数据要素

Data Elements

数据作为新型生产要素，是数字化、网络化、智能化的基础，已快速融入生产、分配、流通、消费和社会服务管理等各环节，深刻改变着生产方式、生活方式和社会治理方式。

数据要素是指以电子形式存在的、通过计算的方式参与到生产经营活动并发挥重要价值的数据资源。在数字经济中，数据被视作与土地、劳动力、资本、技术并列的五种生产要素之一。数据要素是推动数字经济发展的核心引擎，是赋能行业数字化转型和智能化升级的重要支撑，也是国家基础性战略资源。

2023 年正式成立的国家数据局，负责协调推进数据基础制度建设，统筹数据资源整合共享和开发利用，统筹推进数字中国、数字经济、数字社会规划和建设等，不仅体现了对数据资源的战略性管理和规范化利用的需求，也体现了国家层面对数字经济发展和数据治理的重视。

数据要素
Data Elements

大数据技术与5G通信技术融合的应用

文◆中国电信股份有限公司厦门分公司　陈华峰

引言

伴随通信行业的快速发展，大数据技术和5G通信技术的融合应用越来越广泛。这种融合既有助于提升用户数据分析的准确性和深度，还能持续优化通信功能，为用户提供更加丰富多样的功能服务。在商业开发和民生服务方面，这种技术的结合带来了更加多样的选择和机遇。本文对大数据技术和5G通信技术的融合展开详细分析，探讨了其在不同应用场景中的潜力和优势。

1 大数据和5G通信的关键技术

1.1 大数据

1.1.1 数据汇总搜索技术

基于数据汇总搜索技术，可以做到对大量数据的快速汇总和高效采集。结合有关硬件和软件设施，实现对基础信息的整合、归纳工作，为后期数据处理和分析提供可靠基础。

1.1.2 数据存储技术

在数据被成功收集后，安全、高效地存储数据是一项关键工作。在存储过程中，应把数据存储于规定区域，并通过科学、完整的存储管理规划，确保所有数据发挥其应有的作用。现代的数据存储技术常常结合云服务平台，通过数据通信技术将数据直接上传至云端，既方便管理，又能保障数据的安全性和可访问性。

1.1.3 数据挖掘技术

数据挖掘技术主要用于对汇总后的基础信息加以精细化处理。通过挑选、预处理和深度联合分析，从海量数据中提取出具有实际运用价值的数据集，不仅提升了数据的使用价值，还能为具体问题的解决提供重要参考。例如，在城市通信网络维护中，使用该技术来对通话故障数据加以深入分析，绘制故障地形图，逐次评价与研究问题，使后续优化措施更为准确有效。

1.1.4 数据分析技术

通过对挖掘出的有价值信息进行统一分析，可以为决策提供科学依据，确保决策执行符合具体要求。不仅提升了数据管理的控制效果，还为未来的管理决策提供了坚实的基础。通过系统的、科学的数据分析，企业或组织能够更加准确地把握市场动态，制定出更具前瞻性的战略规划。

1.2 5G 通信关键技术

（1）多输入多输出（MIMO）技术。与过往传统的无线通信系统相比，MIMO技术能够做到对多路信号的推送和接收。通过推送分集技术，系统可以在不同路径上同时发送多个信号，提高信号的抗干扰能力。室内空间重复使用技术允许在同一空间内重复使用频率资源，从而提升频谱利用率。波束形成技术则通过智能调整天线的发射和接收方向，提高信号的传输效率和覆盖范围。在5G技术中，MIMO技术依赖于基站收发台中大量天线的使用，这些天线能够同时处理多路信号，极大地提升了无线数据传输与连接的可靠度。

（2）超密集网络技术。凭借宽带网络化、一体性和多元化的特点，超密集网络能够在小范围内部署大量的小基站，从而提高网络的覆盖密度和信号质量。通过创建数据总流量分析管控机制，超密集网络可以实时监控和优化网络资源的使用，降低网络运维的难度。此外，该技术还

【作者简介】陈华峰（1977—），男，福建龙海人，本科，中级工程师，研究方向：5G行业应用。

能有效提高信号传输品质，减少信号覆盖不均匀带来的负面影响，从而提供更加稳定和高质量的通信服务[1]。

（3）自组织网络技术。该技术基于多个移动端口与无线收发设施，构建一个动态、灵活的无线网络体系。相比传统的集中式网络结构，自组织网络能够根据网络环境的变化自动调整资源分配和路由策略，从而大幅提升资源利用效率。每个节点在自组织网络中不仅是数据的终端，还是数据的中继节点，这种结构使网络具有较高的自愈能力和扩展性。

2 大数据技术与5G通信技术的融合应用方法

2.1 提升频谱资源利用率

在大数据技术和5G通信技术的融合过程中，提高频谱资源利用率是一个关键目标。为实现这一目标，应综合分析这两种技术的功能特征及其融合是否协调可靠。在实际操作中，通过合适的方法提升频谱资源的利用效率极为重要。具体来说，可以利用大数据技术来优化无线频率规划。例如，基于MR测量报告数据的大数据技术能够更好地区分无线通信的频谱资源，减少无线干扰，进而满足不同用户提出的稳定性要求。通过大数据分析，联合上下行业务要求、资源流量和通信水平，分析并实现科学的上下行资源分配，既提升了频谱资源的利用效率，还能更加有效地解决通信新要求，提供更优质的用户体验，同时也为商业开发工作提供更好的支撑作用。

2.2 优化系统容量

系统容量在大数据和5G通信技术的融合过程中起着十分关键的作用。应当全面了解系统容量的设置和改进方法，从而改进服务模式，解决传统服务体系下存在的不足之处与冲突问题。在大数据和5G通信技术加以结合的背景下，用户会产生新的要求。因此，应强化对这些需求的合理评估与分析，结合需求特征与系统容量的拓展机制、维护手段，不断提升通信水平。在此方面，有序规划无线网络部署是关键，通过大数据分析结果，可以更好地分配5G通信系统的容量和服务机制。既能够提升音频、视频等大文件的传输速率，还能确保传输过程的可靠性，从而满足用户的多样化需求[2]。

2.3 自动化的无线监控技术

在5G移动平台全面覆盖的基础上，自动化无线监控技术通过实时的全方位监管，保障了用户个人信息的安全性。近年来，盗窃用户个人信息的现象频繁发生，所以，使用自动化无线监控技术显得尤为关键。在智能无线监控技术平台的支持下，可以实现全方位的信息传输安全防范和控制，从而建设形成更为可靠的移动信息分析体系。无线通信系统监控人员基于自动化的远程监控设施，可以精准预判、分析会对用户隐私信息造成威胁的因素，从而全面提高信息保护力度。利用先进的自动化无线监控技术，通信运营商和安全机构能够实时监测网络流量，识别异常活动，并及时采取措施防止非法访问和数据泄露。尤其在5G网络下，数据传输速度极快，自动化监控技术可以通过机器学习和大数据分析技术，迅速识别潜在威胁，确保用户信息在传输过程中不被窃取或篡改。不仅有助于保护个人隐私，还能增强用户对新一代通信技术的信任，促进5G技术的广泛应用和发展[3]。

2.4 高密度异构网络

随着通信产业的快速发展，高密度异构网络在现代通信系统中逐渐占据重要地位。高密度异构网络结构能够应对各种用户群体的不同需求，并通过高频谱复用度和减小蜂窝尺寸来显著增加系统容量，从而提升移动通信网络的整体性能。具体来说，高密度异构网络基于部署更多的小基站，提高了频谱的使用效率，使同一频谱资源可以被更多用户共享。同时，缩小蜂窝尺寸减小了每个基站覆盖的范围，提升了信号质量和传输速度。大数据和5G通信技术的结合，使得即使用户处于偏远地区，也能享受到高质量的通信服务。例如，通过大数据分析，精准预测用户流量需求，优化基站部署策略，确保资源得到最优配置。此外，借助5G网络的低时延和高速率特性，用户在高速移动状态下（如乘车、乘飞机）仍能保持稳定的通信连接。这种全新的网络架构不仅扩展了通信服务的覆盖范围，还大幅提升了服务质量。高密度异构网络通过智能化管理和优化，能够灵活应对各种复杂的通信环境，为用户提供更可靠、更高效的服务体验，也为通信产业的发展提供了广阔的空间和更多的可能性，推动了整个行业的技术进步和业务创新。

2.5 打造多元性信息共享模式

把大数据与5G通信技术有机结合，可以显著优化数据结构的调控，充分展现出5G通信技术的特殊优点，在缩小时延的基础上，还可确保数据整理与综合的可控性和高效性。通过云计算，

可以高效地进行资源管理，满足各种服务与使用要求。基于5G技术的应用，结合网络切片技术，形成更具个性化的服务管理机制，深入探索信息之间的内在联系，确保数据使用以及服务效果达到预期效果。

此外，大数据技术在实际运用、开发和部署方面，能够为Web服务和数据库提供支持，有效发挥出数据的使用价值。在与5G通信技术融合后，凭借云计算工具构建高效且经济的服务机制，满足虚拟化服务需求。这种多元性的信息共享模式，不仅提高了数据处理的效率和准确性，还为用户提供了更加灵活和多样化的服务体验。

3 大数据技术与5G通信技术融合的应用场景

3.1 网络自动驾驶

（1）实时过程分析。通过大数据技术可采集和整合设施连接信息、5G网络连接数据、连接质量和不同端口之间使用质量等信息资料，做到了感知级别的动态驾驶过程状态分析。这种高度实时化的分析机制弥补了传统路测在实时性方面的缺陷，能够即时反映网络状态和用户体验，确保网络优化的及时性和准确性。（2）精准监控。基于对网络设施运作信息的收集，并结合大数据实行动态监管与评估，可以做到对网络运营情况、用户使用效果的高精度监管，为自动化驾驶提供了可靠依凭，使网络能够自动调整和优化，从而提升整体性能和用户满意度[4]。

3.2 智慧城市

（1）交通管理。通过大数据技术采集与分析交通流量和车辆位置等信息，可实现对交通信号灯的智能管控、拥堵路况的准确预测，改善城市交通的顺畅度，缓解交通堵塞问题，提高通行效率和安全性。

（2）环境监控。利用传感器等设施收集环境数据，通过大数据技术加以深入研究与评估，对大气、水体、土壤等关键环境要素做到全面监管，作为开展环保工作的重要依据，帮助城市管理者及时应对环境问题，改善居民生活质量[5]。

3.3 医疗保健

（1）远程医疗。凭借5G的高速度传输以及超低时延的特点，医疗人员可基于高清视频等方式为患者提供实时性的远程诊治和医疗咨询服务，为偏远城市与行动不便的患者改善医疗水平。这种新型医疗模式打破了地域限制，使优质医疗资源得到更广泛地共享。

（2）智能医疗设备。凭借5G的大规模连接功能把医疗设施和互联网平台进行连接，做到不同设施间的信息共享与协同处理，提高了医疗设施的智能化程度，使工作效率大幅提高，提供更准确的诊断和治疗支持，让其应用价值得到更好的开发。

（3）公共卫生。利用大数据技术研究疾病的传播动向、干扰因素，医疗单位可以制定有效的预防对策，提升公共卫生服务质量，为应对突发公共卫生事件提供了有力支持，保障了社会的健康安全。

结语

通过将大数据和5G通信技术有机联合，既可以降低技术应用过程中的漏洞，同时也革新了技术部署计划，从而在长期工作中取得了显著进步。今后的发展过程中，应着眼于进一步优化二者的联合，合理规划两者的运用方向。对于不同地区、行业和项目，应采用专业、合理的方法实行技术革新，严格把握大数据和5G技术的联合需求，同时遵守国家有关规范标准，以确保技术应用的安全性和有效性。

引用

[1] 陈三龙,谢俊.大数据技术在5G通信网络中的应用探析[J].中国新通信,2023,25(23):1-3.

[2] 陈云柯,葛裴.大数据技术在5G通信网络中的应用[J].科技资讯,2023,21(18):25-28.

[3] 吴洁慧.大数据技术在5G网络通信中的应用及前景分析[J].中国高新科技,2023(10):139-141.

[4] 王丽丽,杨昊天.大数据技术在5G通信网络中的应用[J].信息记录材料,2023,24(2):115-117.

[5] 李永钢.大数据技术在5G通信网络中的应用探析[J].电脑知识与技术,2022,18(27):56-58.

数字化转型下的数据脱敏系统研究与实现

文◆福建省农村信用社联合社　陈书宪

引言

近年国家陆续出台了《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等法律法规，银行的监管部门也推出《金融数据安全 数据安全分级指南》《金融数据安全 数据生命周期安全规范》等行业标准，要求商业银行加强数据安全保护[1]。在此背景下，数据脱敏技术成为确保数据安全使用、共享的重要基石。数据脱敏作为数据处理的关键环节，确保在保护敏感信息的同时，使数据可用于分析、建模和业务测试，充分挖掘数据要素的价值，为进行精准营销、风险控制等深度应用提供了坚实基础。

1 数据脱敏概述

1.1 数据脱敏

数据脱敏是指按照一定规则对个人身份证号、姓名、手机号码、密码等敏感数据进行形变、遮蔽等处理，从而实现对敏感和隐私的数据进行保护[2]。数据脱敏按照定义和使用场景不同分为静态数据脱敏（Static Data Masking）和动态数据脱敏（Dynamic Data Masking）[3]。静态数据脱敏是对数据库或文件中存储的原始敏感数据进行一次性处理，将真实数据转化为虚构但保持原有格式和属性的数据的过程。动态数据脱敏则是一种实时数据保护技术，在用户访问数据库时，根据预设的规则和策略，对返回给请求者的敏感数据进行遮蔽或变形，确保只有授权的人员才能看到真实数据。

1.2 数字化转型的脱敏要求

一是要求数据脱敏系统支持对 PB 级别的数据进行抽取、脱敏处理，二是脱敏的速度要足够快以满足数据需求部门对数据的时效要求。商业银行存在大量异构数据，商业银行需要处理的有内部各业务系统的数据，也有来自外部共享的数据，很多情况下是异构数据。因此，要求数据脱敏系统支持异构数据脱敏、同步转换等处理。商业银行数字化转型时，挖掘、分析数据要素价值要求数据脱敏系统对数据处理后应保持数据特征，不能简单遮蔽以确保数据可用、可分析。同时，数据脱敏的范围应符合监管对数据分类分级的标准要求。最后数字化转型要求数据脱敏系统具备较高的自动化水平，减少人工操作提高效率。

2 数据脱敏系统的业务需求

基于数字化转型对数据脱敏的要求进行需求分析，应构建一套满足监管要求、数据处理性能高、脱敏后数据质量高、自动化程度高的数据脱敏系统，切实保护用户隐私和敏感数据。以下对关键需求点进行详细分析。

从脱敏的原数据形式以及异构方面考虑，数据脱敏系统应支持从数据库脱敏到数据库（简称库到库）、从数据库脱敏到文件（简称库到文件）、从文件脱敏到文件（简称文件到文件）等模式。数据库应支持传统的 Oracle、MySQL 数据库以及近年来快速发展的达梦、人大金仓等国产数据库。文件应支持纯文本文件和 Excel、JSON、XML 等格式文件。对于数据源与目标数据是异构的场景，应支持异构数据转换[4]。

从脱敏后的数据质量方面分析，需要满足以下条件。一是脱敏后的数据字节长度保持与原数据一致，以确保数据不会超长从而可以导入测试库，同时更容易复现生产环境出现过的问题。二是脱敏后应保持原数据的数据格式并满足校验规则，如身份证号码、统一信用代码、营业执照号、手机号码等数据都有特定的数据

【作者简介】陈书宪（1982—），男，福建泉州人，研究生，工程师，研究方向：信息系统项目管理、数据安全管理。

格式要求，脱敏后的数据应满足对应的格式特征要求。三是应保持业务关联性，同一个数据项在不同数据库表中脱敏后应保持一致，如一个身份证号码在不同数据库或不同表中脱敏后的值只有保持一致才能确保数据关联性。四是确保数据唯一性，即原数据中不同的数据项在脱敏后也应不同，保持其唯一性。

从数据量和脱敏时效角度考虑，数据脱敏系统须具备大数据量数据处理的能力，支持对 PB 级别的数据进行抽取和脱敏，脱敏的速度至少需要达到 15 万行 / 秒。

自动化程度方面，首先应支持自动识别、发现敏感数据及其敏感类型和敏感级别。在脱敏流程中应支持数据准备完成后自动触发脱敏作业，减少人工干预，提高工作效率。

3 数据脱敏系统架构

3.1 系统总体架构

根据系统的业务需求选定系统方案，对系统总体架构进行设计。为满足大数据量高速脱敏的需求，系统采用集群多节点架构，支持多个脱敏任务并行执行，单个大文件或者大表分片并行脱敏，脱敏计算节点支持横向扩展。同时系统通过工作流引擎实现对脱敏操作进行权限控制、审批和审计。

3.2 系统关键技术和实现

3.2.1 异构数据转换

为了将脱敏后数据存入异构数据库，应做数据转换，数据转换类型一般包括数据类型转换、关系模式映射和数据转换。数据转换中尤其注意空值、时间、日期的转换问题。

本文基于 XML 实现异构数据转换，转换过程如下。首先，应配置源与目标库的映射关系表，映射关系包括字段类型的对应关系、字段长度对应关系等信息。其次，脱敏处理时先将脱敏后的数据存储到一种特定模式的 XML 文档，这个文档分数据（DataXML）和结构（StructureXML）两部分，第一部分存储数据结构信息，如源库的字段名、类型、键、注释等信息，第二部分存储具体数据内容。再次，存储到 XML 文档后根据源与目标的信息从配置好的映射关系表查找到映射信息，根据映射信息将源 XML 文档转换成目标 XML。最后，将目标 XML 数据装载进目标数据库中。

3.2.2 大文件拆分并行脱敏

很多场景数据以文件形式存储，大数据环境下的文件基本上都超过 1TB，传统的顺序读写方式进行数据脱敏无法满足时效性需求。业内很多方案是使用 HDFS 存储大文件，利用分布式计算框架（如 Spark 等）进行脱敏 [5]。这类方案可以有效提高脱敏速度，在 2 台 8 核 32G 服务器实验环境中，测试顺序读写方式速度是 2000 行 /s，用 Spark 框架可以达到 10 万行 /s 以上的速度。但是这类方案存在一个缺点是脱敏完成后合并文件的耗时比脱敏更长，实验环境中，HDFS 的文件块 block 配置为 256MB 的情况下，脱敏一个 50G 的文件，脱敏用时 32min，合并文件耗时 47min。

为了缩短文件合并处理时间，本系统参考 Spark 分布式计算框架，基于 Redis 设计了一套大文件切片并行处理的方案，集群节点之间通过 Redis 共享数据。系统将大文件进行切片，切片的大小根据文件大小和集群服务器数量以及每个服务器支持的线程数进行计算。

切片大小 = 文件大小 /（集群节点数 * 单节点线程数量）

主控节点分出一个切片即可将数据分配给各集群节点的脱敏处理线程进行脱敏处理。单个线程处理完成后将脱敏后数据转交数据合并的线程进行文件合并。文件拆分、脱敏和文件合并的线程可并行处理，从而缩短了最后文件合并的等待时间。

3.2.3 敏感数据发现

敏感数据发现的方式有多种。一是基于关键字，如通过姓氏字典识别中文姓名。二是通过正则表达式以及校验位算法进行识别。识别到敏感数据类型后，还要根据类型对应到数据敏感级别，以帮助用户确定是否脱敏，从而满足监管要求。敏感数据发现是脱敏的前提，系统也需要支持自定义发现规则以识别未经系统定义的数据类型。

3.2.4 脱敏算法

脱敏算法就是将敏感数据变成其他数据的计算方法，业内常见的算法如下 [6]。

（1）随机映射。用随机值替换现有敏感数据，替换后保持原有类型。由于是替换成随机值，每次脱敏结果都不同，因此无法保持数据关联性。

（2）固定映射。用固定值替换现有敏感数据，替换后保持原有类型。相比随机映射，固定映射可以保持数据关联性，但安全性较随机映射的低。

（3）遮掩。用特定字符（如 * 或者 X 等）替换敏感数据中的部分字符，使用遮掩算法一般无法用于业务测试或挖掘分析。

（4）混洗。类似于扑克洗牌，数据行之间数值进行替换以洗掉数据相关性，从而消除统计类数据的偏差。

（5）保留格式加密。保留格式算法 FPE（Format–Preserving Encryption）[7] 是一种能够保持数据格式不变的加密算法，具有特定格式的敏感数据（如身份证号、手机号等）按照规则进行分段，逐段数据进行加密运算，加密后进行相关处理确保符合原数据类型。

在实际应用中，在对数据进行保护的同时，应最大限度保留数据的价值以利于业务测试、数据挖掘模型建立、数据价值挖掘、数据治理应用。因此进行数据脱敏时，不仅需要保证处理后数据安全性，还要保持数据真实的结构特征。为达到这个效果，本系统在对身份证号码、手机号码、统一信用代码证、卡号等具体特殊结果特征的数据采用保留格式加密算法。

3.2.5 脱敏规则

脱敏规则是指针对每一种数据类型确定数据分段以及选择具体的脱敏算法。例如，配置 18 位身份证号码脱敏规则时可以根据脱敏后的数据用途选择对各分段进行脱敏；脱敏数据用于数据分析，且希望保留地区码则可以选择对 8 位出生日期和 3 位顺序码进行脱敏；脱敏算法根据需求可以选择保留格式加密、固定映射、随机映射等算法，根据前 17 位数据按照身份证校验位算法计算出最后一位校验位数值。各数据类型的脱敏规则的集合组合成一个脱敏策略，要进行数据脱敏前选择一个脱敏策略即可。为确保敏感数据都被脱敏，系统应设置一道最后防线，目的是当敏感数据没有匹配到任何敏感类型时也会使用最后的脱敏算法进行脱敏，如按字符位随机脱敏或置空等算法。

3.2.6 脱敏任务管理

脱敏任务应支持选择具体的数据源、目标数据、脱敏的模式（库到库、库到文件、文件到文件等）、脱敏的策略（即每种数据类型的脱敏规则的集合）、脱敏任务调度等。为满足实际使用场景，脱敏任务的调度应支持立即启动、定时启动、外部条件触发等方式。在商业银行环境中，不允许数据脱敏系统直接从生产数据库抽取数据进行脱敏，而是从备份磁带库中恢复出数据或通过数据泵等工具在生产库闲暇时间段卸载出需求方需要的数据，这个过程称为数据准备。通过外部条件触发可以实现数据源准备任务与脱敏任务联动，从而实现数据准备、脱敏的自动化。具体过程是数据准备任务完成后发送一个标识文件到脱敏服务器的轮询目录下，数据脱敏系统每隔一定时间扫描该目录，当扫描到标识文件则启动对应的脱敏任务。

4 数据脱敏系统实现效果

本数据脱敏系统实现了在大数据环境下敏感数据自动发现、数据准备和数据脱敏联动，支持周期性自动发起脱敏任务实现脱敏的自动化。支持丰富的脱敏算法，根据不同数据特征，系统内置了随机映射、固定映射、保留格式加密、遮掩等算法。支持大数据表、大数据文件分片并行脱敏，并分发到集群节点进行脱敏，支持增加集群节点时提高脱敏速度。 在 10 台 16 核 CPU、64G 内存服务器的集群环境下，文件脱敏速度达到 28 万行 /s，表数据脱敏速度达到 42 万行 /s，且不限制脱敏原数据大小，可以满足 PB 数据量级的脱敏速度需求。

结语

数据脱敏系统是保障银行数据安全的重要手段。通过深入研究和实践，不断优化和完善系统的功能和性能，将为商业银行的数据安全和数字化转型发展提供坚实的基础。该系统主要针对结构化数据进行脱敏，下一步将基于人工智能技术实现对非结构化数据脱敏。

引用

[1] 乔梦梦,李彦彪,王嘉源.数字化转型背景下金融数据安全面临的风险及对策建议[J].金融科技时代,2023,31(12):76-80.

[2] 乔宏明,梁奂.运营商面向大数据应用的数据脱敏方法探讨[J].移动通信,2015,39(13):17-20+24.

[3] 王鑫,王电钢,母继元,等.基于机器学习的数据脱敏系统研究与设计[J].电力信息与通信技术,2018,16(1):33-38.

[4] 陈施卫.基于XML的异构数据库数据交换的研究与实现[D].成都:电子科技大学,2012.

[5] 邱岳.大规模集群异常状态检测与数据脱敏技术研究与实现[D].北京:北京邮电大学,2018.

[6] 周猛.数据脱敏及其在央行支付系统业务数据中的应用[J].北方金融,2020(6):87-89.

[7] 王鹏.多类型数据保留格式加密技术的研究与实现[D].北京:北京邮电大学,2017.

