文章编号：2095-4980(2024)12-1400-07

基于改进LDA算法的电力用户咨询文本分类算法

李竹青[1]，侯本忠[2]，曹培祥[1]，王一蓉[3]，李向阳[4]

(1.国网安徽省电力有限公司，安徽 合肥 230061；2.国家电网有限公司，北京 100032；
3.国家电网有限公司大数据中心，北京 100032；4.北京国网信通埃森哲信息技术有限公司，北京 100053)

摘　要：针对目前情感极性分析中电力咨询短文本的准确性较低的问题，提出一种基于改进潜在狄利克雷分配(LDA)算法的电力用户咨询文本分类算法。在分析电力咨询短文本与情感的关联关系基础上，定义了基于情感词共现袋、主题特殊词以及主题关系词的概念；为提高语义分析的质量，设计了改进LDA算法的电力用户咨询文本分类算法执行流程。实验表明，所提模型表现出优异性能，平均精确度和平均召回率为90.91%和85.03%。所提模型可充分发挥多模型集成优势，有效提升模型性能。

关键词：电力咨询；文本分类；主题分析；卷积神经网络；潜在狄利克雷分配

中图分类号：TP393　　**文献标志码：**A　　**doi：**10.11805/TKYDA2023119

Text classification algorithm of power user consultation based on improved LDA algorithm

LI Zhuqing[1]，HOU Benzhong[2]，CAO Peixiang[1]，WANG Yirong[3]，LI Xiangyang[4]

(1.State Grid Anhui Electric Power Co.，Ltd，Hefei Anhui 230061，China；2.State Grid Corporation of China，Beijing 100032，China；3.Big Data Center of State Grid Corporation of China，Beijing 100032，China；4.Beijing State Grid Accenture Information Technology Co.，LTD，Beijing 100053，China)

Abstract：In response to the current issue of low accuracy in sentiment polarity analysis of short texts in power consulting, this paper proposes an improved Latent Dirichlet Allocation (LDA) algorithm-based classification algorithm for power user consulting texts. Based on the analysis of the relationship between power consulting short texts and sentiment, concepts such as sentiment word co-occurrence bags, topic-specific words, and topic relationship words are defined. To improve the quality of semantic analysis, an execution process for the improved LDA algorithm for classifying power user consulting texts is designed. Experiments show that the proposed model demonstrates excellent performance, with an average precision of 90.91% and an average recall rate of 85.03%. The proposed model can fully leverage the advantages of multi-model integration, effectively enhancing the model performance.

Keywords：power consulting；text classification；theme analysis；Convolutional Neural Network (CNN)；Latent Dirichlet Allocation

随着电力市场[1-2]不断发展，高质量的电力服务成为电力企业快速竞争的重要手段。随着网络、大数据、物联网、通信[3-5]等技术日益成熟，电力服务平台积累了大量短文本。这些短文本承载了电力用户的情感需求信息。短文本具有语义稀疏、维数高等特点。为此，迫切需要通过一定手段分析这些短文本，以了解电力用户需求，从而有效提升电力服务质量。

目前，主题分类[6]已成为短文本处理分析领域的热门研究领域之一。主题分类可以发现文档和词之间潜在的语义关系，从而有效地挖掘短文本的潜在语义信息。潜在狄利克雷分配(LDA)[7]是一种主流的文档生成的概率模型。LDA的基本思想是将文档视为隐含主题的混合物，其中每个主题由与主题相关的词的概率分布表示。因此，LDA可用于识别大规模文档集或语料库中的潜在主题信息。文献[8]提出了一种基于主题提示的电力命名实体识别方法。该方法将每个实体类型视为一个主题，并使用主题模型从训练语料中获取与类型相关的主题词。文献

收稿日期：2023-05-06；修回日期：2023-07-26

[9]提出了一种基于LDA模型的电力投诉文本热点话题识别方法。尽管上述文献在主题情感分析和语义提取方面取得了一些突破，但大都集中在LDA主题模型上，对现有电力领域文本情感分类研究较少。为分析高维和稀疏的电力短文本，需提高情感分析的聚类精确度。传统的LDA主题模型只考虑了短文本上下文之间的关系；同时电力短文本复杂，呈现多模态、多维度的特点，因而传统机器学习模型和主流深度学习模型无法有效提取数据特征，情感分类准确性较低。

