机器学习算法在央行内部审计问题管理中的应用

■ 中国人民银行福州中心支行 陈唯源

摘 要：在央行内部审计中，审计人员需对审计发现的问题进行分类管理和归纳总结。随着内部审计的发展，发现的问题逐年增加，需要采用新技术和新方法来提升问题管理的效率。本文探索了分类、预测、聚类和自然语言处理等机器学习算法在央行内部审计问题管理中的应用，实现了对问题标签的自动预测，并提出了闭环式问题管理思路。

关键词：机器学习；内部审计；问题管理

随着人民银行内部审计的发展，审计涉及的部门和领域以及发现的问题逐年增加。《审计署关于内部审计工作的规定》（审计署令第11号）指出：“单位对内部审计发现的典型性、普遍性、倾向性问题，应当及时分析研究。”面对审计作业系统逐年积累的问题，做好内部审计问题管理，有助于审计部门发现普遍性和倾向性的问题，建立健全内控措施，深化审计成果运用。随着机器学习技术的发展，人工智能已渗透到各行各业，基于机器学习算法的问题管理是当前审计工作的新技术和新方法。

一、内部审计问题管理的功效

（一）问题管理的总体功能

内部审计发现的问题通常兼具普遍性和特殊性：普遍性指同类问题可能多次发生；特殊性指具体到某个问题，其相对于同类的其他问题也有所不同。内部审计问题管理是指以“给问题贴标签”的方式对问题进行分类归纳总结，有助于审计人员揭示普遍性的问题。

内部审计问题管理功能通常内嵌在审计作业系统中。以人民银行内审综合管理系统为例，审计人员在录入审计项目及问题时，将每个问题的标签信息一并录入系统。常见的问题标签可包括风险类别、所属职能、问题词条、所属依据等不同维度的信息。风险类别是指问题涉及的风险，如操作风险、法律风险、声誉风险等不同形式；所属职能指问题发生的职能领域，如资产管理、采购管理、法律事务管理等多种领域；问题词条是问题的精要归纳，通常是一个简短的语义单

作者简介：陈唯源（1989-），男，福建福州人，工学硕士，工程师，供职于中国人民银行福州中心支行，副主任科员，研究方向：信息技术审计。

收稿日期：2020-03-26

元，如“预算编制不合理”；所属依据是指该问题违反的依据文件。问题标签的设置即独立也有所重叠，例如不同职能均可能涉及同一种风险，资产和采购管理职能均可能涉及操作风险等。一个典型的审计项目及问题录入形式，如图1所示。

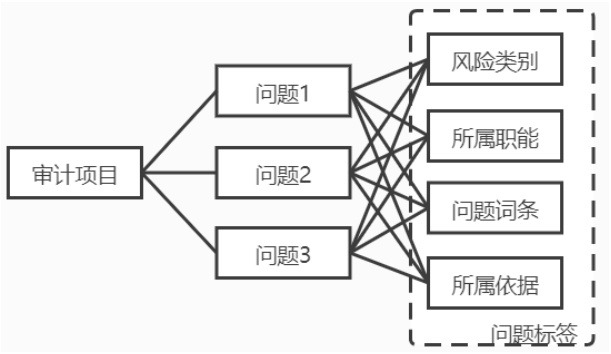

图1 审计项目及问题录入形式

（二）问题标签的闭环管理

问题标签的闭环管理主要指标签的设置和选择构成一个闭环。各类标签可以预先录入到审计作业系统中，供审计人员录入问题时选择，审计人员也可以根据新的问题类型动态地改进或调整标签。机器学习算法可用于问题标签的闭环管理：在标签选择方面，可以借助分类预测算法自动选择问题所属的标签，提高标签选择精度，减少标签录入耗时；在标签设置方面，可以借助聚类算法辅助总结归纳标签，动态优化标签的设置。下文将通过自然语言处理、分类预测和聚类等机器学习算法，对问题标签的闭环管理提出新思路。

二、机器学习算法及其适用性分析

（一）自然语言处理算法

近年来，自然语言处理算法有了长足的进步，常见的自然语言处理任务包括文本分类、文本聚类、摘要自动生成和自动问答等。相关算法在上述任务和领域上逐渐达到或超过人类的水平。转换器输出式双向编码器（Bidirectional Encoder Representations from Transformers，BERT）算法是效果较好的自然语言处理算法，可将文本转化为数值向量，供后续算法进一步处理。本文使用BERT算法处理问题描述文本，生成数值向量，提供给后续分类预测和聚类算法做进一步处理。

