文章编号:2096-1472(2024)01-0068-06
DOI:10.19644/j.cnki.issn2096-1472.2024.001.016

基于 BERT-LDA 和 K-means 聚类的绘画作品价值评估指标体系构建

李天义，刘勤明

(上海理工大学管理学院，上海 200093)
✉ 18437773865@163.com; lqm0531@163.com

摘　要:针对目前绘画领域缺乏标准的价值评估指标体系，提出了基于 BERT-LDA 和 K-means 聚类的绘画作品价值要素挖掘方法。运用超平面法对绘画文献进行了停用词筛选，基于 BERT-LDA 模型构建了包含文本语义信息的融合特征向量，运用 K-means 算法对融合特征向量进行降维可视化，随之构建了绘画作品价值评估指标体系。结果表明，基于 BERT-LDA 模型和 K-means 算法识别的主题及主题词相比传统 LDA 模型的查准率、查全率和 *F* 值分别提升了 28.5%、10%和 21.5%。通过随机森林等算法对指标体系进行验证，验证了构建的绘画作品价值评估指标体系的科学性。

关键词:BERT-LDA;融合特征向量;K-means 聚类;绘画;指标体系
中图分类号:TP18　　**文献标志码**:A

Construction of a Painting Works Valuation Index System Based on BERT-LDA and K-means Clustering

LI Tianyi, LIU Qinming

(*Business School, University of Shanghai for Science and Technology, Shanghai* 200093, *China*)
✉ 18437773865@163.com; lqm0531@163.com

Abstract: This paper proposes a method for mining the value elements of painting works based on BERT-LDA ((Bidirectional Encoder Representations from Transformers-Latent Dirichlet Allocation) and K-means clustering in view of the lack of standard value evaluation index system in the current painting field. The hyperplane method is applied to filter out stop words in painting literature, fusion feature vectors containing textual semantic information are constructed by using the BERT-LDA model, and the K-means algorithm is employed for dimensionality reduction and visualization of the fusion feature vectors. Subsequently, a valuation index system for painting works is established. The results indicate that, compared to the traditional LDA model, the precision, recall, and F-value of the themes and keywords identified by the BERT-LDA model and K-means algorithm increase by 28.5%, 10%, and 21.5%, respectively. The index system is validated by algorithm such as Random forest, verifying the scientific nature of the constructed valuation index system for painting works.

Key words: BERT-LDA; fusion feature vectors; K-means clustering; painting; index system

0　引言(Introduction)

绘画作品价值评估作为一种绘画作品价值衡量方法，可以推动高质量绘画作品的产生和交易。但目前在绘画领域仍存在价值判定标准模糊且复杂、没有标准的价值评估体系等问题，导致对绘画作品的价值评估不准确[1]。因此，如何构建一套标准的、可解释的绘画作品价值评估体系是对绘画作品价值进行准确评估的前提。

传统的绘画作品价值评估方法依赖专家的知识、经验和分析判断能力，普遍存在主观性强等问题[2]。解欣[3]以国画艺术品为例，运用经验法分析了国画艺术品的价值构成，然后构建了包含文化价值、审美价值、艺术价值、艺术家影响力、历史价值、社会认可度六个维度的指标体系。数据挖掘为指标体系构建提供了新的思路。庄穆妮等[4]将双向编码器表征法(Bidirectional Encoder Representation from Transformers，BERT)和隐含狄利克雷分布(Latent Dirichlet Allocation，LDA)主题模型融合，构建了优化主题向量帮助文本主题聚类。刘晋霞等[5]运用 LDA 主题模型，从 2000～2020 年的制氢领域的期刊文献数据中抽取主题，然后从两个维度构建了主题识别的指标体系。智能化构

收稿日期:2023-05-15

建方法可以从大量异构文本数据中挖掘关键的语义信息，具有普遍适用性。

本文采用 BERT-LDA 模型与 K-means 聚类提取绘画价值要素主题及主题词，结果表明，该方法能够提高主题的查准率、查全率和 F 值，构建的指标体系科学、合理，可应用于绘画作品价值的进一步研究。

1　方法框架(Method framework)