为改善上述问题，本文提出一种基于改进LDA算法的电力用户咨询文本分类算法。该算法以LDA模型为基础，结合卷积神经网络(CNN)和K–means等模型，可有效对电力咨询短文本进行分类。

1　问题陈述

1.1 情感词共现袋

由于词性的不同，电力咨询短文本与情感的关联程度不同。通常情况下，最能反映情感的词性包括形容词、动词和副词。这些词用于修饰名词，以便对人物、事件和热门主题的电力短文本进行最终分析。因此，为提取情感词汇，首先应建立情感词共现袋。

令 S_{T} 为短文本词袋，情感词共现袋 $F(S_{\mathrm{T}})$[10]可定义如下：

$$F(S_{\mathrm{T}})=c\left(\sum_{1}^{i}s(a_{\mathrm{dj}})\right)\cup c\left(\sum_{1}^{k}s(a_{\mathrm{dv}})\right)\cup c\left(\sum_{1}^{j}s(v)\right)\cup\sum_{1}^{j}\sum_{1}^{h}s(v+n_{\mathrm{oun}})\cup c\left(\sum_{1}^{n}s(e_{\mathrm{lse}})\right)\tag{1}$$

式中：a_{dj}、a_{dv}、v、n_{oun} 和 e_{lse} 分别为形容词、副词、动词、名词和其他词性；i、k、j、h、n 分别为短文本包中形容词、副词、动词、名词和其他词性的数量；$\sum_{1}^{i}s(a_{\mathrm{dj}})$、$\sum_{1}^{k}s(a_{\mathrm{dv}})$、$\sum_{1}^{j}s(v)$、$\sum_{1}^{j}\sum_{1}^{h}s(v+n_{\mathrm{oun}})$ 和 $\sum_{1}^{n}s(e_{\mathrm{lse}})$ 分别为形容词、副词、动词、名词和其他词性的词袋；$c(\cdot)$ 为共现。需注意，名词袋 $\sum_{1}^{j}\sum_{1}^{h}s(v+n_{\mathrm{oun}})$ 用于表示动词和名词的共现，且该词袋取决于原始的短文本，而不取决于词典。

当删除停止词后，假设这些共现词的表达是相邻的动词和名词，可基于此提取共现词。此外，形容词袋和副词袋的情感极性主要取决于词典中是否有反义词、否定词和转折词。假设 $C_{\mathrm{adj,\,adv}}(x)$ 表示形容词袋和副词袋的词汇限制：

$$C_{\mathrm{adj,\,adv}}(x)=\begin{cases}\sum_{1}^{i}\sum_{1}^{k}s(a_{\mathrm{dj}}+a_{\mathrm{dv}}),\ x=a_{\mathrm{dj}},a_{\mathrm{dv}},\ p=0\\ -\sum_{1}^{j}\sum_{1}^{k}s(a_{\mathrm{dj}}+a_{\mathrm{dv}}),\ x=a_{\mathrm{dj}},a_{\mathrm{dv}},\ p=1\end{cases}\tag{2}$$

式中：p 为句子中是否存在反义词或否定词；“–”为表达极性相反的词。

动词袋的词汇主要取决于词典中是否有扩展的形容词或副词。同理，动词袋的词汇限制 $C_v(x)$ 为：

$$C_v(x)=c\left(\sum_{1}^{j}s(v)\right),R_{\mathrm{oot}}(a_{\mathrm{dj}},a_{\mathrm{dv}})\in s(v)\tag{3}$$

式中 $R_{\mathrm{oot}}(a_{\mathrm{dj}},a_{\mathrm{dv}})$ 为词根。

1.2 主题特殊词

主题特殊词[11]是主题的中心词，用于区分不同主题的特征。在电力文本领域，主题特殊词可理解为主题中最具代表性的词汇。不同的主题具有不同的主题特征。

假设 A_i 为主题 T 的第 i 个特殊词，w 为主题特殊词，则 w 定义如下：

$$s_p(w,A\in T)=\sum_{w\in A_i,w'\neq w}d(w,w')\tag{4}$$

式中：w' 为主题关系词；$d(w,w')$ 为 w 和 w' 的中心权重，且由 w 和 w' 的共现度计算。当一个词与这个主题中的其他词有更高的共现度时，表明该词更具代表性，对这个主题更为重要。

