统数据挖掘算法应用到大数据挖掘,一些研究者对关系代数中标准运算的 MapReduce 映射方法进行了探讨.

设关系模式 $R(A_1,A_2,\cdots,A_n)$,$M(B_1,B_2,\cdots,B_n)$,R 中的单个元组 $t_k(a_1,a_2,\cdots,a_n)$,$k=1,2,\cdots,m$,R 中元组的集合为 T,M 中元组的集合为 T'.

(*a*)选择运算 $\sigma_C(R)$. Map:$T\rightarrow\{t_i,t_i\}$,$t_i\in T$ && $t_i\overset{C}{\Rightarrow}true$. Reduce:$\{t_i,t_i\}\rightarrow\{t_i,t_i\}$.

(*b*)投影运算 $\pi_S(R)$. Map:$T\rightarrow\{t_k'(\{a_i\}),t_k'(\{a_i\})\}$,$a_i\in S$,$t_k'\in R'(\{A_i\})$,$k=1,2,\cdots,m$. Reduce:$\{t',[t',t',\cdots,t']\}\rightarrow\{t',t'\}$.

(*c*)并运算 $R\cup M$. Map:$\{T,T'\}\rightarrow\{t_i,t_i\}$,$t_i\in T\cup T'$. Reduce:$\{t_i,t_i\}\rightarrow\{t_i,t_i\}$.

(*d*)交运行 $R\cap M$. Map:$\{T,T'\}\rightarrow\{t_i,t_i\}$,$t_i\in T\cup T'$. Reduce:$\{t_i,x\}\rightarrow\{t_i,t_i\}$,$x=[t_i,t_i]$.

(*e*)差运算 *R-M*. Map:$T\rightarrow\{t_i,1\}$,$t_i\in T$ $T'\rightarrow\{t_j,2\}$,$t_j\in T'$. Reduce:$\{t_i,x\}\rightarrow\{t_i,t_i\}$,$x=1$.

(*f*)自然连接运算:$R(A,B)$与$M(B,C)$进行连接. Map:对于R中的每个元组(a,b),生成键-值对$(b,(R,a))$,对于S中的每个元组(b,c),生成键-值对$(b,(S,c))$. Reduce:每个键值b会与一系列对相关联,这些对的形式要么是(R,a),要么是(S,c). 基于(R,a)和(S,c)构建所有的对. 该键及其值表的输出结果是一系列键-值对序列,每个值为三元组(a,b,c),其中对应的(R,a)和(S,c)处于输入的值表当中.

目前,大数据挖掘技术主要以机器学习方法为主. 机器学习是人工智能领域的重要分支,它基于概率论、统计学、逼近论、凸分析等多种理论,专门研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能. 一些云计算框架已经集成了机器学习算法库,其中比较典型的有 Spark 平台的 MLlib 库[64]. 机器学习算法按其用途,通常可以分为如表 2 所示.

表 2 机器学习常见算法分类

类别	功能	典型算法
聚类算法	对用例集按照某种距离维度聚成多个“簇”	k-Means 算法、CURE 算法、GRGPF 算法等
回归算法	量化因变量受自变量影响的大小,建立回归方程预测模型	最小二乘法,逻辑回归,逐步式回归,多元自适应回归样条等
决策树学习	根据数据的属性采用树状结构建立决策模型	CART、ID3、C4.5、随机森林、多元自适应回归样条等
贝叶斯学习	基于贝叶斯定理的一类算法,主要用来解决分类和回归问题.	朴素贝叶斯算法、平均单依赖估计、Bayesian Belief Network (BBN)等
基于核的算法	把输入数据映射到一个高阶的向量空间进行求解	支持向量机(SVM)、径向基函数(RBF)、线性判别分析(LDA)等
关联规则学习	通过搜索最能够解释数据变量之间关系的规则来确定有用关系	Apriori 算法、Eclat 算法等
深度学习	过建立具有阶层结构的人工神经网络(Artifitial Neural Networks,ANNs),在计算系统中实现人工智能	受限波尔兹曼机(RBN)、Deep Belief Networks(DBN)、卷积神经网络(CNN)、递归神经网络(RNN)等
集成学习	用一些相对较弱的学习模型独立地对同样的样本进行训练,然后把结果整合起来进行整体预测	Boosting、Bagging、AdaBoost、堆叠泛化、梯度推进机、随机森林等

对于具体的大数据审计数据挖掘分析算法的应用,很多研究者已经进行了深入剖析. 文献[65]中明确指出了银行监管数据挖掘的实施步骤,通过借鉴国外银行的成功经验,阐明数据挖掘技术在数据分析中的应用,并以某市信用社 1998 年至 2005 年不良贷款额与贷款余额两者之间的定量关系为依据进行数据挖掘,从设定对象、目标,到具体实施,最后总结分析给出了一个完整的应用案例. 文献[66]探讨了数据挖掘方法,如数据概化、统计分析、聚类分析、关联分析、预测分析等,在审计业务中的应用场景及实施步骤. 文献[67]通过分析时间序列数据特点,提出了去峰值的显著连续序列数据发现方法,提高可疑数据的发现效率,并基于 HowNet 改进语句语义相似度算法实现审计规则提取. 文献[68]尝试采用模糊神经网络与遗传算法相结合的办法解决审计数据的总体分析及审计规则提取问题. 文献[69]提出一种基于密度的增量式离群点识别算法并应用于社会保障数据审计中. 文献[70]提出一种基于模糊匹配的审计方法,利用字段相似检测算法执行数据表中各字段模糊匹配查询,获得可疑数据. 文献[71]以国家审计保险审计工作为背景,讨论文本特征化表示方法,并结合审计知识库的语义关系,提出基于审计知识库的文本关联分析方法. 文献[72]对于 301 篇涉及审计分析性程序文献的分析,总结了这些文献使用的机器学习算法类型及频度,如图 4 所示.