1.1　研究思路

本文的研究思路如下。

(1)数据获取与预处理。运用基于辅助集的超平面法构建绘画领域停用词集。将获取的与绘画、书法、电影、医学、科技等价值相关的文献摘要的数据划分为目标集和辅助集，通过计算词语与构建的超平面之间的距离，构建绘画领域的特有停用词集。对数据进行分词及词频统计。将绘画领域停用词集与哈工大停用词表合并为本文的停用词集，并对原始数据进行过滤处理，形成结构化数据[6]。为了提高模型的主题提取精确度，对 BERT 提取的词向量、文本向量、位置向量与 LDA 主题特征向量进行特征融合，形成融合特征向量。

(2)K-means 聚类及其可视化。通过 K-means 聚类算法将语义和主题相似的词语分配到相同的类别。基于困惑度选取 K-means 算法的 K 值为 10 个，运用统一流形逼近和投影(Uniform Manifold Approximation and Projection，UMAP)算法实现聚类效果的降维可视化，并构建各个主题的词云图，依据聚类结果构建绘画作品价值评估指标体系。

(3)绘画作品价值评估指标体系的验证。通过雅昌艺术网获取部分绘画作品的交易数据，并对相关指标进行修补，运用随机森林、XGBoost 等机器学习算法对构建的指标体系进行对比验证，选取最适合的模型运用到实际绘画作品价值评估中。

本文的研究框架如图 1 所示。

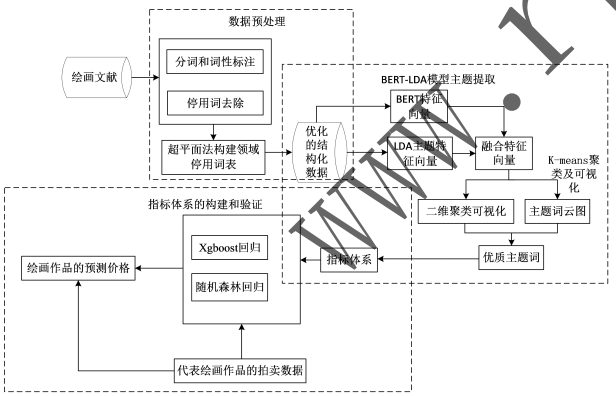


图 1　基于 BERT-LDA 和 K-means 的绘画作品价值评估指标体系构建及验证整体框架

Fig. 1 The overall framework for constructing and validating a painting work valuation index system based on BERT-LDA and K-means

1.2　研究方法

1.2.1 Ncoders 构建的语言预训练模型[7]

BERT 引入遮蔽语言模型(Masked Language Model，MLM)，通过将输入文本序列中的 15% 的单词做 MASK(掩膜)操作，此外为了使模型具备回答问题和语言推理能力，BERT 还引入句子预测(Next Sentence Prediction，NSP)任务增强模型的语义表示能力，添加了前缀符号[CLS]和分隔符标记[SEP]对句子进行分割。BERT 的句子级表示如图 2 所示。BERT 的输入由三个部分组成：即 Token Embedding、Segment Embedding 和 Position Embedding。

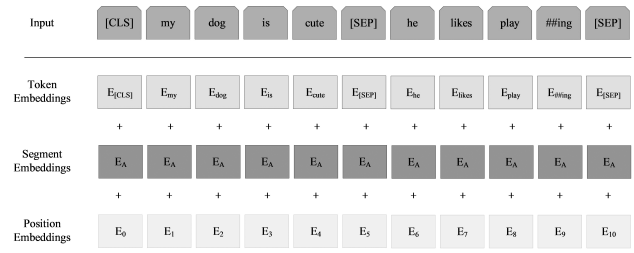


图 2　BERT 的句子级表示

Fig. 2 BERT sentence-level representation

1.2.2 LDA 主题特征向量构建

LDA[8]是一种包含词、主题和文档的三层贝叶斯概率模型，用于对文本文档进行主题建模，它可以从一组文档中自动地发现隐藏在其中的主题，并确定每个文档中每个主题的相对权重，它的工作原理是先将每个单词表示为一个向量，然后通过一系列迭代计算找到每个单词与每个主题的关联度。在此过程中，每个文档都被表示为一个主题分布向量，每个主题都被表示为一个单词分布向量。最终，LDA 主题模型可以输出每个主题的单词分布以及每个文档的主题分布。其中，对词向量进行训练，计算公式如下：

