审计研究报告

第 10 期

审计署审计科研所　　2019 年 3 月 21 日

以审计智能化为基础 构建大数据审计平台*

陈　鑫　吴　蔚

【摘要】人工智能因其强大的数据信息处理能力，正走进和改变着我们的生活，其在颠覆传统审计组织模式的同时，正引领着审计工作新的发展与变革。本文从审计自动化、深度学习、网络蜘蛛等三个方面论述了大数据审计平台的构建及基本实现思路。

* 陈鑫，审计署驻重庆特派员办事处。吴蔚，国家电网重庆市电力公司审计中心。

一、人工智能与大数据审计的概念

（一）人工智能的概念。

人工智能是研究开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的技术科学，它是计算机科学的一个分支，包括机器人、语言识别、图像识别、自然语言处理和专家系统等，运用人工智能的目标是使机器能够胜任一些通常需要人类智能才能完成的复杂工作。

（二）大数据的概念。

什么是大数据？位于美国康涅狄克州的高德纳咨询公司给出了这样的定义，大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力来适应海量、高增长率和多样化的信息资产。麦肯锡全球研究所给出的定义是：一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合。简单地说，大数据是一种信息资产，是巨量数据集合。它具有海量数据规模、快速数据流转、多样数据类型和价值密度低四大特征。就审计工作而言，大数据是指开展审计工作所需、所产生或与被审计对象相关的信息和数据的集合。根据数据来源不同，可分成被审计单位数据、审计工作本身形成的数据，以及国土、海关、公安等外部数据。

（三）大数据审计的定义。

截止目前，大数据审计尚没有明确的概念。2014 年 10 月《国务院关于加强审计工作的意见》提出，推进有关部门、金融机构

和国有企事业单位等与审计机关实现信息共享，加大数据集中力度，构建国家审计数据系统。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。在此基础上，我们可将大数据审计定义为：在大数据环境下，运用新的审计理念和新的数据获取、分析技术，通过对审计大数据进行充分共享和深度挖掘，发现审计线索或生成审计结论，以进一步提高审计工作效率的审计方式。

二、审计工作运用人工智能的意义

人工智能伴随着大数据的发展而发展，一方面，人工智能离不开大数据的支持，比如，分析销售商品的类型、数量、配送时间、配送地点等因素，能够实现商品集中统一、实时配送，大幅度节约商品库存和配送成本；另一方面，人工智能也推动着大数据技术的发展，比如，分析顾客近期浏览商品，可为顾客推荐针对性商品。

审计工作运用人工智能，或者说审计智能化，就是将大数据的理念、技术运用于审计工作，通过将普适性强、成熟度高的数据采集技术、数据整理技术、数据分析技术、数据挖掘技术及审计判断技术等以审计模板的形式固定下来，形成以审计模板为核心的审计软件，审计人员通过点击按钮及输入必要的条件，即可开展审计的技术。审计智能化的基础是大数据、核心是人工智能，审计智能化将大大减小时间、地域等因素对审计项目的制约，将

颠覆传统的审计组织模式，引领审计工作新的发展与变革。主要体现在:

（一）审计智能化有利于实现全面审计。

当前，审计工作受审计时间和人力资源等因素的影响，审计组常采用重点抽查，重点抽查因“总体 > 样本”，存在可能遗漏重大问题的局限性。使用人工智能后，审计软件可以日以继夜、不知疲倦地工作，使审计工作摆脱审计时间、特别是人力资源的制约，从而使全面审计得以真正实现。同时，人工智能还可以避免人为因素的影响，如审计人员判断失误，甚至故意“放水”等导致重大问题的遗漏。

（二）审计智能化有助于审计工作自动化。

利用人工智能，审计软件可以按审计模板（实质是审计人员预先设定的程序）自动进行数据采集、数据整理、数据分析，甚至还可以对被审计单位的常规性问题进行纠错查弊。同时，使用深度学习技术，还可以自动搜索、识别、归纳和总结被审计单位新出现的普遍性、倾向性异常问题，并自动形成可移植、“傻瓜型”的审计模板，其他审计组使用该模板，即可自动对该类问题进行审计。