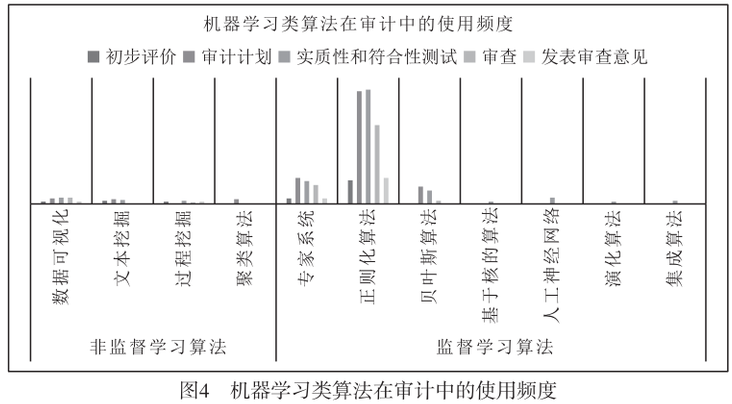

图4 机器学习类算法在审计中的使用频度

由图 4 可以看出,监督学习算法使用的频度占绝对优势,这与审计是针对性任务有关;其次,从审计的各个阶段可以看出,目前使用的主流分析技术是正则化算法和专家系统,对于目前新兴的神经网络类算法使用较少,大数据分析技术在审计中的应用还有很大的探索空间.

3.5　大数据审计可视化技术

验证型审计方式在审计宽度、审计深度方面都面临较大风险,可视化分析技术是实现审计工作向发掘型审计方式转变的可行途径[73]. 数据可视化技术借助图形化的手段,以更简洁清晰的方式表达海量被审计数据信息中内在因素间的关联关系,可以帮助审计人员从海量数据中快速发现审计疑点,提高审计效率. 随着数据量的快速增长,数据可视化技术已经成为人们快速掌握复杂信息的关键技术[74,75]. 目前,大数据审计可视化技术主要可分为文本可视化、网络可视化和时空数据可视化、多维数据可视化等,其主要方法如表 3 所示.

表 3　可视化方法

类别	功能	典型方法
文本可视化	将文本中的词频与重要度、逻辑结构、主题聚类、动态演化规律等语义特征直观地展示出来	标签云方法、Tile Bars 方法、Word Tree、FP-Tree、ThemeRiver、EventRiver 等
网络可视化	基于网络节点和连接的拓扑关系,直观地展示网络中潜在的模式关系	H-Tree、圆锥树、气球图 Balloon View、树图技术 Treemaps、Voronoi 图填充、TreeNetViz 等
时空数据可视化	对时间与空间维度以及与之相关的信息对象属性建立可视化表征,对与时间和空间密切相关的模式及规律进行展示	流式地图 Flow map、时空立方体等
多维数据可视化	对具有多个维度属性的数据变量进行展示	散点图、投影、平行坐标等

文献[76]提出一种基于集合评审技术(PERT)的审计方案可视化建模方法,并在开源软件 Gnome Planner 的基础上实现了该方案. 该方法能够能按照审计方案中各审计任务的内在联系,用箭头来表示其先后顺序,画出一个各项任务相互关联的网络图,以便使用者对全局有一个比较完整的概念,从而找到关键环节和主要矛盾所在,相比较于一般的审计方案的文档描述来看更加直观.

文献[73]在评估某市教育资源库建设项目的使用绩效时,采用可视化技术,从不同维度构建学年访问量变动图,很直观的发现某段时间内访问量明显较低,从而发现了日志被人为关闭的问题. 文献[77]基于 Tableau 展示了数据可视化技术在电子数据审计中的应用及优势. 可视化建模工具可以帮助审计人员对同一数据从网络、时间、空间和统计等多个不同角度建立分析视图. 例如,审计人员利用 IBMI2 的链接分析、路径分析、群集分析、社会网络分析等可视化分析算法和工具,可以直观地展示图形中数据之间显示和隐式的关联关系、时间关系和空间关系,这对寻找审计思路、发掘审计线索意义重大.

3.6　基于区块链的大数据审计

在大数据审计中,被审计单位所提供电子数据的真实性、正确性和完整性是事关具体审计业务走向的基本条件,对电子数据等会计信息进行必要的辨别和界定是审计人员完成审计项目的基本环节,也是确保审计质量的基础. 大数据审计的基本流程是获取必要和充分的信息,建立被审计资料数据库,建立审计中间表、分析数据、延伸查实和审计取证,采集、转换、清理电子数据是开展大数据审计工作的基础性工作,电子数据的质量直接影响审计目标实现. 因此,能够获取完整的、一致性的、可追溯的审计电子数据尤为重要. 区块链技术作为十三五中的关键技术,融汇吸收了分布式架构、块链式数据验证与存储、点对点网络协议、加密算法、共识算法、身份认证、智能合约、云计算等多类技术,具备去中心化、保障数据完整性、透明性、不可撤销性等特征,对于解决审计过程中电子数据的不可靠因素具备独特优势,文献[78]以区块链技术为支撑,探讨审计电子数据采集及可信性确认机制、一致性传输及可追溯存储机制等可靠性与可追溯性加强方法的实现机制. 认为基于区块链的大数据审计主要可从以下几个方面开展研究.