基于 OneNET 的多节点温湿度数据采集的设计

文◆太原师范学院 计算机科学与技术学院 陈 思 屈明月

引言

传统的有线网络由于自身的物理局限性，逐渐被无线网络取代。随着无线网络技术近几年的快速发展，其已成为一项与人们生活密不可分的技术，如煤矿监测、农业环境监测、智能家居等行业都离不开无线传感网络的参与。早期的无线传感器网络大多采用组网技术，节点数据需通过中继节点才能到达上位机，中继器的工作状态直接影响所有节点，且组网范围有限，节点部署直接影响整体网络性能。此外，还需要特定场所用于发送信息、存储信息和处理信息，系统维护、系统升级和系统保护则导致成本过高。随着物联网技术的飞速发展，技术不断革新进步，出现了越来越多的第三方物联网平台。本文论述的多节点温湿度数据采集系统设计中，采用 OneNET 云平台，将每个节点采集的数据直接发送至云平台，无需中继器，且各节点之间的距离没有任何限制，进而数据传输也不再受限，节点部署也不会影响整体网络的性能，系统节点可以随时增加或减少，使系统具有极强的拓展性和极高的灵活性。同时，在云平台方面能够以极低的成本实现数据处理、数据存储和数据转发。

1 系统整体架构设计

系统整体架构图如图 1 所示，主要由传感器节点、OneNET 服务器和手机等移动设备 3 部分组成，每个传感器节点又分成温湿度采集电路、MCU 主控电路和无线通信电路 3 部分。

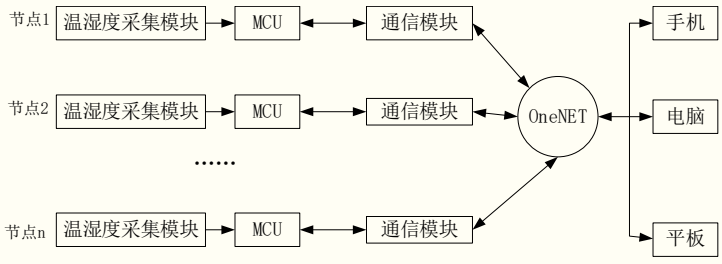


图 1 系统整体架构图

2 系统工作原理

每个节点的 MCU 读取温湿度传感器采集到的数据，在对其进行处理之后发送给通信模块，通信模块负责将数据发送到 OneNET 服务器。在 OneNET 上可以对数据进行处理和存储，并将其发送到手机等移动设备端 [1]。

3 数据采集节点电路设计

通过对节点的组成电路分析，明确各模块功能，设计每个模块硬件电路，基于 OneNET 的多节点温湿度数据采集系统整体电路设计如图 2 所示。

3.1 主控电路设计

本次设计中采用 STM32F103C8T6 作为采集节点的主控。这款单片机属于 STM32 增强型系列，外设丰富，处理能力强大，方便在后期对其进行二次开发。此外，其工作环境要求低，低功耗的工作模式可以实现省电的需求，主控电路设计如图 3 所示。

3.2 温湿度采集电路设计

温湿度传感器采用 SHT30，这款温湿度传感器是 SHT3x 湿度传感

【作者简介】陈思（1988—），女，山西晋城人，硕士研究生，实验师，研究方向：数据挖掘、区块链、计算机应用。

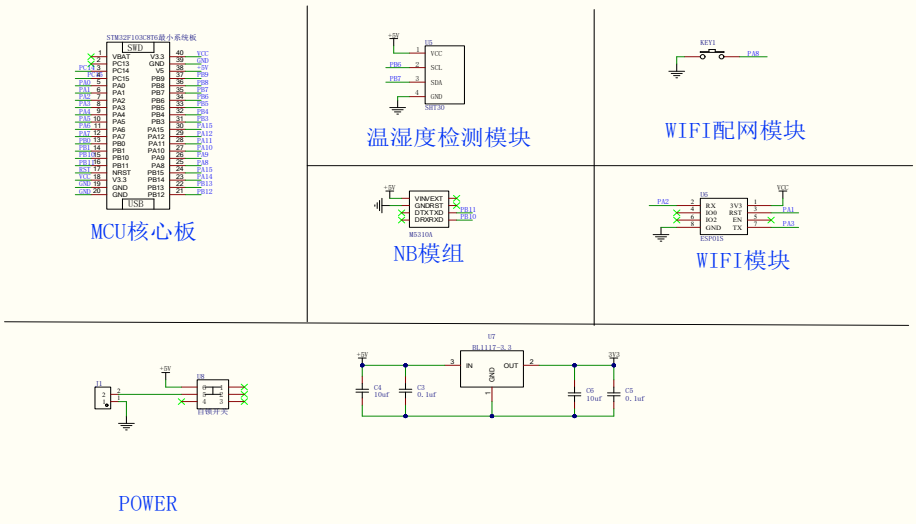


图 2 基于 OneNET 的多节点温湿度数据采集系统整体电路设计

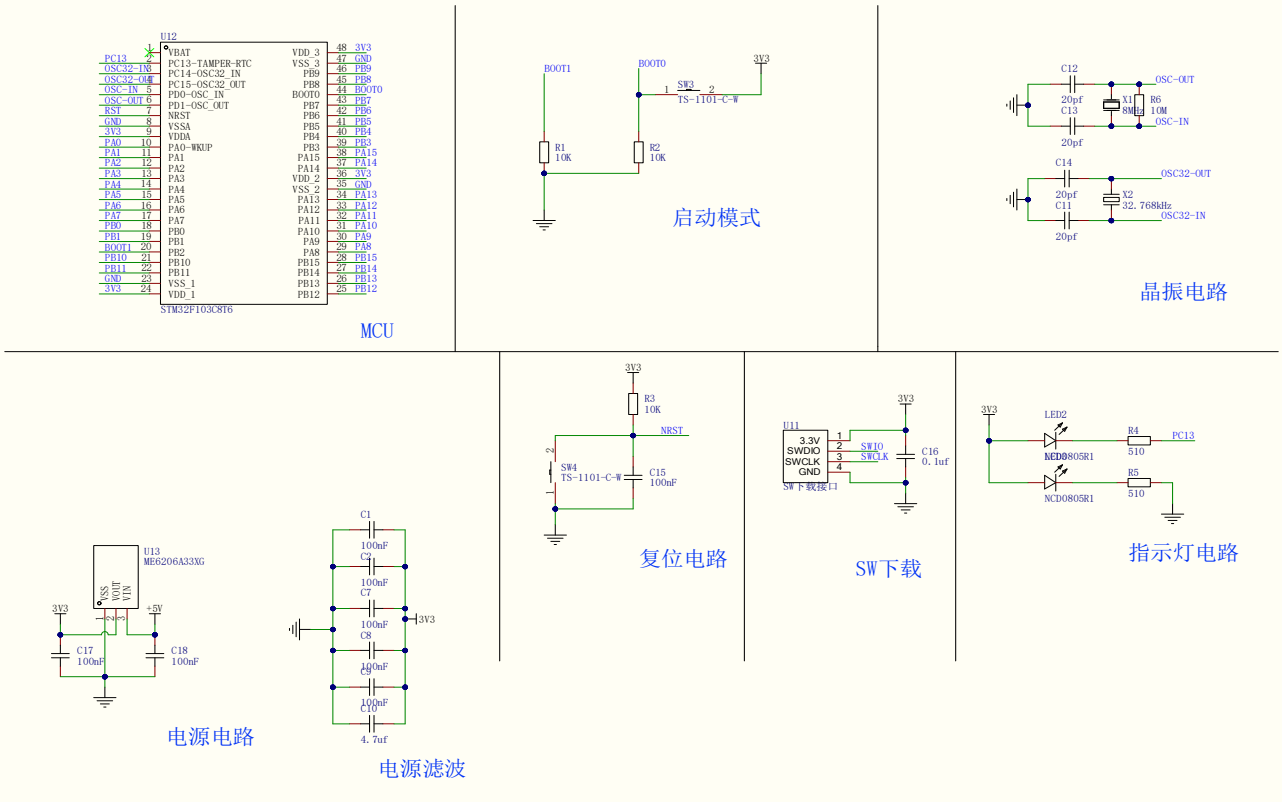


图 3 主控电路设计

器系列的低成本版本，其继承了 SHT20 系列的优点，同时产品的可靠性和精度规格得到了进一步提升。

MCU 主控与 SHT30 温湿度传感器之间采用两线式串行通信协议——标准 IIC 协议，通过对 SDA 数据线和 SCL 时钟线的高低电平时序的控制，可以实现 SHT30 与 MCU 之间的数据传递，温湿度传感器电路设计如图 4 所示。

3.3 通信模块电路设计

MCU 与通信模块之间采用串行异步通信方式，STM32 单片机利用自身的通用同步异步收发器 USAR 将 AT 指令发送至通信模块，通信模块将根据 AT 指令接入 OneNET 服务器并进行数据传递。

通信模块方面根据节点所处环境的不同采用两种方式，一种是用于有 WiFi 覆盖区域的 WiFi 模块 ESP01S，另一种则采用的是 NB 模组 M5210A，通信协议采用一种低开销、低带宽占用的即时通信协议——MQTT 协议 [2]，

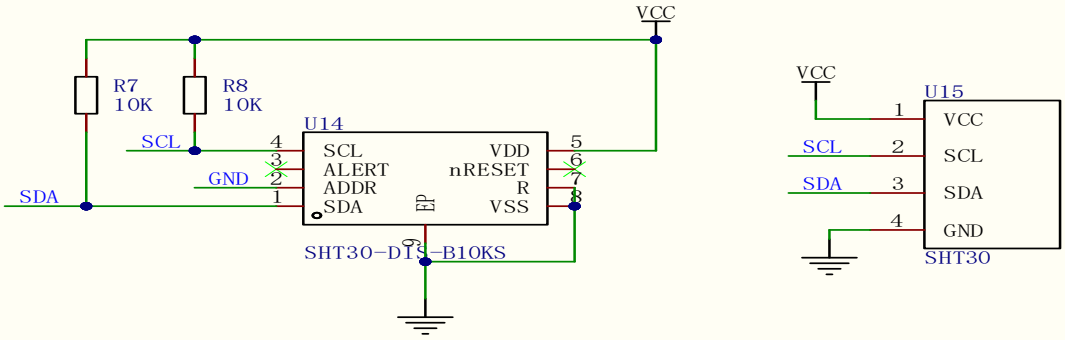


图 4 温湿度传感器电路设计

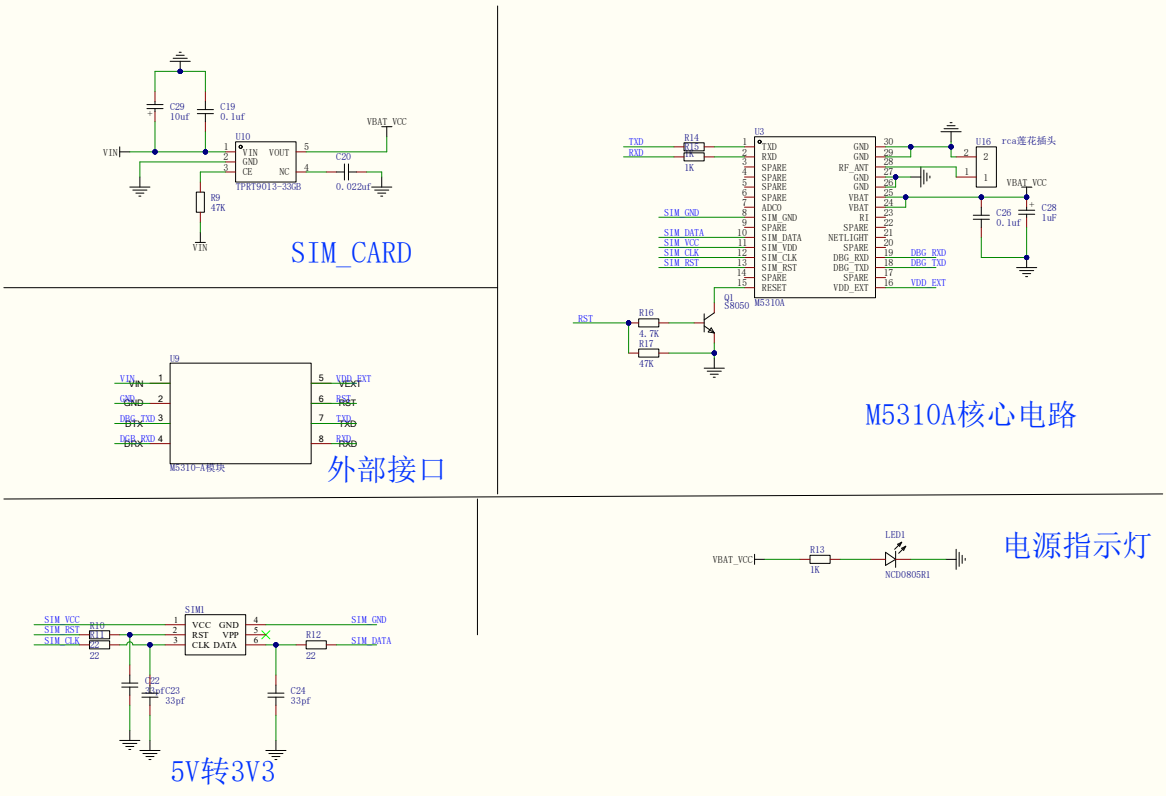


图 5 NB 模组 M5310A 电路设计

NB 模组 M5310A 电路设计如图 5 所示。

结语

为了直观观测系统的工作状态，设计和制作了基于 OneNET 的多节点温湿度数据采集系统。在验证过程中，采用了 3 个节点设备、两个 WiFi 通信设备和 1 个 NB 模组通信设备，并将两个 WiFi 通信设备分别连接到两个不同的 AP 热点。实践证明，本系统在未采用组网技术的前提下，很好地实现了实时采集并监测多个节点温湿度数据的需求。相较于利用组网技术的无线多节点网络，本文论述的系统对节点个数和节点之间的距离没有任何要求。各个节点之间相互独立，运行效率更高，省去了中继节点，降低了设备成本。利用第三方云平台 OneNET，实现数据存储、数据处理和数据转发功能，大幅减少了系统的维护成本。这种方式在工业控制、农业生产、智能家居、医疗监护和车载电子等领域被广泛采用。

引用

[1] 刘福祥,沈甦,邓世英.基于OneNET的数据采集监测系统设计与实现[J].电子制作,2022,30(17):47-49

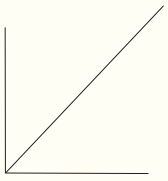
[2] 龙巧玲,牛德雄,林利云.基于OneNET云平台与物联网MQTT协议的智慧节能控制系统[J].计算机测量与控制,2021,29(7):127-130,135.

基于大数据技术的 API 接口自动识别方法

文◆中国移动通信集团广东有限公司　陈智扬

引言

基于大数据技术的 API 接口自动识别是指利用大数据技术和算法来自动识别和管理 API 接口的一种方法。通过分析 API 接口的数据流量、请求频率、参数结构等信息，结合机器学习和数据挖掘技术，实现对 API 接口的自动分类、识别和监控，以提升系统的整体效率和用户体验。



1 API 接口相关概述

数据流量指的是通过 API 接口传输的数据量。在大数据技术背景下，API 接口的数据流量非常庞大，通常用于处理大规模数据集或实时数据流。API 接口可以传输各种类型的数据，选择合适的协议可以优化数据传输效率和实时性。为了减少数据传输的带宽消耗和提升安全性，API 接口通常会使用数据压缩（如 GZIP）和加密（如 TLS/SSL）技术。请求频率表示客户端向 API 接口发送请求的速率。在大数据环境中，请求频率非常高，原因是客户端需要频繁获取或处理大量数据。为了避免服务器过载，API 接口通常会实施请求限流和配额管理。这些策略可以基于时间窗口（如每秒、每分钟最大请求量）、用户身份、API 密钥等进行配置。同时，为了减少重复请求和提升响应速度，API 接口会实施缓存机制。针对大数据处理需求，API 接口支持异步请求处理。客户端发送请求后，可以立即返回任务 ID 或状态，后台异步处理数据并提供结果。参数结构指的是 API 请求和响应中所用到的参数及其组织方式。在大数据技术中，客户端向 API 接口发送请求时，通常会包含必需的参数（如 API 密钥、请求类型、数据筛选条件等）。请求参数的结构应清晰明确，避免冗余和混淆。API 接口返回的数据通常以结构化格式（如 JSON）呈现。响应结构应根据数据类型和应用场景设计，包括标准字段（如状态码、错误信息）和业务数据。为了确保接口兼容性和演化，API 接口会实施版本控制机制，客户端可以根据接口版本号选择性地使用不同的参数结构和功能。

2 大数据技术在 API 接口自动识别中的应用

2.1 数据收集与处理

2.1.1 数据源的选择与获取

在基于大数据技术的 API 接口自动识别方法中，一方面是选择数据源。API 日志是一个重要的数据源，记录了 API 接口的请求信息、响应时间、参数等。通过分析 API 日志，可以获取大量关于 API 接口行为的数据。监控网络流量数据可以帮助了解 API 接口的使用情况和请求频率，也是另一个重要的数据源。一些 API 管理平台提供了丰富的 API 接口数据统计和监控功能，因此可以作为数据源之一。另一方面要采集实时数据。对于实时监控和识别，需要实时采集 API 接口的数据流量和请求信息，以便及时发现异常行为。对于历史数据分析和模型训练，可以通过批量方式获取 API 接口的历史数据，进行离线处理和分析。

2.1.2 数据清洗与预处理

针对数据分析阶段出现的重

【作者简介】陈智扬（1971—），男，广东深圳人，本科，高级工程师，研究方向：网络与信息安全、大数据、人工智能。

复数据记录，应进行去重处理，避免对后续分析造成干扰。检测并处理数据中的缺失值，可以通过填充均值、中位数或使用插值等方法处理缺失数据。利用统计方法或机器学习算法来识别并处理数据中的异常值，并根据具体情况进行处理或剔除。应确保数据的格式和结构统一，使数据在后续处理和分析中更易于操作和理解。同时，应对数据进行转换，如对文本数据进行标准化、编码处理，以便进行后续分析和挖掘。从原始数据中提取有用的特征，通过特征工程的方法选择和构建合适的特征，以提高模型的准确性和效率。对数据进行标准化处理，如将数据缩放到相似的范围，避免不同特征之间的量纲影响模型训练和结果[1]。

2.2 特征提取与分析

2.2.1 API 接口特征的提取方法

常见的 API 接口特征提取方法包括 4 个关键方面。首先是请求频率特征，通过统计每个 API 接口的请求次数，可以反映该接口的使用热度和重要性。此外，分析请求之间的时间间隔有助于识别定时任务或异常请求。其次是分析参数特征，涉及统计不同 API 接口的参数个数及其数据类型，如字符串、整数或浮点数等，有助于理解和分类各种参数的含义及其用途。再次是数据流量特征，通过统计 API 接口的数据传输量来评估接口的数据处理复杂度。分析数据格式（如 JSON、XML 等）则有助于理解传输数据的结构和内容。最后是响应时间特征，通过计算 API 接口的平均响应时间来评估其性能表现。同时，分析响应时间的分布情况可以帮助检测异常响应行为。此外，错误码特征用于统计 API 接口返回的错误码频率以帮助检测异常情况和问题接口。识别不同类型的错误码有助于理解其产生原因以及相应的处理方式。API 返回码设计如表 1 所示。

表 1 API 返回码设计

返回码值	说明
0	成功
99999	系统发生未知异常
10000 ～ 19999	参数校验错误
20000 ～ 29999	A 步骤执行失败
30000 ～ 39999	B 步骤执行失败

2.2.2 特征分析与建模

首先对 API 接口数据进行探索性分析，了解数据的分布、统计特性和相关性。根据业务需求和模型建立的目标，选择合适的特征进行建模。通过特征相关性分析、特征重要性评估等方法选择最具代表性的特征。对原始数据进行特征工程处理，包括特征缩放、特征变换、特征组合等，以提高模型的性能和准确性。通过可视化技术，将特征之间的关系和分布可视化展示，帮助理解特征之间的相互影响和重要性。使用标记的 API 接口数据集对选定的模型进行训练，以学习 API 接口特征之间的模式和关系。通过交叉验证、混淆矩阵等方法对模型进行评估，检验模型的性能和泛化能力。根据评估结果对模型进行调参和优化，以提高模型的准确性和效率。

2.3 模型训练与优化

2.3.1 机器学习算法在 API 接口自动识别中的应用

在 API 接口自动识别中，机器学习算法发挥着重要作用。一是聚类算法，如 K 均值聚类、层次聚类等，可用于将 API 接口进行分组，识别具有相似特征的 API 接口，从而实现自动分类和识别。二是分类算法，如决策树、支持向量机、随机森林等可用于训练模型，将 API 接口分为不同类别，帮助系统自动识别不同类型的 API 接口。三是异常检测算法，如孤立森林、LOF（局部离群因子）等可用于检测 API 接口中的异常行为，帮助系统及时发现潜在的安全风险。四是时序分析算法，如 ARIMA、LSTM 等可用于分析 API 接口的时间序列数据，预测未来的 API 请求趋势，帮助系统做出智能调整和优化。五是特征选择算法，如信息增益、方差分析等可用于选择最具代表性的 API 接口特征，提高模型的准确性和泛化能力[2]。

2.3.2 模型参数调优与性能优化

对于机器学习算法中的超参数（如学习率、正则化参数等），通过交叉验证等技术，选择最优的超参数组合，以提高模型的泛化能力。通过网格搜索和随机搜索技术，在给定的参数空间内搜索最佳的参数组合，以优化模型性能。对于深度学习模型，调整学习率可以帮助模型更快地收敛和取得更好的性能。通过正则化技术（如 L1 正则化、L2 正则化）来控制模型的复杂度，防止过拟合，提高模型的泛化能力。此外，

应进一步优化特征工程步骤，选择更加有代表性的特征，对特征进行更深入的处理，以提高模型的性能和准确性。通过模型集成技术（如投票法、堆叠法），结合多个模型的预测结果，提高整体模型的性能和稳定性。增加训练数据量，帮助模型更好地学习数据的模式和特征，提高模型的泛化能力。利用并行计算技术（如分布式计算、GPU 加速），加快模型训练和推理的速度，提高模型的性能和效率。对于深度学习模型，可以通过模型压缩技术（如剪枝、量化）减少模型参数和计算量，提高模型的推理速度和性能。

3 基于机器学习的 API 接口自动识别方法

3.1 监督学习方法

收集大量的 API 接口数据，包括请求频率、参数信息、响应时间等，对数据进行标记，即为每个 API 接口分配相应的类别标签，作为监督学习的训练目标。从 API 接口数据中提取特征，如请求频率、参数个数、响应时间等作为输入特征。对特征进行预处理和转换，确保数据格式符合监督学习算法的要求。同时，选择适合的监督学习算法，如决策树、支持向量机、随机森林等，用于构建 API 接口自动识别模型。使用测试数据集对训练好的模型进行评估，评估模型的分类准确率、召回率、F1 分数等指标。根据评估结果对模型进行调整和优化，以提高模型的性能和泛化能力 [3]。

3.2 无监督学习方法

基于大数据技术的 API 接口自动识别方法中，无监督学习方法应收集大量的 API 接口数据，包括请求频率、参数信息、响应时间等，无需标记类别信息。对数据进行预处理和特征提取，以便进行无监督学习算法的处理。使用聚类算法（如 K 均值聚类、DBSCAN 等）对 API 接口数据进行聚类，按照 $d(x, y)^2=\sum_{j=1}^{m}\left(x_j-y_j\right)^2=\|x\text{-}y\|_2^2$ 计算。将相似的 API 接口归为一类。同时，应用异常检测算法（如孤立森林、LOF 等）来识别 API 接口中的异常行为，帮助系统发现潜在的安全风险。异常检测可以帮助识别不符合正常行为模式的 API 接口，从而加强系统的安全性。运用关联规则挖掘算法（如 Apriori 算法）来发现 API 接口之间的关联关系，了解 API 接口之间的依赖性和共同特征。利用降维技术（如主成分分析、t–SNE 等）将高维的 API 接口数据转换为低维空间，便于可视化和理解数据特征。

3.3 深度学习在 API 接口自动识别中的应用

3.3.1 神经网络结构设计

在基于大数据技术的 API 接口自动识别方法中，深度学习可以使用卷积神经网络、循环神经网络、长短期记忆网络、注意力机制、深度神经网络等进行深度神经网络结构设计，更好地理解和识别 API 接口的模式和特征，提高自动识别的准确性。通过合理设计神经网络结构，结合大量的 API 接口数据进行训练，深度学习技术可以有效实现对 API 接口的自动识别和分类，提高系统的安全性和性能。

3.3.2 深度学习模型训练与优化

在基于大数据技术的 API 接口自动识别方法中，准备大量的 API 接口数据作为训练集，使用训练集对选定的深度学习模型进行训练，学习 API 接口的特征和类别之间的关系，或者通过反向传播算法和优化器（如 Adam、SGD）来调整模型参数，以最小化损失函数，提高模型的准确性。使用验证集对训练好的模型进行评估，评估模型的性能指标，如准确率、召回率、F1 分数等。而调整学习率可以帮助模型更快地收敛和取得更好的性能，并根据训练情况动态调整学习率。此外，使用批量归一化技术加速模型训练过程，提高模型的稳定性和泛化能力。

结语

基于大数据技术的 API 接口自动识别方法结合了数据分析、机器学习和深度学习技术，为系统提供了自动化、智能化的 API 接口管理解决方案。通过不断优化和改进，可以实现对系统中各种类型 API 接口的准确识别和分类，为系统的稳定运行和安全性提供有力支持。

引用

[1] 欧阳弘毅.基于大数据技术的红色旅游用户画像应用[J].宜春学院学报,2023,45(1):70-74.

[2] 张静辉,刘蔚,侯春梅,等.基于大数据技术优化科技期刊同行评议模式研究[J].中国科技期刊研究,2024,35(1):59-64.

[3] 李均瑶,沈心怡,祝迦洁,等.基于大数据下的城市户外广告牌管理系统的设计[J].办公自动化,2024,29(1):80-82+20.

基于大数据的智能咨询系统设计

文◆广东省电信规划设计院有限公司　陈文锋　赵米桥

引言

随着信息技术的快速发展，大数据已成为重要的战略资源，是国家和企业竞争力的重要支撑。大数据在金融、医疗、政务、教育、交通等领域得到了广泛应用，在为各行各业带来机遇的同时，产生了大量数据。目前，数据存储成本越来越高，数据质量越来越差。而人工智能技术具有“非结构化”和“半结构化”的特点，可以解决这一难题。目前，人工智能技术在金融行业应用已经比较成熟，但在政务、医疗等领域未能被较好地应用。

本文针对基于大数据的智能咨询系统进行设计，首先，分析了系统功能的需求。其次，对系统总体进行设计，包括系统设计目标、系统技术架构、系统功能设计。最后，对系统功能的实现与测试做了详细介绍。本文的研究成果对于智能咨询系统的设计与实现具有一定的参考和借鉴意义。

1 系统功能需求

咨询人员通过移动端、PC 端、语音端和数据库进行咨询。咨询人员在咨询前应对用户进行分类，将用户分为普通用户、VIP 用户、专家用户和高级用户。不同用户有不同的需求，普通用户只需要对普通问题进行回答，VIP 用户可以回答一些复杂的问题，高级用户可以回答更复杂的问题，专家用户可以回答更深入的问题。

本系统主要包括 5 个模块，即智能咨询、智能问答、智能推荐、智能管理和智能客服。系统包括两个数据库，即知识库和智能知识库，知识库分为结构化和非结构化两种类型。

用户通过移动端或 PC 端登录系统，在系统中进行咨询后，系统会将咨询信息以短信或者微信的方式推送给咨询人员。咨询人员可以通过语音、文字、图片等方式回答用户问题，当用户不能独立完成问题时，系统会自动推荐合适的咨询师进行解答。如果需要人工服务，则可以通过智能客服机器人来解决问题。

2 系统总体设计

2.1 系统设计目标

2.1.1 提供智能咨询服务

系统根据用户提问，利用机器学习、自然语言处理和数据挖掘等技术，分析用户问题的语义和上下文，通过语义分析和知识图谱构建，识别用户的问题类型。利用知识库为用户提供智能咨询服务。最后，系统根据用户的需求推荐不同类型的智能咨询服务，以提升用户体验。

2.1.2 提供业务推荐服务

根据用户问题的语义和上下文，分析用户需求和业务特征，并通过语义分析和知识图谱构建，为用户推荐相应的业务产品或服务。根据用户问题和业务特征，使用自然语言处理技术和数据挖掘技术构建知识库。知识库中包括金融知识、政务知识、医疗知识等领域[1]，系统可以根据用户的需求推荐相应领域的咨询产品或服务。利用云计算技术，将大数据平台与云计算相结合，为用户提供一个高性能和可扩展的云计算平台。

2.1.3 提供数据分析功能

对咨询业务进行分析和处理后得到的数据可以用于数据分析和数据挖掘工作，并将分析结果返回给用户作为智能咨询服务的参考依据。利用大数据分析技术对海量数据进行分析和处理后得到的结果可以通过可

【作者简介】陈文锋（1990—），男，广东阳江人，本科，研究方向：大数据。

视化方式展示出来，帮助用户更加直观地了解系统的运行情况和性能指标等信息。

2.1.4 提供风险控制功能

根据用户选择的咨询产品或服务以及其他相关业务规则，对用户所关心的问题进行风险控制处理。

2.1.5 提供实时监测功能

根据用户实时查询条件以及查询历史记录等信息实时监测系统运行状态、智能推荐或人工客服等功能，保证系统在各个环节正常运行，提高用户体验和满意度。

2.1.6 提供用户信息管理功能

根据用户信息定义其访问权限和访问路径，实现用户信息管理功能，并为业务发展提供数据支持和决策依据等功能。

2.2 系统技术架构

系统采用 MVC 设计模式，将业务逻辑和数据存储在不同的组件中。其中，业务逻辑组件负责向用户展示想要咨询的问题，数据存储组件负责将用户的问题数据转换为系统能够处理的格式，并存储在数据库中。

（1）用户界面组件。用户界面由前台和后台组成。前台界面负责展示用户需要咨询的问题，通过搜索框等方式向后台请求答案。后台管理系统是整个系统的核心，将前台界面获取的数据转换为系统能够处理的格式。

（2）数据存储组件。数据存储组件主要是将用户获取的问题数据转换成数据库中可以保存的数据格式，如文本、图片等。系统中有两种类型的数据存储方式，一种是关系型数据库，另一种是非关系型数据库。非关系型数据库可以直接操作数据库中的数据，而关系型数据库需要将数据转化为表来保存。为了提高查询效率和降低存储成本，采用非关系型数据库进行数据存储 [2]。

（3）智能推荐组件。智能推荐组件是将用户咨询的问题数据转换为系统能够处理的格式，并将这些转换后的数据存储在系统中。智能推荐组件主要包括自然语言处理引擎、推荐算法、知识库管理系统等。

2.3 系统功能设计

数据智能服务引擎的三层结构由原子服务层、引擎控制层、应用层组成，下层为分布式架构，核心为引擎中控层。通过 Leader-Worker 架构来保障“集中式”的高可用性与一致性，其内部由功能决策、缓存、融合优化器、状态控制器、逻辑运算器、排序模块、存储服务、溢出模块等构成，以实现对表达式的优化与对表达式服务的优化。最下层为具有 AI 模式推理功能的原子服务层，包括一个负载均衡器和若干业务实例。顶层为应用层，实现表达式生成、实时表达精度检测以及服务管理。

智能咨询系统主要功能包括智能咨询、智能推荐、用户管理、统计分析和数据可视化。

（1）智能咨询功能。智能咨询功能为用户提供实时的咨询服务。根据用户偏好，为用户推荐相关的问题和答案。系统根据用户之前的操作，判断用户遇到的问题和需要咨询的问题。然后，将问题和答案通过 Hadoop 平台进行大数据分析，并将分析结果返回给用户。

（2）智能推荐功能。基于人工智能技术，本系统为用户提供了多个业务模块，如保险推荐、房产推荐、基金推荐、教育推荐等。本系统将大数据技术与传统业务相结合，实现了业务协同和创新发展。

