大数据环境下基于自然语言处理技术的IT治理审计方法

陈伟　李晓鹏　杨建荣

摘　要：本文首先分析了IT治理审计常用审计方法的不足，进而根据目前开展大数据环境下信息系统审计的需要，提出了基于自然语言处理技术的IT治理审计方法，并分析了其原理；在此基础上，采用Python语言实现了基于自然语言处理技术的IT治理审计方法；最后，结合具体案例，分析了基于自然语言处理技术的IT治理审计方法的应用。

关键词：大数据审计；信息系统审计；IT治理审计；自然语言处理

中图分类号：F239.1, C931.6　**文献标志码：**A　**文章编号：**1006-286X(2019) 06-0069-04

IT（信息技术）治理审计是目前信息系统审计领域研究与应用的一个重点，本文结合目前大数据审计与信息系统审计的研究与应用现状，以IT治理审计为例，研究大数据环境下的信息系统审计问题。

一、基于自然语言处理技术的IT治理审计方法简介及可行方法分析

（一）IT治理审计简介

IT治理用于描述一个单位是否采取有效的机制，使得IT的应用能够完成组织赋予它的使命，同时平衡信息技术与过程的风险，确保实现组织的战略目标。良好的IT治理对于有效控制信息系统运行风险非常重要，因此，IT治理审计是信息系统审计的一项重要内容。在开展IT治理审计时，审计人员一般根据IT治理的相关要求，逐项检查被审计单位的IT治理制度是否存在，以及相应的执行落实情况。一般而言，可以关注以下内容：

在高级管理层情况方面，检查是否设立一个由来自高级管理层、信息科技部门和主要业务部门的代表组成的类似于信息科技管理委员会的专门机构；检查信息科技管理委员会是否定期向董事会和高级管理层汇报信息科技战略规划执行、信息科技预算和实际支出、信息科技的整体状况。在信息科技（或信息管理）部门情况方面，检查是否设立或指派一个特定部门负责信息科技风险管理工作；检查信息科技部门是否根据工作内容，制定完整的内部工作流程和内控制度，建立与相关职能部门之间的协调配合机制，保证信息科技工作的有序、高效；检查是否建立与单位业务相适应的信息科技部门，负责信息科技产品的开发、外包、测试、上线和变更，负责相应信息系统的运行、维护和安全，为单位提供信息科技业务产品等。

目前常用的信息系统审计方法，如访谈、现场观察、文档查看、抽样、穿行测试等，是依据相关法律、法规、规章制度，基于审计人员的审计经验对相关问题进行地毯式排查或重点式查找，没有充分利用大数据环境下采来的大数据，不能很好地发现一些隐藏的审计线索，造成存在的一些IT治理问题很难被发现。另外，大数据环境下大量的文本数据使审计人员分析的难度越来越大，传统的浏览和筛选等方法无法满足大数据环境下文本数据等非结构化数据审计的需要，因此，需要探索如何采用大数据的相关技术开展信息系统审计。

（二）面向IT治理审计的大数据审计可行方法分析

1.大数据智能分析技术。大数据智能分析技术以各种高性能处理算法、智

基金项目：国家自然科学基金（71572080）；教育部人文社会科学研究规划基金（14YJAZH006）；江苏高校品牌专业建设工程资助项目

作者简介：陈　伟，南京审计大学教授，江苏省审计信息工程重点实验室副主任；
李晓鹏，审计署金融审计司处长；
杨建荣，审计署驻上海特派办处长。

能搜索与挖掘算法等为主要研究内容，这是目前大数据分析领域的研究主流，它从计算机视角出发，强调计算机的计算能力和人工智能，例如各类面向大数据的机器学习和数据挖掘方法等。目前关于大数据智能分析技术的研究在审计领域的应用仍不成熟，多是停留在理论研究层面。

可用于大数据智能分析的技术很多，如关联规则分析、遗传算法、神经网络、机器学习、社会网络分析、自然语言分析等。本文根据IT治理审计的需要，选择了适合文本数据分析需要的基于自然语言处理的相似度分析方法（Koh，2010；陈伟，2018）。