$$\boldsymbol{x}=\{V(w_1),V(w_2),\cdots,V(w_t)\} \tag{1}$$

其中：$\boldsymbol{x}$ 表示所有词汇的词向量表示形式，$V(w_t)$中的 w_t 表示为该词汇的词向量表示方式，t 代表词汇总数。

本文使用 Python 中 Gensim 包的 LDA 模型构建文本的 LDA 主题特征向量。根据模型训练的结果，取每个文档 d_i 中概率最大的主题 $z_{\max}$，然后再从主题 $z_{\max}$ 中选择前 n 个词(t_1，t_2，…，t_n)和其对应的概率值(p_1，p_2，…，p_n)，将对应的概率值做归一化处理，并将其作为 n 个词语的权重，计算公式如下：

$$q_i=\frac{p_i}{\sum_{a=1}^{n}p_a} \tag{2}$$

其中：(q_1，q_2，…，q_n)分别表示(p_1，p_2，…，p_n)进行归一化处理后的值，代表前 n 个词的权重大小。对词向量($C(t_1)$，$C(t_2)$，…，$C(t_n)$)进行加权求和，得到 LDA 主题向量 $\boldsymbol{d}_z=\sum_{b=1}^{n}q_b\times C(t_b)$。

1.2.3 BERT-LDA 特征向量融合

由于传统 LDA 采用的是词袋法，其原理是将每一篇文档视为一个词频向量，并以此为依据将文本信息处理为数字信息，但词袋法没有考虑到词与词之间的上下文关系，这会导致单词产生歧义、表达能力缺失等问题，因此本文将 BERT 模型的语义特征向量与 LDA 主题特征向量进行融合，形成包括语义信息的融合特征向量再进行主题抽取。

将数据预处理后的结构化数据输入 *BERT* 模型进行词嵌入，构建 BERT 的语义特征向量 $\boldsymbol{d}_m=w_{ij}(\omega+\delta+\rho)$，将 BERT 构建的语义特征向量与 LDA 主题特征向量 $\boldsymbol{\mu}$ 进行向量拼接，形成包含语义特征信息及词义信息的融合特征向量，定义为 $\boldsymbol{d}'_m$，计算公式如下：

$$\boldsymbol{d}'_m=\{\boldsymbol{\mu};\boldsymbol{d}_m\} \tag{3}$$

其中：$\boldsymbol{d}'_m$ 代表融合了 BERT 语义特征向量和 LDA 主题特征向

量的文本向量化表示，“;”为向量拼接符号。BERT-LDA 模型示意图如图 3 所示。

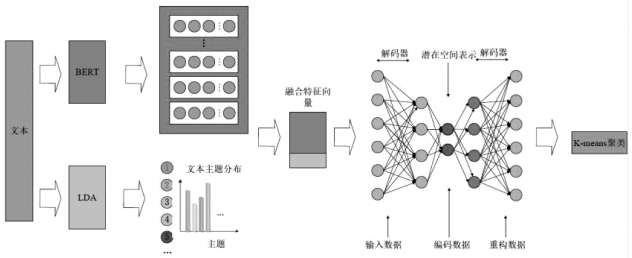


图 3 BERT-LDA 模型示意图

Fig. 3 BERT-LDA model diagram

1.2.4 K-means 聚类及其可视化

将 BERT 词向量与 LDA 主题特征向量拼接的结果称为融合特征向量。融合特征向量既包含了词汇之间的语义信息，还包含了词汇在不同主题下的概率分布。但是，向量拼接后的维度升高了，而且还造成了高维空间的信息稀疏问题，因此本研究引入了 K-means 聚类算法，以便实现将语义和主题相似的词语分配到相同的类别的目标。K-means 算法是一种常见且高效的聚类分析算法，可以根据不同的 K 值点到其所属的类别的距离平方和大小估计聚类的质量，并且只需要输入一个参数类别数 K[9]。本文选用计算困惑度的方法确定最优主题数，并以此作为 K-means 算法的 K 值。运用 UMAP 算法实现各个主题下的主题词的降维可视化，将概率排名为前 10 的主题词及其对应概率进行输出，并以此为基础构建绘画作品价值评估指标体系。