(1)在审计电子数据采集的过程中,因大数据环境下审计电子数据多而复杂,涉及多方主体,采集的电子数据通常夹带了大量噪声,其完整性、真实性很难保证. 基于区块链技术,可以将各个审计节点单位的每笔待审计记录进行自动采集,并将其定时分批加密进行广播. 同时,借助分布式节点自动解密和验证机制,对每笔待审计记录涉及的关联方进行交叉确认. 经过确认的审计记录数据添加时间戳和加密机制后被确认加入区块链,无法再进行修改. 与传统审计数据采集方法相比,基于区块链技术的审计数据采集方法将具有质量高、实时、无法篡改、可追溯的特点,将大大减少审计过程中的低层次重复劳动,解决审计电子数据采集不完整、不真实,难以应用于全覆盖审计分析的问题.

(2)在审计电子数据传输中,可以借助区块链的时间戳机制,以每组数据传输前后的时间戳为主要参数,结合审计电子数据传输方式、数据量、数据重要程度、可恢复能力等数据相关信息,构建数据一致性评估模型,为数据传输过程中的可靠性评估提供依据.

(3)对于存储的审计电子数据的安全性和可靠性,以区块链的多副本共识技术为基础,将审计电子数据以多副本的方式分布式存储,并根据存储的地点、安全等级、管理权限等多方面的因素,对其存储的副本进行本地评估,最后将所有副本的评估结果进行综合,构建该副本当前的存储有效性评估模型.然后以此为基础,对该电子数据的有效性进行评估,然后结合评估值及该数据的使用范围确定可用性,避免错误数据进入审计系统,导致重大审计风险.同时,基于区块链技术对审计数据进行组织,并根据审计电子数据的类型、属性等多个因素,以 B+树等方式构建多级索引.基于该索引及区块链的链式结构,设计数据修改记录的快速追溯跟踪方法,一方面追溯存储不可靠原因,另一方面保障重要数据的可重构性,为实现快速可跟踪审计奠定基础.

3.7 大数据审计技术具体应用

(1)大数据审计辅助蓝天保卫计划

审计场景 蓝天保卫战三年行动计划是中国政府部署的一项污染防治行动计划,旨在持续改善空气质量,为群众留住更多蓝天.习近平总书记在全国生态环境保护大会上强调,把解决突出生态环境问题作为民生优先领域,坚决打赢蓝天保卫战是重中之重.国家审计署也要求各地要聚焦打赢蓝天保卫战,开展生态保护和污染防治相关专项审计.当前,我国大气污染防治形势依然严峻,各级审计机关应从如下多个角度开展大气污染防治审计工作:(*a*)关注各类污染物减排目标及进展情况,推动加强污染物的协同控制;(*b*)聚焦资源能源、重点产业、机动车等领域大气污染防治政策措施落实情况,推动经济高质量发展;(*c*)关注大气污染防治体制机制制度建设情况,推动大气环境治理体系和治理能力现代化;(*d*)关注大气污染与人体健康问题的关联性,推动完善环境与健康政策.

基于大数据技术的审计模型 大气污染物排放清单既是制定污染防治政策的根本依据,也是开展审计工作的重要依据.目前,对主要污染物排放总量、时空分布、行业贡献、减排潜力等信息掌握不足.这不仅影响了大气污染防治政策的制定和实施,也影响着审计思路、审计内容的确定.在大数据背景下,首先对工商部门、发展与改革部、财政、税务等部门的数据进行联网采集,然后对各部门采集的数据进行融合清洗后,建立大数据分析平台,通过关联分析掌握污染型企业清单,并分析得出污染型企业的"数据画像",如图 5 所示.

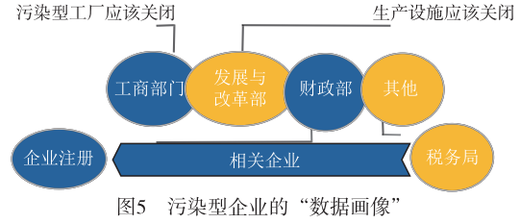

图5 污染型企业的"数据画像"

对污染型企业的"数据画像"进一步的数据分析,从企业注册信息、交税情况、用电度数、员工缴纳社保情况、主要生产设施、财务状况等数据中分析企业的运行情况,从而判断不同区域、不同产业结构且已经被关闭的情况是否仍然继续经营?如图 6 所示.

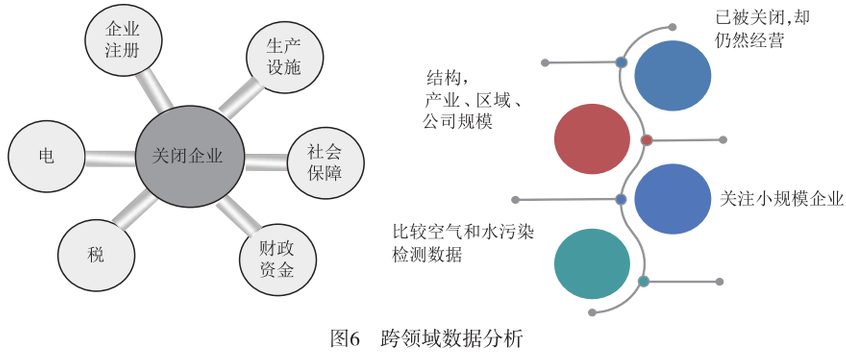

图6 跨领域数据分析

在具体数据分析过程中,主要将电力数据与空气和水污染物的监测数据相结合,如 NO,COD 指数等,通过大数据分析,部分企业的用电量情况异常,夜间耗电量大,而没有排放高污染物,可能有假,存在秘密经营情况.