令 $R(w_{\mathrm{noun}}|w_{\mathrm{v}})$ 和 $R(w_{\mathrm{v}}|w_{\mathrm{noun}})$ 表示名词 w_{noun} 与动词 w_{v} 的相对共现，则有：

$$\begin{cases} R(w_{\text{noun}}|w_{\text{v}}) = \dfrac{f(w_{\text{noun}}|w_{\text{v}})}{f(w_{\text{noun}})} \\ R(w_{\text{v}}|w_{\text{noun}}) = \dfrac{f(w_{\text{v}}|w_{\text{noun}})}{f(w_{\text{v}})} \end{cases} \tag{5}$$

式中：$f(w_{\text{noun}}|w_{\text{v}})$为词$w_{\text{noun}}$和$w_{\text{v}}$在同一主题中一起出现的次数；$f(w_{\text{v}})$和$f(w_{\text{noun}})$分别为词$w_{\text{v}}$和$w_{\text{noun}}$出现在主题中的次数。

假设$c(w_{\text{noun}}|w_{\text{v}})$表示名词$w_{\text{noun}}$和动词$w_{\text{v}}$的共现，则有：

$$c(w_{\text{noun}}|w_{\text{v}}) = \frac{R(w_{\text{noun}}|w_{\text{v}}) + R(w_{\text{v}}|w_{\text{noun}})}{2} \tag{6}$$

如上所述，根据式(5)和式(6)，可以计算两个词之间的共现度，并且可以计算基于情感共现词袋的主题专用词。

1.3 主题关系词

主题关系词[12]是在所有主题中都可以观察到的通用词，这些词表征与其他主题的每个主题词最密切相关的词。

假设B_i为主题T的第i个主题关系词，则主题关系词定义如下：

$$r_{\text{e}}(w', A_i \in T) = \sum_{A_j \in T, A_j \neq A_i} \sum_{w' \in A_j, w' \neq w} d(w, w') \tag{7}$$

2 改进LDA主题分类算法

提出的一种基于改进LDA算法的电力用户咨询文本分类算法的框架如图1所示。

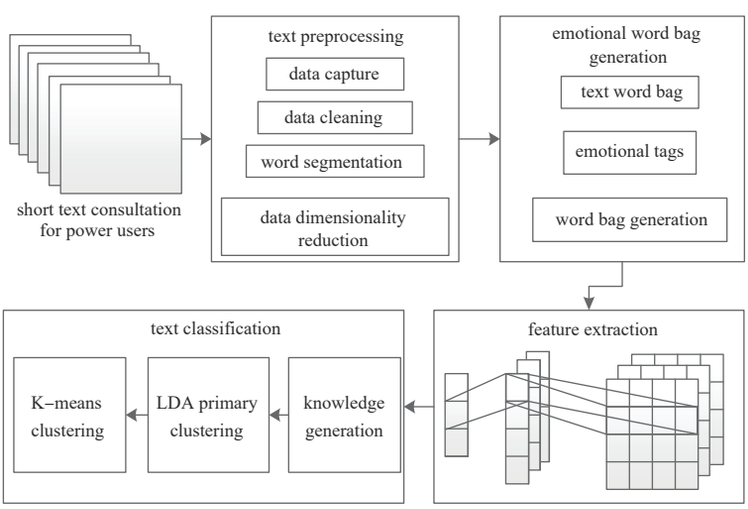

Fig.1 Power user consultation text classification algorithm based on improved LDA algorithm
图1 基于改进LDA算法的电力用户咨询文本分类算法

算法使用预处理的电力用户咨询短文本作为训练集，首先在LDA主题模型中进行训练，缩减维度以获得初始主题集；然后，构建基于情感极性标记的词共现袋，并通过特征处理获得主题专用词集；之后进一步提取相似的主题构造知识，获得一组主题关系词；再将主题特殊词和主题关系词的知识对注入Word2Vec卷积神经网络(Word2vec Convolutional Neural Network，WCNN)模型进行特征提取；接着，基于LDA模型对提取的特征进行初级聚类，生成Top30主题特殊词集。最后，利用K–means算法将Top30主题特殊词集作为K–means聚类的初始聚类中心，并计算主题特殊词的情感聚类结果。