（3）用户管理功能。通过本系统可以实现对用户的权限管理、身份验证和隐私保护等功能，以确保用户信息安全。

（4）统计分析功能。利用统计分析技术对本系统产生的大量数据进行处理，提高了数据分析的效率。

3 系统功能实现与测试

基于大数据技术的智能咨询系统，主要是为了解决传统咨询模式中存在的问题。为了更好地分析用户行为数据，应对数据进行深入分析和处理，包括数据的预处理、特征提取等。

在大数据时代，针对用户行为数据进行分析处理已经成为一种必然趋势 [3]。（1）数据预处理。根据用户的咨询问题和需求，对原始数据进行初步分析和预处理。（2）特征提取。特征提取是指从原始数据中提取有用的特征，用来描述数据集的某些属性或规律。在智能咨询系统中，特征提取可以分为文本分类、图像识别、情感分析等步骤。（3）智能推荐。用户可以通过智能咨询系统的智能推荐功能，从海量的用户行为数据中找出与自己需求相关的信息，然后根据自己的需求进行选择，最终得出最佳解决方案。（4）在线答疑。用户可以通过在线答疑功能向智能咨询系统提问，系统会将问题反馈给专家进行解答。用户也可以通过在线留言功能与专家进行沟通交流。（5）历史记录查询。用户可以通过历史记录查询功能查询自

己之前所咨询过的问题和答案以及自己的特殊需求和偏好。（6）文档管理和分享。智能咨询系统可以对用户提交的文档进行存储和管理，并实现文档共享功能。（7）统计分析和报表生成。用户在智能咨询系统中，系统可以对用户咨询内容进行统计分析以及报表生成等操作，为用户提供更加个性化的服务和帮助[4]。

测试结果表明，基于大数据技术的智能咨询系统能有效帮助用户解决问题，提高问题的解决效率和质量，从而达到提高生活质量和工作效率的目的。

3.1 功能实现

通过对用户行为数据的分析，发现用户的共同需求，这些需求是用户产生问题的根本原因。在智能咨询系统中，通过对用户行为数据的分析，发现用户想要表达的意思。然后将这些意思提取出来，用自然语言进行表达。通过自然语言处理技术将这些意思转化成易于用户理解的语言，并结合知识库进行解答。以问答方式回答用户问题时，需要对问题进行分词处理，然后对分词结果进行相似度计算。在这个过程中需要利用 TF–IDF 算法来计算词和短语在句子中出现的频率。

例如，“我想吃肯德基”与“吃肯德基”这两个词之间的相似度很高，可以利用 TF–IDF 算法将这两个词归类为“肯德基”类。这样就可以根据相似度高的关键词来生成对应的短语和句子。

在系统中还需要设计一个知识库，知识库包含了各种类型问题的解答和知识储备。例如，用户提问“我想知道关于 xx 公司的业务情况”，在知识库中输入这个问题，然后查询该公司有关业务数据并进行分析。

智能咨询系统基于大数据技术实现了对用户问题进行分析和解答。首先，通过对用户行为数据进行分析处理，发现用户的需求。其次，利用自然语言处理技术对问题进行分析和解答。最后，通过知识库对问题进行匹配和知识储备，生成解答句子和回答内容。

3.2 系统测试

目前，互联网上有很多类似的咨询网站，但这些网站主要是针对某个领域的问题进行解答，缺乏全面性[5]。另外，这些网站大多采用人工咨询的方式，效率低、成本高。下面是一个使用该系统对用户咨询的结果展示。首先，在系统中导入相关数据，包括用户浏览信息、咨询问题、问答记录等，并将数据保存到数据库中。其次，将这些数据导入系统中进行分析处理，处理后的数据为用户浏览信息、咨询问题、问答记录等。最后，利用大数据技术对这些数据进行处理，根据大数据技术对这些数据进行清洗、挖掘等工作。

此外，通过编写程序来实现系统功能，编写程序需要注意以下内容。一是应先定义好接口函数；二是在编写程序时应遵循相关技术规范和行业标准；三是编写程序时应注意数据量、数据处理时间等因素的影响；四是在编写程序时应注意代码质量问题。

为了验证本系统的可靠性和实用性，使用该系统对用户咨询进行测试。首先，进行用户登录测试。用户登录后系统会自动弹出问题列表，用户点击问题列表中的任意一个问题后就能得到相应的解答结果。其次，进行用户信息验证测试，在获取到用户信息后，通过对数据进行清洗、挖掘等工作，生成新的数据集。最后，进行问答结果验证测试，使用训练好的模型对用户提出的问题进行解答并生成新的问答结果。

结语

通过对智能咨询系统框架的研究，可以为企业和用户提供更全面、更深入和更系统地咨询服务，提高企业和用户的满意度。本文介绍了基于大数据技术的智能咨询系统框架设计，并详细阐述了该框架中各个模块的具体实现方法。该框架还可以为企业和用户提供更多有针对性、个性化和人性化的服务，使企业能够更好地了解用户需求并做出相应调整，从而提高企业和用户之间的满意度。

引用

[1] 王浩田.面向科技咨询的数据智能服务开发平台研究与实现[D].北京:北京邮电大学,2023.

[2] 智能城市建设与大数据战略研究项目组.智能城市规划建设与大数据战略研究[M].杭州:浙江大学出版社,2018.

[3] 智能城市建设与大数据战略研究项目组.智能城市建设大数据战略研究[M].杭州:浙江大学出版社,2018.

[4] 王敏,汪依帆,黄维.基于大数据的立法智能分析平台[J].电子技术与软件工程,2018(17):194-195.

[5] 王吉鹏.咨询业应主动拥抱大数据[J].企业管理,2018(1):28-29.

数据转型背景下的档案业务过程数字化战略分析

文◆中国石化华北油气分公司　**姜志超**

引言

企业在推动档案业务过程数字化战略期间面临复杂多样的工作任务，审视当前档案管理工作的现实需求，完善数据转型背景下的管理架构，优化管理模块，革新管理方法，提高档案管理水平和效率。

1 企业档案业务过程概述

企业档案业务过程涉及多个环节，是一项综合性的工作，主要目的是保证档案完整、安全、有效。其业务过程主要包括档案收集与整理、立卷与归档、利用借阅、鉴定与销毁。建立完善的档案管理制度是落实各项工作的基础，可通过规范化管控形式加强对档案的循环控制，并同步落实对档案管理人员的教育培训，结合信息化、数字化升级，使档案管理工作更加完整、全面和有效。

2 数据转型背景下企业档案业务过程数字化的重要性

2.1 提高档案完整性

传统的档案收集与整理工作存在较大的主观因素影响，需要人员花费大量精力收集整理。企业在数据转型背景下对档案业务过程进行数字化管控是一项关键性的事务，能够从业务源头着手，按照应归档范围标记业务活动，一旦发生业务活动中存在应归档文件，即开始自动标记跟踪，业务活动完成后，自动分类形成移交文件，收集与整理工作全自动完整完成。

2.2 提高档案业务效率

传统纸质档案管理需要大量人工操作，对档案信息进行分类、存储、检索，这一过程耗时且容易出错，实施数字化转型后，就可以交给数字化系统，实现高效规范完成。

2.3 降低成本

随着企业持续运作，其档案数量持续增长，传统纸质档案会占据越来越多的物理空间，带来较大的成本。此时，企业实施数字化档案管理可将纸质档案信息转化成电子文档资料，减少对物理空间的依赖。

2.4 安全性管理

企业通过数字化档案管理还能够提升数据管理的安全性，由于纸质档案容易受到自然灾害、盗窃、意外破损等风险的威胁，此时企业通过数字化转型，引进安全加密技术、备份系统，对档案信息的安全性和完整性给予保障，及时应对发生的意外，并且在多个场景对档案信息进行恢复，有助于减少损失。

2.5 价值开发

企业结合大数据、区块链技术推动数字化档案业务过程转型，有助于提高档案资料的利用价值，利用大数据、人工智能等技术，挖掘档案信息中的数据价值，将低价值密度的档案资料进行价值提取和开发，使档案信息的功能得到完美释放。引进数字化档案资源能够在互联网终端更加高效传递信息资料，为档案资源借阅、查询等工作提供便利，以此推动档案业务现代化发展。

3 数据转型背景下档案业务过程数字化策略

3.1 存量档案数字化处理与转换

存量档案数字化是将已有档案进行数字化扫描，形成易于读取、检索的可全文识别的电子文件，利用大数据技术，挖掘档案

【作者简介】姜志超（1989—），男，黑龙江大庆人，本科，中国石化华北油气分公司馆员 / 副主任，研究方向：档案管理、档案信息化。

信息量化价值。

开展档案数字化，一是要做好标准建设。按照国家、企业标准结合公司实际，着眼信息技术发展的应用需求，制定一套包括不同载体类型、介质尺寸、印制方式、组卷方式、对电子文件扫描质量、指标的要求，如 OCR 识别率、DPI、去噪、对比度、数据备份要求等。二是做好档案安全管理，通过制定数字化场地规范、设备要求、网络要求、操作规程等一系列措施，保障数字化过程中的档案实体与信息安全保密 [1]。

3.2 数字化档案管理系统建设

3.2.1 架构设计

在数据转型视域下，企业推动档案业务数字化转型应依托一个完整、全面的档案管理系统和平台，在该环节，企业应构建完善的平台系统架构体系，平台架构与档案管理业务体系充分融合，结合模块化设计模式，将各模块的功能进行连接互动，提升各模块的独立性和可扩展性。此外，企业在开发系统架构时应同步提升系统的易用性，以便对系统进行有效操作使用和维护 [2]。

3.2.2 信息收集

完成对系统架构的设计后，应严格参照档案处理流程丰富完善各项功能。首先，在数据采集与集成板块，档案管理系统应支持多源数据采集，包含对纸质档案进行数字化转化后所得到的图像数据以及企业在信息化管理过程中所生成的电子档案和文本信息进行处理，该系统具备数据集成管理能力，可将不同来源、不同格式的数据资料进行整合，形成统一的数字资料库。但在数据采集和存储过程中，系统应同时嵌入制度规范，保障采集到的数据具有规范性，通过系统预制规范，核证数据描述具备准确性和一致性，避免数据冗余或错误。

3.2.3 存储和管理

通过技术手段对档案信息进行存储和管理，确保数字化档案得到安全存储和快速访问，在此背景下，企业应建立完善的数据管理机制，选购适配的硬件设备，构建数据分类、索引、备份、恢复等管理循环，结合数据压缩、加密等技术手段，提高数据存储的安全性和综合效率。同时，企业应注重提升数据的迁移性，以适应不同业务开展的需求。

3.2.4 检索和利用

检索和利用是发挥档案价值的关键功能，系统应提供强大的检索服务，支持多种检索方式。例如，管理人员可通过关键词检索、全文检索、模糊检索，快速得到目标信息和相关文件，并提供数据分析和可视化功能，帮助企业管理者深入挖掘档案信息价值。系统中应融合人工智能分析模块，建立自动生成主题档案数据包、决策支持数据包功能，梳理大事记、发展简史功能。

3.2.5 系统升级

企业应对系统进行维护与升级，建立完善的维护机制，定期对系统进行检查、维护、更新，预留相应的升级端口，以便对系统进行升级和再次改造，适应数据转型背景下开展档案业务的需求。

3.3 推行数据标准化和规范化处理

在数据转型背景下，企业在对档案业务过程进行数字化管控时，应对数据标准化与规范化管理事务进行严格把关控制。企业应制定唯一、统一的数据标准，以确保数字化档案具备一致性和互操作性，相关标准应覆盖数据格式、命名规则、编码方式、元数据扫描等多方面的事项，保证不同来源、不同格式的数据信息能够按照统一标准得到处理和存储 [3]。之后，应对数据清洗与整合标准进行构建，对数据中存在的重复、冗余、错误等问题进行定点消除处理，提高数据质量和准确性，并对于分散在系统中的档案数据进行整合处理，形成统一的数据视图，以便后续进行分析和利用。

在该环节，由于不同系统、不同平台之间的数据格式标准存在一定差异，企业应进行数据映射和转换，制定对应的映射规则，以便将不同来源数据映射到统一标准上，再利用数据转换工具或技术，实现数据格式自动转换，以此提高数据处理的准确性。在元数据管理过程中，应作出相应的定义和规范，明确元数据收集、存储、更新、使用规范，并建立元数据标准体系，以便对元数据的完整性和一致性进行管控。通过对元数据的标准化管理，更加高效地对档案信息进行扫描、分类和检索，以此提高数据的综合利用效率。在对数据实施标准化管理过程中，应引进合适的技术工具和平台，借助先进的软硬件设施，如 ETL 工具，对数据进行抽取、转换和加载，并同步利用数据仓库技术，对数据进行集中存储，利用大数据对海量资料进行快速处理和分析，提高数据标准化与规范化管控品质和效率。

3.4 引进数据备份与容灾机制

企业应同步引进数据备份与容灾机制来提高档案业务数字化管理品

质，提高档案管理工作的容错率。管理者必须根据业务需求和数据重要性选取不同的备份方式和备份周期。将备份好的数据信息存储在安全可靠的地方，考虑数据可访问性和恢复速度，采用磁带库、云储存、离线存储、在线存储等技术，保障备份数据的安全性和可用性。

在完成数据备份后，应同步完善容灾机制，优先完成容灾站点的建设，确保数据中心与容灾站点独立相连，在发生灾难时接管主数据中心业务。

3.5 数据安全与隐私保护

企业推动档案业务过程数字化转型应重点关注数据安全和隐私保护问题，利用数据加密技术，对数据传输、存储、处理过程的安全性问题进行严格把关。常见的加密技术包含对称加密以及非对称加密，可对敏感数据信息进行有效处理，并同步利用去中心化技术，提高数据访问的安全性和可追溯性。同步建立访问控制机制和身份认证机制，保证只有授权人员才能访问数据档案资料，其中包含实施多层次的身份验证，构建密码、指纹、面部识别等多个认证体系，提高数据系统的安全性。针对企业内部的敏感数据信息，引进双重认证机制。

3.6 大数据和人工智能技术的应用

在数据转型生态视域下，应同步引进大数据与人工智能技术，提高数字化档案管理的智能化水平。大数据技术在档案业务过程中的应用可通过数据整合与挖掘，将不同来源、不同格式的档案信息进行集中处理，形成一个统一的数据仓库，提取出有价值的信息，为业务决策提供支持。期间，可利用数据可视化技术，将复杂档案数据以直观、易懂的方式呈现出来，有助于用户更好地理解和利用档案数据信息，提高档案工作的效率，提升档案资源的利用价值。

在人工智能系统板块，企业可利用人工智能技术对档案数据进行智能分类检索，提高档案管理的准确性和效率，如自然语言处理技术、模式识别技术。此时，系统可基于用户行为分析和档案内容评估，提供智能推荐和辅助决策等服务，向用户推荐相关资源，提供个性化服务，并同步利用机器学习技术、训练模型，对档案信息进行自动识别整理，利用深度学习技术，对档案资料进行审核，减轻人工审核负担和压力。

3.7 云计算与云储存技术应用

企业随时会产生大量的数字档案信息，对档案信息进行存储管理是一项复杂且成本高昂的工作事务。企业应从成本、安全、可访问性等多个层面来存储档案资料，引进云计算和云储存技术来解决档案存储管理过程中所遇到的瓶颈问题。期间，应构建起一个稳定、可靠的云计算平台，作为档案业务数字化的基础。相关平台具备良好的可扩展性、灵活性和安全性，能够满足企业档案业务不断增长和变化的需求，在该环节，应将线上云计算平台与线下档案资源库进行动态调控，将档案资源进行集中配置，将本地资源与现场资源进行关联，根据业务需求调整资源配置，以保证业务连贯性和可管控性。

云计算平台中的档案信息应得到高效应用，企业应对档案管理服务进行改造优化，实现对服务快速部署和灵活调用，提升档案服务的效率和品质。云计算平台应提供强大的数据计算功能，能够对存储在云端的档案信息进行深度分析挖掘，并识别风险隐患问题。在云储存技术应用环节，企业应制定严谨可靠的存储方案，将档案数据分散存储在多个节点，实现数据备份和容灾管控。云储存具备自动扩展功能，可随着数据量的增长而增加存储空间。

结语

在数据转型环节，企业推动档案业务过程数字化革新面临较多的困难和挑战，企业应完善现有的技术体系和架构，在推行档案管理数字化模式的过程中完善数据标准，优化数据体系，革新技术手段，以此提高档案管理水平，降低管理成本和风险。

引用

[1] 朱锐鹏.基于平台视角探究商业生态系统协同演化的动力和机制[D].大连:东北财经大学,2015.

[2] 方译翎,曹麒麟,丁蕊.平台商业生态系统演化过程分析——基于S-D logic价值创造视角[J].商业经济研究,2020(15):77-81.

[3] 王怡静.协同演化视角下跨境电子商务平台运行机制研究[J].管理观察,2020(8):40-43.

基于灰色系统理论 GM（1，1）模型的青岛港货物吞吐量预测探究

文◆延边大学　金逢敏　安国山

引言

随着全球经济化的深入发展，港口在全球贸易方面起着重要作用。作为“一带一路”建设的关键节点，港口逐渐成为经济增长的核心力量。在“一带一路”倡议的推动下，在国家政策和战略的支持下，青岛被定义为新亚欧大陆经济走廊主要节点城市和海上战略支点，获得了“双定位城市”的称号。港口货物吞吐量的预测对于港口发展规划和制定指向性战略具有巨大意义。因此，对青岛港货物吞吐量进行科学预测尤为重要。

迄今为止，多位学者已采用不同方法对不同地区港口的货物吞吐量进行了深入且系统地预测研究。这些方法大致可以分为两类，一类是基于历年港口货物吞吐量数据的时间序列预测方法；另一类则是利用港口货物吞吐量的关键影响因素进行因果分析预测。例如，学者黄冲[1]建立分数阶灰色模型来对北部、东部以及南部海洋经济发展示范区内的港口货物吞吐量进行了预测；李广儒等[2]使用具有处理动态信息能力的 Elman 神经网络对宁波舟山港的港口货物吞吐量进行了全面的预测。刘明英等[3]则是基于季节影响，运用一元回归模型对防城港的港口货物吞吐量进行了预测。基于此，本文以青岛港为研究对象。由于影响港口货物吞吐量的因素较多，且在港口吞吐量内外部影响因素难以完全明确、可提供的分析样本有限的情况下，采用灰色系统理论 GM（1，1）模型从港口货物吞吐量历史数据出发，构建青岛港吞吐量预测模型。

1 GM（1，1）模型的实际应用

1.1 样本数据来源

本文数据来源于《中国港口年鉴》，其中选取 2013—2022 年青岛港货物吞吐量作为研究的初始数据，以 GM（1，1）模型作为研究方法来预测青岛港 2023—2028 年的货物吞吐量，青岛港 2013—2022 年货物吞吐量数据如表 1 所示。

表 1 青岛港 2013 年—2022 年货物吞吐量数据

年份	货物吞吐量（万吨）	年份	货物吞吐量（万吨）
2013	45003	2018	54250
2014	46802	2019	57736
2015	48453	2020	60459
2016	50036	2021	63029
2017	51031	2022	65754

1.2 数据处理与计算

1.2.1 准光滑比检验

在运用 GM（1，1）模型进行预测之前，首先需要进行准光滑比检验，检验公式为 $p(k)=\frac{X^{(0)}(k+1)}{X^{(1)}(k)}(k=1,2,3...,n)$。检验公式中，满足 $k>2$，$p(k)<0.5$ 时，可以认为初始数据及其通过一次累加生成的数据通过检验，满足使用 GM（1，1）的条件，数据呈指数递增趋势。假设若数据不满足上述条件，则需要对初始数据进行处理，以确保其满足条件。根据表

【作者简介】金逢敏（2000—），男，朝鲜族，吉林延吉人，硕士，研究方向：港口物流、市场营销。
【通讯作者】安国山（1973—），男，朝鲜族，吉林延吉人，博士，教授，研究方向：港口物流、市场营销。

1 提供的信息，初始数据 $X^{(0)}$ 如下。

$$X^{(0)}=\{45003,46802,48453,50036,51031,54250,57736,60459,63029,65754\}\quad(1)$$

再将原始数列进行一次累加，以下为获得的一次累加数据。

$$X^{(1)}=\{45003,91805,140258,190294,241325,295575,353310,413769,476798,542552\}\quad(2)$$

根据 $X^{(0)}$ 和 $X^{(1)}$ 获得光滑比如下。

$$P(k)=\{1.0340,0.5278,0.3567,0.2682,0.2248,0.1953,0.1711,0.1523,0.1379\}\quad(3)$$

由于该光滑比满足条件：当 $k>2$ 时，$p(k)<0.5$，表明数据通过了准光滑比检验，因此基于 GM（1，1）进行预测具有合理性。

1.2.2 **滑动平均的计算**

滑动平均值公式为 $Z^{(1)}=\frac{x^{(1)}(k)+x^{(1)}(k-1)}{2}$，由此公式获得滑动平均值数列如下。

$$Z^{(1)}=\{68404,116031.5,165276,215809.5,268450,324443,383540.3,445284.2,509675.8\}\quad(4)$$

1.3 模型的应用以及求解

1.3.1 **模型的应用**

基于 GM（1，1）模型灰色微分方程，可以构建出关于数据依次累加数列的一阶白化方程，表达式如下。

$$\frac{dx^{(1)}}{dt}+\alpha x^{(1)}=p\quad(5)$$

最小二乘法估计参数矩阵 $\hat{a}$ 的表达式如下。

$$\hat{a}=(a,p)^t=(B^TB)^{-1}B^TY\quad(6)$$

B 和 Y 的矩阵表达形式如下。

$$B=\begin{bmatrix}-z^{(1)}&(2)&1\\-z^{(1)}&(3)&1\\&\vdots&1\\-z^{(1)}&(4)&1\end{bmatrix}\quad(7)$$

$$Y=\begin{bmatrix}x^{(0)}&(2)\\x^{(0)}&(2)\\&\vdots\\x^{(0)}&(2)\end{bmatrix}\quad(8)$$

将 B 矩阵和 Y 矩阵带入最小二乘法的估计参数矩阵，计算得出 $\hat{a}$，再带到白化方程和灰色微分方程，计算得出的一次累加数据 $X^{(1)}$ 预测值，表达式如下。

$$\hat{x}^{(1)}(k)=(x^{(0)}(1)-\tfrac{p}{\alpha})e^{-ak}+\frac{p}{\alpha},\ k=1,2,\cdots,n\quad(9)$$

将一次累加数据 $X^{(1)}$ 预测值，即式（9）进行累减，就可还原得出计算原始数据序列的预测值，表达式如下。

$$\hat{x}^{(0)}(k+1)=x^{(1)}(k+1)-x^{(1)}(k)=(1-e^{a})(x^{(0)}(1)-\frac{p}{\alpha})e^{-ak},k=1,2,...,n\quad(10)$$

1.3.2 **模型求解**

首先，将滑动平均值式（4）和初始数据列式（1）分别代入式（7）和式（8），计算得出的 B 矩阵和 Y 矩阵如式（11）、式（12）所示。

$$B=\begin{bmatrix}-68404&1\\-116032.5&1\\-165276&1\\-215809.5&1\\-268450&1\\-324443&1\\-383540.3&1\\-445284.2&1\\-509675.8&1\end{bmatrix}\quad(11)$$

$$Y=\begin{bmatrix}46802\\48453\\50036\\51031\\54250\\57735.9\\60458.6\\63029.3\\65754\end{bmatrix}\quad(12)$$

其次，将 B 矩阵和 Y 矩阵代入式（6）和式（8），再计算获得 $\hat{a}$ 的值，表达式如下。

$$\hat{a}=(a,p)^{\mathrm{T}}=\begin{bmatrix}-0.0447\\42873.0082\end{bmatrix}\quad(13)$$

最后，由估计参数矩阵 $\hat{a}$ 和原始数据序列的预测值式（10），求得原始数据序列的预测值如下。

$$\hat{X}^{(0)}=\{345003,45905.16,48005.22,5020137,52497.98,54899.65,57411.19,60037.63,62784.22,65656.47\}\quad(14)$$

1.4 对预测数据的检验、预测

本文通过计算平均相对误差和进行关联度验证，旨在检测预测数据的有效性。为此，首先应计算青岛港 2014—2022 年货物吞吐量的实际值和预测值的残差，其计算公式如下。

$$\sigma_k=x^{(0)}(k)-\hat{x}^{(0)}(k)\quad(15)$$

再计算相对误差，其计算公式如下。

$$\delta_{(k)}=\left|\frac{x^{(0)}(k)-\hat{x}^{(0)}(k)}{x^{(0)}(k)}\right|*100\%\quad(16)$$

关联系数的计算公式如下。

$$\xi_{(k)}=\frac{\min\limits_i\min\limits_k\left|x^{(0)}(k)-\hat{x}^{(0)}(k)\right|+\beta\max\limits_i\max\limits_k\left|x^{(0)}(k)-\hat{x}^{(0)}(k)\right|}{\left|x^{(0)}(k)-\hat{x}^{(0)}(k)\right|+\beta\max\limits_i\max\limits_k\left|x^{(0)}(k)-\hat{x}^{(0)}(k)\right|}*100\%\quad(17)$$

（β=0.5，k=2，3，…，n），由此计算结果见表 2。

由表 2 计算出来的相对误差，再计算平均相对误差，求得其平均误差的计算公式如下。

$$\bar{\delta}=\frac{1}{n}\sum_{k=1}^{n}\left|\frac{x^{(0)}(k)-\hat{x}^{(0)}(k)}{x^{(0)}(k)}\right|*100\% \quad (18)$$

经计算得出预测数列的平均误差为 1.004%，因此预测精度高达 98.996%。平均关联系数计算公式为 $\xi=\frac{1}{n}\sum_{k}^{n}\xi_k$，经计算得出预测数列的平均误差为 71.90%，大于 60%。平均相对误差和关联度验证均表明本文的预测具有较高的精度，因此可以合理地将结果代入式（10）中，对 2023—2028 年青岛港未来 6 年间的货物吞吐量进行预测，且具有合理性，预测结果见表 3 和图 1。

2 结论及展望

本文以灰色系统理论 GM（1，1）作为理论基础，选取青岛港货物吞吐量为研究对象，并将其 2013—2022 年的货物吞吐量作为初始数据，预测 2023—2028 年青岛港货物吞吐量。预测结果的结果均通过平均相对误差和关联度验证，证明了预测数列具有较高的精度，因此运用 GM（1，1）进行预测具有合理性。由表 2 和图 1 可以看出，2016—2018 年的预测值对比实际值偏大，其余时间段的预测值对比实际值偏小；青岛港货物吞吐量会在未来持续增加并预测在 2024 年到达 7.1 亿吨，2027 年到达 8.2 亿吨。以上表明，随着我国建设海洋强国战略的不断深入和海上丝绸之路的不断拓展，作为我国北方地区的重要港口，青岛港拥有广阔的发展前景。

表 2 实际值和预测值的相对误差以及关联度

年份	实际值	预测值	残差	相对误差	关联系数
2014	46802	45905.16	896.84	1.92%	50.97%
2015	48453	48005.23	447.77	0.92%	70.35%
2016	50036	50201.37	-165.37	-0.33%	92.45%
2017	51031	52497.98	-1466.98	-2.87%	37.76%
2018	54250	54899.65	-649.65	-1.19%	60.08%
2019	57736	57411.19	324.76	0.56%	78.53%
2020	60459	60037.63	421	0.69%	71.98%
2021	63029	62784.23	245.04	0.39%	84.93%
2022	65754	65656.47	97.53	0.14%	100%

表 3 2023—2028 年青岛港货物吞吐量（单位：万吨）

年份	2023	2024	2025	2026	2027	2028
预测值	68660.12	71801.18	75085.93	78520.95	82113.12	85869.62

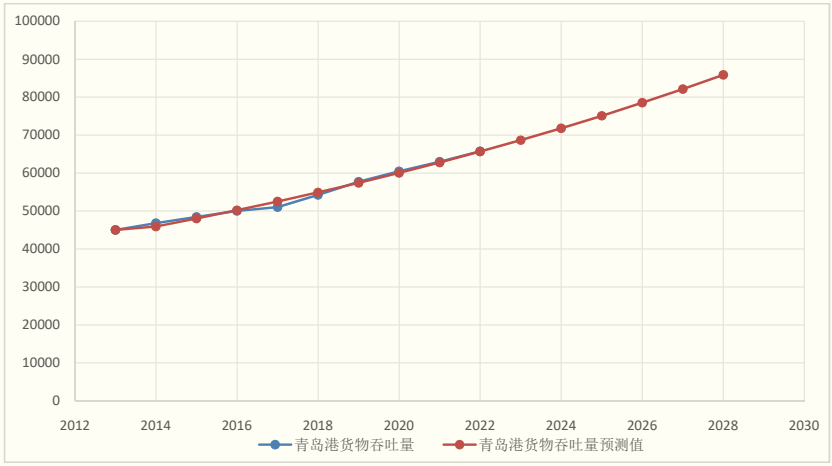


图 1 2012—2028 年青岛港货物吞吐量实际值与预测值以及趋势

在未来的研究中，可以基于本文的内容进行更深入地研究。本研究虽然建立了 GM（1，1）模型来对青岛港货物吞吐量发展的趋势进行预测，但并未深度挖掘除时间因素以外的其他因素对货物吞吐量的影响。因此，在接下来的研究中，应进一步挖掘其他影响因素与货物吞吐量之间的关系以及年度范围出现的季节性波动影响。此外，虽然 GM（1，1）模型在短期趋势预测中具有一定的优势，但对于中长期预测的效果并不稳定。因此，后续的研究中，应通过将多种预测模型结合使用，进行对比预测，从而提高预测的准度和效果。

引用

[1] 黄冲,张凯.海洋经济发展示范区港口建设水平分析及未来趋势预测[J].中国海洋经济,2022,7(2):39-55.