审计结果 通过多部门大数据分析最终得出,截止 2017 年共四百多家应该关闭的企业没有关闭仍在秘密经营.

(2)大数据审计辅助精准扶贫

审计场景 中共中央发布了一系列扶贫政策法规目标;到 2020 年,所有极端贫困人口都将摆脱贫困,享受必要的社会服务.在过去的五年里,中国政府已经帮

助了6800万人,现在,正努力在未来几年内每年帮助1000万以上的人摆脱贫困.国家审计署提出“关于进一步加强扶贫审计的意见”,所有审计机关都应高度重视扶贫审计,根据国务院的统一安排部署,审计署对全国各地区、各部门贯彻落实国家重大政策措施情况进行跟踪审计,一项重要内容就是对各地区实行精准扶贫情况进行跟踪审计,持续关注了“实施精准扶贫、精准脱贫”重点项目推进情况、资金统筹使用情况、相关政策落实情况,揭示和反映了不作为等问题.

基于大数据技术的审计模型　扶贫审计的重点在于精准识别贫困人口,重点关注贫困人口建档识卡情况、扶贫资金使用管理情况、减贫目标实现情况.首先对民政部门、工商部门、机动车辆部门、公务员管理部门、教育管理部门、社会保障部门等扶贫信息数据进行采集,经过清洗、融合与抽取分析后建立统一的扶贫大数据平台,如表4所示.

表4　扶贫相关数据采集分析表

数据分类	数据信息
民政部门	记录人员基本信息,如姓名,年龄,收入,家庭住址,教育程度等.
工商部门	工商行政管理部门的企业注册数据
机动车辆部门	机动车辆部门的汽车数据
公务员管理部门	公务员管理部门的公务员数据
教育管理部门	教育部门的学生数据
社会保障管理部门	社会保障管理部门的社会保障数据

然后进行大数据分析,建立大数据审计模型,如图7所示.

(*a*)对贫困人员基本信息进行“建档识卡”.在民政部门登记在册的农村五保户信息中,重点抽取人均纯收入2736元以下人员信息,建立贫困人员“数据画像”.

(*b*)关联工商部门数据,分析金融支持力度是否加强?

(*c*)关联车辆部门数据,分析是否有房有车有公职的人员列为贫困人员?

(*d*)关联生态环境部门数据,分析招商引资企业是否造成污染?

(*e*)关联金融部门数据,分析非贫困人员享受扶贫小额贷款?

(*f*)关联教育部门数据,分析贫困人口是否失学辍学、应免未免、应补未补?

(*g*)关联社会保障部门数据,分析社保兜底扶贫,保障贫困人口是否全部纳入新农合保障范围?贫困人口的危房改造资金是否享受?

审计结果　通过对大数据审计分析得出,有两千多人不符合扶贫建档立卡标准,其中还存在不少人已经购买了汽车、住房以及个体工商户等不同类型人员.

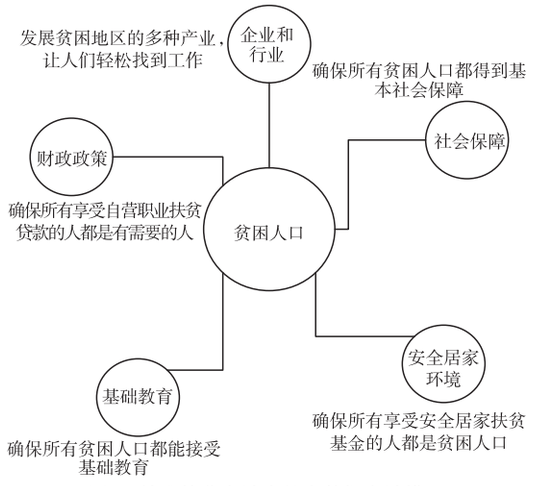

图7　精准扶贫审计中的大数据审计模型

4　国家审计大数据未来趋势

4.1　大数据审计发展中面临的挑战

从大数据技术在审计领域的理论研究和应用实践角度,大数据审计面临如下几个方面的挑战.

(1)尚未形成一套基于大数据技术审计理论体系

近年来,有部分学者在联网审计基础上,提出一些大数据技术的审计发展的影响、建议、规划和方法等[3,79,80].但是这些研究在发展规划上较为宽泛、技术方法上较为零散,尚未形成一套系统科学的研究方法论.

(2)审计大数据存储与共享机制尚不健全

在大数据环境下,为获取全面可靠的审计证据,不但需要采集被审计单位的财务报表、文档资料等数据,还需要采集其他部门的相关数据,甚至需要从互联网上获取公开数据.例如在社保基金使用审计中,不仅需要从人社处采集参保信息,还需辅以公安机关关于人员死亡情况等信息,以便更全面真实地反映审计问题.但是这些数据资源往往较为敏感机密,不宜公开.而采用租赁普通云平台存储管理的方式,若遭到网络恶意攻击将会造成严重后果.因此,需要针对大数据审计的特定需求设计对应的共享和权限管理机制,以有效管理和保护审计数据采集、传输、存储、使用、维护、更新、销毁等全生命周期的安全性和可用性,降低审计风险.