2.1 文本预处理

对电力用户咨询短文本语料库进行预处理。首先使用爬虫对电力用户短数据进行抓取，同时，对抓取后的数据进行数据清洗，删除词干、词尾、文档频率低的词；然后，采用分词软件对中文短文本进行分词；最后，利用LDA主题模型进行数据降维处理。短文本语料库数据预处理的目的是对数据进行降维和去噪，并存储语料

库中每个词的初步知识。

2.2 情感词共现

经过短文本预处理后，基于情感词共现的词袋算法在电力用户咨询短文本词袋中添加词性标记，得到情感词袋，用于情感特征提取。通过提取主题特殊词集和主题关系词集执行从知识集的特征提取，同时，分析与主题相关的词性，从而确保提取有用的知识。基于情感词共现的词袋算法流程如下：

算法 1　情感词共现的词袋算法流程

输入：短文本词袋 S_{T}；

输出：情感词袋 $F(S_{\mathrm{T}})$；

//执行过程

1　初始化

2　for w in S_{T} do

3　if $w==a_{\mathrm{dj}}$ **or** $w==a_{\mathrm{dv}}$ **or** $w==(v,n_{\mathrm{noun}})$

4　根据式(2)和(3)判断极性

5 保存并更新 S_{T}

6 end if

7 end for

8　输出 $F(S_{\mathrm{T}})$

算法中，最重要的环节是判断词汇和极性。当输入的词性是形容词、动词、副词或动词–名词共现对时，将保存并更新情感词共现袋 $F(S_{\mathrm{T}})$。最终，算法输出电力用户咨询短文本的情感词袋。

2.3 特征提取

特征提取的主要功能是提取文本的最小信息，降低向量空间的维数，从而提高文本处理的速度和效率。本文基于改进的WCNN模型对电力短文本进行特征提取。

WCNN主要基于CNN模型构建。输入层中引入Word2Vec预训练情感词包。在卷积层，使用不同大小的多个卷积核并行学习文本特征，最终在输出层中生成文本特征。图2为WCNN模型结构。

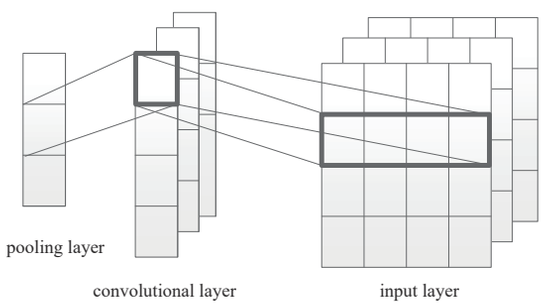

Fig.2 Structure of WCNN model

图2　WCNN模型的结构

2.3.1 输入层

为充分提取文本特征，在输入层使用Word2Vec训练每个情感词袋。特定的词嵌入通过指定相应的参数获得，包括词嵌入的维度、迭代次数和上下文窗口的大小(即每个窗口中的字符数)。每个出现多次的词都会扩展到 $m\times k$ 维，其中 m 为词袋中的词数，k 为训练期间指定的词嵌入维度。

假设电力文本情感词袋 x 包括 n 个词，则 x 表示为：

$$x_{1:n}=x_1\oplus x_2\oplus\cdots\oplus x_n \tag{8}$$

式中：$\oplus$ 为连接操作符；$x_i\left(i\in[1,n]\right)$ 为情感词袋中的词。

2.3.2 卷积层

使用具有不同大小的多重卷积核的并行卷积层学习电力文本特征；同时，设置多个卷积核全面获取情感词袋表达中的特征。卷积层包括3个尺度的卷积核，分别设置为：$h_1\times k$、$h_2\times k$、$h_3\times k$。其中，k 为整数且为词嵌入的维数，$h_i\left(i\in[1,3]\right)$ 为滑动窗口每次移动时滑动的词数。卷积核生成的特征 c_i，可根据式(9)计算：

$$c_i=f(w\bullet x_{i:i+h-1}+b) \tag{9}$$

式中：w 为卷积层的共享权重；$x_{i:i+h-1}$ 为词嵌入的连接，即来自情感词袋中任意 i+h–1 个词；b 为偏置；f 为一个非线性函数，本文选取ReLu函数。因此，有：

