DOI：10.13718/j.cnki.xdzk.2020.10.001

“大数据与行业应用”专题

随着科学技术的发展，各行各业的数据都出现了大规模的增长. 这些海量数据中隐藏着与民生相关的重要信息，利用不同的数据分析工具，挖掘数据中的潜在应用价值，可以为行业问题的解决提供量化的支持. 本期“大数据与行业应用”专题，刊载了 2020 年第八届中国计算机学会大数据学术会议（CCF BIGDATA2020）推荐的 5 篇论文，利用人工智能算法，对电力、医疗、隐私等相关领域的行业数据进行深度挖掘与分析，在促进相关理论研究发展的同时，为相关领域的科学决策提供有效的见解.

面向电力审计领域的两阶段短文本分类方法研究

赵雅欣[1]，　郑明洪[1]，　石林鑫[1]，
向　菲[2]，　江金洋[2]，　尹　心[2]

1. 国网重庆市电力公司，重庆 404100；2. 国网重庆市电力公司 电力科学研究院，重庆 401123

摘要：为解决电力审计领域中将审计发现问题按标准问题定义进行归类的现实需求，提出了一种两阶段短文本分类方法. 该方法包括粗分类和细分类两个阶段. 粗分类阶段通过对审计报告文本中的主观问题定性，对国网电力审计问题库文本中的审计问题的一级类目等粗粒度特征进行模糊匹配，实现审计发现问题预分类. 细分类阶段通过对审计报告文本中的问题进行描述，对国网电力审计问题库文本中的审计问题的二、三级类目等细粒度属性进行文本特征抽取、特征向量化，并将粗分类阶段的结果用于文本特征和向量化的权值调整和相似性度量的对象范围选择，实现基于非结构化文本的审计问题分类. 国网重庆市电力公司 2016 年审计发现问题汇总数据的实验结果表明，该方法能够有效提升审计问题分类性能，能为电力审计行业规范审计报告、提高审计效率提供有力支持.

关　键　词：电力审计；短文本；两阶段分类器；文本特征提取；加权 Word2vec；模糊匹配

中图分类号：TP391.1　　**文献标志码：A**　　**文章编号：**1673－9868(2020)10－0001－07

在电网审计领域中，如何根据所发现问题的自然语言描述对审计问题进行规范化的定性和归类是当前尚未解决的关键问题. 一方面，随着信息化的快速发展，电网各企业积累了大量的非结构化审计报告文本，文本中包含了审计人员人工记录的审计问题描述、审计问题定性和问题分类. 受审计人员个性化语言表达和主观判断的影响，这些文本具有非特征性和歧义性的显著特点，主观的问题分类与国网电力审计问题库的问题类目标准定义存在较明显的不一致. 因此，如何基于电力审计问题报告文本对审计问题进行科学规范的分类是提升审计管理质量和效能的现实需求. 另一方面，电网企业的审计报告和国网电力审计问题库

收稿日期：2020－08－15
基金项目：国家电网公司总部科技项目（1300201999440A0000）；国网重庆市电力公司电力科学研究院科学技术项目（SGCQDK00NYJS2000111）.
作者简介：赵雅欣(1977－)，女，高级会计师，主要从事大数据审计分析的研究.

大多是短文本，信息有限，同时又具有鲜明的行业特征(例如：文本相似度高，分类边界模糊等[1-2])，这些特征给电力审计领域短文本分类带来一定阻碍，有待进一步研究.