[2] 李广儒,张新,朱庆辉.基于Adaboost的改进Elman神经网络港口吞吐量预测方法[J].重庆交通大学学报(自然科学版),2021,40(5):1-5.

[3] 刘明英,翁世洲,贺依婷.基于季节分解法的防城港港货物吞吐量预测及分析[J].水运管理,2019,41(2):15-17+32.

基于国产数据库的大量设备日志管理功能设计与实现

文◆中国电子科技集团公司第三十研究所　李开文　陈德勇

引言

2009 年，某研究所项目组开发设备日志管理系统时，用户提出需要设备日志导出功能，以离线方式查看设备日志，并实现日志归档。基于此需求，项目组使用开源框架 PHPExcel 实现了设备日志导出功能。

由于受限于当时的硬件环境，系统提供的设备日志管理性能非常有限并且耗时较长。但随着服务器硬件突飞猛进的发展，硬件成本降低，性能越来越好。此外，国产数据库最近 10 年发展迅猛，国产数据库的稳定性和性能明显提高，与国外主流数据库的差距越来越小。无论硬件改善还是软件提升，日志管理功能经过长期迭代后不应停留在之前的指标上。

基于国产数据库，首先，根据系统提供的查询条件和国产数据库的特点，通过优化索引将数据库查询的响应时间控制在极短时间内。其次，使用“PHP+ZendFrame”技术，结合系统的代码优化、技术重构，以分页的方式实现大量数据的快速浏览。最后，导出时将日志数据写入多个 CSV 文件，压缩成一个文件供用户下载。优化后的设备日志管理不仅实现千万数据级的高效管理，还极大地降低了系统资源消耗，提升了系统性能。

1 数据库优化设计

设备日志数据库涉及 3 个表，即设备基本信息表、日志事件表和设备日志表。设备基本信息表存储设备的基本信息，如设备编号、设备名称等；日志事件表用于存储事件信息，如事件 ID、事件名称等，设备上报日志时只上报事件 ID，不上报事件名称，以节约网络流量；设备日志表用于存储设备上报的日志信息，包括日志对应的事件 ID、日志产生时间、结果、日志内容等。设备经过长年累月运行必将产生大量日志，因此，设备日志表的设计应考虑存储大量数据的情况。

经过反复尝试和验证，并结合国产数据库的特点，设备日志表中的数据达到百万后，通过表关联进行联合查询时效率极其低下，难以在数据库层面快速获取查询结果。若实现秒级响应，只能使用单表查询。由于设备基本信息表字段较多，且设备名称、IP 地址等存在编辑修改的情况，因此，设备基本信息表不能合并到设备日志表。

日志事件表主要存储事件 ID 和事件名称的对应关系，设备上报日志时使用对应的事件 ID。若设备上报日志时直接使用事件名称，则无需日志事件表。如此，设备上报日志时将占用更多网络带宽，但对于动辄千兆甚至万兆的带宽来说，增加的几个或十几个字节完全可以忽略，且设备可以随意扩展事件名称。

经过优化后，日志数据库只保留设备基本信息表和设备日志表，各表设计如下。

1.1 设备基本信息表设计

设备基本信息表（TBL_EQUIPMENT）如表 1 所示，主要存储设备的基本信息，包括设备编号、设备名称、生产厂商、软件版本、出厂号等。设备基本信息表存储的设备信息，一般不超过 10 万条数据，表中的各个字段和类型相对简单。

1.2 设备日志表设计

设备日志表如表 2 所示，主要存储设备上报的日志信息，包

【作者简介】李开文（1981—），男，河南周口人，学士，工程师，研究方向：密码管理。

括日志编号、设备编号、日志结果、发生时间、详细描述等。此表保存上千万条数据，因此应合理设计表中的各个字段和每个字段的数据类型，避免空间浪费。设备日志表的日志名称用于存储日志的简要描述，如设备开机、设备关机、设备远程重启等，一般 32 个字符即可。设备编号和设备基本信息表中的设备编号存在对应关系，二者必须一致。日志级别、日志结果都是有限数量的枚举类型，使用 Tinyint 即可满足需要。日志的详细描述用于日志的详细信息，是失败原因或其他扩展信息，应设计为 Text 类型，确保数据不丢失。

1.3 外键和索引设计

系统接收设备上报的日志时需记录设备 ID，以便区分不同的设备，因此，设备日志表中的设备 ID 应设计成外键，对应设备基本信息表的设备 ID。但设备基本信息删除后，一般需保留其日志信息，若要保留删除设备的日志信息，那么设备日志表中的设备 ID 不能和设备基本信息表有外键关系 [1]。

通常情况下，设备日志表中的设备 ID 不作为外键，而是作为索引，用于快速定位某一设备的日志 [2]。

系统提供设备名称、日志名称、发生时间、日志结果等查询条件，因此应将设备基本信息表中的设备名称，设备日志表中的日志名称、日志结果、发生时间设计为单列索引。

经验证，以上设计在人大金仓数据库中能够完成快速查询 [3]，但在达梦数据库中，因为存在全表遍历 [4]，查询效率依然特别低下，必须创建包括设备 ID、日志名称、日志结果、发生时间的多列索引才能实现快速查询 [5]。

表 1 设备基本信息表

字段描述	字段名称	字段类型	说明
设备编号	EM_ID	int	主键
设备名称	EM _NAME	varchar(64)	
生产厂商	EM_PRODUCT	varchar(64)	
软件版本	EM_SOFTV	varchar(32)	
出厂号	EM_MANUNO	varchar(32)	
设备 IP	EM _IP	varchar(15)	

表 2 设备日志表

字段描述	字段名称	字段类型	说明
日志编号	LG_SEQ	int	主键
设备编号	LG_EMID	int	
日志名称	LG_NAME	varchar(32)	
日志结果	LG_RST	tinyint	
日志对象	LG_OBJ	varchar(32)	
发生时间	LG_TIME	datetime	
详细描述	LG_CONTENT	text	

综合以上，设备基本信息表的设备名称作为索引列，设备日志表的设备 ID、日志名称、日志结果、发生时间等 4 个字段作为索引列，同时创建包含设备名称、日志名称、日志结果、发生时间的多列索引 [6]。

2 代码优化实现

处理大量数据，应慎重考虑数据库查询中的多表关联查询以及业务逻辑处理中的嵌套循环。任何问题经过百万倍放大后都会成为系统的性能瓶颈。日志管理功能主要包括数据库查询、数据处理、数据列表显示，并提供以文件方式将日志导出。

2.1 更新程序控件

经过反复验证，PHPExcel 处理 Excel 时非常耗时，6 万条数据从优化后的数据库中查询出来耗时 5s，而生成文件需要 3min；导出 10 万条数据时，系统各种异常、卡顿，无法正常导出。

经过对比和反复测试发现，使用 PHP 自带的 fputcsv 函数 [7] 将文件保存成 CSV 格式可以大幅提升生成文件的效率。使用达梦数据库，从 3500 万条设备日志中导出符合条件的 111 万条数据并生成文件，全程耗时可压缩到 1.2min。

2.2 取消表关联

设备日志表关联设备基本信息表后，查询效率极其低下，即使建立索引，当数据达到百万、千万级别时，第一次查询数据库就需要 45s 左右。取消关联后，从设备基本信息表获取设备 ID 和设备名称的对应关系，在遍历日志列表时根据设备 ID 和设备名称的对应关系，将设备 ID 转换成设备名称。简化后，只需要从设备日志表中查询数据，所有的查询条件都在设备日志表中。

经验证，取消表关联后数据库查询效率大幅提高，从 3000 万数据中筛选出 100 条数据仅需 1s 左右。

2.3 压缩导出的文件

当导出的数据过多时，由于单个 CSV 文件最多只能存储 104 万条记录，应将数据存放在不同的文件中，以防数据丢失。为方便用户下载，应将文件打包成一个文件，并且，文件压缩后既节约大量磁盘空间，又提高了下载效率。

在测试环境中，不同的日志数量，导出 CSV 文件压缩前后对比如表 3 所示。

表 3 导出 CSV 文件压缩前后对比

日志数量（万）	压缩前（MB）	压缩后（MB）
19.5	24.4	1.28
48	60	3.22
261	188	12.6
324	406	26.5
642	642	43.3

3 数据验证和对比

为验证以上数据库设计和功能实现效果，搭建测试环境，对达梦数据库、人大金仓数据库、MYSQL 数据库进行对比分析，内容分为无索引和有索引。

配合验证的环境包括 3 台相同配置的长城服务器，安装银河麒麟操作系统 V4.0，数据库软件分别使用达梦数据库 V8.1，人大金仓数据库 V8.2, MYSQL 数据库 V5.7.26，且均使用默认配置。

3.1 查询统计

日志总数量为 1100 万。使用谷歌浏览器登录日志管理系统页面，设置查询条件为日志级别为调试、日志结果为异常，设备名称包含“测试网络”，日志内容包含“test”，日志时间从 2024–01–01 开始，结束时间分别为 2024–01–31、2024–02–29、2024–03–31、2024–04–30、2024–05–31。分别查询出 6600，12700，19100，25600，29000 条数据，每页加载 500 条，页面加载最后一页的时间对比。无索引复杂查询数据对比如表 4 所示，有索引复杂查询数据对比如表 5 所示。

表 4 无索引复杂查询数据对比 （单位：s）

MYSQL	24.54	25.62	25.12	25.66	26.64
达梦	7.66	8.35	9.15	9.98	10.35
金仓	31.20	32.00	32.42	33.33	33.94

表 5 有索引复杂查询数据对比 （单位：s）

MYSQL	1.09	1.15	1.16	1.16	1.13
达梦	1.95	3.45	5.05	6.68	7.61
金仓	3.79	4.83	3.71	3.71	3.76

3.2 数据分析与总结

根据以上测试数据，在执行复杂的查询语句时，达梦、人大金仓仍落后于 MYSQL，但在某些方面，达梦、人大金仓的表现要优于 MYSQL。创建索引可以提高查询效率，但需要根据查询条件并结合国产数据库特点配置索引。

结语

本文基于国产数据库，通过数据库优化设计、代码逻辑优化、系统配置优化等实现了大量设备日志查询和导出功能，将设备日志管理系统的性能指标提升到千万级别，同时降低了系统资源消耗和硬件成本，系统响应时间提升明显。

引用

[1] 姚世军.Oracle9i 数据库原理与应用教程[M].北京:国防工业出版社,2004:188-191.

[2] 王珊,萨师煊.数据库系统概论[M].北京:高等教育出版社,2014: 231-235.

[3] 北京人大金仓信息技术股份有限公司.Kingbase产品手册:4.2阅读执行计划[DB/OL].[2023-10-31].https://help.kingbase.com.cn/v8/perfor/sql-optimization/sql-optimization-8.html#index-scan.

[4] 武汉达梦数据库股份有限公司.DM8系统管理员手册:查询优化[DB/OL].[2023-10-31].https://eco.dameng.com/document/dm/zh-cn/pm/query-optimization.html.

[5] 武汉达梦数据库股份有限公司.DM8 SQL语言使用手册:数据查询语句[DB/OL].[2023-10-31].https://eco.dameng.com/document/dm/zh-cn/pm/check-phrases.html.

[6] 甲骨文公司. MySQL 5.7 Reference Manual:13.1.14 CREATE INDEX Statement[DB/OL].[2023-10-31].https://dev.mysql.com/doc/refman/5.7/en/create-index. html.

[7] www.php.net.fputcsv[CP/OL].[2023-10-31].https://www.php.net/manual/zh/function.fputcsv.php.

基于雪茄用户画像的雪茄共享仓模式分析综述*

文◆云南省烟草公司丽江市公司　李祥华　孙　锐　张辰初　马文广
成都鹅掌楸科技有限公司　郭俊松
青岛海大新星软件咨询有限公司　宫会丽

引言

雪茄用户在社会化网络平台上对雪茄的认知可以反映出国内雪茄消费特征与趋势，有助于国产雪茄企业了解国内雪茄消费者的消费偏好。目前对于雪茄用户画像的构建缺乏系统性方法，本文利用文献研究法，对雪茄用户画像数据抓取方式、数据处理、建模路径进行研究，将获取的部分数据进行相关处理验证，进而获取雪茄用户画像建模路径。通过 LDA 主题模型获取雪茄用户关键词，将获取的关键词利用社会网络分析构建雪茄用户偏好社交网络，进而生成雪茄用户画像，并且利用 SnowNLP 对雪茄用户情感进行分析，获取雪茄用户群体在不同时间段关注雪茄的情绪变化，再应用 LDA 主题偏好模型构建雪茄用户画像，解决雪茄共享仓储模式中关键性指标问题，构建数据驱动、平台支撑、组织重构的新型雪茄共享仓体系，加速雪茄仓储数字化、网络化、智能化转型升级，为国产雪茄共享仓储开发提供新的考量依据。

1 研究背景

移动互联网时代，网络用户快速扩增，社交化网络平台成为用户发表观点、表达个人情绪以及结交朋友的载体，网络用户因相似属性聚集，建构起丰富多样的虚拟社群。研究发现，微博用户线上分享的内容对用户特征研究有着重要意义，可以准确发现用户在关注社会热点、情感强度、态度倾向等方面的特征，对于深入研究用户群体行为具有重要作用。

用户画像（User Profile）是基于特定使用情境下真实用户数据所形成的描述用户属性及其特征的标签集合[1]。用户画像在预估用户行为、发现用户潜在需求[2]等方面均发挥了重要作用，成为管理政策执行的有效依据。随着大数据应用的不断革新，目前众多学者对用户画像进行了相关研究。王洋等[3]将用户浏览行为日志以及爬虫数据作为相互补充构建用户画像，并基于 Hadoop 大数据平台对用户数据进行清洗、规范化、分析与处理，分析出用户兴趣偏好，改进 K-means 算法，用户画像准确率得到了较大提升。张亚楠等[4]采用一种考虑全局和局部信息的科研人员科研行为画像方法，引入机器深度学习算法，借助深度学习自动从数据中提取高度抽象特征的特点，构建科研人员的立体精准画像。林燕霞等[5]基于社会认同理论，用户使用微博过程中做出的行为特征、展现的态度与他们感受自己所属群体成员身份紧密相关。马超[6]通过基于 4 种不同的评价标准构建主题模型，并基于主题模型的半监督学习算法框架能够将部分已知的用户属性信息作为监督信息加入模型中，提升用户画像的准确率。张烨等[7]利用网络爬虫技术获取 Quora 中高血压主题下的用户社交行为数据，构建不同类型用户社交网络，分析用户社交网络结构和用户社交类型的特征，进而生成面向高血压主题的用户画像，直观地展现用户兴趣特征、情感丰富度、态度倾向、消费趋势，有助于进一步正确深入研究社会化网络平台用户的群体特征行为。

国内外对用户画像的构建研究集中在只从单一某个方面进行偏好特

*【基金项目】云南省烟草专卖局科技项目“基于数字化的雪茄管养技术研究与应用”（2022530700242005）
【作者简介】李祥华（1975—），男，云南丽江人，本科，助理经济师、高级物流师，研究方向：市场营销、物流管理。
【通讯作者】马文广（1972—），男，云南玉溪人，硕士研究生，研究员，研究方向：烟草品种选育及种子科学。

征挖掘或只基于静态特征偏好进行研究，分析的维度单一，缺乏实时性，忽略了用户兴趣偏好的相关性、动态性、情境性以及潜在的社交网络关系。在新的消费环境下，情境成为连接信息资源和用户的桥梁，只有极少数的研究将用户某个情境带到用户的推荐系统中，无法精确了解用户所处的真实情境。张超等[8]在研究中融入了用户的情境属性要素，将用户画像维度划分为自然属性、社交属性、兴趣属性和能力属性，提高了用户画像准确性。

2 雪茄用户画像构建路径

本文以 Python 程序语言开发爬虫程序，抓取微博雪茄用户发布的社会化网络分享信息，将数据合并后进行数据清洗，包括去重、提取目标文本、文本分词等，形成雪茄用户群体基础数据库，应用雪茄用户画像 LDA 主题模型和社会化标签分析，分别获取雪茄用户偏好特征和雪茄用户社交网络图，从而建立雪茄用户画像，雪茄共享仓构建路径图如图 1 所示。

3 雪茄用户数据抓取

3.1 数据采集时间及社交媒体平台选择

微博在 2016—2020 年是国内社交媒体的热点平台，是当时国内最突出的虚拟社交群体之一。根据新浪微博财报显示，2016—2020 年是微博用户最为互动频繁的时期，在此期间，国内雪茄消费人群增长较快，且雪茄用户偏好特征具有一定的代表性。用户在微博平台以文本和图片形式分享兴趣偏好以及情绪表达，用户信息清晰明确，有利于描述用户画像。

3.2 微博雪茄用户群体数据抓取方法及流程

孙锐等[9]在雪茄关注热点的时空演化中利用 Python 程序抓取雪茄用户文本信息。本文采用自编爬虫语言抓取 2016—2020 年间包含雪茄关键词的新浪微博网页的文本数据，采集后的数据经过分词词典处理、词语词频统计、无效数据过滤，以标准化格式进行存储。将处理后的基础数据以用户为单位，按照时间和场景维度序列化方式存储。

4 构建雪茄烟用户 LDA 主题模型

将处理存储的数据按照格式化存储的用户“文档—词项矩阵”输入，然后对输入的“文档—词项矩阵”进行 Gibbs 抽样迭代，得出雪茄用户主题类型，根据相关性度量指标，用来判断词汇与特定雪茄主题的相关性程度。C.Sievert 等提供一个相关性度量指标，计算公式如式（1）所示，用来判断不同的词汇对特定主题的相关性程度大小，最后得到该主题的主要内容。

$$relevance\left(\frac{\varpi}{\tau}\right)=\lambda\times\rho\left(\frac{\varpi}{\tau}\right)+(1-\lambda)\times\rho\left(\frac{\varpi}{\tau}\right)/\rho(\varpi) \quad (1)$$

式（1）中，$relevance(\frac{\varpi}{\tau})$是特定词语在某个主题下的相关度，$\rho(\varpi)$是特定词语在某个主题下出现的概率，$\rho\left(\frac{\varpi}{\tau}\right)$是特定词语出现的总率。

5 构建雪茄烟群体用户社会网络

基于社会认同理论，根据雪茄用户群体多种行为标签的内在相互关联形成共词矩阵，进行用户社会网络构建，并基于社会网络分析指标对雪茄用户标签网络的权重进行计算，将以“雪茄”

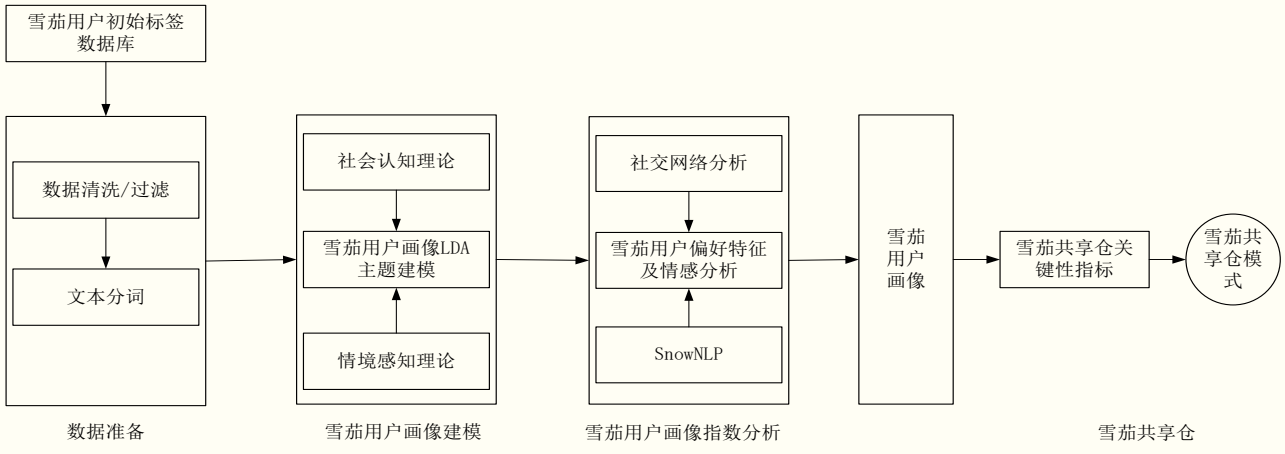


图 1 雪茄共享仓构建路径图

为关键词进行表达的社交媒体使用者的社会关系进行凝聚化表达，形成“用户—资源—标签”的多维度标签体系，即凝聚子群。从形成社交网络的连通性、稀疏性和凝聚性刻画雪茄关键词社交网络之间联系紧密，从而反映其用户群体的偏好之间的联系。

雪茄用户画像场景社会网络图如图 2 所示，图中圆圈的大小表示度中心度的大小，具体计算过程如式（2）所示，度中心度的大小表明其在雪茄群体用户不同偏好重要程度，连线上的数字表明雪茄与其两者之间的联系强弱。

$$C_D(N_i)=\sum_{J=1}^{g}x_{ij}(i\neq j) \quad （2）$$

用圆点大小来表示词语的关联度大小，即与该词有连线的词语数量。社会网络图使用的是无向网络图，关联度也能表示点的中心性，即核心地位大小，因此，词语的关联度越大，圆点越大，词语也就越重要。

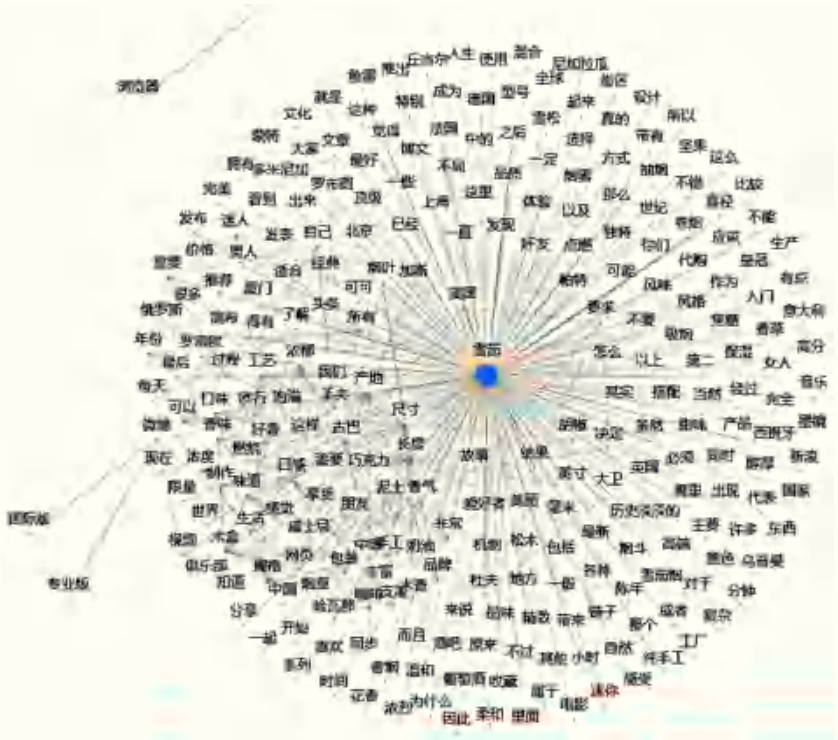


图 2 雪茄用户画像场景社会网络图

结语

本文基于社会认同理论，对社交平台微博 2016—2020 年雪茄用户进行数据爬虫，提取雪茄用户关键词，并利用情境感知理论，将时间与情境因素融合到雪茄用户群体画像的构建中，解决了用户画像动态性，同时通过标签社会网络得到雪茄用户群体在雪茄风味、尺寸、消费场景、配饮偏好的主题内容以及偏好间关系，解决了新消费环境下雪茄用户群体潜在的社会网络关系，提升了雪茄用户群体画像的准确性。基于雪茄用户画像开发雪茄共享仓模式，发挥雪茄烟管养“平台化”作用，将“共享仓储”引入雪茄烟管养的仓储运行，结合数字化应用，集中资源开发雪茄烟管养技术，并结合烟草商业公司的“官方”信任背书，有利于消费者以及零售渠道接受雪茄烟管养醇化后雪茄烟市场价值，打通雪茄烟市场培育的品质提升、信用评价以及数字化服务。

引用

[1] 陈烨,王乐,陈天雨,等.基于社会网络分析的社会化问答平台用户画像研究[J].情报学报,2021,40(4):42-46.

[2] 马费成,周利琴.面向智慧健康的知识管理与服务[J].中国图书馆学报,2018,44(5):4-19.

[3] 王洋,丁志刚,郑树泉,等.一种用户画像系统的设计与实现[J].计算机应用与软件,2018(3):8-14.

[4] 张亚楠,黄晶丽,王刚.考虑全局和局部信息的科研人员科研行为立体精准画像构建方法[J].情报学报,2019,38(10):1012-1021.

[5] 林燕霞,谢湘生.基于社会认同理论的微博群体用户画像[J].情报理论与实践,2018,41(3):142-148.

[6] 马超.基于主题模型的社交网络用户画像分析方法[D].合肥:中国科学技术大学,2017.

[7] 张烨,王乐,陈天雨,等.基于社会网络分析的社会化问答平台用户画像研究[J].情报学报,2021,40(4):414-423.

[8] 陈晶.移动数字图书馆用户画像模型及其情境化推荐方法研究[J].图书馆,2018(4):19-23.

[9] 马文广,孙锐,郭俊松,等.基于7Es组合的中式雪茄烟体验营销战略分析[J].中国市场,2022(16):133-135+138.