$$c_i=\max(0,w\bullet x_{i:i+h-1}+b) \tag{10}$$

当执行完卷积后，输出特征为：

$$\boldsymbol{C}=[c_1,c_2,\cdots,c_{n-h+1}] \tag{11}$$

2.3.3 池化层

池化层旨在从先前的特征图中提取最大值，表示最重要的信号。本文将非线性下采样的最大池化应用于特征图上的区域，并将最大值作为特征图输出的特征：

$$\hat{c}=\max(\boldsymbol{C}) \tag{12}$$

式中$\hat{c}$为经过池化层后的输出特征。

2.4 文本分类

将提取出的电力文本特征通过K–means算法(预设为K类)进行聚类。执行完特征提取后，文档之间的相似度问题转化为特征向量之间的相似度问题，本文基于余弦相似性实现特征之间的相似性度量。与距离度量相比，余弦相似度更关注2个向量之间的方向差异，而不是距离或长度。2个向量之间的角度越小，其相似性越高。令2个特征向量为$\boldsymbol{a}$和$\boldsymbol{b}$，其相似性为：

$$S(\boldsymbol{a},\boldsymbol{b})=\frac{x_1x_2+y_1y_2}{\sqrt{{x_1}^2+{y_1}^2}\sqrt{{x_2}^2+{y_2}^2}} \tag{13}$$

式中：(x_1,x_2)为向量$\boldsymbol{a}$的横坐标和纵坐标；(y_1,y_2)为向量$\boldsymbol{b}$的横坐标和纵坐标。

在相似性度量完成后实现知识对的提取。在知识对提取过程中，使用提取的主题特殊词和主题关系词生成每个主题的领域知识。生成的知识对由(A_i,B_i)组成，其中A_i表示主题特殊词集，B_i表示主题关系词集。知识对提取模型如图3所示，其中δ为一个超参数，Ψ为δ的概率分布，X_i为吉布斯采样，W_n为知识范围，V_n为采样范围。将通过主题特殊词和主题关系词计算的知识对带入LDA进行初级聚类，从而找到Top30重要词，即主题特殊词集和主题关系词集的前30个重要词集。最后，将Top30重要词带入K–means聚类，最终输出文本聚类结果。

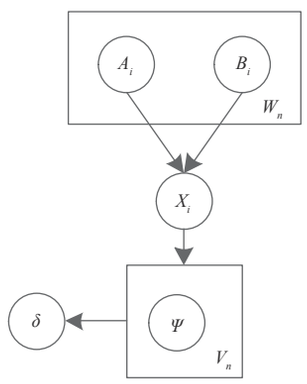

Fig.3 Knowledge pair extraction
图3 知识对提取

基于主题知识对的K–means聚类算法的执行过程如下：

算法2 主题知识对的K–means聚类算法的执行过程

输入：特征向量$\boldsymbol{T}_j=(A_1,A_2,\cdots,A_{30})$；聚类数$K$；最大迭代次数$T_{\max}$；迭代终止条件$\varepsilon$

输出：K主题特殊词聚类

//执行过程

1 根据式(13)对$\boldsymbol{T}_j$进行相似性度量

2 提取知识对

3 根据LDA获取Top30重要词

4 K主题特殊词聚类

5 计算与知识对中主题距离

6 计算每个类的标准度函数E

7 判断迭代条件是否满足

8 满足则输出；否则，执行步骤2

9 输出K主题特殊词聚类

算法中，标准度函数E定义为：

$$E=\sum_{n=1}^{k}\sum_{X\in C_n}|X-\bar{X}_n|^2 \tag{14}$$

式中：$\bar{X}_n$为聚类的中心主题；C_n为聚类的类别。

3 仿真与分析

3.1 仿真与分析

仿真所用数据集为中国某电力公司提供的电力用户咨询数据，数据类型包括：语音、短信信息、微博信息、调查报告、网站留言等。首先，对数据进行预处理：对文本数据，移除无用元素(如特殊符号)、分词、词性标

记、命名实体识别、虚假信息过滤；对语音数据，最终转化为文本信息。最终生成的文本数据集共包含11 606个数据，其中正面情绪和负面情绪分别为5 803个，按8:2分为训练集、测试集。

实验时仿真环境设置如下：硬件为Intel Core i9−7900X CPU 3.30 GHz、32 GB RAM和Ubuntu 18.04操作系统；算法由python3.7编写，并基于pytorch1.7搭建特征提取网络。