近年来，短文本的分类问题得到广泛关注，已对此提出多种方法. 其中，通过将短文本向量化来构建文本表示模型是当前的主流方法. 该类经典方法包括向量空间模型(Vector Space Model，VSM)、Word2vec模型、Doc2vec模型等. VSM模型最初由文献[3]提出，这种模型将单词以向量的形式表示出来，因此也常被称为词袋模型(Bag of Words，BoW)[3]. 词袋模型灵活简单，在处理少量文本时能达到不错的效果，但由于词袋模型和词汇数量密切相关，当处理的数据较大时，容易造成维度灾难. 文献[4]提出的Word2vec模型是当下比较流行的文本建模方法. 这种方法将单词放入神经网络中训练得到词向量，该方法不受词汇量的约束，能够人为设置向量维度，可与其他特征提取方法结合计算词向量，这种模型结构克服了词袋模型的缺点. Doc2vec模型[5]是基于Word2vec模型提出的句向量算法，该算法能从句子、段落等较长的文本中学习得到固定长度的特征表示，从而实现文本向量化. 目前，Word2vec模型、Doc2vec模型已应用到中文文本的分类问题中，并取得了较好的分类效果. 文献[6]提出了基于Word2vec模型的中文短文本分类方法，将词性作为特征计算加权向量. 这种方法适用于词性对分类结果有很大影响的问题，比如情感分析领域中，对分类结果有较大影响的形容词和副词应具有较高权重，但不太适合于以反映客观事实为核心的审计报告类公文. 文献[7]在Word2vec模型以及信息增益的基础上，考虑了同义词集以及词语位置信息，并结合特定规则调整词语的特征权重，以决策树作为分类器对审计文本进行了分类. 文献[8]提出了改进的Doc2vec模型，并结合专业词典增强文本特征以实现专业领域文本分类. 但是，这些方法均不太适用于分类标签多，分类标签具有层次结构，且训练数据集有限的电力审计分类问题.

由于审计报告文本存在审计问题主观定性、审计问题详细描述等粗、细粒度不同的属性，国网电力标准问题库文本也存在审计问题一、二、三级类目等不同粗、细粒度特征，可考虑对不同粒度属性分阶段处理实现分类. 基于类似的分阶段分类思想，一些现有研究将粗、细(coarse-to-fine，CTF)策略[9]应用于分类过程，提出两阶段分类模型. 文献[9]采用粗、细两阶段分类法对医学细胞图像实现了分类，粗粒度分类阶段采用聚类算法对原始数据集进行了粗粒度的划分，细粒度分类阶段采用卷积网络对粗粒度数据集进行了进一步分类. 文献[10]提出了一种基于两阶段学习的半监督支持向量机分类算法，该算法第一阶段采用标签传播算法去除数据噪声，第二阶段采用支持向量机对去噪数据进行分类. 但是，这些方法并未针对粗、细粒度属性进行分阶段分类的设计，不太适用于电力审计分类问题.

本文从电力审计文本存在的多种粒度属性共存、文本相似度高、文本特征类间分布差异较大等特点出发，提出一种两阶段电力审计分类方法. 具体来说，本文针对两类审计文本展开研究. 第一类审计报告文本数据包含待分类的所有问题样本，每个样本包括主观问题分类、主观问题定性、主观问题事实描述3项. 根据这3项属性的层次结构关系，本文将主观问题分类视为粗粒度属性，将主观问题定性和主观问题事实描述视为细粒度属性. 第二类国网电力审计标准问题库文本数据包括一级类目、二级类目、三级类目3项. 同样基于不同级别类目之间的层次结构关系，比如一级类目“人力资源管理”下可包含“福利保障管理”等多个二级类目，每个二级类目下又可包含类似“超计划、超范围、超标准劣质福利”等多个三级类目，本文将一级类目视为粗粒度属性，将二级类目和三级类目视为细粒度属性. 本文所提方法将对粗、细粒度属性分成两个阶段进行分类处理. 将本文提出的方法应用到2016年国网重庆市电力公司审计问题汇总数据，结果表明，该方法能够提高审计问题的分类准确性和计算速度，实现非结构化电力审计问题的标准化归类.

1 两阶段电力审计短文本分类方法

本文提出的两阶段电力审计短文本分类方法包含粗粒度分类、细粒度分类两个阶段(如图1所示). 在第一个粗粒度分类阶段，针对粗粒度属性进行处理，采用不考虑语义信息的模糊匹配方法[11]，通过字符串近似查找将审计报告中的审计问题预归类到标准问题库中的一级类目. 在第二个细粒度分类阶段，首先将粗粒度分类的结果用于词频—逆文档词频(Term Frequency - Inverse Document Frequency，TF-IDF)算

法[12]的文本特征项权重计算，然后将获得的关键词权重与 Word2vec 方法结合进行文本特征向量化，最后采用余弦相似度计算审计报告中的审计问题和粗分类结果中所包含的二、三级类目的对应细粒度属性的相似程度，从而得到最终分类结果.