(3)大数据在审计数据分析方面的应用力度不足

尽管大数据分析技术,如自然语言处理、深度学习、人工智能、模式识别等发展如火如荼,新理论、新方法、新框架、新软件不断被提出和开发.但是审计是一

门实践性较强的学科,在现有的审计方法和实践中,最常使用的还是如聚类、关联规则挖掘、异常点检测、回归分析等数据挖掘技术,而更智能化的文本分析、深度学习、知识图谱等方法尚未广泛应用于审计工作中.

(4)云计算技术大数据审计平台还处于探讨阶段

大数据时代下的审计数据采集、存储、分析等环节及信息资源共享均依赖大数据信息平台. 全国各地逐步实现审计信息化,建立起信息管理和分析平台或软件. 大数据时代,一方面,基于关系型数据库的数据分析系统并不适合审计大数据的多源异构特性;此外,数据规模大、共享机制和安全保护机制也对平台提出了分布存储、并行计算、统一调度、安全可控等方面的要求. 而云平台为上述需求提供了解决方案. 目前商用云平台,如亚马逊云、阿里云等已被广泛使用. 但是由于审计数据保密性要求,特别是国家审计中涉及的数据往往与国民经济发展现状甚至国家安全密切相关,依托于云计算供应商的商用云环境使得审计单位和被审计单位都不能清楚自己的数据会被存在什么地方、国内还是国外、如何保护,因而直接购买或租赁这样的云平台回给大数据审计造成潜在风险. 另一方面,目前为止,我国针对审计云平台的构建还处于理论探讨阶段,尚未建成专用于审计领域的云平台. 因此,如何针对我国审计信息化发展现状、未来发展规划,设计和开发适用于审计业务的整体可控的云平台还有待解决.

4.2 未来发展趋势探讨

当前在上述几个方面的研究工作都面临着大数据带来的新问题,也意味着每个方向都有不少新挑战. 展望未来,在大数据环境下,以下几个方面的研究将是问题的核心.

基于大数据技术变革审计实施的方式方法 审计是一个确定性领域,现有的审计方式方法在数据采集、存储、分析应用以及可视化方面都具有相对固定的模式和特点. 在大数据环境下,审计实施的方式方法将会出现较大变革. 如在数据采集阶段,因为审计难以预测,因此传统的过程具有反复而持续的特点,需要根据审计过程获得的数据不断迭代提取新的数据,这不但耗时,也可能因为其中某个过程出现纰漏而导致大量的采集工作无效. 但在大数据的支持下,可以通过多方因素联合分析,快速获得初步的审计结果,以便能够根据审计结果的反馈再次采集针对性的数据,从而减少迭代过程,一次性采集尽可能全而准确的电子证据,有效提升数据采集的效率和质量.

基于大数据技术提升审计的预警监督作用 审计的最终目的是防微杜渐,减少甚至避免不必要的损失. 限于数据的缺失和分析能力的不足,传统的审计主要以事后审计为主. 随着大数据技术的方法应用,能够获得的审计数据将越来越全,数据分析手段和能力将大幅度提升,这将必然促使事前审计的快速发展,审计的预警监督作用将越来越明显.

基于大数据技术提升审计的智能化 传统的审计都是先指定审计目标,然后审计人员根据经验逐步搜集相关证据进行验证复核. 随着大数据技术的发展,审计的智能化程度将越来越高,审计的决策层能够利用大数据分析,更快捷的找到重点问题,制定更有效的审计目标. 审计的执行者能够利用大数据分析技术,制定更符合审计目标的审计流程,更快更好的获得审计结果.

基于大数据技术的审计云平台建设 大数据环境下的审计特性使得传统的审计数据采集、存储、分析等技术方法不再适用. 面对审计大数据时,首先需要构建满足大数据存储和分析的平台架构,实现远程存储、服务弹性、数据整合、信息按权限共享、安全可控、平台/软件高效好用等基本目标. 基于此,设计一个通用的系统架构和一套对应的标准化规范是大数据审计技术发展的首要问题. 为将大数据最新技术服务于审计工作,需要加强对审计数据分析通用模型和审计软件的开发,通过梳理归纳出审计业务的一般流程和核心分析模块,抽象其中共性部分并以软件编程方式实现固化,提高技术泛化应用能力. 为满足数据安全性和保密性需求,构建的审计平台应具有数据安全可控、防止单点故障、抗恶意攻击、容灾恢复、保障各应用系统不间断运行等能力.

5 总结

大数据技术在各行各业的广泛应用给国民经济高速发展和社会和谐稳定带来深刻影响,也是实现审计全覆盖的必由之路. 本文基于我国现阶段审计信息化发展现状,梳理了电子数据审计发展脉络、总结了大数据技术在审计电子数据的采集、存储、分析和可视化工作中的相关研究,探讨了大数据环境下电子数据审计面临的新机遇、新挑战,并展望了未来的主要研究方向. 总之,与计算机辅助审计和联网审计相比,大数据审计在思维模式、技术方法等方面都有显著差异. 尽管目前已有一些探索性研究工作,但总体来说,大数据审计的研究还很年轻,尚有诸多问题亟待解决.

参考文献

[1] Alali A F,Pan F. Use of audit software:Review and survey [J]. Internal Auditing,2011,26(5):29 - 36.