2.大数据可视化分析技术。大数据可视化分析技术是目前大数据审计应用中比较成熟和主流的内容。常见的大数据可视化技术包括直方图、散点图、气泡图、热力图以及文本可视化技术标签云等。将文本数据中的内容或规律以视觉符号的形式展示给审计人员，有助于审计人员利用视觉感知的优势来快速获取大数据中蕴涵的重要信息，从而发现审计线索。本文根据IT治理审计的需要，选择了适合文本数据分析需要的基于自然语言处理的标签云技术（Koh，2010；陈伟，2018）。

二、基于自然语言处理技术的IT治理审计方法原理分析

为了便于审计人员从整体上把握被审计单位的信息系统风险等情况，快速发现与IT治理相关的可疑问题，提高审计效率，实现“集中分析、分散核查”的审计方式，基于自然语言处理技术的IT治理审计方法原理可概述为：根据对被审计单位的调查，在访谈和现场观察等基础上，采集被审计单位的内外部IT治理相关信息如业务办公会会议纪要、风险报告、审计报告、相关IT治理管理制度等非结构化数据，以及从外部相关监管部门网站或其他网站上公开数据源采集来的相关文本数据等非结构化数据，如监管部门出具的相关警示函等，在审计大数据预处理的基础上，采用自然语言处理技术实现标签云分析，并对相关警示函、业务办公会会议纪要等相关非结构化数据进行建模和整体分析，审计人员结合自己的审计背景知识，通过对可视化的分析结果进行分析和观察，快速从被审计大数据信息中检查是否存在与IT治理相关的可疑问题情况，发现异常数据，获得审计线索。同时，根据需要交互改变标签云分析的参数设置，改变输出的可视化图形和图像大小，从不同方面获得对被审计数据的理解。在此基础上，对相关IT治理管理制度等非结构化数据进行相似度分析，检查被审计单位的相关制度是否随着机构变更而更新。通过对这些异常数据做进一步的延伸审计和审计事实确认，从而发现被审计单位IT治理方面的风险线索，最终获得审计证据。综上分析，基于自然语言处理技术的IT治理审计方法原理如图1所示。