（三）审计智能化将大大减轻审前知识储备。

随着经济的发展，社会分工将越来越细，同时，法律法规的完善也意味着其内容将更加丰富与多样。为更好地履行职责，作为“经济守夜人”的审计人员需学习的法律法规、行业背景等知识也

将越来越多、越来越细。为此，传统审计模式下审计人员在审前阶段不得不花大量时间学习相关法律法规及行业背景等。运用人工智能，计算机可以实时更新现行法律制度，自动搜索行业背景和专业知识，且不会出现记不清、记不准、记不牢，更不会出现注意力不集中、精神疲惫等情况。

（四）审计智能化将大大解放审计人员。

审计智能化不仅可以缩短审计工作所需的时间，还可将审计人员从繁琐、重复的简单劳动中解脱出来。比如，传统审计模式下，审计组对某银行进行审计，一般需派七至八名审计人员专门对电子数据进行采集、整理，运用人工智能后，其工作均可由审计软件自动完成，因此，审计智能化相当于增加了审计力量。这样，审计人员就可以进一步发挥人的主观能动性，专心关注审计软件难以完成的重大事项，如查处重大违纪违规问题及从体制机制等角度分析问题产生的原因，并从更高层面、更广范围、更加综合的视角提供具有前瞻性的审计建议。

三、构建大数据审计平台的关键技术

构建大数据审计平台是项较为复杂的系统工程，包括网络、数据库、审计平台软件等技术工程，归纳和总结审计方法和审计经验形成的审计模板，以及技术工程与审计模板的集成。从技术上看，构建大数据审计平台以大数据为基础，并主要依靠审计自动化、深度学习、网络蜘蛛等三种技术，其中审计自动化是基础，深度学习和网络蜘蛛是对审计自动化技术的延伸与拓展。

（一）审计自动化。

审计自动化是指审计软件在没有审计人员干预或较少干预的情况下，按照审计模板设定的审计程序，通过数据采集和数据整理自动化、数据分析和数据挖掘自动化及审计问题查处自动化，实现审计目标的过程。审计自动化的关键是审计程序的设定（或审计模板的固化）。其工作原理如图 1 所示:

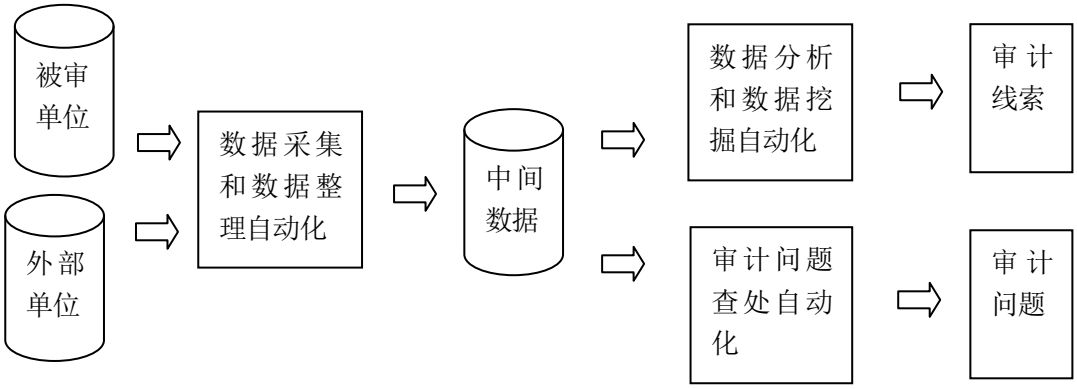

图 1　审计自动化工作原理

1. 数据采集和数据整理自动化。该技术是在摸清大部分被审计单位信息系统（如现有的数据库类型、数据结构等）的基础上，根据审计标准数据接口的要求，对传统的数据采集整理技术进一步进行归纳总结，最终形成数据采集整理模板。每到设定的数据采集周期（如一季度采集一次或半年采集一次），审计软件将自动接收相关单位的原始数据，形成审计标准数据，并存入指定的数据库。虽然形成模板的工作量较大，但一旦形成就能持续使用；同时，运用该技术，审计软件平时自动完成数据采集和整理，避免了原始数据提供慢、中间数据形成时间晚等问题。