构建高质量医院数字化管理信息平台的研究

文◆中山大学附属第一医院　林　琳　王宇翀　申芷晴

引言

国家卫健委印发“十四五”全民健康信息化规划的通知称，到2025年，初步建设形成统一权威、互联互通的全民健康信息平台支撑保障体系，基本实现公立医疗卫生机构与全民健康信息平台联通全覆盖。随着信息技术的不断发展，数字化管理已成为医院管理的重要趋势和必要手段。因此，构建高质量医院数字化管理信息平台，有利于实现医院各项业务的全面覆盖和高效管理，同时还对提升医院运营效率和医疗服务质量具有重要的现实意义和深远的社会影响。

1 医院数字化管理信息平台概述

1.1 信息平台在医院管理中的作用

1.1.1 提升医院的管理效率

传统的医院管理模式依赖纸质文档和人工操作，不仅耗时耗力，还容易出现信息错漏、丢失等问题。数字化管理信息平台通过集成各类管理系统，可以实现信息的实时录入、查询和共享，提高了管理流程的自动化和智能化水平[1]。以病历管理为例，数字化平台可以实时更新患者病历信息，医生只需通过平台即可快速查阅，极大地节省了时间成本。同时，平台还可以对医疗资源进行合理配置，如手术室等医技的预约、药品耗材的库存管理等，有效避免了资源的浪费和闲置。

1.1.2 优化患者就医体验

在现代社会，人们对医疗服务的要求越来越高，便捷、高效、个性化的服务成为患者关注的焦点。数字化管理信息平台整合各个患者服务系统，并统一对外提供在线预约、挂号、缴费等服务，简化了患者的就医流程，减少了排队等待的时间。同时，平台还可以根据患者的历史就诊记录和健康状况，提供个性化的健康管理和建议，使患者能够更好地了解自己的身体状况，推动其积极参与自我健康管理。

1.2 信息平台的功能需求与业务需求

近年来，国家愈发重视医疗行业的信息化、服务化、便捷化发展，出台了一系列政策推动医疗行业的数字化转型。例如，2023年颁布的《国家卫生健康委属（管）医院分院区建设管理办法（试行）》，该文件明确指出要求委属（管）医院利用信息化手段提升医疗服务质量和效率。医院作为医疗服务的主要提供者，亟须构建一个高质量的数字化管理信息平台，以满足日益增长的医疗需求和管理挑战[2]。

1.2.1 功能需求

（1）数据采集治理平台。面对患者多元化的医疗需求与差异化的身体信息数据，医院应拥有一个数据采集治理平台，以实现对患者各类医疗信息的全面、准确、高效采集和治理。该平台应具备强大的数据整合能力，能够对接各类医疗设备、系统，实现数据的实时同步和标准化处理。通过数据采集平台支撑数据中台数据的分布式采集、采集任务可视化编排、数据集中汇聚、运维监控可视化维护，以便于解决数据中台数据采集中多租户管理问题、协议适配问题、采集组件二次开发问题，保障数据高效、准确、及时地汇聚到数据中台。

（2）数据开发治理平台。随着医院业务的不断扩展和复杂化，数据开发治理平台成为数字化管理信息平台的核心组件。该平台提供了数据建模、数据处理、数据分析等一系列功能，帮助医院

【作者简介】林琳（1982—），男，广东潮州人，硕士研究生，工程师，研究方向：信息工程。

从海量数据中挖掘有价值的信息，为决策支持、业务优化等提供有力依据。通过数据加工治理平台支撑数据中台数据准备，以解决数据血缘不可见、业务逻辑不可管的问题。

（3）数据资产管理平台。数据资产管理平台是医院数字化管理信息平台的重要组成部分，主要负责全面管理和维护医院的数据资产。通过该平台，医院可以清晰地了解自身数据资产的分布、状态和价值，从而更加科学地进行数据治理和决策。数据资产管理平台应具备数据目录管理、数据质量管理、数据安全管理和数据价值挖掘等功能。通过数据目录管理，实现对各类数据资源的统一登记、分类和检索的可视化运营管理，以此提高数据资源的可见性和可访问性。

（4）数据质量管理平台。数据质量管理平台主要负责监控、评估和提升医院数据的质量。在医疗行业中，数据质量直接影响诊断的准确性、治疗的效果以及患者的安全。因此，建立数据质量管理平台对于提高医疗数据的质量至关重要。通过数据质量管理平台支撑数据中台数据库全量数据的自定义规则配置、自定义规则集配置、数据分布式稽核调度、数据监控预警、数据质量报告，可以保障数据中台数据质量可管、可控、可评估。

1.2.2 业务需求

医院数字化管理信息平台的业务需求主要包括以下内容。（1）提升医疗服务质量。平台应支持医疗服务质量的持续改进和提升。通过实时监测医疗服务过程和结果，发现服务中的不足和问题，以便及时采取改进措施，提高医疗服务质量和患者满意度[3]。（2）优化资源配置。平台应根据医院的实际需求和运营状况，合理配置医疗资源。通过对搜集到的相关数据进行分析和预测，优化医疗资源的分配和使用，提高各类资源的利用效率和效益。（3）加强医患沟通与合作。平台应建立便捷的医患沟通渠道和合作机制。通过在线咨询、健康管理等功能，加强医患之间的沟通和互动，以此增强患者对医院的信任感和归属感。

2 数字化管理信息平台的设计

2.1 设计原则与框架

2.1.1 设计原则

保障数字化管理信息平台的设计科学合理，能够达到各项需求的预期效果，必须在设计平台时遵循平台的自动化原则、智能化原则以及可持续性原则[4]。自动化原则指的是在设计数字化管理信息平台时，应实现各项业务流程的自动化处理，减少人工干预，以提高处理效率。智能化原则要求平台能够利用人工智能、大数据等技术，对医疗数据进行深度挖掘和分析，为医院管理决策提供有力支持。可持续性原则强调在设计平台时，应充分考虑其未来的可扩展性和可升级性，以适应医疗行业不断变化的需求和技术发展。

2.1.2 平台框架

根据设计原则，将数字化管理信息平台的设计框架划分为 3 个层次，即数据采集层、数据治理层和数据应用层。

从应用系统与支撑系统的详细界面进行划分，该系统的主服务被设计为一个单独的程序，其余的子应用（如患者 360 视图、元数据订阅、资源目录、CDR、应用开发、服务管理、服务目录、数据地图等页面）将作为单独的模块，同时每一个子应用都可以单独使用。子应用的全局数据将以状态管理库存储在内存中，并且存储在本地存储中当作数据的持久化功能。

2.2 模块划分与功能

从系统的内部详细界面进行划分，各个模块的交互将以架构顶部的子应用分类来切换，用户可以根据自己想使用的功能对系统页面的进行点击操作，从而快速进入主要大类进行各模块的交互。

整个系统覆盖的模块有数据资产、360 视图、元数据、资源目录、共享目录、数据标准、数据质量、数据采集、数据工厂、服务目录等。同时，系统要求 Chromium 内核的现代浏览器，如 Edge 浏览器或 360 浏览器，保证用户能够流畅地访问和使用平台上的各项功能。

以数据资产页面为例，该页面主要负责展示医院的数据资产情况，包括数据的来源、类型、存储位置、使用情况等。通过数据资产页面，医院管理者可以清晰地了解医院的数据资源情况，为数据治理和决策提供支持。该页面应该具备以下功能。

（1）数据资产目录展示。展示医院所有数据资产的目录，包括数据表、视图、数据库等。按照数据来源、数据类型、数据状态等维度进行筛选和排序，以便用户快速定位所需的数据资产。（2）数据资产详情查看。当用户选中某个数据资产时，可以查看该数据资产的详细信息，包

括数据定义、数据结构、数据来源、使用情况等。同时，系统还应提供数据预览功能，让用户能够直接查看数据内容。（3）数据资产权限管理。为确保数据的安全性和完整性，系统应对数据资产进行权限管理。只有经过授权的用户才能访问和操作特定的数据资产。同时，系统还应记录用户的操作日志，以便进行审计和追溯。除了数据资产页面外，其他模块也应根据各自的功能需求进行设计和实现。通过科学合理地设计和实现数字化管理信息平台的各个模块和功能，为医院提供全面、高效、便捷的数据管理和决策支持服务，有助于提升医院的服务质量、加强医患沟通与合作，推动医院数字化转型的深入发展。

3 医院数字化管理信息平台的关键技术与应用

3.1 病人数据采集与处理技术

医疗数据采集与处理技术是数字化管理信息平台的核心技术之一。该技术涉及从各种医疗设备、信息系统和病人直接输入等多个渠道，实时、准确地采集病人的各种医疗数据。数据包括但不限于病人的基本信息、病史、检查结果、治疗方案、手术记录、护理记录等。采集到的数据需要经过审核、整合、转换等一系列处理，以保证其质量和可用性。在采集数据时，通过 API 接口、数据抽取工具、ETL 工具等方式，实现与医院现有信息系统的无缝对接，以确保数据的完整性和准确性。在实际应用过程中，该技术能够让医护人员通过平台随时查看病人的基本信息、病史、检查结果等，从而帮助该医护人员快速做出诊断并制定治疗方案 [5]。

3.2 信息安全与隐私保护技术

信息安全与隐私保护技术是医院数字化管理信息平台的重要保障。随着医疗数据的不断增加，如何确保数据的安全性和隐私性成为医院面临的重要问题。数字化管理信息平台采用先进的访问控制技术等手段，确保了病人信息的安全性和隐私性。例如，在访问控制方面，设置访问控制列表（ACL）和控制表权限的方式限制数据库访问，该控制列表只允许授权用户或应用程序访问数据库，进而完成对数据的保护。在过去，由于缺乏有效的技术手段，医疗数据的安全性和隐私性常常受到威胁。而现在，通过数字化管理信息平台，医院可以更加严格地控制数据的访问和使用，有效防止了数据泄露和滥用的情况发生。

3.3 病人的疾病风险评估技术

病人的疾病风险评估技术是医院数字化管理信息平台的重要应用之一。通过对病人的各项数据进行综合分析，平台可以评估出病人的疾病风险，为医护人员提供更加精准的诊疗建议。不仅可以提高治疗效果，还可以降低医疗成本，为患者带来更好的就医体验。例如，在疾病风险评估方面，数字化管理信息平台的应用能够使医院更加精准地把握病人的病情和治疗需求。以往由于信息的不完整性和不准确性，医生难以对病人的疾病风险进行准确评估。现在通过数字化管理信息平台，医生可以更加全面地了解病人的病情和病史，如潜在的并发症、慢性病风险、药物相互作用等都可以得到更为精确的预测和评估。不仅可以提高医生的治疗精准度，还可以帮助医院更好地管理病人的健康档案，提高医疗质量和效率。

结语

构建高质量医院数字化管理信息平台，不但符合国家卫健委的要求，还是提升医院运营效率和医疗服务质量的关键。通过病人数据采集与处理等核心技术的应用，数字化管理信息平台能够为医院提供全面、高效、精准的数据管理和决策支持服务。不仅有助于医院优化资源配置、提高服务质量，还能推动医院数字化转型的深入发展，实现医疗行业的持续创新与发展。

引用

[1] 陈红,潘淮宁,戚建伟,等.医院综合档案数字化管理的应用和体会[J].中医药管理杂志,2010,18(9):851-853.

[2] 李虹.医院人事档案数字化管理有效路径探析[J].兰台内外,2024(8):4-6.

[3] 郭丽媛.浅析数字化在医院档案管理中的应用[J].信息记录材料,2021,22(4):98-99.

[4] 张敏.大数据时代,医院数字化档案资源发展的特点分析[J].兰台内外,2023(16):16-18.

[5] 刘艳军.医院综合档案信息化建设的应用及体会[J].办公室业务,2019(16):99-100.

数据要素综合服务中心对推动地区经济高质量发展的意义

文◆内蒙古国信数据要素综合服务有限公司 **刘博伟**
北京国信钧元科技有限公司 **王杰伟**

引言

随着数字经济的高质量发展，数据资源不断积累，数据资产价值不断提升，数据要素价值挖掘路径愈加丰富。数据资产登记评估作为数据要素流通与增值的前置条件，对政府、企业（特别是国有企业）具有战略性价值。本文以内蒙古数据要素综合服务中心为案例，详解介绍政府和企业共同开展区域数据要素市场建设的先行先试经验。

【作者简介】刘博伟（1993—），男，河北保定人，本科，研究方向：企业数据要素资产化、政府数据要素市场化体系建设及运营。
【通讯作者】王杰伟（1987—），男，安徽淮北人，工程硕士，工程师，研究方向：大数据、人工智能、数据要素与产业、经济的融合研究。

内蒙古数据要素综合服务中心由呼和浩特新城区政府和北京国信钧元科技有限公司联合设立，其立足呼和浩特市，是覆盖内蒙古自治区的数据要素综合服务机构。

内蒙古数据要素综合服务中心以数据资产登记中心和数据要素产业园为载体，协同会计师事务所、律师事务所、资产评估机构、数据评价机构等上下游机构，形成数据资产登记评估服务生态体系，为企业提供一站式服务。以数据资产登记评估作为数据要素有序流通和价值释放的重要抓手，推进数字经济与实体经济融合，推动地区经济高质量发展。

1 建设背景

党的十九届四中全会以来，随着数据促进经济高质量发展的作用日益凸显，数据成为一种生产要素的观念逐渐深入人心。党中央陆续出台一系列政策文件，旨在通过数据要素市场建设，建立数据参与生产分配长效机制，促进数据价值持续释放，带动社会经济高质量增长。

对政府而言，坚决响应党中央对数据要素的相关政策规定，遵循国家标准，对所辖区域的数据要素进行认定评估，以数据资产评估为抓手，先行先试，将数据资产纳入政府对区域生产总值的统计维度，进而围绕数据资产评估发展区域专业评估机构和专业评估师认定，带动区域数据要素配套数据服务产业发展，提升区域数字经济占比，实现新旧动能转化。

对企业而言，在双资监管的框架下，遵循国家标准，将国有企业之间、国有企业与社会企业（特别是互联网企业）之间既有的数据流通标准化，科学界定数据要素的流通价值，将数据资产纳入双资监管的体系，确保在企业数据资产入表、企业投融资、企业拆分兼并等领域数据要素的保值增值。因此，数据资产登记评估成为构建区域数据要素市场高标准发展的重要一环。

2 整体介绍

为整合并充分发挥呼和浩特市的区位以及标杆优势，新城区政府和北京国信钧元科技有限公司联合设立了立足呼和浩特市、覆盖内蒙古自治区的数据要素综合服务机构——内蒙古数据要素综合服务中心（见图 1）。

图 1 新城区政务服务中心

2.1 以数据资产登记中心和数据要素产业园作为业务载体

内蒙古数据要素综合服务中心位于新城区数字经济产业园，在 B 座政务服务中心大厅设立数据资产登记中心，以 4 个数据要素综合服务窗口（见图 2）为载体，为企业提供数据资产登记和数据要素综合咨询服务。在 C 座建设数据要素产业园，产业园一层用于接待、展示和部分业务工作，产业园二层为办公区，设立数据资产评估和入表服务区域、数据知识产权服务区域、数据流通交易服务区域、数据应用场景开发服务区域、数据资产金融服务区域等，聚集律师事务所、会计师事务所、资产评估机构、数据评价机构、数据治理公司、应用场景开发公司、金融服务机构等上下游机构，形成数据要素服务生态体系，为企业提供一站式服务。

2.2 以数据资产登记、评估、运营及流通交易平台作为数据要素基础设施

内蒙古数据要素综合服务中心以数据资产登记、评估、运营以及流通交易平台作为数据要素基础设施，提供从数据登记、评价、治理、质量等级认定、评估到数据运营流通的一体化能力，构建全生命周期的数据资产服务体系。平台建设以数据资产价值流通为导向，以区块链、数据安全和数据智能等技术为支撑，建立安全、可追溯的数据资产全生命周期服务生态链，提供链上数

据资产登记、质量评价、质量等级认定、价值评估等服务以及对指标 / 规则和算法知识等进行运营管理，促进政府、企业、行业和社会等数据资产的量化评估和价值实现，是保障数据要素综合服务中心正常运营与业务开展的基础平台。

2.3 多措并举营造数据要素市场化氛围

内蒙古数据要素综合服务中心协同国家信息中心、中国电子技术标准化研究院等专业机构，组织行业专家围绕数据要素形成区域地方制度、标准、管理规范、业务指南，为数据要素市场体系建设提供制度保障。组织专业培训，积极培育政府、企业和行业全领域的数据资产评估服务能力，建立规范化的数据资产评估的方法和流程，培养数据资产评估人才。开放政府公共数据授权运营、国有企业数据资产登记、评估入表等场景试点示范，引导区域性保险、金融等机构参与数据要素市场化落地实践。

图 2 内蒙古数据要素综合服务中心

2.4 探索区域数据资产评估多场景创新

内蒙古数据要素综合服务中心结合地区优势产业，探索城市数据资产评估产业先试先行。创新应用场景示范包括以下 4 个种类。（1）开展政务数据资产评估示范，评价政务数据质量与价值，为政务数据资产运营提供支撑与服务。（2）开展国有企业数据资产评估示范，为国有数据资产保值增值提供评价支撑，有效防止国有数据资产流失。（3）数据生产要素统计核算试点，探索将数据要素价值纳入国民经济核算体系，推动数字经济跨越式发展。（4）探索数据资产质押、数据资产入股等创新示范，为企业提供创新融资方法，为当地数字经济注入新活力。

2.5 培育数据要素市场综合服务产业生态链

内蒙古数据要素综合服务中心围绕数据要素市场，引入和培育一批数据要素生产、确权、流通、应用、经纪人等数据运营服务创新型企业集群，建立数据要素和数据资产评估产业人才培训体系，联合各类银行、金融、基金等融资机构，打造涵盖数据资产质押、数据资产凭证等融资服务，打造自治区数据要素综合服务产业高地。

结语

内蒙古数据要素综合服务中心以数据资产登记和评估作为数据要素有序流通与价值挖掘的重要抓手，推动新城区、呼和浩特市、内蒙古自治区数据要素市场体系化建设发展。旨在充分发挥数据要素乘数效应，以数据要素为重要引擎发展新质生产力，推进数字经济与实体经济深度融合，培育新产业新业务新模式，推动经济高质量发展。

锻造网络和数据安全专业能力服务制造强国高质量建设

文◆中国软件评测中心（工业和信息化部软件与集成电路促进中心） 唐 刚 杨晓琪 张德馨 黄英男 路 红 刘亚娟

引言

在数字时代，数字产业化和产业数字化加速演进，云计算、大数据、人工智能、算力网络等新技术新应用推动了工业互联网等新业态的转型发展，对我国制造业产生了巨大影响。目前，ERP（企业资源规划）、MES（制造执行系统）、DNC（分布式数字控制）、MDC（制造数据采集）、PDM（产品数据管理）等先进信息化管理系统已经在制造企业得到广泛应用，有效提升了制造业的工作效率和质量。

随着制造业企业数字化程度不断提升，网络和数据安全对制造业的影响不断加大，建立稳固安全的网络基础设施和完善的数据安全保护机制已经成为保障制造业高质量长远发展的基石。同时，扎实可靠的网络和数据安全保障能力也将有力推动制造业的创新发展和国际竞争力的提升。中国软件评测中心安全事业部着力把握网络和数据安全在制造强国中的重要角色和关键作用，发挥其在安全领域的研究能力、测评服务能力和产业引领作用，从政策标准、技术创新、产业生态角度提出发展建议，进一步锻造网络和数据安全专业能力，保障制造业的稳定运行和持续发展，为加快推进制造强国建设、新型工业化发展和中华民族伟大复兴贡献力量。

1 网络和数据安全在制造强国建设中的重要角色

1.1 网络和数据安全是制造强国建设的重要保障

制造业是国家经济命脉所系，是立国之本、强国之基。数字技术正以新理念、新业态、新模式全面融入人类经济、政治、文化、社会、生态文明建设的各领域和全过程，给制造业带来了深远的影响。在智能制造加速普及的形势下，网络攻击和数据泄露对企业生产流程构成严重威胁。恶意软件会破坏生产设备或系统，导致生产线停滞，影响生产计划和交付时间。数据泄露则会泄露商业机密信息，损害企业竞争力和品牌声誉。

随着企业数字化进程加速，忽视网络和数据安全问题会导致巨大经济损失，甚至威胁企业生存。因此，加强安全意识培养、投资安全防护措施、强化企业安全管理，对制造业企业至关重要。通过建立稳固的网络基础设施和完善的数据保护机制，企业能够严密守护生产数据、研发成果以及商业机密等核心信息，有效抵御黑客攻击和数据泄露的风险，确保企业的竞争力和市场地位，增强合作伙伴和客户的信赖，为制造业高质量长远发展奠定了安全基石。

1.2 网络和数据安全是推动制造业创新发展的重要驱动力

网络和数据安全在制造业网络化发展中占据重要地位，是推动创新发展的关键因素之一。在数字化转型和智能制造的背景下，网络和数据安全为制造业创新发展提供了技术支撑和驱动力。制造业对网络技术和大数据分析的依赖日益加深，许多创新活动都需要依托网络和数据技术来实现，如远程监控、智能预测、个性化定制等。网络和数据安全为制造业创新构建了坚实的信任基石，确保了知识产权的安全和数据的完整性，促进了制造业的协同创新和资源共享，为技术进步和产品创新注入了新动力。

在智能制造、工业互联网等新兴业态中，网络和数据安全发

【作者简介】唐刚（1981—），男，四川德阳人，硕士，高级工程师，研究方向：网络与数据安全、数字化转型。
【通讯作者】张德馨（1981—），男，湖北蕲春人，博士，高级工程师，研究方向：网络与数据安全。

挥着不可或缺的作用。只有在网络和数据安全得到充分保障的情况下，制造业才能敢于创新、大胆尝试，从而推动产业转型升级和高质量发展。为了推动制造业的高质量发展，必须坚持创新驱动，提高自主创新能力，集中科研力量，实施一批新兴产业技术攻关项目和重大基础研究、应用基础研究项目，力求在核心技术瓶颈上取得突破，推动制造业的可持续发展和转型升级。

1.3 网络和数据安全是提升制造业国际竞争力的关键因素

全球化背景下，网络和数据安全成为制造业企业保持竞争优势的关键。随着技术发展和知识产权价值提升，制造业面临保护核心技术和商业机密的巨大挑战。在国际竞争中，数据跨境流动易导致用户数据泄露、滥用，甚至使国家战略信息被恶意挖掘。因此，网络和数据安全能力建设为企业构建了强大的防线，有效抵御潜在威胁，确保技术创新和商业机密的安全。

网络和数据安全还为企业合作提供了可信基础。在全球供应链中，制造业需与各方紧密合作。健全的网络和数据保护机制树立了信任，促进了数据共享和联合研发，加速了创新进程，提升了产品质量。完善的网络基础设施和数据保护体系为企业在国际舞台上提供了稳固支撑，不仅保障了技术信息的安全，还促进了与合作伙伴之间的信任与合作，推动数字化转型和智能化制造，保持全球竞争领先地位。

2 立足安全主责主业，赋能制造业数字化转型

2.1 持续提升安全测评服务能力，锤炼主业基本功

网络安全服务业是网络安全事业前行的基础，持续提升安全测评服务能力，为制造强国目标提供了安全保障。立足中国软件评测中心安全测评基础业务，优化完善安全测评服务流程，严格依照标准化的操作流程进行，确保安全测评的全面性和可操作性。强化商务、实施等工作人员的职业培训，提升业务专业性，在实践中锤炼自身本领。同时，与时俱进，保持不断学习和研究的态度，紧跟技术的发展趋势，通过参加行业会议、研讨会和培训课程等方式，研究学习前沿的网络与数据安全解决方案和技术，引进先进的安全测评技术设备，不断提升自身的技术水平，锤炼形成高效高质量的测评业务能力。

2.2 拓展安全领域新业务新模式，锻造核心竞争力

我国实施创新驱动发展战略，格外重视自主创新和创新环境建设，致力于提升我国产业水平和实力，推动我国从经济大国向经济强国、制造强国转变。根据服务对象、国家法律法规和相关行业标准的特点，结合中国软件评测中心业务的实际情况，积极建立具有全面性、稳定性、及时性的网络和数据安全领域服务保障体系。随着人工智能、大数据、云计算等新技术快速应用发展，及时扩展安全领域新业态新模式，转变思路，创新模式，推动项目有序开展。同时，积极构建安全领域发展平台，为 5G、人工智能、大数据等技术提供深度融合的安全屏障。将安全领域的创新成果与经济社会各领域深度融合起来，产生化学反应、放大效应，为客户提供更为优质、高效的服务，为全面推进新型工业化、建设以先进制造业为骨干的现代化产业体系提供有力的安全保障。

2.3 引领数据安全产业稳步发展，培育优质新动能

高质量发展需要新的生产力理论来指导，而新质生产力在实践中形成并展示出对高质量发展的强劲推动力、支撑力，可以用以指导新的发展实践。数据作为一种新型生产要素，在数字经济中的关键作用愈发凸显，但随之而来的数据安全风险挑战也不断加大，数据安全已成为事关国家安全与经济社会发展的重大问题。数据安全产业是为保障数据持续处于有效保护、合法利用、有序流动状态提供技术、产品和服务的新兴安全业态。在数据安全产业发展过程中，一方面，加强技术研发和创新，提升数据安全产业的核心竞争力，推动数据安全技术的不断突破和升级。另一方面，加强产业协同，构建数据安全产业生态，推动产业链上下游企业之间的紧密合作，形成优势互补、协同发展的良好局面。

促进数据安全产业高质量发展是推进落实重要国家发展战略的有力举措，也是实现制造强国和新型工业化的重要组成部分。围绕数据安全产业政策支撑研究、协同机制构建、标准体系建设、评价能力建设、创新载体建设、专业人才培育、产业生态完善等方面，中国软件评测中心以产业研究、产业标准、人才培养、能力评价、成果转化、产业活动为

执行抓手，调动产学研用积极参与，凝聚各方智慧和力量，有力推动数据安全产业高质量发展。

3 激发安全领域活力，助推新型工业化进程

3.1 强化政策标准引导，坚持促进发展和依法管理相统一

2023 年 9 月，全国新型工业化推进大会召开，推动新型工业化发展迈向新征程。随着国家数字化进程的加速推进，工信领域数字化、网络化、智能化也在加速提质升级，但与此同时，伴随而来的网络和数据安全事件频发，风险日益突出。《网络安全法》《数据安全法》《工业和信息化领域数据安全管理办法（试行）》等法律政策陆续出台，为工信领域网络和数据安全监管与保护工作提供了指导和依据。网络和数据安全产业发展相关指导性文件的发布，也为产业发展指明了方向。市场需求加速供给释放，产业供给侧规模不断壮大，企业技术产品研发提速，产业各方在人才培养和标准体系建设等方面共同发力。

优化政策指引，坚持发展与管理协同并进。完善网络和数据安全法律法规配套政策，健全重要数据识别备案、风险信息报送共享、检查评估和投诉举报等重点管理机制，并落实安全审查、安全监测、数据交易等制度的细则，进一步完善安全保护规则体系。贯彻落实好产业发展相关政策，从发展核心技术产品、壮大安全服务、推进标准体系建设、推广技术产品应用等方面入手，加大促进发展的驱动力，完善我国安全产业生态体系，全面提升产业供给能力，支撑保障数据要素有序开放和深度利用，为我国网络和数据安全产业发展创造良好环境。

3.2 提升技术攻坚能力，坚持安全可控和开放创新相统一

制造业的核心是创新，是掌握关键核心技术，因此，需要自力更生、自主创新。在网络和数据安全产业需求牵引下，供给侧企业在新竞争常态中积极求存求发展，通过自研、合作、并购、重组等手段延伸产品线、丰富业务线。目前，针对传统需求的供给已较为成熟稳定，而智能化、定制化则成为研发重点方向。同时，网络和数据安全技术与人工智能、大数据、区块链等新兴技术的交叉融合创新也取得了良好进展，且在 5G、工业互联网、物联网、车联网等领域的融合应用不断深入。

打好网络和数据安全关键核心技术攻坚战，加强科技创新和产业创新深度融合，以科技创新推动产业创新。开展重点行业典型技术、产品、服务以及应用场景遴选，加强对通用性强、效率高的技术产品创新研发。利用国家项目加强关键核心技术攻关，利用“揭榜挂帅”“赛马”等机制，推动技术产品创新突破。鼓励企业、高校、科研院所等主体积极承接重大科技项目，充分利用国家自然科学基金等申请相关课题，对前沿技术、重点产品开展攻关研究。

3.3 深化产业多方合作，坚持共同参与和各显神通相统一

传统产业改造升级和战略性新兴产业培育壮大需要依托产业集群。推动网络和数据安全产业集聚发展，离不开产业各界通力合作，协同发展。引导供给侧企业集聚，发挥地区产业基础优势，规划布局建设国家级网络和数据安全产业园，打造产业发展集群，促进安全企业、资本、人才、技术等要素汇聚。引导应用企业集聚，打造一批具有带动性、引领性的创新应用示范区，集中示范推广技术产品。构建产学研用良性生态，鼓励企业开展协同创新，支持中小微企业与龙头骨干企业分工协同、合作共赢，支持企业与高校、科研院所等开展战略合作，形成市场需求推动科学研究，科研成果助力市场发展的增益效果。协调产业要素合理配置，研究利用财政、金融、土地等各类政策工具支持数据安全市场主体培育、技术攻关、创新应用、人才培养，推动各类社会资本向数据安全领域集中。

结语

近年来，我国制造业创新发展步伐加快、提质增效成效显著，制造强国战略得到了全面部署和推进，国家聚力实现从制造大国向制造强国的转变。在这个过程中，新一代信息技术与制造业深度融合，数字产业化和产业数字化协同发力，推动了制造业数字化、网络化、智能化快速发展，网络和数据安全的重要性也日益凸显。制造业是国家安全的重要领域，网络和数据安全是制造业数字化转型、网络化协同、智能化改造的重要技术支撑和保障。因此，要把握好网络和数据安全在制造强国中的重要角色和关键作用，充分发挥中国软件评测中心在安全领域的研究能力、测评服务能力和产业引领作用，加强网络和数据安全领域服务保障体系建设，促进网络和数据安全产业高质量发展，推动网络和数据安全技术研发和应用，全面提高网络和数据安全保障能力，确保制造业的稳定运行和持续发展，为加快推进制造强国建设、新型工业化发展和中华民族伟大复兴贡献力量。

大数据时代做好税收数据治理的思考

文◆国家税务总局深圳市税务局　王晓明　肖胜球

引言

大数据时代，信息技术的创新成果与经济社会各领域深度融合，“数据已成为国家基础性战略资源”“以数据流引领技术流、资金流、人才流、物资流，突破传统资源要素约束，提高全要素生产率”。数据从资源转换为蓬勃的生产力，激发出许多新模式新业态，用数据说话、用数据决策、用数据管理、用数据创新已成为趋势。税收涉及社会生产、销售、流通、分配各个环节，与经济社会发展联系非常紧密，税收大数据也因此具有覆盖经济活动领域最全、展现经济活动内容最真实等天然优势。国家税务总局局长胡静林在《高质量推进中国式现代化税务实践》一文中指出[1]，税务部门要在税收治理中掌握主动、提高效能，必须积极顺应数字时代发展浪潮，把海量数据资源和强大算法算力作为核心竞争力和新质生产力，坚持“用数据、找问题、堵漏洞、防风险、增效能”工作策略，以数据的深度应用推动税务管理、税费治理的科学化精细化智能化。由此可见“以数治税”的重要意义。