表1 电力情感词统计结果

Table1 Statistics of emotional words in electric power

	number of positive emotion keywords	number of negative emotion keywords
purchase electricity	15	14
package	11	9
electricity sharing	21	23
charge	16	14
security	10	11
electricity meter	15	14

3.2 实验设置

将预处理后的电力咨询文本数据带入特征提取网络进行训练，并提取特征向量。然后，应用基于主题知识对的K−means聚类算法对特征进行处理；最终输出K个聚类结果。

根据分析结果，主题词聚类可分为：购电、套餐、电力共享、收费、安全、电表等。其中，不同主题词包含不同的情感主题词，如购电中包含：方便、省事、满意等具有正面情感的词；同时，也存在部分负面情绪主题词，如太贵、退火、不切实际等。电力情感词统计结果如表1所示。

为验证所提模型有效性，选取k平均精确度(mP_k)和k平均召回率(mR_k)指标，分别与随机森林(Random Forest，RF)、支持向量机(Support Vector Machines，SVM)、LDA、递归神经网络(Recurrent Neural Network，RNN)、长短时记忆(Long Short Term Memory，LSTM)等模型进行对比，结果如表2所示。指标mP_k和mR_k计算如下：

$$mP_k=\frac{1}{q}\sum_{i=1}^{q}p_k \tag{15}$$

式中：q为查询数；p_k为前k个评估结果所占相关目标比例。

$$mR_k=\frac{1}{q}\sum_{i=1}^{q}R_k \tag{16}$$

式中R_k为前k个评估目标中发现相关目标的比例。

表2 不同模型在测试数据集上的平均预测结果

Table2 Average prediction results of different models on the test dataset

model	mP_{30}	mR_{30}
RF	0.792 1	0.053 5
SVM	0.680 5	0.324 4
BPNN	0.685 3	0.551 9
LDA	0.581 4	0.649 5
RNN	0.803 5	0.681 9
LSTM	0.852 8	0.780 7
proposed model	0.909 1	0.850 3

3.3 对比与分析

表2为不同模型在测试数据集上的平均评估结果。从表中可以看出，所提模型mP_{30}为90.91%，mR_{30}为85.03%，性能优异，说明所提模型可充分发挥多模型集成优势，有效提升模型性能。此外，RF和SVM的mR_{30}指标明显较低，表明这2个模型出现过拟合问题。原因是电力短文本复杂，呈现多模态、多维度的特点，传统机器学习模型和主流深度学习模型无法有效提取数据特征。

4 结论

本文对电力咨询短文本分类进行了研究与分析，设计了一种混合计算智能的电力咨询短文本分类模型。该模型可基于WCNN提取文本特征，并基于LDA和K−means实现文本分类。该模型为电力服务行业发展提供了一定的借鉴作用。

参考文献：

[1] 杨争林,曾丹,冯树海,等. 电力市场实验能力建设面临的挑战及关键技术[J]. 电力系统自动化, 2022,46(10):111−120. (YANG Zhenglin,ZENG Dan,FENG Shuhai,et al. Challenges and key technologies of experiment capability construction for electricity market[J]. Automation of Electric Power Systems, 2022,46(10):111−120.) doi:10.7500/AEPS20210820001.

[2] 向德军,周睿,黄志生,等. 基于混合云计算平台的电力市场交易平台关键技术的研究[J]. 山东农业大学学报(自然科学版), 2021,52(4):704−708. (XIANG Dejun,ZHOU Rui,HUANG Zhisheng,et al. Study on key technologies of electricity market trading platform based on hybrid cloud computing platform[J]. Journal of Shandong Agricultural University(Natural Science Edition), 2021,52(4):704−708.) doi:10.3969/j.issn.1000−2324.2021.04.031.

[3] 周戈,谢妮娜,潘宇晨. 物联网电力通信运维架构系统设计及关键技术[J]. 系统仿真技术, 2022,18(1):12−17. (ZHOU Ge,XIE

Ni'na, PAN Yuchen. Research on design of power communication operating maintenance architecture system and its key technology[J]. System Simulation Technology, 2022,18(1):12-17.) doi:10.16812/j.cnki.cn31-1945.2022.01.002.