（二）分类和预测算法

分类和预测算法可以根据历史数据，对新数据的类别或数值进行预测。常见的分类和预测应用包括文本情感分类、垃圾短信文本识别和新闻自动分类等。本文通过K近邻、Softmax分类器和深度神经网络（Deep Neural Networks，DNN）3种算法实现了标签信息的自动预测选择。

（三）聚类算法

聚类算法可以将相似的数据自动归类，常见的聚类应用包括文本自动聚类、消费群体聚类等。本文采用层次聚类对问题描述文本进行自动聚类，并提取文本摘要帮助生成分级的问题词条等标签，从而实现问题标签的闭环管理。

利用上述算法，基于机器学习的问题管理流程如图2所示，其中标签的生成、选择及预测构成了闭环管理。

该算法流程首先使用BERT算法将问题描述文本生成向量，然后通过聚类算法辅助生成标签信息供分类算法进行选择，并根据问题文本向量预测最合适的标签信息，从而实现标签的闭环管理。

三、数据处理和算法应用

从审计作业系统中抽取人民银行某省所有分支机构近3年发现的问题，主要包括审计年度、问题描述文本、风险类别、所属职能、问题词条等标签的数据。一个典型的问题信息见表1所列。

（一）问题文本向量生成

使用BERT算法将每个问题描述文本转化为向量，向量维度设置为768维。例如，对表1中“未及时修订

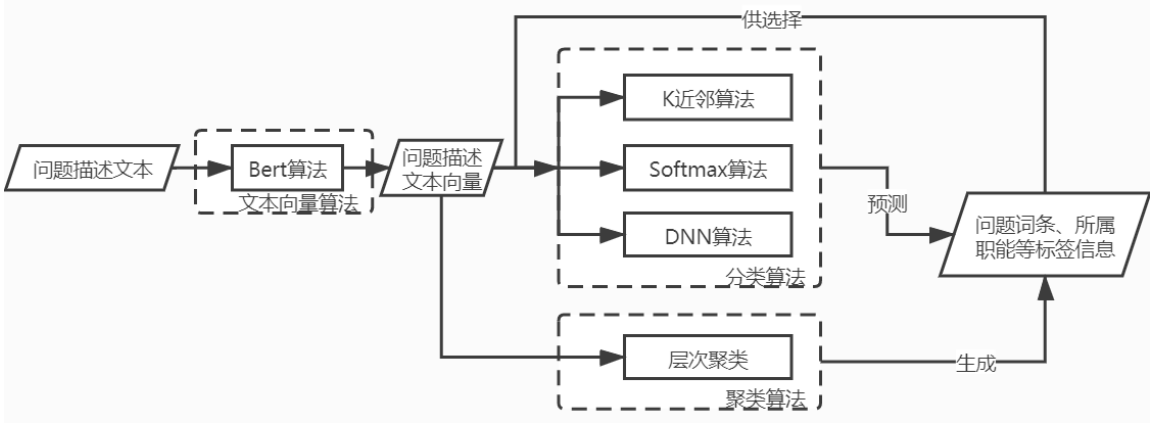

图2 内部审计问题管理算法流程

表1 问题数据示例

年度	问题描述文本	风险类别	所属职能	问题词条
2019	未及时修订财务管理相关制度；**财务费用管理规定中的采购程序等内容与**规定不一致，未及时修订	操作风险	预算管理	会计基础工作不规范—财务制度不健全

财务管理相关制度”的问题描述文本，生成的向量v为[0.223,0.045,…−0.158,−0.029]，该向量维度为768维。

为检验所生成的问题描述文本向量的准确性，可以通过计算向量相似度的方式找到和给定问题最相似的问题，以检查文本向量的精度。定义数据集中第i条和第j条文本向量的相似度为向量余弦相似度：

$$\mathrm{sim}(i,j)=\cos(v_i,v_j)=\frac{v_i,v_j}{|v_i||v_j|}$$

在计算出所有问题描述的向量后，对于给定的某个问题，可以通过计算该问题向量和其余问题向量的余弦相似度，得到相似度最高的问题。例如，与表1中的问题最相似的为“制度修订不及时，《**财务费用管理规定》与《**办法》规定不一致，未对《**财务费用管理规定》中的限额进行相应的调整”，相似度计算结果为8.286，算法求得的向量较为准确地反映了文本的语义。