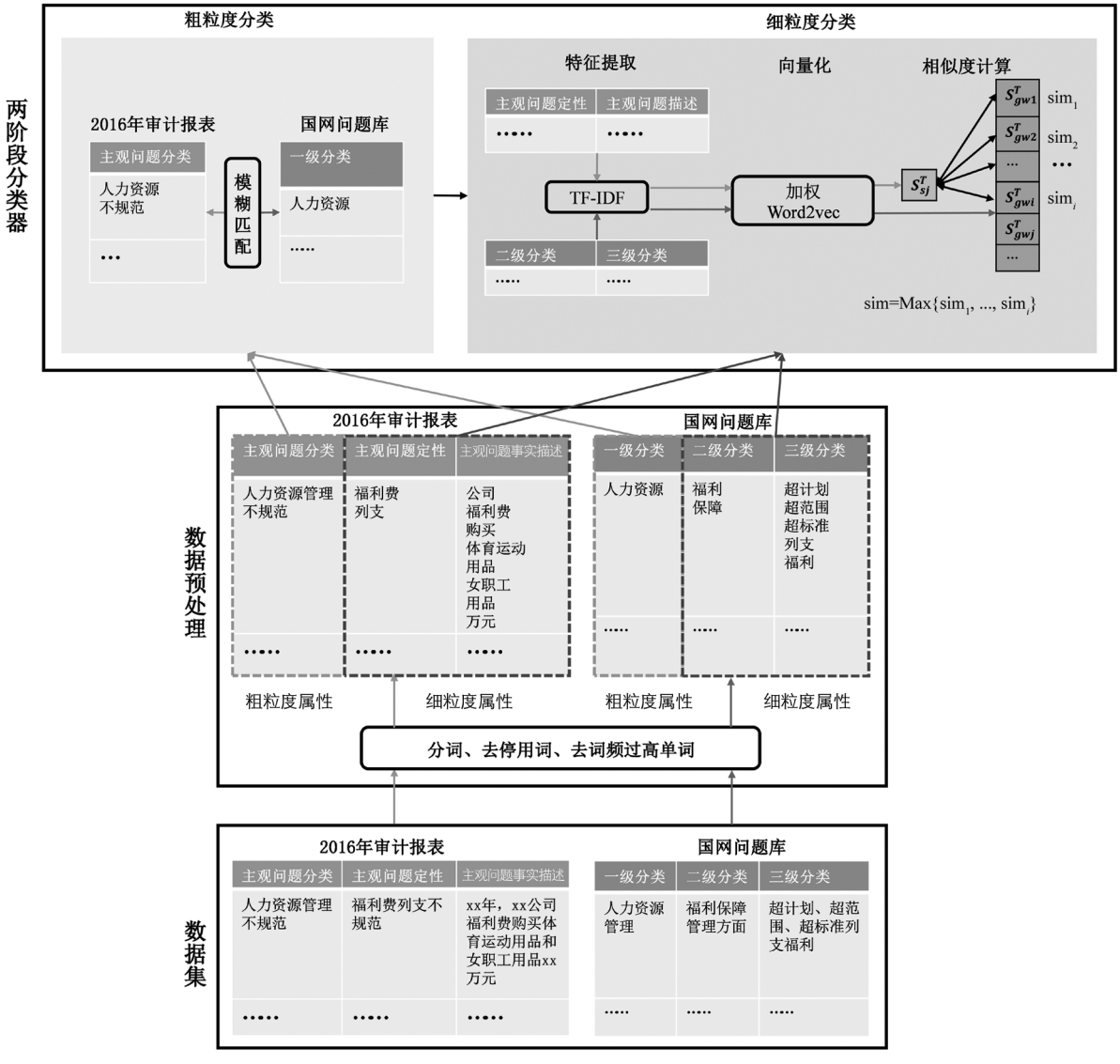

图 1　两阶段电力审计短文本分类方法示意

1.1　阶段一：基于模糊匹配的粗分类

由于审计问题报告中的粗粒度属性(即审计人员主观判定的审计问题分类和审计标准问题库中的问题一级类目)具有较强的相似性，因此可采用模糊匹配方法[13]对审计报告中的问题进行粗粒度分类. 模糊匹配适用于不考虑语义信息，通过计算字符串相似性的方式来衡量两个文本的匹配程度，匹配程度越高表明两个文本字符串越相似，越可能属于同一类. 编辑距离算法(Levenshtein Distance 算法)[11]是一种经典的模糊匹配方法. 编辑距离是指两个字串之间，由一个转成另一个所需的最少编辑操作次数. 这里的编辑操作包括将一个字符串转换成另一个字符串，插入一个字符，删除一个字符. 通常编辑距离越小，两个字符串的相似度越大.