2 实验设计(Experimental design)

2.1 数据收集

本文选择中国知网平台、雅昌艺术网作为本次研究数据的来源。利用 Python 语言的网络爬虫功能分别对中国知网平台(https://www.cnki.net/)中有关绘画、书法、电影、医学、科技等价值的相关文献和雅昌艺术网中的绘画作品交易数据进行采集，在中国知网采集到各类文献的题目、摘要、关键词等信息和 2 233 篇各类期刊文献的相关信息，经过去重及人工筛选后，剩下 1 998 篇文献的信息作为本研究的数据集；在雅昌艺术网采集了绘画作品的作家、售卖价格、作评尺寸等信息，经过筛选后将 187 幅国画的交易数据作为验证绘画作品价值评估指标体系的数据集。

2.2 数据预处理

由于绘画领域的文献内容具有发散性，直接采用普通的通用停用词对绘画作品文献数据进行去噪声处理所得到的结果中仍包含大量与绘画作品价值无关的关键词，因此需要构建领域停用词表并剔除领域停用词[10]。ALSHANIK 等[11]提出了一种基于辅助集构建超平面的方法自动获取数据集中的部分或全部领域停用词，该方法利用单词与超平面的距离大小进行领域停用词的选取。通过计算单词与超平面的距离，可以获得不同类型的停用词，单词与超平面的距离越小，该单词是领域停用词的概率就越大。这种方法不仅可以提高数据处理效率，而且能够更好地实现文本分类等任务。

本文首先收集了与绘画价值相关的文献作品作为目标集(A 类)，还采集了书法、电影、医学、科技等数据集(B 类)，并以此构建与绘画价值相关度高的辅助集 1 及与绘画价值相关度低的辅助集 2。构建 A、B 两类数据的超平面。其中，词语与构建的超平面之间的距离可以通过公式(4)至公式(8)计算：

$$wx+b=0 \tag{4}$$

$$w=Center(A)-Center(B) \tag{5}$$

$$b=-wx_0 \tag{6}$$

$$x_0=\frac{w}{2} \tag{7}$$

$$d=\frac{|wx_i+b|}{\|w\|} \tag{8}$$

其中：w 为两类文献向量化 x 的质心之差，b 为平面上词向量与平面斜率的偏移量，x_0 代表两类文献质心连线的中间点，x_i 表示任意单词的向量。

本文采集数据的时间跨度均为 2000 年 1 月 1 日至 2023 年 1 月 1 日，针对各类文献采集的数量均为 1 998 篇，将各个类别文献的摘要作为文本进行价值要素挖掘。目标集和辅助集数据分布如表 1 所示。

表 1 目标集和辅助集数据分布

Tab.1 Distribution of target set and auxiliary set data

数据集类型	类别	文本量/篇
目标集	绘画	1 998
辅助集 1	书法	1 998
	电影	1 998
辅助集 2	医学	1 998
	科技	1 998

对表 1 数据集中常见的通用停用词去除、分词及筛选词性，并选取词长度不小于 2 的名词作为绘画作品价值要素词。因为数据集是期刊类文献，因此哈工大停用词表更适合此类文献的停用词过滤处理[12]。通过对数据进行预处理，得出了 4 396 个非重复的词语。经过多次测试分析，首先基于辅助集 1 构建的超平面将距离最近的 89 个词作为领域停用词，其次对绘画作品价值文献进行过滤处理，得到最终优化后的结构化数据。