[2] Robert L B,Harold E D. Computer-assisted audit tools and techniques:Analysis and perspectives[J]. Managerial Auditing Journal,2003,18 (9):725 - 731.

[3] 陈伟,SMIELIAUSKAS Wally. 大数据环境下的电子数据审计:机遇、挑战与方法[J]. 计算机科学,2016,43(1):8-13,34.
Chen Wei, Smieliauskas Wally. Opportunities, challenges and methods of electric data auditing in big data environments[J]. Computer Science,2016,43(1):8-13,34. (in Chinese)
[4] 陈伟. 联网审计技术方法与绩效评价[M]. 北京:清华大学出版社,2012. 20-30.
[5] Lambrechts A J, Lourens J E, Millar P B, et al. Global Technology Audit Guide: Data Analysis Technologies [M]. FL:The Institute of Internal Auditors,2011. 15-27.
[6] Alexander K, Ephraim F S, Miklos AV. Continuous online auditing:a program of research [J]. Journal of Information Systems,1999,13(2):87-103.
[7] 陈耿,景波,陈圣国,冯国富. 计算机审计[M]. 大连:东北财经大学出版社,2012. 82-83.
[8] 唐琳,付达杰. 大数据审计内涵特征、现实困境与发展对策[J]. 西部财会,2017,(7):70-72.
[9] Staff, S. Dealing with data. Challenges and opportunities. Introduction[J]. Science,2011,331(6018):692.
[10] 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.
Wang Yuanzhuo,Jin Xiaolong,Cheng Xueqi. Network big data:Present and futrue[J]. Chinese Journal of Computer,2013,36(6):1125-1138. (in Chinese)
[11] Li Xuelong, Gong Haigang. A survey on big data systems [J]. Scientia Sinica Informationis,2015,45(1):1-44.
[12] Earley C E. Data analytics in auditing: Opportunities and challenges [J]. Business Horizons, 2015, 58 (5): 493-500.
[13] Yoon K, L Hoogduin, Li Z. Big data as complementary audit evidence[J]. Accounting Horizons, 2015, 29(2): 431-438.
[14] Zhang J, Yang X, Appelbaum D. Toward effective big data analysis in continuous auditing[J]. Accounting Horizons,2015,29(2):469-476.
[15] D A Appelbaum, M A Vasarhelyi. Big data and analytics in the modern audit engagement:Research needs[J]. Auditing A Journal of Practice & Theory, 2017, 36(4): 1-27.
[16] 刘碧湘. 如何利用大数据推进计算机审计[J]. 科技信息,2013,(21):107-107.
[17] 何琰. 大数据技术在审计中的应用[J]. 郑州轻工业学院学报(社会科学版),2016,17(3):67-71.
[18] 秦荣生. "互联网+"时代的审计发展趋势研究[J]. 中国注册会计师,2016,(1):84-88.
[19] 郑伟,张立民,杨莉. 试析大数据环境下的数据式审计模式[J]. 审计研究,2016,(4):20-27.
[20] 陈伟,居江宁. 大数据审计:现状与发展[J]. 中国注册会计师,2017,(12):81-85.
[21] 王光伟,邬华琼,苏莉民,等. 计算机辅助审计常用的数据分析模型[J]. 中国管理信息化,2016,19(5):44-45.
[22] 王业刚. 利用 Excel 进行审计数据回归分析的探讨[J]. 农场经济管理,2015,(9):34-37.
[23] 陈伟,孙梦蝶. 基于网络爬虫技术的大数据审计方法研究[J]. 中国注册会计师,2018,230(7):78-82.
[24] 陈琦,陈伟. 一种基于 C#的审计数据采集方法的设计与实现[J]. 中国管理信息化,2015,(17):37-39.
[25] 赵华,闵志刚. Oracle 审计数据的采集与转换[J]. 审计与理财,2015,(3):17-18.
[26] 董海韬,田静,杨军,等. 适用于网络内容审计的 SSL/TLS 保密数据高效明文采集方法[J]. 计算机应用,2015,35(10):2891-2895.
Dong Haitao,Tian Jing,Yang Jun,et al. Efficient plaintext gathering method for data protected by SSL/TLS protocol in network auditing[J]. Journal of Computer Applications,2015,35(10):2891-2895. (in Chinese)
[27] 赖春林. 如何有效利用审计采集的数据? -以医保基金审计为例[J]. 审计与理财,2016,(11):17-18.
[28] 王志之. 审计数据预处理探析[J]. 中国经贸,2017,(16):260-261.
[29] Colombo T, Fröning H, Garcìa P J, et al. Optimizing the data-collection time of a large-scale data-acquisition system through a simulation framework[J]. The Journal of Supercomputing,2016,72(12):4546-4572.
[30] 卢学英. 计算机审计中如何获取真实完整的电子数据[J]. 价值工程,2017,36(20):205-206.
[31] 朱作付,徐超,葛红美. 基于 DES 和 RSA 算法的数据加密传输系统设计[J]. 通信技术,2010,43(4):90-92.
Zhu Zuo-fu,Xu Chao,Ge Hong-mei. Design of DES and RSA-based data encryption transmission system design [J]. Communications Technology,2010,43(4):90-92. (in Chinese)
[32] Lee K M, Sang H L. Remote data integrity check for remotely acquired and stored stream data[J]. Journal of Supercomputing,2017,4(9):1-20.
[33] Fan W, Geerts F, Yu W, et al. Conflict resolution with data currency and consistency[J]. Journal of Data & Information Quality,2014,5(12):1-38.
[34] 杜岳峰,申德荣,聂铁铮,等. 基于关联数据的一致性和时效性清洗方法[J]. 计算机学报,2017,(1):92-106.
Du Yuefeng,Shen Derong,Nie Tiezheng,et al. A cleaning method for consistency and currency in related data[J]. Chinese Journal of Computers,2017,(1):92-106. (in Chinese)
[35] Papenbrock T, Ehrlich J, Marten J, et al. Functional dependency discovery: An experimental evaluation of seven

algorithms[J]. Proceedings of the Vldb Endowment, 2015,8(10):1082 – 1093.