针对两个给定的字符串，首先计算这两个字符串的编辑距离. 由于编辑距离和相似度成反比，因此可采用如下公式将编辑距离转换为相似度值 sim：

$$\mathrm{sim}=\left(1-\frac{ld}{\mathrm{Max}\{len(str_1),\ len(str_2)\}}\right)\times 100 \tag{1}$$

其中，ld 是字符串 str_1 和 str_2 的编辑距离，sim∈[1，100]. 通常，如果 sim 值大于某个相似度阈值(通常设为 60)[11]，则判定两个字符串模糊匹配成功；反之，匹配失败.

采用模糊匹配方法将审计问题与标准问题库中一级类目进行匹配，可以得到多个候选分类结果，从而缩小分类的计算空间. 比如，在审计报告中，审计人员将某个审计问题主观定性为“工程管理不规范”，采用模糊匹配在标准问题库中得到“工程项目管理”和“工程财务管理”两个可能结果. 这样，只需要在这两个一级类目下去进一步匹配二级、三级类目，得到最终结果.

1.2 阶段二：基于模糊匹配结果的细分类

在细分类阶段，为了在粗分类结果下进一步细化分类结果，得到定性更准确的二、三级分类结果，需要进一步比较审计报告文本和审计标准问题库文本的相似性. 在这一阶段，本文首先采用 TF-IDF 算法对审计问题报告和审计标准问题库的细粒度属性进行特征提取，然后采用加权 Word2vec 模型，通过词向量化和句向量化两个子步骤实现文本特征向量化，最后在粗分类结果的基础上，采用余弦相似度计算审计报告中的审计问题描述和审计标准问题库中二、三级类目对应问题描述的相似性，从而得到最终分类结果.

1.2.1 基于 TF-IDF 的文本特征提取

文本的核心信息常常由某些关键单词传达，即可通过抽取关键词得到文本特征. 文本特征提取主要根据一定规则抽取文本的核心信息，从而达到降维、简化计算的目的. TF-IDF[12]是一种经典的特征提取方法，该算法实现简单，易于理解，且克服了使用词频衡量单词重要性的缺点，已在文本分析中广泛应用. TF-IDF 通过计算特征的词频(Term Frequency，TF)和逆向文本频率(Inverse Document Frequency，IDF)两个指标对特征集合中的每个特征(即单词)进行统计评估，使得每个特征获得一个权值，权值越大的特征通常更加重要，并具有更好的分类能力. TF-IDF 的计算公式为

$$\mathrm{TF}_{dw}\text{-}\mathrm{IDF}_{dw}=\mathrm{TF}_{dw}\times\mathrm{IDF}_{dw} \tag{2}$$

其中：TF_{dw} 是单词 w 在文本 d 中的词频；IDF_{dw} 是单词 w 在文本 d 中的逆文档词频，$\mathrm{IDF}_{dw}=\log\left(\frac{N}{n_w}+1\right)$，单词 w 的 IDF 值越大通常说明 w 具有更好的类别区分能力；N 为语料库中文本数量总和；n_w 为语料库中出现单词 w 的文本数量；$\mathrm{TF}_{dw}\text{-}\mathrm{IDF}_{dw}$ 为单词 w 在文本 d 中的权重.

经典 TF-IDF 算法未考虑特征项在不同类的分布差异. 如果某一特征项在某个类别中大量出现，而在其他类中出现较少，这样的特征项就可能具有较强的分类能力，应给予较高的权重. 因此，本文将考虑审计文本特征在一阶段粗分类结果上的分布差异，通过增加在某一类中频繁出现的特征项的权重来扩大类间区分度. 改进后的 IDF 计算公式为

$$\mathrm{IDF}_{dw}=\log\left(\frac{m_{c,w}}{n_w}\times N\right)=\log\left(\left(\frac{m_{c,w}}{m_{c,w}+k_w}\right)\times N\right) \tag{3}$$

其中：$m_{c,w}$ 代表粗分类结果中类 c 出现单词 w 的文本数量；n_w 代表所有文档中包含单词 w 的数量；k_w 代表除类别 c 外，出现单词 w 的文本数量；N 为文档总量.