2.3 BERT-LDA 模型提取结果分析

本研究基于 Google 的 BERT 基本预训练模型，利用绘画领域的期刊文献摘要等语料库对预训练的 BERT 模型进行微调，并构建了基于 BERT 的语义特征向量，将其和 LDA 的主题特征向量进行拼接，然后对融合特征向量进行聚类生成绘画领域的主题。选用基于困惑度的方法来估计与绘画作品价值相关的主题数量。主题困惑度是一种常用的用于确定主题模型最优数的方法，其计算方法如公式(9)所示：

$$Perplexity=e^{\frac{-\sum\log(p(w))}{N}} \tag{9}$$

其中：$p(w)$代表每个词汇在测试集的概率，N 代表测试集总长度。

运用 BERT-LDA 主题模型选取不同的主题数的困惑度变化曲线如图 4 所示，从图 4 中可以得出，主题数越大，困惑度就越小，在主题数为 10 时，困惑度出现了比较明显的拐点，在主题数小于 10 之前，困惑度下降趋势明显，在主题数大于 10 之后，困惑度的变化幅度逐渐变小并慢慢趋于稳定状态。虽然主题数越大，困惑度就越小，但是随着主题数增大，提取的主题噪声也会随之增多。因此，本研究将模型最终主题数选定为 10 个。

数据降维可视化工具 UMAP 算法可以将 BERT-LDA 模型提取出的关键价值相关主题进行可视化，聚类结果如图 5 所示。

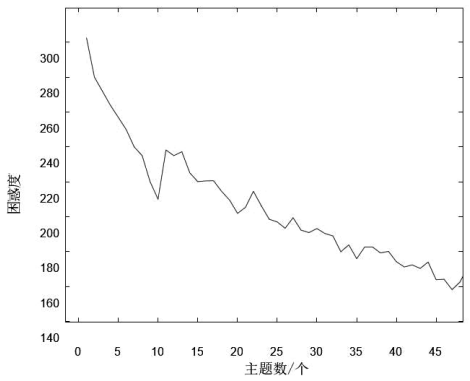


图 4　不同主题数下 BERT-LDA 模型的困惑度变化曲线

Fig. 4 Perplexity variation curve of the BERT-LDA model under different numbers of themes

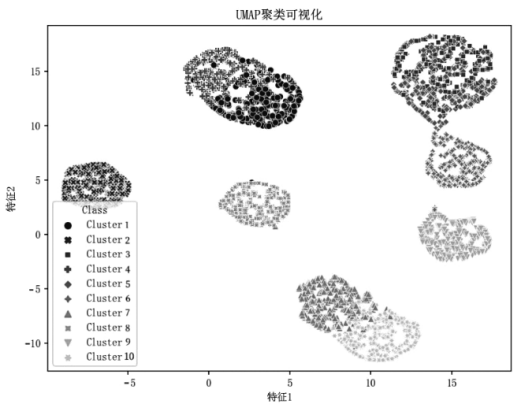
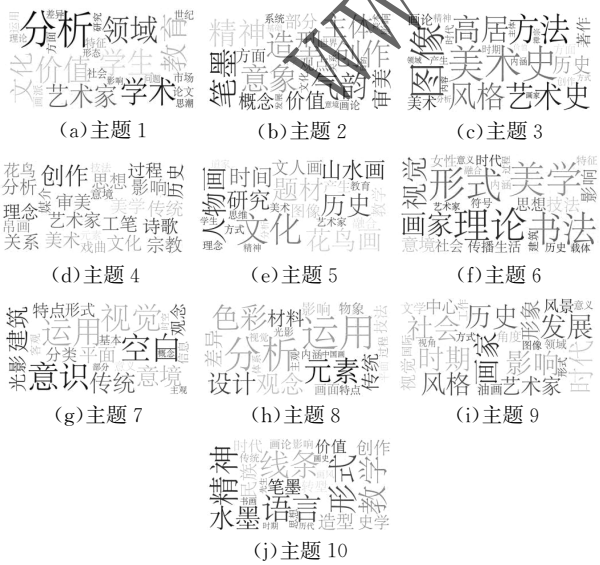


图 5　基于 UMAP 的二维聚类可视化聚类结果

Fig. 5 Visual clustering results of two-dimensional clustering based on UMAP

在 BERT-LDA 模型识别出的与绘画作品价值相关的 10 个主题的基础上，选择每个主题的前 25 个主题词进行可视化分析，以构建 10 个主题的词云图，绘画作品价值要素主题词云图如图 6 所示。