2. 数据分析和数据挖掘自动化。数据分析和数据挖掘是审计自动化的核心，其目的是获得审计线索。传统审计模式下，数据分析和数据挖掘依靠审计人员手动进行，分析成果很大程度上取决于审计人员的业务能力，分析技术强、业务能力好，获得的线索就多、质量就高；反之，线索就少、质量就差。目前，审计署既懂数据分析和挖掘技术又懂审计业务的两栖人才不多，懂审计业务的数据分析技术往往不太好，反之亦然。数据分析和数据挖掘的自动化，就是将成熟的、行之有效的数据分析方法、数据挖掘技术以分析模板的形式固定下来。对审计人员而言，具体的分析方法就类似于“黑匣子”——拿来即可用，而不用了解具体的分析过程。数据分析和数据挖掘自动化技术实质是将审计人员的成功经验和专业知识，以分析模板的形式透明地、无阻碍地分享给其他审计人员，达到减轻技术依赖和工作强度的目的。

3. 审计问题查处自动化。审计问题查处自动化是在数据分析和数据挖掘自动化的基础上，根据审计判断模板设定的程序和条件，判断分析发现的审计线索是否符合现行法律法规并直接获得审计问题的技术。本质上，审计问题查处自动化技术是数据分析和数据挖掘自动化的技术升级。相同的是：模板都是核心，模板质量很大程度上决定了审计软件能否尽可能地发现被审计单位存在的问题、决定了发现的问题是否是“真正违反现行的法律法规”；不同的是，审计问题查处自动化获得的是审计问题，进一步减少了审计人员的直接参与，但它对人工智能的要求也更高，需要审

计软件具备人的判断力，而不仅是按预先设定程序执行某些操作。

4. 审计模板。审计模板是审计自动化的核心，模板质量的好坏是审计自动化成败的关键，模板质量包括内容质量和技术质量两个方面。内容质量指模板内容（即设定的程序）符合相关的审计要求、审计程序及法律法规，其质量依赖优秀审计人员对其自身审计经验的归纳和总结。技术质量有四个指标：通常性、稳定性、移植性、可扩展性。通常性是指审计模板适用的广泛性，是适用于几类审计情况，还是某一个或几个特定情况；稳定性是指审计模板对审计环境的适应性，即使审计环境发生较小变化仍能正常工作；移植性是指对某类审计模板进行适当修改（或审计软件自动修改），即可形成新的审计模板；可扩展性是指审计模板的兼容性好，添加、删除新模板简单方便，且模板的增减不影响正常工作。一般地，通常性强、稳定性好、移植性高、可扩展性大的模板，属技术质量好的模板。但内容质量是模板质量的基础，只有合格与不合格两个指标，技术质量再好的模板，一旦内容质量存在偏差，均属不合格模板。

（二）深度学习技术。

深度学习的概念源于人工神经网络的研究，由 Hinton 等人于 2006 年提出，它是一门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性能的学科。深度学习是机器学习中一种基于对数据进行表征学习的方法，它是通过组合低层特征形成更加抽象的

高层表示属性类别或特征，以发现数据的分布式特征表示。

1. 深度学习简介。深度学习也称为深度结构学习、层次学习或者是深度机器学习，它是一类算法集合，是机器学习的一个分支。它是在多层神经网络上运用各种机器学习算法解决图像、文本等各种问题的算法集合。如有篇“关于某银行向某企业发放贷款”的纸质会议记录，要让计算机写出该会议记录的摘要（即主要内容）。第一层神经网络通过扫描，得到该会议记录的图片；第二层神经网络对图片上的符号（即文字、图像、表等）进行识别，并把纸质会议记录的图片内容转换为计算机本身能够识别的文字、图像、表；第三层神经网络对该会议记录的内容进行拆分及归纳、总结，最后得到该会议记录的摘要。