1.2.2 基于加权 Word2vec 的文本特征向量化

Word2vec 模型[4]是一种经典的基于神经网络的文本特征向量化模型. Word2vec 通过重构语义上下文，将词汇空间映射到一个高维实向量空间. 但是经典 Word2vec 模型主要考虑词和词之间的关系，没有考虑单词在文本中的重要程度，因此向量化后的文本特征可能不具有很好的分类效果. 针对这一问题，本文将 Word2vec 模型训练的词向量与 TF-IDF 获得的权值结合计算单词的加权词向量，计算方式为

$$\boldsymbol{word}_{\boldsymbol{vec}_i}=\boldsymbol{vec}_i\times \mathrm{e}^{\mathrm{TF}_{dw}-\mathrm{IDF}_{dw}(i)} \tag{4}$$

其中：$\boldsymbol{vec}_i$ 代表采用 Word2vec 模型训练的第 i 个单词的词向量；$\mathrm{TF}_{dw}\text{-}\mathrm{IDF}_{dw}(i)$ 是通过 TF-IDF 计算的第 i 个单词的权值.

Word2vec 包括两种计算模型：Skip-gram 模型[4]和 Continuous Bag of Words (CBOW)模型[4]. Skip-gram 模型采用中心词预测上下文的策略，而 CBOW 模型采用上下文预测中心词的策略. 由于 CBOW 模型

在处理小样本数据集上的表现优于 Skip-gram 模型，因此本文采用 CBOW 模型训练词向量.

句向量是句子的向量化表示，可通过每个句子中所有单词的词向量的求和平均作为该句的句向量. 计算公式为

$$\boldsymbol{sen}_{vec}=\frac{\sum_{i=1}^{m}\boldsymbol{word}_{vec_i}}{m} \tag{5}$$

其中：$\boldsymbol{sen}_{vec}$ 代表当前句向量；m 表示当前文本中的单词个数；$\boldsymbol{word}_{vec_i}$ 代表第 i 个单词的加权词向量.

1.2.3　基于余弦相似度的相似性计算

余弦相似度计算是计算文本相似度的常用方法，通过计算向量空间中两个向量夹角的余弦值衡量两个个体间差异的大小[14]. 本文基于审计问题和标准问题库细粒度属性的句向量化结果，采用余弦相似度计算二者的相似度，相似度最高的类目即为最终分类结果. 余弦相似度的计算方式为

$$\cos\theta=\frac{\boldsymbol{sen}_{vec_i}\cdot\boldsymbol{sen}_{vec_j}}{\|\boldsymbol{sen}_{vec_i}\|*\|\boldsymbol{sen}_{vec_j}\|} \tag{6}$$

其中：$\boldsymbol{sen}_{vec_i}$ 是审计问题文本中细粒度属性“主观问题定性”和“主观问题事实描述”的加权句向量；$\boldsymbol{sen}_{vec_j}$ 是审计标准问题库中二、三级类目的加权句向量.

2　实验及结果

本文基于国网重庆市电力公司 2016 年审计发现问题汇总报告和国网电力审计标准问题库两个文本进行实验. 其中：国网重庆市电力公司 2016 年审计发现问题汇总报告是待分类文件，包含 21 个主观问题分类项、242 个主观问题定性项和 309 个主观问题事实描述；国网电力审计标准问题库是标签文件，包含 16 个一级分类项、66 个二级分类项和 276 个三级分类项. 基于这些文本数据，将本文提出的两阶段电力审计短文本分类方法与 Doc2vec 模型、Word2vec 模型、加权 Word2vec 模型等 3 类模型的分类效果进行对比实验. 由于审计发现问题汇总报告和国网电力审计标准问题库中包含多种形式的非结构化数据，且存在噪声，因此首先对文本数据进行分词、降噪等预处理.

在数据预处理阶段，首先采用哈工大语言技术平台(Language Technology Platform，LTP)分词工具[15]对审计发现问题报告和审计标准问题库文本进行中文分词. 分词过程采用 LTP 内部分词表和外部分词表结合的方式. 在实验分析过程中，外部分词表根据电力审计领域专家知识和审计领域常见词集自行建立. 接着采用哈工大的停用词表[16-17]来剔除不相关、无语义信息的副词、虚词等. 然后对经过分词和去停用词操作后的词集进行词频统计，剔除掉类似“管理”、“方面”、“不规范”等在绝大多数记录中出现，即对分类贡献较小的高频词汇.