[36] Wang J, Tang N. Towards dependable data repairing with fixing rules[A]. ACM SIGMOD International Conference on Management of Data[C]. USA:ACM,2014. 457 – 468.

[37] 杨东华,李宁宁,王宏志,等. 基于任务合并的并行大数据清洗过程优化[J]. 计算机学报,2016,(1):97 – 108.
Yang Donghua, Li Ningning, Wang Hongzhi, et al. The optimization of the big data cleaning based on task merging[J]. Chinese Journal of Computers,2016,(1):97 – 108. (in Chinese)

[38] 冉德彤,游宏梁. 一种基于数据一致性的记录比较方法[J]. 电子设计工程,2018,26(1):66 – 69.
Ran Detong, You Hongliang. A consistency based record compare method in entity resolution[J]. Electronic Design Engineering,2018,26(1):66 – 69. (in Chinese)

[39] Samtani S, Yu S, Zhu H, et al. Identifying supervisory control and data acquisition (SCADA) devices and their vulnerabilities on the internet of things (IoT): A text mining approach[J]. IEEE Intelligent Systems, 2018, (99):1 – 1.

[40] 付艳艳,张敏,陈开渠,等. 面向云存储的多副本文件完整性验证方案[J]. 计算机研究与发展,2014,51(7):1410 – 1416.
Fu Yanyan, Zhang Min, Chen Kaiqu, et al. Proofs of data possession of multiple copies[J]. Journal of Computer Research and Development, 2014, 51(7): 1410 – 1416. (in Chinese)

[41] McKusick, Kirk, Quinlan S. GFS: Evolution on Fast-Forward[M]. GFS: Evolution on Fast-Forward. 2010. 90 – 91.

[42] 黄晓云. 基于 HDFS 的云存储服务系统研究[D]. 大连:大连海事大学,2010. 34 – 45.

[43] Ateniese G, Burns R, Curtmola R, et al. Provable data possession at unfrosted stores[A]. Proceedings of the 14th ACM Conference on Computer and Communications Security[C]. New York:ACM,2007. 598 – 609.

[44] Ateniese G, Kamara S , Katz J . Proofs ofstorage from homomorphic identification protocols[J]. Lecture Notes in Computer Science,2009,(5912):319 – 333.

[45] Bowers K D, Juels A, Oprea A. Proofs of retrievability: Theory and implementation[A]. Acm Workshop on Cloud Computing Security[C]. New York: ACM, 2009. 43 – 54.

[46] Ateniese G, Pietro R D, Mancini L V, et al. Scalable and efficient provable data possession[A]. Proceedings of the 4th International Conference on Security and Privacy in Communication NetWorks[C]. New York: ACM. 2008. 1 – 10.

[47] Erway C, Papamanthou C, Tamassia R. Dynamic provable data possession[J]. ACM Transactions on Information and System Security,2015,17(4):1 – 29.

[48] Juels A, Kaliski B S. Pors: Proofs of retrievability for large files[A]. Proceedings of the 14th ACM Conference on Computer and Communications Security[C]. New York: ACM,2007. 584 – 597.

[49] Shacham H, Waters B. Compact proofs of retrievability[A]. Proceedings of the 2008 International Conference on the Theory and Application of Cryptology and Information Security[C]. Berlin: Springer,2008. 90 – 107.

[50] Wang C, Wang Q, Ren K, et al. Ensuring data storage security in cloud computing[A]. Proceedings of the 2009 17th International Workshop on Quality of Service[C]. Piscataway, NJ: IEEE,2009. 1 – 9.

[51] Wang C, Wang Q, Ren K, et al. Privacy-preserving public auditing for data storage security in cloud computing[A]. Proceedings of the 29th IEEE International Conference on Computer Communications[C]. Piscataway, NJ: IEEE,2010. 1 – 9.

[52] Wang Q, Wang C, Ren K, et al. Enabling public auditability and data dynamics for storage security Transactions on Parallel&Distributed. in cloud computing[J]. IEEE Systems,2011,22(5):847 – 859.

[53] Zhu Y, Ahn G J, Hu H, et al. Dynamic audit services for outsouroed storages in clouds[J]. IEEE Transactions on Services Computing,2013,6(2):227 – 238.

[54] Liu C, Chen J, Yang L T, et al. Authorized public auditing of dynamic big data storage on cloud with efficient verifiable fine-grained updates[J]. IEEE Transactions on Parallel&Distributed Systems,2014,25 (9):2234 – 2244.

[55] Liu C, Ranjan R, Yang C, et al. MuR-DPA: top-down levelled multi-replica Merkle hash tree based secure public auditing for dynamic big data storage on cloud[J]. IEEE Transactions on Computers,2015,64(9):2609 – 2622.