2. 文字识别技术。在探讨深度学习在审计工作中的运用前，先解释一下文字识别技术。计算机表示文字和图像的方式是不同的，我们输入文字时，计算机用其能识别的二进制编码表示，因此，在计算机内部，每个字都二进制编码，只是在屏幕上显示时，计算机用点阵技术将二进制编码转换成我们人眼看到的一个字。但图像的表示就不同了，计算机先把一幅图像分解成一个个很小的像素，每个像素再用二进制编码表示。无论是以英语为代表的字母文字，还是以汉字为代表的笔画文字，每个字（或字母）本质上都是一幅图画。

因此，文字识别技术其实就是将一个个像素转变成二进制的过程。识别时，首先，计算机通过扫描或摄像等方式得到一幅图

画；其次，根据图片的像素规律（比如有文字的地方是黑色，文字间是白色或灰色等）将图片拆成一个或数个由很多个像素组成的小图像；再次，计算机根据图像文字匹配库，结合每个图像的特征（如字母 A 有个尖、字母 P 有个圈、汉字“十”是两条垂直交叉的线等），将每个图像识别成文字，并用二进制编码表示。至此，计算机完成文字识别。但实际工作中文字识别要复杂的多，比如简单的“十”字，有人写成行书、有人写成草书、还有人写成行草，即使是同一个人，有时写的大一些，有时写的小一些。因此，在图像文字匹配库中，一个文字往往对应多个图像，同时，由于文字书写的不规范，计算机在识别文字时，还需用到模糊匹配技术，即找出一个最像的图像等等。

3. 深度学习技术。审计工作运用深度学习技术主要包括两个方面，一是计算机根据识别出的文字做出某种判断。仍以“关于某银行向某企业发放贷款”的会议记录为例，计算机识别出纸质版会议记录后，根据参会发言人员的权重、发言的内容等，做出该笔贷款的发放是经集体决策、还是个人违规决策的判断。其中，发言人员的权重可根据会议记录上参会人员的方式和先后次序来判断，如参会人员优于列席会议人员，前面人员优于后面人员；发言的内容可从内容是否包含“同意”、“没意见”或“不同意”、“保留意见”等关键词语进行判断。该例中，若大部分人的发言都包含了“不同意”、“保留意见”等关键字，则属违规决策。当然，要生成一个好的审计判断模板，远不是判断发言中是否包

括上述那几个关键字那么简单，但核心仍是尽量模拟审计人员的审计判断过程，模拟程度越高，模板内容质量就越好。

二是自动生成审计模板。该技术就是软件根据设定的程序、在无审计人员参与的情况下，自动对数据进行分析，若发现新情况新问题，则提示审计人员是否需要关注该异常、该分析过程是否需生成模板或对原有模板进行修改。为介绍新模板的自动生成，以“近两年医药代表在全国各地注册公司，并向药企大量开具咨询费、服务费发票”为例。首先，计算机根据全国工商数据，统计各年度各类公司（如按公司类型、注册地址、注册金额等）的注册数量，并判断注册数量是否突然大量增加，假设设定增长率为 50%，超过 50%可视为突然大量增加。第二，判断公司是否存在异常。公司注册方面，可用某特定关键字（如公司注册地址、股东身份证号、办理人员身份证号、预留电话号码等）对公司数量进行分类统计，若同一身份证号（或同一预留电话）注册的公司数较多，则存在批量注册公司情况。注销方面，可用某一地区为关键字，分类统计公司存续时间为某特定时段（如一年、两年等）的公司数，若这类公司大量存在，则属部分人群为避免特定事项而注册（或注销）公司。第三，以上述异常公司名称为关键字，在税务系统数据中统计这类公司开具的发票总额，同时，分类统计发票对方单位包含某特定关键字（如药、酒店等）的金额，若包含“药”的金额占发票总额超过 60%，则可认定为：这些公司主要向药企开具发票，以及涉及公司数、药企数、发票金额等。