基于预处理后的数据，本文所提方法与其他 3 个短文本分类方法在分类准确率、召回率和调和平均数等分类性能指标的对比结果如图 2 所示. 结果表明，本文所提的两阶段电力审计短文本分类方法在各个分类评价指标上均是最优值，表明本文方法较大地提升了审计问题分类性能.

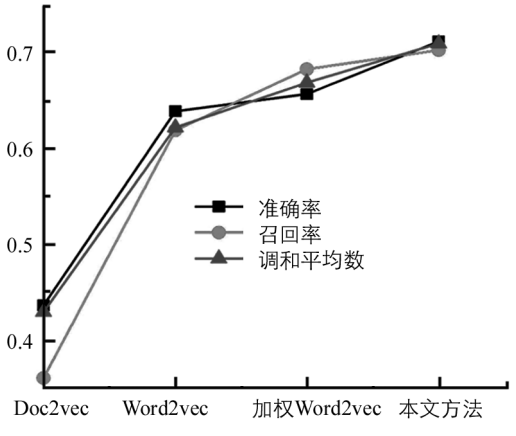

图 2　对比实验结果

为进一步分析本文方法的有效性，对审计发现问题汇总数据涉及的“人力资源管理”、“物资管理”、“招投标管理”、“工程财务管理”、“财务资产管理”和“工程项目管理”等 6 个审计标准问题一级类目上的分类性能进行了对比，结果如图 3 所示. 结果表明，与其他单阶段的分类方法相比，本文提出的两阶段分类

方法在所有一级类目上的分类准确率都优于其他方法. 表明基于模糊匹配的粗粒度分类能在一定程度上弥补句子向量化后语义信息不充分、特征不明显的问题，从而提升分类效果.

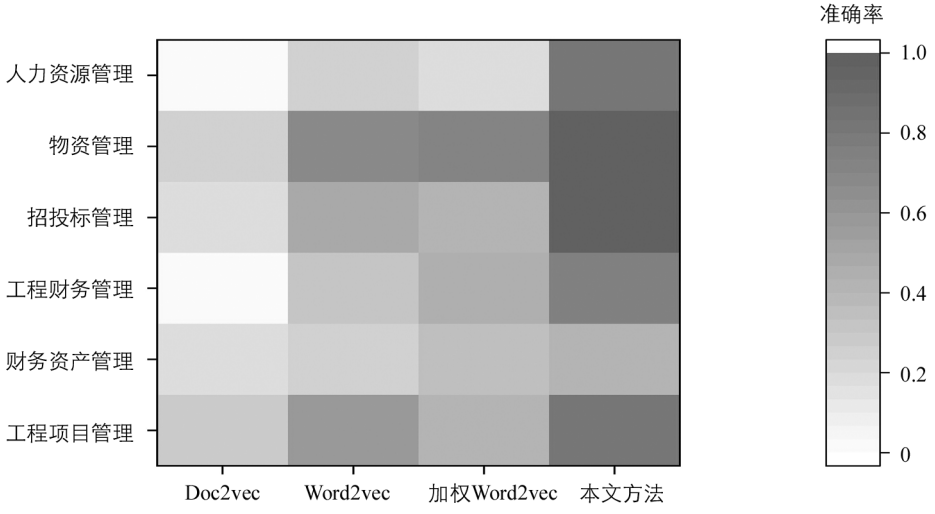

图3 一级分类实验结果

3 结 论

本文针对电力领域审计问题标准化归类的现实需求，从电力审计文本信息有限、部分内容相似程度较高、分类特征不明显的特点出发，提出了一种两阶段电力审计短文本分类方法. 该方法针对审计文本中存在的粗、细粒度属性，综合运用模糊匹配、TF-IDF、加权 Word2vec 等技术并对权重计算方式进行了重定义. 基于模糊匹配的粗分类结果能在一定程度上弥补传统 Word2vec 方法文本特征抽取不充分的问题. 细分类阶段重定义的加权文本特征提取方式通过调整单词权值，提高单词的类别区分度. 基于真实电力审计数据的实验结果表明，本文提出的方法具有更好的分类效果，能够为审计问题的定性和分类提供有力支持，有利于提升审计工作效率，并支持后期审计管理人员的深入分析.