（二）分类预测问题标签信息

对分类预测任务，使用2017年−2019年6月的问题作为训练集，2019年7−12月期间的问题作为测试集，训练集和测试集均为表1中的数据格式。该任务假设在2019年7月对训练集进行训练，之后对测试集中的每个问题，在仅给定问题描述文本及所有候选标签的情况下，通过算法预测标签信息，并通过人工核对检查预测结果。

分类预测使用K近邻、Softmax分类器和DNN3种备选算法，参数设置如下：K近邻算法选取和给定问题描述最相似的K个问题，将出现频次最多的标签作为预测值，参数K设置为10；Softmax分类器将回归用在多分类问题上，不涉及参数设置；DNN算法使用深度神经网络的架构，对输入的问题文本向量进行分类预测，神经网络中包括输入层、隐层及输出层，输入层维度设置为文本向量维度768维，隐层设置为512维，输出的维度为对应候选标签的个数。在上述参数设置下，测试集的预测准确度见表2所列。

表2显示，风险类别和职能的预测准确度较高，问题词条由于候选的词条标签数量较多，因此准确率低于其他标签的预测准确度。

（三）聚类辅助生成标签信息

在上述数据集中，标签信息均在审计前就已设置

表2 测试集预测准确度

算法	风险类别	所属职能	问题词条
K近邻	82%	75%	55%
Softmax	89%	80%	68%
DNN	91%	77%	65%

完成。由于问题词条标签较为复杂，在审前较难考虑到所有潜在的问题，可能存在某些问题无法找到对应词条的情况。为解决问题词条设置不完善的问题，可以在录入问题后，对问题进行聚类，生成需要的问题词条标签。在得到问题文本向量后，使用相似度衡量问题文本间的相似性，使用层次聚类进行聚类，聚类结果如图3所示。

图3中，横轴为问题，层次聚类将最相似的问题两两合并，之后不断地向上层进一步合并，生成问题树。多个问题聚为一类，见表3所列。

该类问题代表的词云如图4所示，通过聚类及词云辅助，审计人员可归纳出该类问题主要是“财务管理

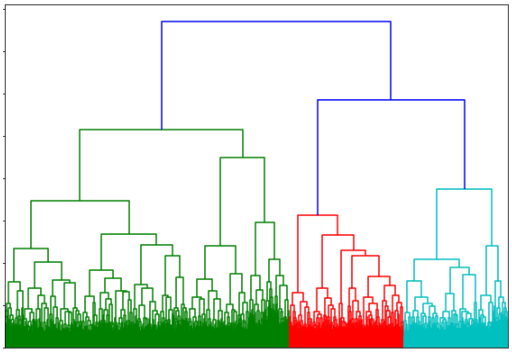

图3 问题的层次聚类结果

表3 层次聚类问题集合

问题描述
未及时修订财务管理相关制度。**财务费用管理规定中的采购程序等内容与**规定不一致，未及时修订。
制度修订不及时，《**财务费用管理规定》与《**办法》规定不一致，未对《**财务费用管理规定》中的限额进行相应的调整。
个别财务管理规定未及时修订更新。《**财务管理规定》有关审批权限的条款未更新。
《**差旅费管理实施细则》有关住宿标准的内容未根据相关通知文件进行调整。
…

图4 聚类结果的词云分布

和采购等制度未及时修订完善”。通过不断重复上述过程，审计人员可以无遗漏地归纳出所有问题词条标签，从而建立树型的层次级问题词条。

（四）小结

通过上述算法，可计算得出较为准确的文本向量，实现了文本聚类和词云生成，辅助审计人员归纳问题标签；实现了标签预测，辅助审计人员根据问题描述文本自动选择问题标签。该算法可以减少传统人工方式归纳和选择问题标签的工作量，提高工作效率。为进一步提高标签归纳的效率，后续研究还可进一步借助文本摘要和文本生成算法自动生成合适的文本标签。

四、总结和展望

随着人民银行内部审计的发展，审计发现的问题逐年增加，做好内部审计问题管理，有助于改善问题分析、促进问题整改。随着人工智能和机器学习等技术的发展，计算机理解、处理及运用人类语言的能力将逐步提高，因此面对审计作业系统逐年积累的问题，将机器学习融入问题管理，可以帮助审计人员更好地管理问题、总结问题，从而提升审计工作效率，深化审计成果运用。FTT

参考文献：

[1]张磊. 文本分类及分类算法研究综述[J].电脑知识与技术，2016(34):225-226.

[2]杨熳. 如何有效发挥内部审计问题词条功效[J].中国内部审计，2019(10):58-61.