计算机得到该结果，发出提示：是否关注该事项；若审计人员确认“需关注”，则进一步提示：是否需生成新的审计模板；若审计人员确认“需生成”，则计算机根据上述分析的 SQL 语句自动生成审计分析模板。

（三）网络蜘蛛。本质上，互联网是一个分配式、开放性、海量的大数据平台。其中，分配式是指互联网信息存储在多个服务器（也即数据库）；开放性是指互联网提供的服务是开放的、透明的，用户无需任何权限、也无需知晓服务器地址就可免费获得大部分资讯；海量是指互联网提供的资讯非常庞大，用户使用任一关键字通过搜索引擎均可获得数万条、甚至数百万条资讯。

网络蜘蛛即 Web Spider，如果说互联网是一个蜘蛛网，那么网络蜘蛛就是在网上爬行的蜘蛛。网络蜘蛛通过网页的链接地址来寻找网页，从网站某一个页面开始，读取网页的内容，找到在网页中的其它链接地址，然后通过这些链接地址寻找下一个网页，这样一直循环下去，直到把这个网站所有的网页都抓取完为止。网络蜘蛛除抓取网页外，还能对网络信息内容进行深度分析，形成每条信息的关键字和摘要，当用户用关键词进行搜索时，网络蜘蛛根据摘要筛选出相关信息，并将用户最需要的信息放在最前面。审计工作使用网络蜘蛛运用互联网资讯，主要集中在以下两个方面：

1. 建立和更新审计对象的行业背景库和专业知识库。行业背景是指这个行业具体工作是什么，及这个行业涉及的部门、人群、

产品、市场等。比如金融审计，金融行业可分成银行、证券、保险三大类，其中，银行又可分成中央银行、政策性银行、大型商业银行等；在此之下，每家银行下还可以标注其功能、职责、发展定位等。在国家金融政策板块，可按时间、金融政策类型（即货币政策、利率政策和汇率政策）建立国家当时的金融调控政策行业背景库。网络蜘蛛还可以建立行业专业知识库，如主营业务包括资产业务、负债业务、中间业务、国际业务等，其中资产业务包括信用放款、抵押放款、保证担保放款、贷款证券化等。建好行业背景库和专业知识库后，当审计某一银行时，审计人员便可通过其了解相关行业背景和行业专业知识。同时，网络蜘蛛在网上不停地搜索，并对相关内容进行更新。

2. 提供审计线索或核实审计情况。网络蜘蛛因其强大的搜索功能，可以在审前阶段和审中阶段为审计工作提供便利。如某省生态环保审计，审前阶段，审计人员可通过网络蜘蛛搜索自己希望得到的资讯。如果审计人员希望搜索到该省生态环保遭到重大破坏的例子，可以在搜索栏中输入关键字“某省&生态环保&破坏”，便能搜索到生态环保遭到破坏的相关信息。为进一步提高搜索命中率，还可将关键字修正为“某省&废水|土壤|河流|空气&污染|破坏|”。

审中阶段，审计人员可通过网络蜘蛛核实某些特定信息。如审计人员需核实“张三”是否为某县政府公职人员，可以在搜索栏中输入关键字“某县&张三”进行搜索，若互联网搜索出张三参

加某项目开工仪式的信息，同时，再结合张三的年纪等信息，基本上就能确定需核实张三是否系该县公职人员。再比如，审计人员怀疑某公司在某自然保护区内搞房地产开发，可用关键字“某公司&自然保护区名称&项目”进行搜索，若互联网提供了该项目在某自然保护区的相关网页，则说明该事项基本上属实。

报：署领导。

送：各省、自治区、直辖市和计划单列市、新疆生产建设兵团审计厅（局），署机关各单位、各派出审计局、各特派员办事处、各直属单位，驻审计署纪检监察组，南京审计大学。

发：本所所领导、各处室，存档。　　共印 40 份

编辑：纪晗　　审核：李春兰　　签发：姜江华