(a)主题 1　(b)主题 2　(c)主题 3

(d)主题 4　(e)主题 5　(f)主题 6

(g)主题 7　(h)主题 8　(i)主题 9

(j)主题 10

图 6　绘画作品价值要素主题词云图

Fig. 6 Word cloud for value-related elements in paintings

基于 BERT-LDA 模型提取出的与绘画作品价值相关的 10 个主题是由绘画作品价值研究领域内出现概率较高的特征词构成的集合。

基于困惑度的 BERT-LDA 主题提取过程如下：首先基于困惑度确定最优主题数的 BERT-LDA 模型，然后从所选绘画文献数据集中抽取 10 个与价值相关的主题及主题词，最后通过人工选取的方式选出与绘画作品价值相关主题词对应概率最大的前 10 个主题词。基于困惑度 BERT-LDA 模型提取的前 10 个主题词如表 2 所示。

表 2　基于困惑度 BERT-LDA 模型提取的前 10 个主题词

Tab.2 Top 10 keywords extracted from the BERT-LDA model based on perplexity

序号	主题	主题词
1	作品教育教学	研究、理论、分析、问题、影响、学术、方面、文人、价值、领域
2	作品精神理念	意象、气韵、创作、造型、精神、笔墨、主体、价值、审美、概念
3	作品的经典性	美术史、图像、方法、艺术史、高居、风格、历史、美术、著作、画论
4	创作者影响力	艺术家、影响、思想、文化、美学、审美、创作、美术、理念、传统
5	作品类型	题材、文化、历史、花鸟画、人物画、山水画、时间、研究、文人画、图像
6	作品创作风格	美学、形式、理论、书法、画家、视觉、意境、技法、影响、思想
7	作品创作内涵	意识、意境、传统、建筑、平面、空白、光影、运用、视觉、特点
8	作品设计布局	色彩、元素、设计、观念、材料、分析、传统、差异、运用、影响
9	作品创作时代环境	画家、时期、影响、发展、历史、社会、时代、风格、艺术家、形象
10	作品技法表现	线条、语言、形式、精神、教学、水墨、造型、民族、创作、笔墨

对表 2 中的结果进行分析可知，BERT-LDA 主题模型提取的有效主题为 10 个，10 个主题类型分别为作品教育教学、作品精神理念、作品的经典性、创作者影响力、作品类型、作品创作风格、作品创作内涵、作品设计布局、作品创作时代环境、作品技法表现。其中：作品教育教学、作品精神理念、创作者影响力、作品创作风格、作品创作内涵、作品创作时代环境、作品技法表现、作品设计布局、作品类型 9 个主题与专家总结出的绘画作品的主题数相同，剩余 1 个主题与绘画作品的价值联系较低，故判定其为无效主题。

2.4　主题提取效果对比

为了验证基于 BERT-LDA 模型的有效性，本文将基于 BERT-LDA 主题向量融合的特征向量的主题提取结果与传统的 LDA 模型主题提取结果进行对比，并分别对比了基于共词分析法和困惑度选取的不同最优主题数对结果的影响。选取查准率、查全率和 F 值对提取结果进行评价，各指标的计算公式如公式(10)至公式(12)所示[13]：

$$P=\frac{T_{\text{correct}}}{T_{\text{extract}}} \tag{10}$$

$$R=\frac{T_{\text{correct}}}{T_{\text{standard}}} \tag{11}$$

$$F=\frac{2PR}{P+R} \tag{12}$$

其中：P 为查准率，R 为查全率，T_{extract} 代表提取到的主题中有效的主题数量；T_{correct} 代表提取到的主题中包含专家总结主题的数目；T_{standard} 代表专家总结出的主题数。BERT-LDA 和传统 LDA 两种模型的主题提取效果对比结果如表 3 所示。

表3 不同模型主题抽取效果对比结果

Tab.3 Comparison result of theme extraction effectiveness in different models

模型	主题数/个	T_{extract}/个	T_{correct}/个	T_{standard}/个	查准率/%	查全率/%	F 值/%
BERT-LDA	10	10	9	10	90.0	90	90.0
传统 LDA	13	13	8	10	61.5	80	69.5