参考文献：

[1] 宁亚辉，樊兴华，吴 渝. 基于领域本体的短文本分类 [J]. 计算机科学，2009，36(3)：142-145.

[2] 邓丁朋，周亚建，池俊辉，等. 短文本分类技术研究综述 [J]. 软件，2020，41(2)：141-144.

[3] SALTON G. A Vector Space Model for Automatic Indexing [J]. Communications of the ACM，1975，18(11)：613-620.

[4] MIKOLOV T，SUTSKEVER I，CHEN K，et al. Distributed Representations of Words and Phrases and Their Compositionality [J]. Advances in Neural Information Processing Systems，2013，2013：3111-3119.

[5] LE Q，MIKOLOV T. Distributed Representations of Sentences and Documents [C] //Proceedings of the International Conference on Machine Learning. Beijing：JMIR，2014：1188-1196.

[6] 汪 静，罗 浪，王德强. 基于 Word2Vec 的中文短文本分类问题研究 [J]. 计算机系统应用，2018，7(5)：211-217.

[7] 伍 洋，钟 鸣，姜 艳，等. 面向审计领域的短文本分类技术研究 [J]. 微电子学与计算机，2015，32(1)：5-10.

[8] 陈 平，匡 尧，胡景懿，等. 增强领域特征的电力审计文本分类方法 [J]. 计算机应用，2020，40(S1)：109-112.

[9] 张梦倩，张 莉. 粗-细两阶段卷积神经网络算法 [J/OL]. 计算机科学与探索，(2020-07-23) [2020-08-05]. https://kns.cnki.net/kcms/detail/11.5602.TP.20200723.1653.018.html.

[10] 陶新民，曹盼东，宋少宇，等. 基于两阶段学习的半监督支持向量机分类算法 [J]. 信息与控制，2012，41(1)：7-13.

[11] LI Y J，LIU B. A Normalized Levenshtein Distance Metric [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2007，29(6)：1091-1095.

[12] ZHANG W，YOSHIDA T，TANG X. A Comparative Study of TF* IDF，LSI and Multi-Words for Text Classification [J]. Expert Systems with Applications，2011，38(3)：2758-2765.

[13] 陈开渠，赵　洁，彭志威. 快速中文字符串模糊匹配算法 [J]. 中文信息学报，2004，18(2)：59-66.
[14] 黄承慧，印　鉴，侯　昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法 [J]. 计算机学报，2011，34(5)：856-864.
[15] CHE W X，LI Z H，LIU T. LTP：A Chinese Language Technology Platform [C] //Proceedings of the 23rd International Conference on Computational Linguistics. Beijing：Association for Computational Linguistics，2010：13-16.
[16] 陈沈焰，吴军华. 基于本体的概念语义相似度计算及其应 [J]. 微电子学与计算机，2008，25(12)：96-99.
[17] 李生琦，天巧燕，汤　承. 基于《知网》词汇语义相关度计算的消歧方法 [J]. 情报学报，2009，28(5)：706-711.

A Two-Phase Short-Text Classification Method for Classifying Audit Problems in Power Grid Companies

ZHAO Ya-xin[1]，　ZHENG Ming-hong[1]，　SHI Lin-xin[1]，
XIANG　Fei[2]，　JIANG Jin-yang[2]，　YIN　Xin[2]

1. State Grid Chongqing Electric Power Company，Chongqing 404100，China；
2. State Grid Chongqing Electric Power Research Institute，Chongqing 401123，China

Abstract：Automatic classification of problems that are found in the process of power-grid auditing into standard problem categorizations is of great practical needs. In order to address this issue，a two-phase short-text classification method is proposed in this paper. This classification method contains a coarse classification phase and a finer classification phase. In the coarse classification phase，the fuzzy matching method is usedto classify the auditing problems based on their coarseness features such as a general specification of auditing problems in the auditing problem reports. Then the results obtained from the coarse classification phase are utilized for text feature extraction and vectorization to determine the class candidates in the finer classification phase. The final class of an auditing problem depends on the similarity between its finer features，i. e. a short-text based problem description，and those of class candidates. The experimental results based on the 2016 auditing data provided by State Grid Chongqing Electric Power Company show that our proposed method can improve the classification performance for power-grid auditing problems.

Key words：power-gird audit；short text classification；two-phase classifier；textfeature extraction；weighted Word2vec；fuzzy matching

责任编辑　廖　坤