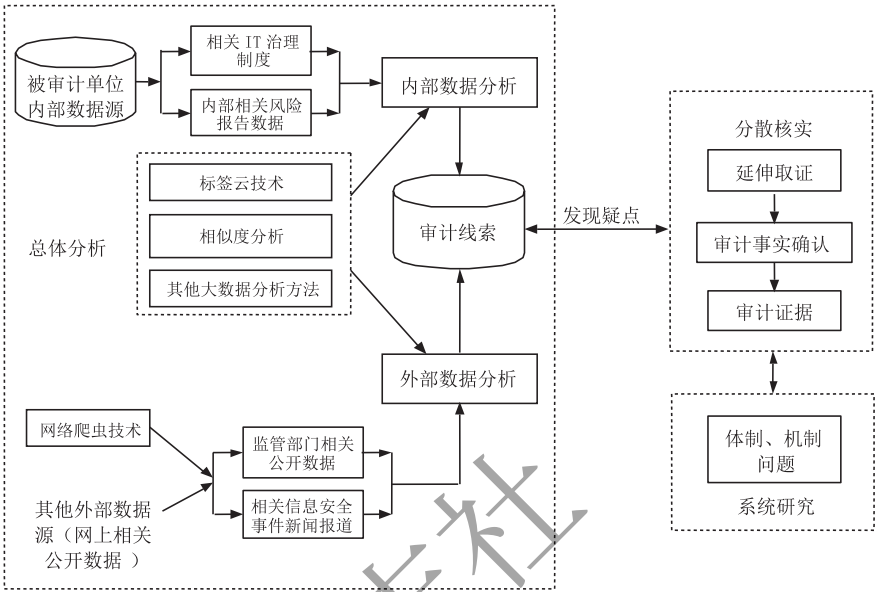

图1　基于自然语言处理技术的IT治理审计方法原理

基于图1，大数据环境下IT治理审计所需的主要数据包括：

1.被审计单位主要内部数据。大数据环境为IT治理审计提供了全方位分析的相关数据，审计人员可以从被审计单位采集被审计单位的相关业务介绍、部门年度工作总结、风险分析报告、审计报告等相关文本数据，通过这些文本数据，审计人员可以了解目前被审计单位的相关业务情况、相关风险等，便于审计人员开展相关审计工作。比如，通过分析被审计单位的相关风险分析报告等文本数据（如被审计单位的“公司交易系统故障事件总结报告”等），可以判断该单位的信息系统建设、运行等工作中是否发生过相关风险；通过分析被审计单位的业务办公会会议纪要等文本数据，可以判断该单位的会议纪要中都关注哪些问题，是否重视信息系统建设、运行以及风险管理等方面的工作；通过分析被审计单位的相关IT治理管理制度等文本数据，可以判断该单位是否有相关IT治理管理制度，以及相关IT治理管理制度的内容是否合理等。

2.被审计单位主要外部数据。一是

图2　外部相关警示函数据标签云分析示例

图3　公司内部相关交易系统故障事件总结报告标签云分析示例

相关监管部门网站警示函数据。通过分析从外部相关监管部门网站上采集来的相关文本数据，如监管部门出具的相关警示函等，可以判断被审计单位是否存在信息系统建设、运行等方面的风险。二是其他网站上公开的信息安全事件的新闻报道数据。通过分析从外部其他相关网站上公开数据源采集来的相关文本数据，如相关网站关于被审计单位的相关信息安全事件的新闻报道等，审计人员可以判断被审计单位是否存在信息系统建设、运行等方面的风险。

三、基于自然语言处理技术的IT治理审计方法在商业银行的应用

为了全面防范化解金融风险，目前在金融审计中需要关注被审计金融机构的重大风险隐患情况，比如：各业务板块存在的风险隐患，主要包括金融类企业不良资产状况和化解处置情况以及互联网金融等创新业务中存在的风险隐患；非金融类企业在经营过程中存在的投资、运营、财务、创新等风险隐患。信息系统建设、运行管理中存在的风险隐患，主要包括信息系统的安全性、有效性、经济性等，揭示信息系统管理和运行中的风险隐患。

根据前文对基于自然语言处理技术的IT治理审计方法原理的分析，本案例以某商业银行审计为背景，以IT治理审计为例，探究该商业银行目前的信息科技运行与风险情况（注：本案例中相关数据已经过脱密处理）。

（一）文本数据总体分析

1. 通过外部公共数据分析获得相关审计线索。为了发现相关信息系统建设与运行风险，审计人员首先采用网络爬虫技术获得相关监管部门网站或其他网站上的相关信息，如监管部门出具的相关警示函等，自动搜索网上关于被审计单位一些公开报道的风险信息等。在此基础上，采用标签云技术等进行综合分析，审计人员借助分析结果，辅助判断被审计单位的相关IT治理情况（示例结果如图2）。

由图2可以发现，该被审计单位存在一些信息系统风险，表明在IT治理方面存在一定的不足。

2. 通过内部相关数据分析获得相关审计线索。审计人员通过采集被审计单位的“公司交易系统故障事件总结报告”等文本数据，然后采用标签云技术进行综合分析，从而发现相关审计线索（示例结果如图3）。