[4] 文耀宽,王献军,王峻,等. 基于随机森林算法的电力计量大数据分析平台研究[J]. 计算机技术与发展, 2021,31(6):216-220. (WEN Yaokuan,WANG Xianjun,WANG Jun,et al. Research on big data analysis platform for electric power measurement based on random forest algorithm[J]. Computer Technology and Development, 2021,31(6):216-220.) doi:10.3969/j.issn.1673-629X.2021.06.038.

[5] 钟建栩,余少锋,廖崇阳,等. 基于云计算的电力设备智能监测系统[J]. 云南师范大学学报(自然科学版), 2022,42(3):37-41. (ZHONG Jianxu,YU Shaofeng,LIAO Chongyang,et al. Research on power equipment condition monitoring system based on cloud computing[J]. Journal of Yunan Normal University(Natural Science Edition), 2022, 42(3): 37-41.) doi: 10.7699/j. ynnu. ns-2022-034.

[6] 关菁华,刘鑫,刁建华. 基于词嵌入的微博谣言主题分类研究[J]. 软件导刊, 2019,18(4):1-3,8. (GUAN Jinghua,LIU Xin,DIAO Jianhua. Research on the topic classification of Weibo rumors based on word embedding[J]. Software Guide, 2019,18(4):1-3, 8.) doi:10.11907/rjdk.191169.

[7] 过云燕,李建中. 分布式潜在狄利克雷分配研究综述[J]. 智能计算机与应用, 2021,11(9):200-205. (GUO Yunyan,LI Jianzhong. A survey of distributed latent Dirichlet allocation[J]. Intelligent Computer and Applications, 2021,11(9):200-205.) doi:10.3969/j.issn.2095-2163.2021.09.040.

[8] 康雨萌,何玮,翟千惠,等. 基于主题提示的电力命名实体识别[J]. 计算机系统应用, 2022,31(9):272-279. (KANG Yumeng, HE Wei, ZHAI Qianhui, et al. Electric power named entity recognition based on topic prompt[J]. Computer Systems & Applications, 2022,31(9):272-279.) doi:10.15888/j.cnki.csa.008750.

[9] 许睿,龙丹,刘佳,等. 基于LDA模型的电力投诉文本热点话题识别[J]. 云南大学学报(自然科学版), 2020,42(S2):26-31. (XU Rui,LONG Dan,LIU Jia,et al. Identification of hot topics in power complaint text based on LDA model[J]. Journal of Yunnan University(Natural Sciences Edition), 2020,42(S2):26-31.)

[10] 刘德喜,聂建云,万常选,等. 基于分类的微博新情感词抽取方法和特征分析[J]. 计算机学报, 2018,41(7):1574-1597. (LIU Dexi,NIE Jianyun,WAN Changxuan,et al. A classification based sentiment words extracting method from microblogs and its feature engineering[J]. Chinese Journal of Computers, 2018,41(7):1574-1597.) doi:10.11897/SP.J.1016.2018.01574.

[11] 张书谙,王曦,代继鹏,等. 基于关键词共现网络的主题词提取算法[J]. 复杂系统与复杂性科学, 2023,20(1):74-80. (ZHANG Shu'an,WANG Xi,DAI Jipeng,et al. Subject words extraction algorithm based on keyword co-occurrence network[J]. Complex Systems and Complexity Science, 2023,20(1):74-80.) doi:10.13306/j.1672-3813.2023.01.010.

[12] 马瑛超,张晓滨. 基于主题关系的中文短文本图模型实体消歧[J]. 计算机工程与科学, 2023,45(1):154-162. (MA Yingchao, ZHANG Xiaobin. Entity disambiguation of Chinese short text using graph model based on topic relations[J]. Computer Engineering and Science, 2023,45(1):154-162.) doi:10.3969/j.issn.1007-130X.2023.01.018.

作者简介：

李竹青(1969-)，男，硕士，高级会计师，主要研究方向为审计管理、数字化审计.email:13722983081@163.com.

侯本忠(1976-)，男，本科，高级工程师，主要研究方向为数字化审计.

曹培祥(1984-)，男，本科，高级经济师，主要研究方向为数字化审计、营销审计.

王一蓉(1979-)，女，博士，教授级高级工程师，主要研究方向为数字信号处理、电力信息通信.

李向阳(1991-)，男，本科，工程师，主要研究方向为审计管理、企业数字化转型.