凡事具有两面性，随着数据的爆炸式增长，劣质数据也随之而来。以机器学习、数据挖掘为基础的高级数据应用、分析技术，其结果的可信任度和价值，在很大程度上依赖数据治理水平，税收数据资产价值的发挥与创造应靠数据治理来保障。在可预见的未来，数据治理将始终是数据界的核心关注问题之一，在税收大数据背景下如何加强数据治理成为新时代的重要课题。

1 税收数据治理概述

数据治理是综合利用现代信息技术对数据进行收集、整理、存储、处理、分析和应用的过程，其目的在于发挥数据要素的价值与作用，让海量数据转换成新质生产力。税收数据治理是在“依法治税，从严治队”工作方针基础上的税务信息化治理，主要包括数据架构和数据标准管理、数据质量管理、数据安全管理、数据开放与共享管理以及数据应用管理。

（1）数据架构是明确如何获得数据，合理存储、加工和使用数据，保证系统能够有效调取和利用数据，明确数据来源、数据存储和数据用途。数据架构管理包括元数据管理、数据源、数据存储、数据模型管理等，其中元数据管理是对数据进行定义的基本数据，主要包括记录数据的数据字典和记录表信息的数据表等[2]。（2）数据标准是保障数据内外部使用与交换的一致性和准确性的规范性约束。以一个由于数据标准不一致而导致乘坐火车要变换轨道的事件为例，中国境内的大部分铁轨轨道长度是国际标准铁轨，这个数据标准是指国际铁路协会在1937年制定1435mm，但俄罗斯采用5英尺宽轨，由于铁轨数据标准不同，所以北京到莫斯科的火车中间需要变换轨道宽度才能到达目的地，这就是数据标准不一致导致的直接后果。数据标准以税收数据的定义、采集、交换、加工和分析为主线。主要包含税收数据元标准、税收数据代码标准、外部数据交换标准、税收数据采集规范以及税收数据加工规范等内容。数据架构和数据标准管理是做好数据开放与共享的必要条件，如果没有明确数据源，没有统一的标准，那么不同系统间数据的共享、分析、挖掘等将是空中楼阁。（3）数据质量是数据满足使用需要的合适程度。常用数据的真实、准确、完整、及时、一致等指标来度量。数据质量既是各方面工作的基础，又是工作质量的综合反映，数据质量已成为

【作者简介】王晓明（1974—），男，广东深圳人，硕士，研究方向：税收信息化。

数据治理的重要环节。（4）数据安全是指根据国家法律法规要求，结合税务电子数据治理要求，实现税务电子数据从创建、使用、传输、存储、销毁的全生命周期中的保密性、完整性、可用性、可控性和不可否认性的安全保障。高度集中的数据，使纳税人个人隐私和商业秘密的法律保护工作对数据安全保护提出了更高的要求。另外，大数据处理采用分布式计算方法，在数据传输、信息交互等环节，保证数据不泄露、信息不丢失也使数据安全的重要性更加凸显。（5）数据开放是指向公民、法人和其他组织提供数据。数据共享是指政府部门内部各单位之间因履行职责需要使用其他单位数据或者为其他单位数据分析和应用提供数据的行为。数据共享分为两种类型，即无条件共享和有条件共享，无条件共享的数据是不设门槛值，提供给各单位和个人有需要时使用；有条件共享的数据主要根据是否符合条件值来限制使用。（6）数据应用是指基于海量涉税数据，运用管理学、经济学、统计学等的理论、方法，进行深入研究，挖掘税收事务内在的本质和必然的联系，助力提升税收征管效率和税费服务水平，为税收工作的方向性、全局性、战略性和前沿性等问题提供辅助决策。

2 美国数据治理的借鉴研究

作为世界上信息技术领先国家，美国的数据治理一直走在前列。美国政府制定了《数据质量法》，又制定了《联邦信息质量指南》，从多层面、多角度来规范美国的数据治理。美国政府对数据质量控制范围进行了规范，数据质量主要包括“关联度、可信度、准确度、一致性、及时度、支持材料获得难易性、后续理解和应用难易度和可追溯度”。这种全方位的定义，把数据质量提到了相应的高度。

在数据标准管理方面，美国商务部成立了国家标准与技术研究院，是国家最高层次的数据标准管理机构，下设计算机科学技术研究所，为公众提供数据国家标准等有关服务，在美国具有较高的地位。在数据开放方面，美国政府 2009 年提出政府数据开放，并先后发布多项政令，不断推进美国政府数据的开放与共享。2016 年，美国参议院和众议院分别通过了《开放政府数据法案》。该法案明确所有数据集应按照开放许可协议被公布，法律禁止的除外，如果无法提供适当的开放协议，则应当被归入全球公共领域供自由使用，充分开发数据的经济潜力和社会价值。该法案增强了数据的易获取性，通过对大数据的有效利用，提高政府管理水平。在数据应用方面，美国政府明确提出“大数据技术是未来的新石油”，美国政府在科学研究、教育、生物医药研究等方面投入巨额资金进行大数据技术突破。大数据技术在公共领域以政府为主导，和企业等多元化合作将为政府和企业带来共赢。

3 加强税收数据治理的思考

税务大数据时代，为使税务数据更好地服务实现税收现代化，加强数据治理成为一个非常重要和有价值的课题。与此同时，数据治理是一项全面性、长期性的重要工作，在国家税务总局数据治理顶层设计框架下，各省和直辖市、计划单列市结合实际创造性地开展数据治理工作，实现数据治理顶层设计与底层自主创新的有效结合。建立涵盖数据规划、标准、采集、存储、加工、共享、质量、分析、应用全环节的统一管理机制，落实数据架构和数据标准、数据质量、数据安全、数据开放与共享和数据应用的各类规范，形成创新、高效、多样化的数据应用格局。

3.1 以加强数据架构和数据标准管理为基础

加强数据架构管理，对现有税务各系统的数据架构进行汇总和分析，形成数据总体架构图，以明确数据从哪里来、数据放在哪里、数据有什么用，为开展数据分析和应用奠定基础。

理想的数据治理模式是在设计各种信息系统之前就对各行各业的数据进行“数据治理”，并形成各行各业的国际数据标准、国家数据标准、行业数据标准，而不是等到无数个软件系统产生了无比海量的不标准、不规范的数据之后再治理数据。面对非理想的现实，应强化数据标准管理意识，从国家税务总局层面制定税收数据标准管理办法，尽量采用国家标准，如核心征管系统行业分类从 2018 年 1 月起采用国民经济行业分类明细对照版 2017 版，这对数据统计分析有较大帮助。强化制度执行，从总局到各省市税务局，对新建或在建应用软件项目，从设计开始就要严格按照税收数据标准执行；对在用项目，应遵循“分类评估，逐步达标”的原则，进行改造升级，使之符合税收数据标准，形成“建标准、管标准、用标准”的良性机制。

3.2 以提升数据质量为核心

数据质量管理建立需求与供

给之间的双向制约机制，促进数据需求质量和数据供给质量双提升，确保数据的真实、准确、完整、及时和一致。大数据时代，大量新数据每时每刻都在产生，提升数据质量只有起点，没有终点。提升数据质量应建立税务总局、省税务局上下联动，各级税务机关内部横向协同的岗责质量保障机制，建立科学合理的数据质量专业管理机构，制定完善的数据质量管理办法，保障各级税务局以及税务局内部各部门间高效沟通。加强数据质量事前事中管理，严谨的过程定义是数据质量保障的基础，数据的存贮质量、数据的传输质量、数据的使用质量等都要严格把控。根据不同情况，可对数据进行时效性、完整性、原则性、模糊性校验。根据数据质量检查的结果，得出问题数据集，提出数据质量整改要求，下发各数据质量负责单位进行整改。各数据质量负责单位对问题数据清理后相同的问题进行监控再修复，促进数据质量提升。总之，要在不断拓宽数据来源、提高数据质量的同时，用好用足现有数据资源，不断提高“以数治税”能力和水平。

3.3 以加快数据开放与共享为重点

数字化时代，数据是新时代的“黄金和石油”，数据的开放与共享是数据赋能、推动经济社会向前的关键。税务部门可通过数据开放与共享，实现决策和治理的科学性和有效性。税务数据资源开放与共享遵循统筹规划、逐步推进、便捷高效、依法管理、安全可控的原则。应大力推动税务部门以及与工商、海关、公安、统计等其他政府部门、银行、企事业单位的数据共享与交换，加快建设实用高效的数据统一开放平台。实施数据共享开放时，应注重保守国家秘密、商业秘密，保护个人隐私。

3.4 以强化安全管理为关键

税收数据安全遵从谁产生谁负责、谁使用谁负责以及授权使用原则。随着自助办税终端、移动办税业务、电子税务局、移动办公、网络发票等技术应用越来越多，各税务局在不断探索如何通过信息技术创新来不断丰富创新纳税服务手段，但由于这些应用在设计、部署和上线过程中考虑数据安全问题还存在漏洞，所以带来了新的安全问题。税务业务专网与互联网的边界越来越多，边界防护难度越来越大，突破边界进入业务专网的风险越发严重。为了做好数据安全工作，应建立统一的数据安全治理体系、技术体系和管理体系。以资产保护为核心，以风险管理为主线，以政策法规为准绳，建立税务系统数据信息安全风险管理体系，做好数据创建、数据使用、数据传输、数据存储、数据销毁等工作，落实国家信息安全等级保护相关政策，基于增强网络空间安全防护和安全事件识别能力，把数据安全措施落到实处。

3.5 以做好数据应用为根本

建立科学便捷的数据分析应用大数据平台，建立一套行之有效的税收数据分析应用管理机制，包括设置科学合理的数据分析应用工作机构，建立严谨详实的数据分析与应用工作制度和流程。在 AI 技术的加持下，建立科学的预测模型，对税收大数据进行分析，发现纳税人的典型行为特征，结合数据间的交叉比对和预警指标等，进行评估选案，提供问题纳税人清册和预警提示，加强自动化预警监控和提示，提高工作效率。集中多个关键业务系统数据，提供跨系统、跨业务查询服务，统一管理和发布税收业务统计以及预测分析报表，为领导提供决策支持。根据纳税人信用等级和行为预测，对纳税人进行洞察分析，精准实现纳税人分类服务，降低纳税成本和税收流失率，提高应收尽收率和纳税遵从度。深挖税收大数据价值，不断研习提升算法算力，以税收分析服务经济高质量发展。

结语

大数据时代，税收大数据是税务部门的“金山银矿”，要下大力气做好数据治理，特别注重数据标准和安全，提升数据质量，做好数据开放与共享，实现高质量的数据分析与应用。集聚优势，充分发挥数据要素的乘数效应和倍增效应，赋能税收事业的税收征管和税费服务，推动税收事业高质量发展，早日实现数据智慧税收，通过建设税收现代化服务中国式现代化。

引用

[1] 胡静林.高质量推进中国式现代化税务实践[N].学习时报,2024-07-05(001).

[2] 赵眸光,赵勇.大数据 · 数据管理与数据工程[M].北京:清华大学出版社,2017.

大数据背景下档案数字化管理实践策略分析

文◆中共四川省委党校　唐利红

引言

随着信息技术的快速发展，大数据已经成为推动社会进步的重要力量。在档案管理领域，大数据技术的应用不仅改变了传统的档案管理方式，还带来了前所未有的机遇和挑战。大数据技术的引入使档案数据的收集、存储、处理、分析和利用变得更高效、便捷和安全。因此，如何在大数据背景下实现档案数字化管理，成为当前档案管理领域的重要议题。基于此，本文分析大数据背景下档案数字化管理实践策略，以供相关从业人员参考。

1 大数据技术的概述

1.1 定义

大数据技术是指用于处理和分析大规模、高复杂性数据集的一系列技术和方法。由于数据集体积庞大、增长速度快以及种类繁多（包括结构化数据、半结构化数据以及非结构化数据），传统数据处理软件无法有效处理。大数据技术的核心在于其能够从这些海量数据中提取有价值的信息，支持决策制定、业务优化和创新。这些技术通常包括数据采集、存储、管理、分析和可视化等多个方面[1]。其中，分布式存储系统和 NoSQL 数据库能够存储和管理大规模数据集；数据处理框架提供了快速、分布式的数据处理能力；机器学习和人工智能算法则能够从数据中学习模式和趋势，实现精准预测分析。

1.2 特点

1.2.1 数据体量巨大

大数据技术显著的特点之一在于其处理的数据体量巨大。传统的数据处理系统主要处理 GB 或 TB 级别的数据，而大数据技术则通常面对的是 PB、EB 甚至 ZB 级别的数据量。这种数据量的激增主要得益于互联网、物联网、社交媒体、移动设备等技术的快速发展，其产生了大量的数据，包括用户的点击流、交易记录、传感器数据、社交媒体互动等。

1.2.2 数据类型多样

大数据技术的另一个显著特点是数据类型的多样性。随着技术的发展，非结构化和半结构化数据的比例不断增加，这些数据类型具有不同的格式和结构，传统的数据处理方法难以应对。针对不同类型的数据，应采用不同的处理方法。例如，文本数据需要自然语言处理技术提取信息；图像和视频数据需要计算机视觉技术识别对象。数据集成因此变得更复杂，企业需要整合来自不同来源、不同格式的数据以便进行综合分析。为了实现这一目标，通常需要使用数据湖等技术，能够存储原始的、未经处理的多种类型的数据[2]。

2 传统档案管理中存在的问题

2.1 存储空间与物理损坏

纸质档案需要占用大量的物理空间进行存储。随着时间的推移，档案数量的增加会使存储空间逐渐变得紧张。此外，纸质档案容易受到环境因素的影响，如温度、湿度、虫害等，这些不利因素都会导致档案的损坏和变质。在存储空间方面，随着档案数量的不断增长，传统的档案管理方式难以提供足够的存储空间。不仅增加了档案管理的成本，还影响了档案的保存和利用。而物理损坏则更直接地威胁到档案的安全性和完整性。一旦档案受损，其中的信息将无法恢复，就会造成不可挽回的损失。

2.2 检索效率低下

由于纸质档案需要人工查找

【作者简介】唐利红（1975—）女，四川内江人，本科，工程师，研究方向：计算机应用、数字化教学。

和翻阅，因此检索过程耗时较长，效率较低。在档案数量较多的情况下，这种检索方式甚至变得不可行。检索效率低下不仅影响了档案的利用效率，还增加了档案管理人员的工作负担。在需要快速查找档案的情况下，传统档案管理方式难以满足需求。此外，由于纸质档案需要人工管理，存在一定的错误率。如果档案管理人员在检索过程中出现错误，就会导致档案误用或者丢失。

2.3 信息共享与利用困难

由于纸质档案需要人工管理和传递，难以实现远程访问和共享，限制了档案信息的利用范围和价值发挥。在信息共享方面，传统档案管理方式无法满足现代社会对档案信息的高效需求，无法提供充足的档案信息资源和便捷的访问方式，导致档案信息资源的浪费和闲置[3]。此外，在档案利用方面，传统档案管理方式也存在一定的局限性。由于纸质档案需要人工管理和传递，无法实现自动化处理和利用，限制了档案信息的深入分析和挖掘，使档案信息的价值无法得到充分发挥。

3 大数据背景下档案数字化管理的实践策略

3.1 建立全面的数据收集与整合系统

利用大数据技术，构建一个具备自动抓取、收集纸质档案的数字化扫描文件，并支持电子档案直接导入的系统。同时，该系统能够从其他业务系统、数据库或公共平台中获取与档案相关的数据，如人员信息、项目资料等。通过全面的数据收集方式，确保档案信息的完整性和准确性。在收集大量数据后，系统需要整合相关数据，消除信息孤岛，实现档案信息的全面共享和高效利用。通过数据整合，使档案管理人员更便捷地查看、管理和利用档案信息，提高档案管理的效率和质量。利用数据清洗技术去除重复、错误或无效的数据，利用数据标准化技术统一数据格式和转换标准以及利用数据关联技术发现数据之间的关联性和规律[4]，确保数据整合的准确性和高效性。随着档案数量的不断增加和业务的不断发展，系统需要灵活地适应新的数据需求和技术挑战。因此，在设计和实现系统时，应考虑其可扩展性和可维护性，确保系统能够长期稳定运行。

3.2 实施智能化数据分类与索引

传统的档案分类方法依赖人工操作，不仅耗时耗力，还容易出错。为应对这些问题，利用大数据技术，构建的系统应具备自动对档案数据进行分类的功能，将相似的档案信息归为一类，便于用户快速找到所需的档案信息。同时，根据用户的需求和偏好进行个性化推荐，提高用户的满意度。此外，系统应通过深度分析档案数据，发现数据之间的关联性和规律，并建立相应的索引机制。当用户需要查找某个档案信息时，系统可以快速定位到相关数据，并展示给用户。基于数据关联性的索引机制不仅提高了检索速度，还帮助用户发现更多相关的档案信息，提高档案信息的利用率。利用文本挖掘技术提取档案数据中的关键词、主题等信息，利用自然语言处理技术对文本进行分词、词性标注等操作以及利用机器学习算法对数据进行分类和聚类等处理。通过这些技术手段的应用，确保了数据分类和索引的准确性和高效性。

3.3 构建高效的数据存储与管理平台

利用大数据分布式存储技术，构建一个既满足大规模数据存储需求，又保证数据高可用性和易维护性的平台。分布式存储技术通过将海量数据分散存储在多个节点上，实现数据的负载均衡和高效访问。这种技术显著提高了数据存储的效率和可靠性，避免因单点故障导致的数据丢失或服务中断。对于档案数据而言，可靠性至关重要。任何数据丢失都会对档案管理工作造成严重影响。随着档案数量的不断增加和业务需求的变化，平台需要灵活扩展以满足不断增长的数据存储需求，通过增加存储节点、提高存储密度或采用更先进的存储技术实现。在数据存储平台中，设置数据备份和恢复机制不可或缺，通过定期备份数据确保在发生意外情况时能够迅速恢复数据，减少损失。此外，数据备份也可以作为灾难恢复的重要手段，保障数据的持续可用性。

3.4 实现数据可视化与决策支持

通过数据可视化，档案管理人员可以更深入地理解数据，发现其中的问题和规律，做出更明智的决策。数据可视化技术将大量的档案数据以图表、图像等形式展现出来，使数据更直观易懂。例如，通过柱状图、折线图等，档案管理人员可以清晰掌握档案的借阅情况、存储量等关键指标；热力图、关系图等有助于揭示档案之间的关联性和规律。这些可视化图表不仅帮助档案管理人员快速了解档案数据的整体情况，还能帮助他们发现数据中的异常和趋势，为后续的档案管理工作提供指导[5]。基于大数据的分析结果，数据可视化技术为档案管理工作提供决策支持。通过对档案数据的深入分析，可以洞察借阅习惯、使用频率等

信息，为档案的布局和排列方式提供优化建议。根据借阅频率和使用习惯，可以将常用档案放置在更容易获取的位置，提高档案的利用效率。此外，根据分析结果，可以预测未来的档案需求趋势，为档案采购和存储计划提供参考。

3.5 加强数据安全保障

首先，采用先进的加密算法，加密存储和传输档案数据，确保数据在传输过程中不被非法截获和破解。在数据存储环节，同样采用加密技术保护数据的机密性，防止数据被未授权人员访问。其次，设定合理的访问权限和身份验证机制，确保只有经过授权的人员才能访问和使用档案数据。再次，建立审计机制，记录和监控数据的访问和使用情况，便于及时发现和应对异常行为。此外，定期备份数据，在发生意外情况时迅速恢复数据，减少损失。为验证备份数据的有效性和完整性，定期进行恢复测试，确保在需要时能够成功恢复数据。最后，加强对档案管理人员的安全教育和培训，增强其安全意识和防范能力，减少因人为因素导致的数据泄露风险。

3.6 推动技术创新与人才培养

随着大数据技术的快速发展，新的技术和管理理念不断涌现，为档案管理工作带来了更多的机遇和挑战。因此，档案管理机构应紧跟时代步伐，积极引入新的技术和管理理念，推动档案管理工作的创新和发展。为了推动技术创新，档案管理机构可以加强与高校、科研机构等单位的合作与交流，共同开展技术研究和应用推广。引入先进的大数据技术和管理理念，实现对档案数据的深度挖掘和分析，发现更多的价值和规律，为档案管理工作提供更科学、高效的解决方案[6]。随着大数据技术的广泛应用，档案管理人员需要具备一定的技术素养和数据分析能力，更好地应对数字化管理带来的挑战。因此，档案管理机构应加强对档案管理人员的培训和培养工作，提高其专业素养和技能水平。特别是在大数据分析和应用方面，培养一批具有专业技能和实践经验的管理人才，为档案管理工作的创新和发展提供有力支持。

3.7 建立档案数字化管理评价体系

档案数字化管理评价体系应覆盖档案管理的各个方面，包括数据收集与整合、数据存储与管理、数据安全保障、数据可视化与决策支持等。通过设立一系列的评价指标和评价标准，全面衡量档案管理工作的质量和效率。评价指标的确定应基于档案管理工作的实际情况和需求，避免主观臆断和偏见。评价方法应科学、合理，能够真实反映档案管理工作的实际情况。为了确保评价结果的客观性和公正性，引入第三方评价机构或专家进行评价。通过收集档案数字化管理过程中的各项数据，如数据收集量、数据存储量、数据访问量等，更直观地了解档案管理工作的实际情况。反馈和分析评价结果，及时发现档案管理工作中存在的问题和不足，并采取相应的措施进行改进。

结语

随着信息技术的飞速发展，大数据已经成为推动社会进步和经济发展的重要力量。在这一背景下，档案管理也迎来了数字化转型的关键时期。大数据背景下的档案数字化管理是档案事业发展的必然趋势，通过有效的数字化管理策略可以显著提升档案工作的效率和质量，实现档案资源的优化配置和高效利用。数字化管理还为档案安全提供了新的保障手段，有助于构建更安全、可靠的档案管理体系。未来，档案管理工作应继续深化对大数据技术的应用，加强人才培养和团队建设，完善数据治理机制，确保档案数字化管理的持续健康发展。档案管理部门应积极拥抱变革，不断探索和创新，适应数字化时代的要求，为社会提供更优质的服务。

引用

[1] 刘梦洁.大数据时代档案数字化转型的思考[J].办公室业务,2022(20):186-187.

[2] 郑晓红.大数据下档案管理和数字化建设研究[J].内江科技,2022,43(5):3-4.

[3] 霍燕兵.浅谈大数据时代档案管理工作的挑战与机遇[J].数字技术与应用,2022,40(5):112-114.

[4] 柯江云.基于大数据的档案数字化及管理研究[J].工程与建设,2022,36(1):256-257.

[5] 陈晓.大数据背景下档案管理与数字化建设思考[J].城建档案,2021(11):64-65.

[6] 代春娜.大数据视角下档案数字化工作优化策略探析[J].黑龙江档案,2021(4):174-175.

新媒体环境下数字媒体技术的运用分析

文◆广州城建高级技工学校　袁　楠

引言

信息技术加快了新媒体发展，我国已迈入新媒体时代，陆续衍生了很多新型媒体技术，对加速媒体行业的现代化进程起到了关键作用。新媒体环境下数字媒体技术广受关注，且相应的技术被应用到了更多领域，凸显了数字化的优势，但数字媒体技术在应用过程中还存在一系列问题。未来有关人员应结合数字媒体技术的发展现状，继续优化和改进技术，拓展技术应用范围。基于此，本文探究了新媒体环境下数字媒体技术的发展和应用情况，以期为实际工作提供参考与借鉴。

1 新媒体与数字媒体的相关概述

1.1 新媒体

新媒体时代的到来是技术进步的表现。总体来看，数字技术与新媒体之间相互作用，数字技术出现后，新媒体也拥有了更多的技术支持，同时新媒体也为数字技术的普及和应用带来了新的可能。相关研究认为，新媒体技术指的是全部现存媒体由电脑转化为数字的数据、照片、动态形象、文本等进行计算生成电脑数据的技术路径。我国新媒体在历年来高速发展，陆续出现了抖音、快手、微视等短视频平台，短视频平台作为新媒体阵地，能够吸引更多媒体人、社会大众加入其中，观看、讨论热点话题，使每一用户都能成为潜在的内容创作者。新媒体的发展为数字媒体技术带来了良好的条件，特别是在短视频与直播方面，数字媒体技术必不可少。

1.2 数字媒体

数字时代尚未到来之前，主要为传统媒体，人们通过收听广播、观看电视、阅读报纸等方式获取信息。随着媒体渠道的多样化，人们不仅能从传统媒体中获取信息，还能从微博、微信、抖音等新媒体平台获取信息，获取途径更便捷，能够获取更多有用的信息。

现阶段媒体数字化趋向明显，越来越多的数字媒体取代了传统媒体。数字媒体与传统媒体最为本质的区别在于其内容能以多种形式存在，视频、文章、音乐等媒体内容能以虚拟现实、数字化等多元方式存在[1]。总之，数字媒体的存储、传输过程和要素都具有数字化特征，在正式传输之前应进行数字化处理，在特定的数字线路或者卫星条件下发送，最后由设备接收后再重新转化，变为最初的声音、文字、视频等。

2 新媒体环境下数字媒体技术的发展现状

2.1 数字信息获取与输出技术

2.1.1 图像信息获取与处理

新媒体环境下常常涉及大量图像信息，在获取和处理图像信息时应利用数字媒体技术。具体来说，获取图像信息时应配备专用设备，利用该设备将目标转化为数字图像信息，本质是为了让计算机可直接处理图像信息。时代进步过程中计算机图形技术越发多样、先进与规范，利用该技术不仅能按照规定处理图像，还具有自动识别模式。总之，获取与处理图像信息时，计算机系统应调整图像的大小，并进行适当的压缩，从图像信息中分辨出关键内容。

2.1.2 声音信息采集与处理

新媒体环境下涉及大量声音信息，在采集与处理声音信息时应合理应用数字媒体技术。数字媒体在采集、处理声音信息时涉及信号的转化过程，首先应获取模拟信号，再将此信号转化为设备可识别的数字音频信号，在有需要的情况下可直接调用和传输转化后的信号。新媒体设备

【作者简介】袁楠（1985—），女，广东揭阳人，本科，讲师，研究方向：视觉设计。

采集到的声音信号难以第一时间被识别和利用，应引入数字音频压缩编码技术对采集到的声音信号进行编码、压缩。基于声音的声学参数来编码，可从根本上减少数据总数；基于听觉特性来编码，可保持较高的编码压缩率[2]。

2.1.3 人机交互

人机交互为当下广受关注的技术，在工业生产等方面有一定的应用。综合来看，人机交互实际上反映的是系统与用户之间的关系，使二者之间保持便捷、高效的信息沟通与交流，其中，系统指的是机器、计算机的系统和软件，具体工作中应设置人机交互界面，方便用户操作。

2.1.4 三维显示技术

虚拟现实为新技术，伴随着计算机信息技术而出现，其中融合了三维图形、多媒体、仿真、显示等多种理论与技术成果，能通过计算机模拟一个相对真实的环境，让人体感官体验事物、环境等情况。当前很多领域都应用了三维显示技术，具体工作中应由计算机生成一个或多个实时动态三维立体图像，利用该模型模拟显示的环境，借助听觉、力学与视觉等感知和参与者保持良好的互动关系，在此过程中，由计算机自动识别和处理参与者的互动信息，并给出一定的反馈。

2.2 数字传播技术 / 数字通信技术

新媒体数据种类多、数量大，存在格式等区别，在处理数据时需要多台计算机保持高度协同。数据传播技术能为新媒体数据在不同用户、渠道之间的传输提供技术支持，但在传播数据时应考虑双方所传输信息的类型和传输路径等，采取必要的安全措施。近年来，网络通信技术越发多样，出现了很多新型技术，因此，在全社会范围内建立了相对完善的新媒体网络系统，构建了新媒体通信环境，为远程会议、电子邮箱、网络教学等带来了诸多便捷。

新媒体应用系统建成后，采集、存储、传输媒体信息成为重点和难点，因为每种媒体信息各有其特点，在处理要求等方面存在显著区别，常规的计算机不具有强大的处理功能。因此，为在生产生活领域合理利用新媒体技术，必须关注上述问题，大力开发新技术，利用新技术扩大设备的存储容量、通信带宽，如 CD、ROM、可擦写光盘系统、云储存技术等[3]。