由图3可以发现，该被审计单位存在一些信息系统风险，表明IT治理水平存在一定的不足。

3. 相关办公会会议纪要等文本数据分析。通过前两部分的分析，审计人员可以推断被审计单位的信息系统存在部分风险，但尚不能确定是哪一方面的风险。

为了做进一步的分析，采集了被审计单位的“业务办公会会议纪要”等文本数据，然后采用标签云技术对业务办公会会议纪要相关会议主题进行综合分

图4　近五年业务办公会议题分析结果示例

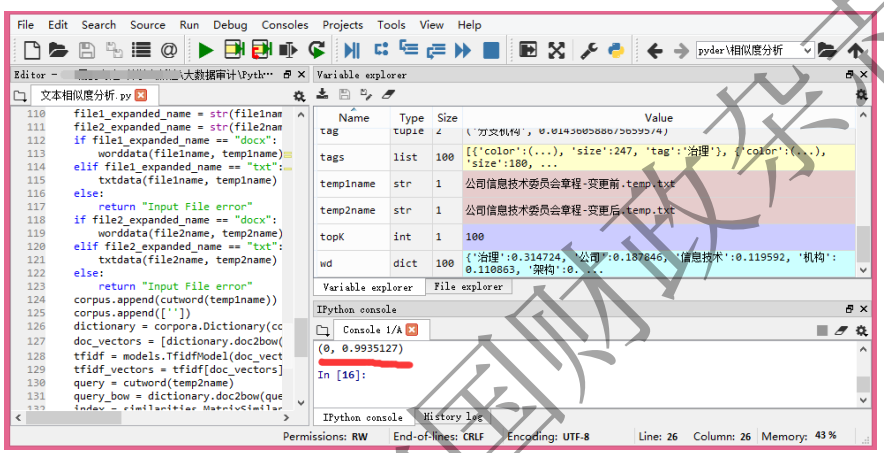

图5　信息技术委员会章程变更前后内容相似度分析过程及结果示例

析（示例结果如图4）。由图4可以发现，该被审计单位的业务办公会会议大多和信息系统无关，表明该公司在IT治理方面不重视，信息技术委员会方面的会议议题很少。

4. 相关IT治理管理办法内容分析。由前文发现该公司在IT治理方面不重视，向被审计单位调取相关IT治理管理办法后，发现尽管该单位为了提高IT治理水平，成立了相关机构信息技术委员会，但自成立以来，没有定期召开相关会议，信息技术委员会效果和作用不明显，该信息技术委员会的功能由业务办公会顺便执行，即被审计单位的IT治理机构发生了变化。为了了解IT治理机构变更后相关管理制度是否进行了同步更新，采用文本相似分析了技术综合分析信息技术委员会的《公司信息技术委员会章程》和变更后的公司办公会的《公司信息技术委员会章程》等文本数据，其分析过程和结果如图5所示。

由图5的分析结果发现，二者的内容相似度为0.9935127，相似度极高，也就是说，尽管被审计单位的IT治理机构发生了变化，但该被审计单位的信息技术委员会章程基本没有变化，值得审计人员进一步关注。

（二）分散核实

根据以上线索，通过进一步的延伸调查最终发现，该被审计单位曾撤消信息技术委员会，造成该信息技术委员会最终以失败而告终。信息技术委员会撤消后，由业务办公会顺便履行其职责，相关制度未及时更新，IT治理实施执行与已有制度描述不符。

（三）系统研究

综上分析发现，由于被审计单位IT治理制度的不足，导致目前IT治理结构、各部门间职责划分不明确，影响IT治理的有效性，进而给本公司的信息系统建设、运行管理造成潜在风险隐患，从而发生了上述监管部门警示函和被审计单位内部系统故障事件总结报告中提及的信息系统故障。因此，被审计单位今后应加强IT治理方可从根源上解决上述信息系统建设、运行管理中存在的潜在风险隐患。

责任编辑　武献杰

主要参考文献

[1] 陈伟. 大数据技术在扶贫审计中的应用[J]. 财务与会计，2018，(15)：69-71.

[2] 陈伟. 审计信息化[M]. 北京：高等教育出版社.2017.

[3] 陈伟，高杰，居江宁. 基于大数据技术的经济责任审计[J]. 财务与会计，2018，(21)：51-53.

[4] 陈伟，居江宁. 基于大数据可视化技术的审计线索特征挖掘方法研究[J]. 审计研究，2018，(1)：16-21.

[5] 陈伟，Smieliauskas W. 大数据环境下的电子数据审计：机遇、挑战与方法[J]. 计算机科学，2016，(1)：8-13，34.