[56] 缪俊敏,冯朝胜,李敏,刘霞. 面向公有云的数据完整性公开审计方案[J]. 计算机应用,2018,38(10):2892 – 2898.
Miao Junmin, Feng Chaosheng, Li Min, Liu Xia. Public auditing scheme of data integrity for public cloud[J]. Journal of Computer Applications,2018,38(10):2892 – 2898. (in Chinese)

[57] Dean J, Ghemawat S. MapReduce: A flexible data processing tool[J]. Communications of the ACM,2010,53 (1):72 – 77.

[58] Requeno J I, Merseguer J, Bernardi S, et al. Quantitative analysis of apache storm applications: The newsasset case study[J]. Information Systems Frontiers,2018,21(1):1 – 19.

[59] Gounaris A, Kougka G, Tous R, et al. Dynamic configura-

tion of partitioning in spark applications[J]. IEEE Transactions on Parallel & Distributed Systems,2017,28(7):1891 - 1904.
[60] Rodek L,Poulsen H F,Knudsen E,et al. A storage model of equipment data based on HBase[J]. Applied Mechanics & Materials,2015,713 - 715(2):2418 - 2422.
[61] Chinnachamy A. Instant Redis optimization how-to learn how to tune and optimize Redis for high performance[J]. Work Employment & Society,2013,28(2):305 - 322.
[62] Chodorow K,Dirolf M. MongoDB:The Definitive Guide[M]. USA:O'Relly Media,2010. 27 - 46.
[63] Lim H S,Kim J S. LevelDB-Raw:Eliminating file system overhead for optimizing performance of LevelDB engine[A]. International Conference on Advanced Communication Technology[C]. Pyeongchang, korea: 2017. 777 - 781.
[64] Meng X,Bradley J,Yavuz B,et al. MLlib:machine learning in apache spark[J]. Journal of Machine Learning Research,2015,17(1):1235 - 1241.
[65] 朱文博. 数据挖掘技术在银行监管工作中的应用[J]. 华南金融电脑,2007,(5):12 - 16.
[66] 吕新民,王学荣. 数据挖掘在审计数据分析中的应用研究[J]. 审计与经济研究,2007,22(6):35 - 38.
[67] 谢岳山. 数据挖掘技术在联网审计中的应用与研究[D]. 长沙:中南大学,2013. 34 - 57.
[68] 常法亮. 基于数据挖掘的智能审计系统的设计与实现[D]. 成都:电子科技大学,2010. 43 - 62.
[69] 黄少滨,吕天阳,迟荣华,夏勇,一种增量离群点识别算法及其在社会保障审计中的应用[A]. 基于互联网的商业管理学术会议[C]. 武汉:2010. 820 - 824.
[70] 陈伟. 大数据环境下基于模糊匹配的审计方法[J]. 中国注册会计师,2016,(11):84 - 88.
[71] 周振煜. 基于审计知识库的文本关联分析研究[D]. 哈尔滨:哈尔滨工程大学,2012. 36 - 53.
[72] Appelbaum,Deniz A,Alex Kogan,Miklos A. Vasarhelyi. Analytical procedures in external auditing:A comprehensive literature survey and framework for external audit analytics[J]. Journal of Accounting Literature,2018(40):83 - 101.
[73] 李强,谢汶莉. 大数据审计中的可视分析[J]. 中国内部审计,2016(2):79 - 86.
[74] 任磊,杜一,马帅,等. 大数据可视分析综述[J]. 软件学报,2014,25(9):1909 - 1936.
Ren Lei,Du Yi,Ma Shuai,et al. Visual analytics towards big data[J]. Journal of Software,2014,25(9):1909 - 1936. (in Chinese)
[75] 陈谊,甄远刚,胡海云,等. 一种层次结构中多维属性的可视化方法[J]. 软件学报,2016,27(5):1091 - 1102.
Chen Yi,Zhen Yuan-Gang,Hu Hai-Yun,et al. Visualization technique for multi-attrbute in hierarchical structure[J]. Journal of Software,2016,27(5):1091 - 1102. (in Chinese)
[76] 阮泓科,张懋生,姚益平. 基于 PERT 的审计方案可视化建模方法[J]. 电子科学技术,2014,1(1):92 - 96.
Ruan Hongke,Zhang Maosheng,Yao Yiping. Visualization modeling method of audit plan based on PERT[J]. Electronic Science & Technology,2014,1(1):92 - 96. (in Chinese)
[77] 陈伟,Wally S. 大数据环境下基于数据可视化技术的电子数据审计方法[J]. 中国注册会计师,2017,(1):85 - 89.
[78] 徐超,姜国标,陈勇. 区块链技术支持下电子数据保障方法探究[J]. 软件导刊,2019,18(5):7 - 10.
Xu Chao,Jiang Guobiao,Chen Yong. Research on electronic data guarantee methods supported by block chain technology[J]. Software Guide,2019,18(5):7 - 10. (in Chinese)
[79] 吴鹏,林敏. 论大数据时代下审计的机遇与挑战[J]. 金融经济,2016,(8):106 - 107.
[80] 秦荣生. 大数据、云计算技术对审计的影响研究[J]. 审计研究,2014,(6):23 - 28.

作者简介

徐　超　男,1980 年出生,湖北红安人,博士,教授,研究方向为大数据审计、区块链技术与应用.
E-mail:xuchao@ nau. edu. cn

陈　勇(通信作者)　男,1986 年出生,湖南娄底人,博士,高级工程师,研究方向为大数据审计,软件可靠性,嵌入式系统优化.
E-mail:cyong1000@ 163. com