2.3 新媒体管理系统

为科学管理新媒体数据，很多企业都开发了新媒体操作系统，该系统兼具多种功能，不仅能存储信息，还能根据多种需求处理信息。现阶段应用相对较多的是对已有 PC 机、工作站的操作系统进行升级，以提高数据处理的实时性。另外，操作系统为解决新媒体应用问题，还必须支持新媒体计算机的计算功能模型。以 PC 机为例，其 Windows 操作系统经历了多次变化，从最初的 WindowsXP 到 Windows10，功能越发多样。

3 新媒体环境下数字媒体技术的具体应用

3.1 虚拟现实与数字媒体的融合

3.1.1 融合媒体

传统媒体行业中融入现代信息技术等即可促进媒体融合，这是当前和未来的一大方向。在现阶段技术发展的过程中，电视、广播、报纸、新媒体等融合趋势明显，已经具有了一定规模。目前，各广播电视台都开始建立融媒体中心，以期转变媒体工作方式，大型广播电视台因掌握了先进的理念、技术和方法，在融媒体工作中具有高度的竞争优势。虚拟现实与数字媒体技术的融合是当前和未来媒体工作领域的发展方向，且此过程并非一蹴而就，而是循序渐进，保障融合效果，使不同媒体之间共享资源。

虚拟现实与数字媒体技术融合后，融合媒体方面的内容制作更具创新性和创意性，内容可选择性增多，且可采用更多的新技术，如虚拟拍摄、数字虚拟主播等，这些都是数字媒体与虚拟现实相结合的产物。虚拟现实与数字媒体技术融合后可带动诸多工作变革，不仅体现在媒体领域，还在其他领域也有一定的创新应用。例如，目前广受年轻人关注的游戏领域，许多年轻人享受元宇宙的游戏形式，在其中应用了虚拟现实、增强现实、混合现实、拓展现实、全息影像等技术，用户在其中不仅可享受游戏的快感，还能体会游戏中的场景、角色等，享受全新的社交氛围；在新闻报道领域，传统的工作模式下为真人报道，现阶段在虚拟现实技术下可打造虚拟主播，由虚拟人物播报新闻并与用户展开一定的互动。虚拟人物具有趣味性形象，其在播报新闻时能以生动的方式抓住人们的眼球，实现了媒体领域的工作变革，体现了虚拟现实与数字媒体技术的融合优势，从而带动广播电视、网络视听的业态变化。

3.1.2 演艺娱乐

目前，演艺娱乐产业逐渐应

用虚拟现实与数字媒体的融合技术，如歌手开线上演唱会时，应建立线上演唱会场景，观众只需要登录平台即可观看直播，不再需要线下观看，从而突破了地点限制，宣传和推广效果相对理想。当前很多综艺节目中都引入了虚拟现实与数字媒体的融合技术，在这些技术下创造了很多新型的艺术场景和要素，带给观众强烈的沉浸之感。例如，中央广播电视总台的“中国诗词大会”节目，在场景设计中应用了“XP+AR”的技术，营造三维立体空间，使观众更好地融入情境内。在舞台设计方面将虚拟现实与数字媒体技术相融合，改变原先的舞台表演环境，丰富戏曲舞台的表演形式，让观众沉浸在表演中，与表演者产生情感共鸣。

3.1.3 文化旅游

新媒体环境下数字媒体技术可与许多技术相结合，改变之前的媒体渠道。在文化旅游方面也成功应用了虚拟现实与数字媒体相融合的相关理论与技术，如在文化展馆、旅游场所、特色街区等利用虚拟现实与数字媒体融合技术开发虚拟现实数字化产品、虚拟融合导航、艺术品展览等产品和服务。很多大型博物馆都开放了云展览，用户足不出户即可观看博物馆中的陈列。博物馆云展览中主要利用了虚拟在线展览、三维全景虚拟现实等技术，可 360° 全景、3D 虚拟在线展览。浙江博物馆的“丽人行”主题在线展览活动中，相当于建立了一个云展厅，用户只需要在手机等终端登录网站即可参与到其中，不仅能观看展厅中的内容，还能实时互动。

3.2 声像编辑领域

新媒体环境下很多场景都需要进行声像编辑，如制作电视节目、动画片、广告性电影、电视剧、声像艺术作品等。为保障声像编辑的便捷性，应结合实际需求应用数字媒体技术。建立新媒体声像编辑系统，在该系统中大量储存声像和其他多媒体信息，按需编辑和回放，重新排列并删减，由系统自动生成多媒体文档、报告或出版物。专业的声像艺术作品的制作流程较为复杂，主要包含剪接影片、编排文本、制作音响、画面等关键步骤。但在新媒体环境下，因为有数字媒体技术的支持，声像编辑更为便捷，利用相应的系统或软件，很多工作都可自动完成，编辑与制作的效率更高、效果更好。

3.3 短视频平台

从 2018 年开始，短视频以其独特的优势收获了大批用户，抖音、快手等直到今天也依旧受到人们的欢迎。短视频的“病毒式营销”，内容大众、特效多样，成为吸引用户的关键因素。以抖音为例，其平台中的内容相对较多，每一个用户都能从中找到自己感兴趣的内容，如美食美妆、奇闻趣事、新闻热点、体育赛事等，再加上短视频的时间短，在有限的时间内使用各种视频特效，可制作出创意性内容，赢得更多受众的关注。抖音中的视频特效实际上就利用了数字媒体技术。首先，内容发布者在分享日常生活时，可在其中添加花字元素，与网络综艺节目中的花字内容高度一致，不仅能够瞬间抓住人的眼球，还能辅助叙事，丰富视频内容。其次，在应用各种特效场景时同样有数字媒体技术。短视频中只需要点击其中的“特效”按钮，用户即可根据自身的需求、兴趣来选择各种特效，并且用户在选择特效时可享受“选择”的过程，沉浸其中。

3.4 直播

近年来我国直播经济稳步发展，市场上出现了多种专业的直播 App，甚至诞生了很多新媒体直播平台，在这些直播软件、平台中数字媒体技术必不可少。与短视频高度一致，直播画面中可通过数字媒体技术创造许多特效，但特殊在于直播特效种类多、效果好，能满足更多的个性化需求。例如，直播间的 GIF 动图，能让主播处于屏幕的任何位置，构成一种有意遮挡的条件，并保持这一过程中的卡通化与趣味性；直播间的送礼物活动，每一种礼物的形象都经过了精心设计，且进行了数字特效处理；直播平台的美颜功能，可自动修饰主播的脸型、妆效等。

结语

数字媒体技术是新时代发展的产物，已逐渐应用到各个领域。新媒体环境下数字媒体技术虽有所发展，得到了相对成功的应用，但相关技术在未来还有较大的发展空间，因此应立足现状，创新技术形式，保持技术深度。

引用

[1] 韩晓娟.基于数字媒体技术的图像融合算法分析[J].电子元器件与信息技术,2022,6(12):21-24,29.

[2] 高婧.探析新媒体环境对数字媒体技术发展的影响[J].文化产业,2022(16):16-18.

[3] 孙云涛.人工智能在数字媒体技术中的应用[J].电视技术,2022,46 (9):166-168.

实践探索

Practical Exploration

近年来，我国信息化事业不断发展，在“全面提高信息化水平，推动信息化和工业化深度融合，加快经济社会各领域信息化”等战略方针的指引下，做出了大量的实践与探索，延伸到了社会、经济、政治、文化、军事等各个领域。移动互联网、云计算、大数据等技术的不断发展和普及，为信息化实践的推进提供了有力支持。此外，智慧城市与数字强国的建设需求也推动了信息化的探索进程，符合时代的要求与趋势。

信息化对经济发展的作用一直是较为重要的课题，在电子政务、电子商务、企业数字化转型、大数据分析应用等各个领域均取得了重大成果。同时，信息化实践探索的工作一直秉承着突破性与创新性齐头并进的态势向前发展。未来将继续以信息化实践与探索为基石，不断推动国民经济和社会信息化事业发展，加快释放信息化发展的巨大潜能，以信息化驱动现代化，全面建成社会主义现代化强国。

实践探索
Practical Exploration

人工智能时代大学生科技伦理失范现象分析

文◆南京审计大学图书馆　常　青

引言

近年，以 ChatGPT 为代表的人工智能自动内容生成技术（也称生成式 AI 技术）在大学校园快速普及，在推进高校教育模式智慧化的同时也引发了诸多大学生的科技伦理失范现象。本文旨在通过对人工智能时代大学生科技伦理失范现象的深入剖析。了解大学生科技伦理素养的现状和问题，深入分析其成因，为引导大学生形成正确的科技伦理观提供参考。

1 AIGC（人工智能生成内容）技术发展现状概述

1.1 AIGC 技术的成熟与应用

随着数字信息技术的进一步发展，世界正迈入人工智能时代。以 2022 年 11 月 OPENAI 公司正式发布的 AIGC 应用“ChatGPT”为时间节点，“人工智能生成内容”技术（或称生成式 AI）广泛且迅速地应用到了社会各个行业，极大地推动了社会的进步，同样也给高等院校的人才培养、科研模式带来了巨大的机遇和挑战。

从技术迭代的角度进行分析，目前的人工智能时代是由互联网时代与大数据时代叠加进化后的新的数字信息时代，当前的主要技术应用表现为 AIGC（人工智能内容生成，Artificial Intelligence Generated Content），该技术应用是借助尝试性多模态学习模型和神经网络模型等内容生成算法，通过大数据技术和自然语言处理技术，针对用户所发出的内容请求，通过预训练系统，实现从海量标记或未标记的数据中进行抽取和整合，生成为用户所需要且具有独特价值的内容。当前人工智能应用中的自然语言处理技术，能够更好地模拟出人类的语言和行为，使人机互动更加贴近人类的交流方式，这也是人工智能应用产品得以迅速普及的重要原因。

1.2 AIGC 应用推动高校人才培养模式的升级

人工智能应用产品日益丰富、功能日趋强大。就高等教育而言，人工智能内容生成技术的各种场景化应用将极大推动当下高等教育人才培养模式的升级。

2 人工智能内容生成技术出现后的大学生科技伦理失范现象综述

在 AIGC 技术出现之前，大学生出现的诸多科技伦理失落现象主要以运用数字信息技术、网络技术进行科研成果复制、学术论文代写、成果造假等行为为主。

2022 年底，AI 技术取得了突破性的进步，人工智能内容生成技术成熟并实现了诸多行业运用嵌入，极大提高了工作和学习效率。但如同任何事物的两面性一样，人工智能应用的强大功能和使用的低门槛化同样也成了大学生科技伦理失范现象产生的温床。2023 年 1 月，全美一项调查表明，美国有近 89% 的大学生都在使用 ChatGPT 完成作业，针对这一现状，纽约地区在教育系统内全面封杀了 ChatGPT 的使用[1]。尽管如此相关监管部门认为“凭着大学生的‘聪明才智’，根本无法禁掉人工智能工具的使用”。2023 年 7 月，英国大学生媒体 The Tab 调查了英国 114 所大学，发现有近 40% 的大学生使用人工智能应用软件进行“学术作弊”，经过调查发现有 100 多名的在校大学生被评估为有“学术

【作者简介】常青（1973—），男，江苏丰县人，硕士研究生，馆员，研究方向：信息素养教育、智慧图书馆建设。

不端行为”[2]。2023 年 4 月，澳大利亚新南威尔士大学（UNSW）的一起 ChatGPT 论文代写案引发了全澳大学对 ChatGPT 的“封杀”，澳洲许多高校已经明令禁止学生使用这类人工智能的应用。同样国内高等学校的大学生运用人工智能工具所引发的学术乱象也层出不穷，2023 年 2 月《每日经济新闻》记者对北京、上海、四川等地重点大学一线老师进行走访调查发现，很多高校教师称已经遇到有学生运用人工智能内容生成工具撰写论文并提交的情况，并称这些学生所提交的“论文水平甚至和老师一样高”。上海某高校教授对记者表示，“国内的著名查重软件无法对使用 ChatGPT 生成的论文进行查重”。2023 年 7 月，北京警方破获一起中国人民大学在校学生运用人工智能工具盗用多位女生个人信息，并制作成颜值打分系统供网友打分的案件，这是一起典型的借助人工智能工具进行科研造假和信息剽窃的案件。类似以上的大学生科技伦理失范事件在近两年来频繁发生，有的明显违背了正确的价值观和伦理道德，有的甚至引发了国家层面对相关法律条款的修订。例如，2023 年 8 月，十四届全国人大常委会第五次会议审议了《学位法》草案中对于“利用人工智能代写学位论文”的条款，对于此类学术不端行为，将由学位委员会审议决定，并由学位授予单位撤销其学位证书[3]。大学生这一系列失范事件的发生都与当下人工智能工具的应用有关，具有更强的隐匿性特点，使用者的年龄层次也呈低龄化发展趋势。

3 人工智能时代大学生科技伦理失范现象及成因分析

通常意义上的“伦理”指的是人伦道德之理，即社会体系中人与人相处的各种道德准则。不同于“法律”由国家强行制定并强制执行，“伦理”主要依靠人们的内心信念和特殊手段来维系。传统意义上的“科技伦理”泛指涉及科学研究、技术开发、科技协作和利用等方面的伦理要求、伦理准则、伦理规约，并在此基础上形成新的伦理关系，其是调整人与人之间以及个人与社会之间科技关系行为规范的总和[4]。科技伦理结构的内容包含两个方面和 3 个层次。两个方面首先是主观方面，即个人在从事科学技术等活动中以心理活动形式所体现出的道德观念、情感、行为和品质，称为个人科技道德；客观方面则是在社会科技活动中，人和人之间的关系，包括反映这类关系的科技行为规范和准则。3 个层次则是指科技道德意识、科技道德关系和科技道德活动。

当前的科研生态和人工智能应用本身存在的问题等是造成大学生科技伦理失范产生的主要原因，其中主要包括以下 4 点。

（1）人工智能场景应用加速了大学生虚拟生活与实现生活失衡。当今大学生对网络媒介的依赖是一个不争的事实，在数字化虚拟世界中，网络的无所不达逐渐弱化了大学生的创造性和积极性，已经成为支配和控制大学生的异己力量。而人工智能技术的成熟和广泛应用使大学生在虚拟生活与现实生活的失衡现象进一步加剧。数字化虚拟空间弱化了大学生实际生活中的人际交往能力，产生自我认同的危机。疏于对自身虚拟空间的道德自律，进而加剧了科技伦理的失范现象。

（2）人工智能应用在高校人才培养中的教育缺失。国内教育体系中从基础教育阶段便存在“重理轻文”的现象，高考制度也造成了一定程度上文理学科的割裂和对立。学生进入高等教育阶段后，这种情形并未得到有效遏制。大学生的就业压力迫使高校专注培养学生的“学科专业知识和应用技能”。从教学形式而言，高等学校比较侧重于对学生科学知识的灌输式教育，普遍重视从自身科研能力的提升来推动高校的学科专业建设水平，忽略了对大学生人文素质的熏陶。过于实用主义和理性主义的作祟致使高等教育只关注大学生“未来生存能力”的培养而忽略了“未来生活浸润”，因此长久以来高等教育培养的人才普遍缺乏人文素养，导致科技伦理失范现象的出现。

针对这一现状，在 2022 年 3 月由中共中央办公厅和国务院办公厅印发了《关于加强科技伦理治理的意见》（以下简称《意见》）。在《意见》中明确提出要“重视科技伦理教育”“在高等学校开设科技伦理教育课程……，教育青年学生树立正确的科技伦理意识……”[5]，但高等学校在实际教学环节中往往忽略了这一点。

（3）高校对大学生人工智能应用的科技伦理监管不到位。对大学生科技伦理进行监管不同于对科技伦理管理，对科技伦理的监管包括监督和治理，监管体现了监管主体、监管工具和监管价值的变化。刘志辉撰文指出，将科技伦理的治理理解为包括政府、高等院校、科研机构、科研人员、大学生等主体在内的多元主体，运用不同的治理工具对科技伦理进行约束和规范的行动。而对大学生的科技伦理监管体系是指多元主体在参与科技伦理监管过程中所形成的体系[6]。

目前，我国正在经历向智慧

型社会转型的过程，人工智能技术的应用无疑将极大地推进转型的速度。大学生科技伦理的监管模式应形成由政府主导和高等学院、科研院所等多主体深度参与的模式。长期以来，由于对科技伦理监管主体的界定模糊，高等学校对科技伦理教育认识上的不足，导致对大学生科技伦理监管的缺位，这也是大学科技伦理失范现象产生的重要原因。

（4）当前人工智能应用的自身设计与运营的缺陷。由于尝试性多模态学习模型、神经网络模型等内容生成算法的介入，人工智能的实际应用已经由初期的内容演绎阶段进化到了目前的归纳输出阶段。对内容的自动生成已经由过去的“他律化”过渡到了需要人工智能应用系统“自律化”的阶段，人工智能应用逐渐具有的“自主”意识也是其有别于其他计算机“智慧”应用的主要标志。由此带来的诸多社会问题也包括伦理问题。以下是人工智能应用所存在的共性缺陷。

第一，人工智能应用的法律主体问题。人工智能算法本身无法成为法律意义里的主体，因此当智能算法得出的结果出现问题，责任主体的不明确使监管部门无法对责任主体实施监管，成为人工智能应用恶意使用者的漏洞。

第二，人机协同性问题。人机协同是指人与计算机系统通过有效地互动来共同完成任务的形式。目前大学生在使用人工智能应用工具时，自然语言的组织和对系统输出信息采纳的双向互动之间普遍存在问题。例如，当大学生面对多个具有相同吸引力的内容输出时所出现的判断困难，造成对结果采纳的迟缓或误判。这样低效或错误的人机协同容易导致大学生科技伦理失范现象的出现。

第三，人工智能应用的算法透明度与可解释问题。目前基于深度学习的人工智能应用环境中，大学生能够通过相应的人工智能应用获得相应内容结果，但是对人工智能得出这样的结果的原因却知之甚少，因为基于深度学习的决策系统缺乏透明度和可解释性，整个决策过程是一个“黑盒人工智能”，所以会让大学生质疑人工智能生成内容的可信度，并且对人工智能生成内容中出现的问题不知从何入手加以纠正，更无法对出现问题的责任进行划分。

第四，人工智能应用的大数据采集、标记行为所带来的版权和隐私侵占问题。由于生成式人工智能的数据训练需要向算法模型提供巨量的数据，而数据的收集和分析涉及大量个人信息，如聊天记录、搜索历史等，因此存在个人信息被滥用的风险。另外，生成式人工智能技术如果用于生成受版权保护的内容，则容易造成侵权行为。

结语

大学生是时代发展和进步的重要力量，处在全新的人工智能信息环境中，他们自觉或不自觉地暴露出很多信息科技伦理素养方面的问题，如对科技伦理原则和规范认识不足、缺乏对科技伦理问题的正确判断和解决能力等。因此，对大学生科技伦理失范现象进行分析，探讨其产生的原因和影响，具有重要的现实意义和理论价值。

引用

[1] 杨欣.基于生成式人工智能的教育转型图景——ChatGPT究竟对教育意味着什么[J].中国电化教育,2023(5):1-8+14.

[2] 蒋里.AI驱动教育改革：ChatGPT/GPT 的影响及展望[J].华东师范大学学报(教育科学版),2023,41(7):143-150.

[3] 王萍.学位法草案初审:完善学位管理体制保学位授予质量[J].中国人大,2023,(18):43-44.

[4] 曹漪那,杨珊,王慧敏,等.冲突与转向:大数据时代的传播伦理研究[M].成都:四川大学出版社,2017.

[5] 汪晨波,王惠敏.探求科技伦理价值,内化科研伦理规范[J].教育家,2023(26):21-22.

[6] 刘志辉,孙帅.大科学时代我国科技伦理中待解决的问题——以“主体-工具-价值”为框架的分析[J].中国高校科技,2020(11):69-73.

基于改进 TF-IDF 与 N-Gram 的恶意 Web 请求智能识别

文◆国家计算机网络与信息安全管理中心西藏分中心 安 渊 杨 琴 杨 雄

引言

目前，我国互联网基础资源高质量发展，数字化平台成为经济增长的新引擎，以在线办公、网络娱乐、公共服务和信息共享等为代表的 Web 应用平台呈现大规模增长趋势。随着 Web 应用平台的大规模增长，Web 应用平台可用性和数据安全问题越来越受到人们的关注，而恶意 Web 请求是导致 Web 应用平台可用性和数据安全问题的主要威胁。针对网络上的恶意 Web 请求，本文提出了一种融合改进版 TF-IDF 和 N-Gram 的智能识别方案，辅以逻辑回归函数优化，旨在提供一种高效精准的恶意 Web 请求识别手段。

1 研究背景

根据奇安信发布的《2023 年全网漏洞态势研究报告》和 OWASP

【作者简介】安渊（1993—），男，甘肃定西人，本科，工程师，研究方向：渗透测试。

组织发布的《Top 10 API Security Risks – 2023》显示，2023 年公开漏洞数量持续攀升，其中最频繁利用的漏洞类型包括代码执行、信息泄露和拒绝服务攻击，此类型的漏洞均通过构造恶意 Web 请求来进行攻击，具有很高的危险性。

传统的恶意请求识别一是分析恶意用户行为，通过孤立森林模型对恶意攻击源进行标记，但这种方法会导致误判，无论请求合法与否，都会被阻断。二是通过深度学习进行恶意流量检测。在恶意流量检测以及攻击识别方面，梁朕齐 [1] 提出利用自适应提升法和长短期记忆网络相结合的方法实现更高的检测准确率和攻击识别率，但该算法依赖数据集标注和样本平衡。高新成等 [2] 提出了一种基于“CNN+CRU”算法的网络异常流量检测模型，解决了神经网络模型梯度消失问题，大幅提高了准确率和检测效率，但存有过度拟合风险。李道全等 [3] 提出小样本恶意流量分类的实验，但实验结果显示实验准确率和样本数量存在较强的依赖关系。

陈春玲等 [4] 提出融合类别特征扩展与 N-gram 子词过滤的短文本分类，可以提升短文本分类模型的性能和质量。鉴于恶意 Web 请求常通过 Fuzzing 测试方式提交，由此本文创新整合了 TF-IDF[5-7] 与 N-Gram[8-10] 技术，以增强恶意请求识别的精确度。N-Gram 技术捕捉请求中的关键子串频率，而 TF-IDF 则衡量这些子串在整个文档集合中的相对重要性，两者的结合构建了一个有效的特征矩阵，为后续逻辑回归分类提供了坚实的数据支持。识别模型如图 1 所示。

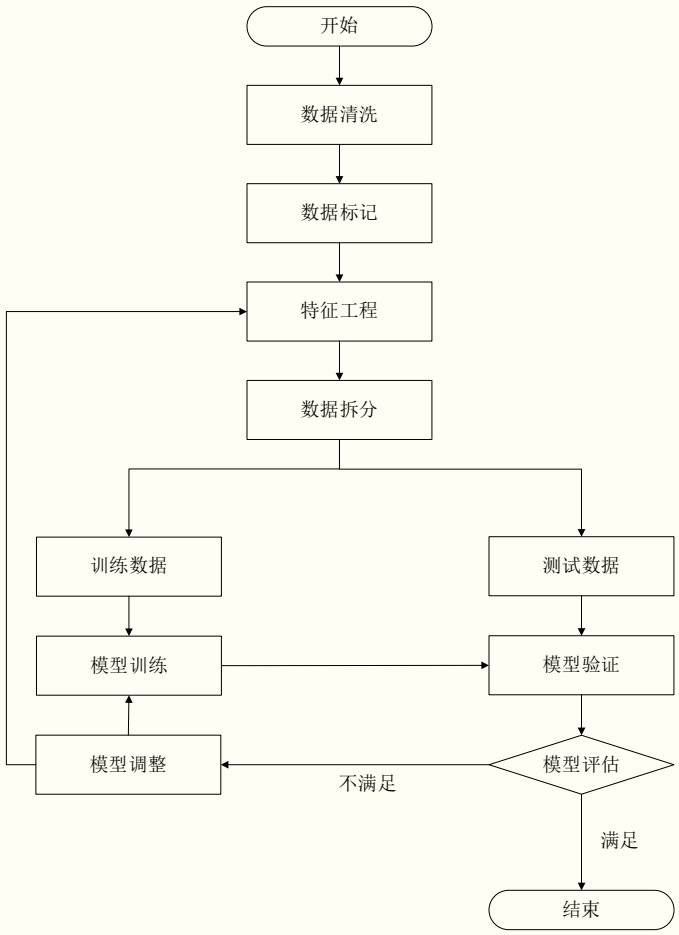


图 1 识别模型

2 数据特征提取

考虑恶意 Web 请求数据的稀疏性和难以量化词语间联系的特点，N-Grams 与 TF-IDF 的联合应用能够有效构建特征矩阵，解决这些问题。

2.1 TF-IDF 技术原理

TF-IDF 技术是一种用于评估词语在文档集合中重要性的统计方法。其原理是基于数理统计，用于评估词语对于一个文档集或语料库中的某篇文档的重要程度。词语的重要性取决于其在单个文档中出现的频率，同时考虑其在整个文档集中的罕见程度。*TF-IDF* 的计算公式如式（1）所示。

$$TF\text{-}IDF = TF * TDF \quad (1)$$

式（1）中，*TF* 表示词频，*IDF* 表示倒文本词频。*TF-IDF* 值越大表示该特征词对这个文本的重要性越大。其基本思想是将文本转换为特征矩阵，并且降低常用词的权重，从而更好地表达一个文本的价值。

2.2 N-Gram 原理

N-Gram 是一种在自然语言处理中常用的统计模型，主要用于预测序列中下一个出现的项。基于概率论和统计学的原理，广泛应用于文本预测、自动补全信息检索等领域。

基本原理是将语料库里面的内容按照字节进行大小为 *N* 的滑动窗口操作，形成了长度为 *N* 的字节片段序列，对所有字节片段出现的频度进行统计，并且按照事先设定好的阈值进行过滤，形成文本的向量特征空

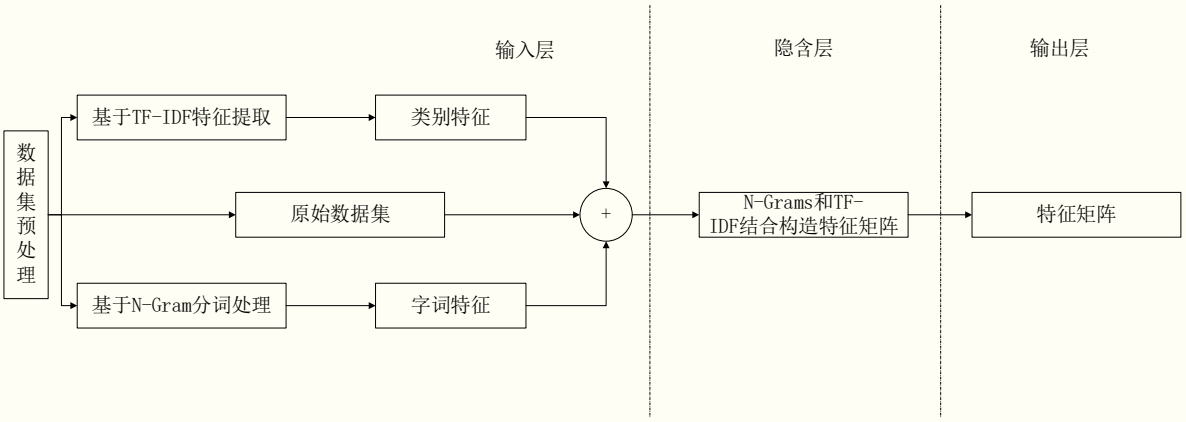


图 2 恶意 Web 请求识别模型结构

间。数学公式如式（2）所示。

$$P(\omega_N \mid \omega_{N-1},\omega_{N-2},\ldots\omega_1)=\frac{count(\omega_N,\omega_{N-1},\ldots\omega_1)}{count(\omega_{N-1},\omega_{N-2},\ldots\omega_1)} \quad (2)$$

式（2）中，$count(\omega_N,\omega_{N-1},\ldots\omega_1)$是序列 $\omega_1,\omega_2,\ldots\omega_N$ 在语料库中出现的次数，$count(\omega_N,\omega_{N-1},\ldots\omega_1)$是序列 $\omega_1,\omega_2,\ldots\omega_{N-1}$ 在语料库中出现的次数。得出如下结论，第 N 个词语的出现只与前面 N–1 个词语具有关联性，而与其他任何词语都不相关。由于流量中的恶意请求词语与词语之间相互独立，不需要考虑它们之间的顺序，于是引入 N–Gram 的概念，可以较好地获取到词语的前后信息。

3 模型构建

选择长度为 3 的 N–Grams 对每个文档（网络请求）的内容进行分词处理，并使用 TD–IDF 来作为文本的特征，以数字矩阵的形式进行输出。过滤掉无关紧要的词语，保留影响整个文本的重要词语。恶意 Web 请求识别模型结构如图 2 所示。

3.1 数据集预处理

从数据集中对日志请求进行清理，移除 HTML 标记、URL、表达式等非文本元素，并进行解码，整理后的数据集包含 56860 条正常请求和 1265112 条恶意请求。数据量如表 1 所示。

3.2 N–Grams 和 TF–IDF 矩阵

当 N=3 时，对文档中的请求内容分词，并使用 TD–IDF 来作为文本的特征，输出 TOP5 特征矩阵。恶意请求 3–Grams 和 TF–IDF 矩阵 TOP5 如表 2 所示，正常请求 3–Grams 和 TF–IDF 矩阵 TOP5 如表 3 所示。

表 2 恶意请求 3-Grams 和 TF-IDF 矩阵 TOP5

关键词	php	sleep	echo	passwd	etc
TD-IDF 权重	0.27	0.16	0.13	0.085	0.085

表 3 正常请求 3-Grams 和 TF-IDF 矩阵 TOP5

关键词	php	jsp	html	index	admin
TD-IDF 权重	0.18	0.035	0.024	0.023	0.022

表 1 数据量

请求类型	数据量（条）
正常请求	56860
恶意请求	1265112

3.3 模型训练

首先，读取数据集中正常请求和恶意请求，使用 3–Grams 对文档中的请求内容分词后，将请求转换为 TF–IDF 特征向量。其次，将数据集分为训练集和测试集，使用逻辑回归模型进行训练。由于攻击者每次提交恶意请求具有随机性，并提高恶意请求识别的效率，使用随机搜索来调整逻辑回归模型的超参数。逻辑回归是一种用于分类的线性模型，它通过一个 *Sigmoid* 函数将线性组合的输出映射到概率空间。模型的损失函数通常是二元交叉熵损失，公式如式（3）所示。

$$L(\theta)=-\frac{1}{N}\sum_{i=1}^{N}[y_i\log(\sigma(z_i))+(1-y_i)\log(1-\sigma(z_i))] \quad (3)$$

式（3）中，$z_i=\theta^T x_i$ 是输入 x_i 和权重向量 θ 的点积，$\sigma(z_i)$ 是 *Singmoid* 函数，$\sigma(z)=\frac{1}{1+e^{-z}}$，$y_i$ 是样本 i 的真实标签（0 或 1）。为了防止过拟合，加入 L2 正则化项，其公式如式（4）所示。

$$L_{reg}(\theta)=L(\theta)+\lambda\|\theta\|^2 \quad (4)$$

其式（4）中，λ是正则化系数。在随机搜索中，从定义的超参数分布中随机抽取参数值来训练模型，并通过交叉验证评估模型性能。对于逻辑回归，主要的超参数是正则化系数λ（在Scikit-Learn中通常表示为C的倒数，即$\lambda=\frac{1}{C}$）。通过随机搜索，寻找能最小化验证集损失函数的λ值。

4 实验结果

4.1 结果计算

计算得到最佳参数C=2.8323。

4.2 实验验证

实验在配置为Windows10、CPU2.2GHz、内存8GB和磁盘500GB的硬件环境下展开，软件环境采用Python3.12.2。通过对海量日志请求进行预处理，包括去除HTML标签、解码和噪声消除，构建了包含56860条正常请求和1265112条恶意请求的数据集。在N=3的条件下，利用TF-IDF对请求内容进行特征提取。最终，通过模型训练与多轮测试，验证了所提方法的准确性超过99%，并在引入更大规模测试集后，保持了95%以上的稳健性。随机抽取有效性验证结果如表4所示。

表4 随机抽取有效性验证结果

script>alert(1)</script>	恶意请求
name=admin' or 1=1'	恶意请求
/admin.php	正常请求
svg onload=confirm(1)>	恶意请求
/etc/passwd	恶意请求
?version=1331749579	正常请求
scripting.nasl</script>.idc	恶意请求
?rev=2|less /etc/passwd	恶意请求
javascript:alert(1)">	恶意请求
<svg onload=location='//p0.al'>	恶意请求

结语

本研究通过融合N-Gram和TF-IDF构建的特征矩阵，结合逻辑回归函数的优化，成功开发出一种高精度的恶意Web请求识别模型。未来的研究方向将聚焦于进一步提升模型的泛化能力和实时响应速度，以适应不断演变的网络安全威胁。

引用

[1] 梁朕齐.基于深度学习的恶意流量检测及攻击识别研究[J].信息:技术与应用,2023(12):89-91.

[2] 高新成,魏壮壮,王莉利,等.基于CNN+GRU的网络恶意流量检测算法[J].计算机仿真,2024(3):388-394.

[3] 李道全,李玉秀,任大用.小样本下基于决策树-SNN的恶意流量检测方法[J].计算机工程与应用,2023,59(21):258-266.

[4] 陈春玲,吴凡,余瀚.基于逻辑斯蒂回归的恶意请求分类识别模型[J].计算机技术与发展,2019(2):124-128.

[5] 马莹,赵辉,李万龙.结合改进的CHI统计方法的TF-IDF算法优化[J].计算机应用研究,2019(9):2596-2598,2603.

[6] 叶雪梅,毛雪岷,夏锦春.文本分类TF-IDF算法的改进研究[J].计算机工程与应用,2019(2):104-109,161.

[7] 许甜华,吴明礼.一种基于TF-IDF的朴素贝叶斯算法改进[J].计算机技术与发展,2020(2):75-79.

[8] 章宗美,桂盛霖,任飞.基于N-gram的Android恶意检测[J].计算机科学,2019(2):154-160.

[9] 姜如霞,黄水源,段隆振.基于规则和N-Gram算法的新词识别研究[J].现代电子技术,2019(4):174-178.

[10] 龚永罡,田润琳,廉小亲.基于MapReduce的三元N-gram算法的并行化研究[J].电子技术应用,2019(5):70-73,77.

基于大数据环境的网络安全计算机信息技术探讨

文◆广西水利电力职业技术学院　陈成龙

引言

大数据时代引领不同行业走向智能化发展，计算机信息技术的应用随之广泛。5G 发展趋势下，网络安全性成为一大问题。为避免发生信息泄露等网络安全问题，国内制定了相关管理条例来维护环境安全。在此基础上，网络安全计算机信息技术的创新研发尤为重要。基于该环境，网络安全计算机信息技术更应实现创新与研发，为社会营造良好网络环境。通过分析大数据环境以及网络安全计算机信息技术现存风险，结合实际问题提出网络安全计算机信息技术在大数据环境中的具体应用。

1 大数据环境分析

尽管国内的信息技术一直处于发展阶段，但部分传统应用软件功能存在一定运行问题，如信息采集不全面、信息孤岛等问题，均会影响数据真实性和可靠性，存在较大网络应用安全问题。行业相关人员通过不断研发与创新，利用大数据技术替代传统软件，以避免网络应用安全问题。

大数据技术具有诸多优势，一是获取信息速度快，二是应用价值较高，三是数据存储容量大。同时，大数据技术是信息应用、信息提取等相关领域的重要介质，因此在各行各业中的应用较为广泛 [1]。大数据环境下，信息提取与信息处理效率显著提升，大幅提升了综合信息资产的决策能力。基于大数据环境的行业发展，更具高效率性和高经济性，如正确使用大数据技术，充分发挥其应用价值，有利于加快社会经济发展，提高国民生活质量以及生活水平。

2 大数据环境下的网络安全重要性

大数据环境更新速度极快，网络安全成为一大问题。网络安全指的是在应用网络系统时，需要经过一定信息传输与处理，在此过程中，存在信息泄露或者信息窃取的问题 [2]。基于此，网络运营中应引进网络安全计算机信息技术来保护重要资料与信息，避免因黑客入侵给社会发展造成巨大损失；对于个人而言，则可利用网络安全防护技术防止病毒入侵，保证网络环境安全可靠。就目前的网络环境而言，计算机信息技术大幅提升了产业生产效率、人民生活水平与质量，同时也催生了更多新兴产业，带动国内经济实现新发展。但是更多网络犯罪也在这样高速发展的环境下衍生，造成网络环境混乱。针对当下面临的不同网络安全问题，研发与更新网络安全计算机信息技术既是必然趋势，也是重中之重。

3 大数据环境下的网络安全风险分析

3.1 网络病毒

网络病毒在网络运转期间发生率较高，导致病毒产生的原因为计算机网络内部的系统代码运行时发生一定偏差，在某种信息偏差下，产生网络病毒，影响网络环境。大数据技术涵盖多种计算机技术，这些新型技术的信息采集、信息查询、数据分析、信息传播以及信息质量均得到显著提升，同时，病毒发生率和传播率也在网络环境中逐步增加。一旦运行代码失误，网络病毒将会快速复制，并对网络环境产生反向影响，造成计算机运行不稳定、不安全。

3.2 安全漏洞

安全漏洞也会影响网络环境

【作者简介】陈成龙（1993—），男，广西南宁人，本科，助理工程师，研究方向：图书馆系统技术及阅读推广。

的安全性，存在一定运行风险。在使用计算机设备时，一般需要访问互联网，并获取其他系统支持。此时，若网络系统本身存在稳定性差的特点，则极容易出现安全漏洞，导致信息泄露、信息窃取等。计算机设备运行过程中，其采用的外部操作系统可分为两类，一类为硬件，另外一类为软件，这种系统构成会产生安全稳定问题。例如，外界病毒入侵后，仅仅借助计算机核心算法无法解决病毒问题，还需要通过现存病毒问题展开一定分析。目前来看，计算机网络均存在动态 Bug，这些动态 Bug 通常会在一定范围内，系统会在此期间利用不同方式修补漏洞。如果修补时产生动态 Bug，那么会出现较大风险性和危害性。一旦漏洞被网络不法分子发现，将会造成恶劣影响。

3.3 管理不规范

计算机网络安全出现问题，很大程度上和管理不规范有着密切联系。专业人员在使用计算机时，会及时下载安全软件以确保设备运行环境安全稳定，如防火墙等。利用安全软件可以有效避免网络运行问题，保证用户使用网络的隐私性和保密性。但部分用户在使用网络时存在侥幸心理，拒绝安装安全软件或者安装盗版软件，造成隐私泄露。

4 大数据环境下的网络安全计算机信息技术应用

4.1 计算机信息技术类型

大数据环境下，相关人员应对网络计算机信息技术类型进行充分了解，提升技术专业储备能力，针对性地处理网络应用中遇到的各类安全问题。技术类型可分为以下 3 种。

4.1.1 加密技术

加密技术较为常见，且有良好的防护作用，可有效保护信息安全。与其他网络计算机信息技术相比，加密技术专业性更强，且防御功能更高 [3]。这种技术产生于密码学之上，其原理为加密，如果细细划分，则可将其加密方式分为 3 类，即密钥、链路和节点。

4.1.2 防火墙技术

防火墙技术在网络通信系统中是一种极为重要的安全防护技术。防火墙技术可以与网络完成快速融合。当前防火墙技术特点主要有 3 个方面，一是应用层。防火墙技术的核心内容为计算机网络安全系统应用层。二是过滤性。防火墙技术对于网络环境的防护，主要利用基础信息完成，信息包含数据包使用端口、源地址、目的地地址等。如果通过数据包判定，那么代表着环境信息处于安全状态。三是智能化。防火墙技术的核心支撑为统计学和概率算法，可以有效识别数据，并根据计算量更新，实现最终简化。相关人员如果合理应用该技术，那么可以快速解决网络信息风险问题，实时识别以及阻断网络病毒，及时控制网络环境中出现的恶意数据流量，从而实现信息保护。

4.1.3 身份认证技术

身份认证技术主要是合理应用身份认证功能，对相关用户的网络数据信息进行安全防护，防止黑客入侵，窃取私人信息。计算机网络信息环境中，某一种特定数据信息可以直接反映某位用户的身份信息，计算机终端处理平台可以针对性识别用户数字身份，并将用户数字身份信息进行授权与圈定，以确保身份认证准确。

4.2 建立信息安全防护系统

大数据技术不仅可以发现隐藏信息，还可以及时提取有效信息，从而实现数据互联网之间的联通和共享。数据流在网络中的增长速度不断提高，其覆盖范围也随之扩大，导致网络环境从简到繁，网络攻击与信息泄露成为普遍现象。这种复杂性对信息安全防护系统产生一定阻碍影响，同时也造成防护不及时、信息不可靠等问题，破坏网络环境安全。就此而言，应建立更加全面可靠的信息安全防护系统，结合大数据环境的特点，允许核心信息区安全防护系统完成安全分析。构建系统时，计算机网络信息威胁检测应放于首位，该功能主要检测网络环境中是否存在恶意活动，一旦发现相关 Bug，则可自动保护系统安全，保证用户网络使用的安全性。

大数据环境下，网络安全管理至关重要。因此在建设核心信息区安全防护系统平台时，应不断更新系统，贴合大数据环境实际发展需求，以此保护用户数据保密性和完整性。例如，大数据技术中有一种分析方法被称为多变量大数据分析方法，这种方法可以高效处理异构数据源，一旦检测到异常情况，则可提供与之对应的日志信息。由于系统对攻击原始信息识别特异性较高，其识别率基本接近 100%。系统建设后，应配备充足人员，提高系统平台应用水平。

4.3 增强从业人员信息安全意识

传统网络信息安全系统对网络环境的处理相对较为狭隘，无论是

设备还是功能均有一定限制性。与此对比，大数据技术在网络信息中的安全管理则能更有效解决相关问题，该技术具有以下优势。一是可以针对网络环境的变化，采取不同应对策略，确保环境安全性。二是数据传输过程更具私密性，加密传输的信息安全度更高。三是计算机网络传输数据防篡改性能更强。为了充分发挥大数据环境下计算机网络安全技术的实际效果，相关从业人员的网络安全意识应受到高度重视。系统与技术背景相同时，人员信息安全意识的高低程度更能影响网络安全管理效果。安全意识提升主要可从以下方面入手。第一，引进计算机网络安全技术专业人才。从高校专业招纳贤士，保证薪资待遇与员工福利，以稳定人才资源库，避免基础保障问题造成人才流失。第二，统一组织和培训现有计算机网络安全技术人才，实时跟进网络安全最新技术研发方向，求新求同，保证培训与时俱进，从实际上帮助现有人才强化专业能力。第三，强调大数据环境下，计算机网络安全技术对网络安全的重要性以及必要性。第四，构建当下最为完善的数据信息管理制度，确定大数据技术可以渗入信息管理的每一个环节，提高信息管理水平。

4.4 优化计算机网络安全技术

计算机网络安全技术中，数据信息采集是必不可缺的环节。采集过程中将会涉及大量用户数据信息以及隐私信息，用户对相关信息的保护意识极高。为满足用户网络完全应用需求，应不断优化计算机网络安全技术。例如，匿名化技术便是技术优化中的一项重要技术，从匿名化保护措施方面来讲，该技术可分为两种类型，一是点匿名，点匿名指的是相关用户在发布数据信息时，可以完全隐藏自己的信息内容；二是边匿名，边匿名也被称为关系匿名，其指的是既要隐藏用户数据信息，又要隐藏其与其他用户之间的关系信息。除了匿名化技术外，还有一种技术名为水印技术，保护用户知识版权是这种技术的主要作用。最开始应用该种技术时，常见领域为保护网络用户图像信息，至后期使用后，才用来保护视频数据、声音数据和其他数据。大数据技术应用越来越广泛，该类技术逐渐成为计算机网络安全技术中保护数据信息的重要技术之一。

优化计算机网络安全技术的同时，也需对数据存储功能进行优化更新。目前大数据技术收集到的数据信息均会存储在计算机服务器中。关于数据存储，则可使用两种技术，一是加密，这种技术可分为对称算法和非对称算法。对称算法更具备高效特征，但这种方法一旦发生密钥丢失的问题，将会产生恶劣影响，造成信息损失。非对称算法安全性以及非公开性功能更高，但是又因其非公开性特点，发件人数据信息无法被确认。二是硬件安全。硬件安全会对用户硬件设施进行有效保障，如果需要存储大量信息，涉及高标准计算技术以及大容量硬盘，因此在硬件运行时，应定期安排和维护专业技术人员，保证硬件存储安全。混合技术加密过程如图 1 所示。

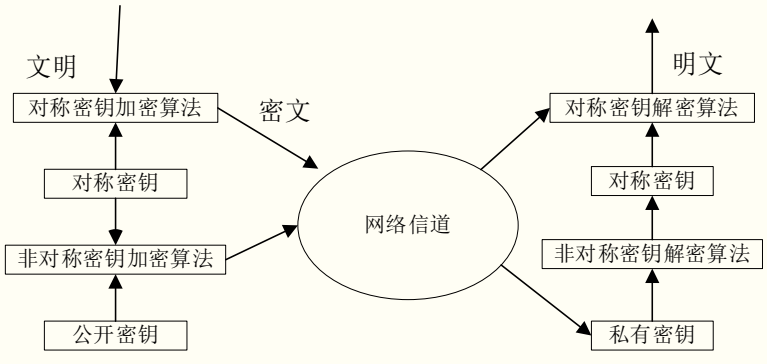


图 1 混合技术加密过程

结语

大数据环境下，计算机网络安全技术应用存在必要性。多元复杂的网络问题不断滋生，更需要科学合理的技术来强化网络环境的安全性与隐秘性。无论是网络安全技术的更新与优化，还是人员网络安全意识与专业能力的强化，均应在业界中受到高度重视，通过技术升级与人才储备，营造和构建更加安全的新时代网络环境。

引用

[1] 徐立溥.大数据环境下计算机网络安全技术实践研究[J].软件,2023,44(9):148-151.

[2] 王彬.大数据背景下计算机信息技术在网络安全中的运用[J].中国新通信,2023,25(12):119-121+223.

[3] 孙学胜.大数据背景下计算机信息技术在网络安全中的应用分析[J].信息记录材料,2023,24(11):142-144.

大数据背景下基层供电企业领导队伍建设研究

文◆广东电网公司广州供电局　陈伟健

引言

为了适应瞬息万变的现代市场经济，企业必须积极改革，提高现代市场经济适应性。然而，大多数基层供电企业仍采取传统的经营管理模式，已经无法适应大数据时代。因此需要发挥领导干部的带头作用，对基层供电企业的领导队伍建设进行优化，从而改变基层供电企业的经营模式，引导基层供电企业与大数据接轨，促进传统能源企业转型升级。

1 大数据背景对基层供电企业领导队伍建设的重要价值

1.1 优化基层供电企业领导队伍构成

基层供电企业的领导人员可以借助信息技术对企业的经营管理与制度安排重新进行审视，通过将信息技术应用到日常工作中，提升工作质量与工作效率。信息技术的优化升级促进了基层供电企业队伍构成的优化升级。例如，基层供电企业的财务管理工作，依托信息技术，整合基层供电企业的财务资源，减少财务沟通成本，提高财务工作效率，拓宽财务工作的广度与深度。除此之外，信息技术的优化升级可以减少财务工作中计算与数据整理方面的内容。因此，凭借信息技术优化升级，可以对基层供电企业财务领导队伍的构成进行优化升级。既然信息技术可以帮助解决基础财务工作，做到对财务工作人员职位与数量的优化，那么在财务工作人员总数减少的基础上，财务领导队伍的配置也不应过于冗杂，避免出现“三个领导管两个员工”的情况。

1.2 提升人才质量

在大数据背景下，高等院校对学生的教育培养过程也逐渐注重大数据信息利用。通过对学生的各项信息数据进行分析，客观认识到学生现阶段的优点与不足，为引导学生全面发展提供数据支撑，并基于对经济情势的分析，及时调整人才培养方向与内容。为基层供电企业提供更高质量的人才。将高质量人才输送至基层供电企业的各岗位，为基层供电企业的领导队伍提供强有力的人才储备。运用大数据技术培养的人才发展更加全面，综合素质更高。同时依靠大数据技术，大学生可以轻松接触到本专业最前沿的知识与信息。对任何感兴趣的知识，大学生都可以凭借大数据技术轻松获取。因此，依靠大数据技术培养的人才在专业能力上更先进，可以更好地满足对基层领导干部工作能力的要求。领导干部掌握了先进的知识，拥有全面发展的素质对促进企业治理体系现代化，推动企业信息化和现代化转型具有重要意义，也为企业在大数据时代获得竞争力提供技术支持。

1.3 提高经济管理效率

大数据背景下，基层供电企业可以通过大数据技术对企业的经营要素进行分析，找出其中的不适应部分，并对其进行优化，从而提高基层供电企业的经济管理效率，并达到降低成本的目的。通过对市场进行大数据分析，在基层供电企业领导干部对企业内外数据信息进行分析的基础上，引导基层供电企业的管理进行合理安排，加大基层供电企业的市场适应性，提高基层供电企业的营收能力，优化基层供电企业的服务流程。

【作者简介】陈伟健（1987—），男，广东增城人，硕士研究生，研究方向：企业管理。

2 大数据背景下基层供电企业领导队伍建设面临的挑战

2.1 基层供电企业领导队伍人才选拔上存在不足

一些基层供电企业在人才选拔上存在不足。首先，在人才招聘上缺少吸引力。导致人才招聘成效低，无法为基层供电企业丰富人才储备。进而导致一些基层领导干部岗位结构偏重，不利于基层供电企业可持续性发展。其次，选聘机制不健全。有些基层供电企业的领导干部提拔完全凭上级领导的个人取向，无法组成适当有效的领导班子，无法发挥领导班子在基层供电企业管理与发展方面的积极作用。再次，一些基层供电企业的领导队伍受制于固有思维，对大数据背景下信息数据的价值认识不足，没有认识到信息数据对增强企业竞争力的重要战略意义，缺少大数据专业人才储备方面的重视，使企业在大数据背景下发展疲软，无法适应大数据时代，拉大基层供电企业与其他企业之间的差距。

2.2 基层供电企业在领导队伍思想建设方面存在的问题

部分基层领导习惯旧的工作模式和思维习惯，对新的管理理念和工作方法接受度不高，阻碍了基层供电企业创新和改革政策的推进。在电力体制改革和市场化变革中，一些基层领导缺乏足够的改革意识和主动性，对新政策和新形势的适应能力不强。当在基层供电企业改革的过程中面临新问题时，存在等待上级指示和任务分配的情况，缺乏主动发现问题和解决问题的积极性。此外，思想政治工作在新形势下面临多样化和复杂化的挑战，需要更加科学化和精细化地开展。从思想上纠正基层供电企业领导干部的工作作风，引导基层供电企业组建现代化的领导队伍。

2.3 基层供电企业在领导队伍绩效考核方面存在的问题

科学的绩效考核机制对基层供电企业领导队伍建设具有重要意义。目前基层供电企业领导队伍在绩效考核方面存在着诸多问题。首先，领导人员的绩效考核目标设定不明确，领导人员缺少明确和可量化的绩效目标，导致难以衡量和评估实际工作成果。当前绩效考核机制过于僵化，过度依赖量化指标，过分强调量化的 KPI 指标，而忽视了领导在团队建设、创新和战略规划等难以量化的方面的表现。其次，缺乏有效的反馈机制，绩效考核结果没有及时反馈给领导或反馈不具体和缺少建设性，导致无法有效促进领导能力的提升。最后，绩效考核在设置时忽视团队协作因素，过于强调个人业绩，忽视了团队协作和集体贡献的重要性。

3 大数据背景下基层供电企业领导队伍建设策略分析

3.1 领导队伍建设要坚持党的领导

基层供电企业要坚持以党的政治建设为统领，任何时候任何情况下都要坚持以党的旗帜为旗帜、以党的方向为方向、以党的意志为意志。学习贯彻习近平新时代中国特色社会主义思想。保持自己的政治纯洁性，用纯洁的党性抵御工作过程中面临的风险与挑战。坚持人民主体思想，把为人民提供更好的供电服务作为工作目标。时刻保持与人民群众的密切联系，做到从群众中来到群众中去。坚守初心，发扬伟大的奉献精神。积极推动基层党建与企业经营发展的有机融合，使党建思想成为供电企业经营发展的重要导向 [1]。

3.2 转变领导队伍选拔的思维模式

基层供电企业的领导队伍选拔应转变传统思维，顺应大数据时代背景，牢固树立数据化、信息化的领导队伍组织思维。重视数据信息对领导队伍选拔的重要作用，重视大数据技术对提升企业核心竞争力的重要意义。大数据不仅指数据量增加，更重要的是如何从海量数据中提取有价值的信息 [2]。运用大数据方式进行领导队伍选拔，通过对基层供电企业领导队伍现状进行科学分析，对基层供电企业领导队伍所缺人才进行专项选拔。同时优化人才数据信息的收集方式，加强对人才数据信息收集的关注，建立基层供电企业的人才数据信息库。提高基层供电企业的信息收集与整理能力，让基层供电在大数据时代背景下拥有更多的人才资源储备。同时，基层供电企业的领导应注意对数据信息的运用，运用科学的技术对领导队伍数据信息进行综合分析，并科学对待分析结果。

3.3 创新领导干部能力分析

将大数据信息应用在对领导干部能力的分析方面，要让基层供电企业在经营过程中增强数据信息应用能力，促进基层供电企业信息化、现代化转型。通过大数据对领导干部的各项能力进行分析，优化领导干部的组织构成，增强领导干部队伍的工作能力，并对领导干部构成人员的工作能力在大数据分析的基础上进行专项培养，提高基层供电企业领导队伍的综合能力。

3.4 提高领导队伍的执行力

基层管理人员在企业组织构架中充当桥梁角色，同时也决定

了各项基层事务的完成效果，使各项战略规划最终落到实处[3]。要进行基层供电企业信息化、现代化转型升级，就必须提高领导队伍的执行力。一方面，应增强基层供电企业领导人员对上级传达的各项决策和工作任务的执行落实能力。另一方面，应增强基层供电企业领导人员在基层供电企业信息化、现代化建设中各项决策的执行落实能力。同时，要求基层供电企业领导人员拥有对企业信息化、现代化建设过程中出现的各种问题及时解决的执行能力。这一过程可以通过大数据技术进行优化，既可以通过大数据技术对各项决策进行创造性表达，又可以通过大数据技术解决遇到的问题。基层供电企业信息化、现代化建设的过程不会一帆风顺，要求基层供电企业领导人员拥有强大的执行力，不拖延，不推诿，一面落实建设策略，一面根据实际作出修改，不断优化基层供电企业信息化、现代化建设道路。

3.5 推动领导队伍决策力的提升

基层供电企业信息化、现代化转型升级应以科学合理的决策为依托，应以大数据广泛和系统的应用为基础，通过挖掘数据要素的应用潜力，将大数据与企业产业要素相融合。将大数据技术以及数据信息引入基层供电企业决策体系。改变传统的以个人经验为基础的决策体系，引导基层供电企业的领导队伍科学、理性地作出决策。首先，决策制定应建立在对数据信息科学分析的基础上。通过对企业的数据信息进行科学分析，把握企业的经营现状，使做出的决定更贴合基层供电企业的经营实际。其次，运用大数据技术对决策制定后的发展脉络进行推演，事先识别决策带来的经营风险，并对决策进行修改，趋利避害。再次，在决策实施过程中可以利用大数据技术不断监控决策实施的相关信息，对决策实施有整体把握能力，并及时对决策预期产生的效果与决策的实际结果进行偏差对比，注重经验总结，在下一次决策制定的过程中避免发生同类过失。

3.6 优化领导队伍构成

基层供电企业信息化、现代化转型升级对企业人力资源管理方面提出了新要求。基层供电企业的领导队伍，尤其是分管人力资源管理的领导，应注重通过对当前领导干部队伍进行分析，寻找领导干部缺口，并进行相关招聘工作。为基层供电企业信息化、现代化转型升级提供智力支撑，转变基层供电企业的人员难以与当前先进的信息化技术相适应的情况。通过大数据专业分析，为领导队伍构成提供专业意见。为基层供电企业信息化、现代化道路的落实提供专业支持。除此之外，应注重为基层供电企业领导干部提供大数据专业知识培养。增强基层供电企业领导干部整体的大数据专业能力，为基层供电企业信息化、现代化建设注入新活力。

结语

大数据对基层供电企业经营管理的作用，并不只是简单地数据信息收集和整理，更重要的是如何基于企业实际的发展状况，结合大数据时代背景，通过对数据信息的分析运用推动基层供电企业信息化、现代化升级。而在大数据背景下如何对基层供电企业领导队伍进行建设，则决定了能否顺利推进基层供电企业信息化、现代化转型升级。

引用

[1] 邓红英.企业管理中的执行力问题及对策[J].企业改革与管理,2021(14):25-26.

[2] 周士魁.浅析大数据对企业管理模式的影响[J].现代营销(上旬刊),2024(6):141-143.

[3] 杨茗程.供电企业基层管理人员执行力培养分析[J].现代企业文化,2023(35):153-156.