对表 3 中不同模型识别的结果进行分析可知，基于 BERT-LDA 主题模型在准确率、查全率和 F 值均比传统的 LDA 模型优秀，准确率、查全率和 F 值分别提升了 28.5%、10% 和 21.5%。BERT-LDA 模型提取的主题类别更接近专家总结的 10 个主题，识别出的主题类别更加清晰、准确，有利于准确地构建绘画作品价值评价指标体系。

3 绘画作品价值评估指标体系构建及验证(Construction and verification of the painting works valuation index system)

3.1 绘画作品价值评估指标体系构建

基于 BERT-LDA 模型和 K-means 聚类对作品教育教学、作品精神理念、创作者影响力、作品创作风格、作品创作内涵、作品创作时代环境、作品技法表现、作品设计布局、作品类型等 9 个主题及各个主题所包含的词语进行了识别，最终归纳总结得到绘画价值评估指标体系如表 4 所示。

表4 绘画作品价值评估指标体系

Tab.4 Valuation index system for painting works

一级指标	二级指标	三级指标
文化价值	艺术价值	作品的艺术内涵 作品的技法水平体现
	精神价值	作品是否对教育是否有影响 作品对精神文明的贡献度
社会价值	创作者属性	创作者是否为名家 创作者拍卖作品的数量 创作者作品最高成交额
	作品属性	作品材质 创作年代 尺寸大小 作品形制
经济价值	作品收益	作品展览权收益 作品拍卖成交额 作品复制权收益

3.2 绘画作品价值评估指标体系的验证

为了验证基于 BERT-LDA 模型构建的绘画作品价值评估指标体系的通用性和实用性，本研究从雅昌美术网收集了 187 幅潘天寿和张大千两位国画大师的画作，剔除不合格信息后，得到 12 个属性，这些属性包括作者、创作年代、尺寸、作品风格、作者名家地位、作品稀有性、艺术家及作品获奖、作品收藏者的知名度、创作者所有拍卖作品中最高的成交额、创作者总拍品的数量以及绘画作品最终的成交价。

将上述属性的数据引入哑变量等进行量化处理，得到可量化的指标数据。采用 Xgboost 回归算法和随机森林回归算法对数据进行训练，并预测绘画作品的成交价格为 p_1，将作品交易实际价格记为 p_2，则可以采用公式(13)计算模型预测的准确率，不同模型的准确率如表 5 所示。

$$Accuracy=1-\frac{\sum_{i=1}^{n}\frac{|p_1[i]-p_2[i]|}{p_2[i]}}{n} \tag{13}$$

公式(13)中的 $p_1[i]$和 $p_2[i]$分别表示第 i 幅绘画作品的预测价格和真实价格。

表5 不同模型的准确率

Tab.5 Accuracy of different models

训练集和测试集的比例	Xgboost 准确率/%	随机森林准确率/%
7∶3	67	69
8∶2	73	78
9∶1	79	81

从表 5 中得到的结果可以看出，随机森林回归模型的训练效果最好，该模型的预测精度可达 81%，把误差降低至 20%以内，在实际交易过程中的误差可以接受。后续可以通过扩大样本数据量进一步提升模型预测准确率。

4 结论(Conclusion)

本文提出的基于 BERT-LDA 模型和 K-means 聚类的要素挖掘的指标体系构建方法，能够从大规模的绘画作品价值评估文献中挖掘与绘画作品价值相关的要素，成功建立了绘画作品价值评估指标体系。本文基于辅助集的超平面法构建了绘画领域停用词表，经过去除停用词处理后，将 BERT 的词向量与 LDA 主题向量拼接形成特征融合向量。本文运用 K-means 算法将语义与主题相似的词语分配到相同类别，然后引入 UMAP 算法进行降维可视化，并依据 BERT-LDA 模型的抽取结果构建了绘画作品价值评估指标体系。本文开展的评估模型性能的实验证明，基于 BERT-LDA 模型提取的主题更准确、全面，相比传统 LDA 模型，查准率提高了 28.5%，查全率提高了 10%，F 值也提升了 21.5%。本研究通过雅昌艺术网的部分交易数据对指标体系的准确性进行了验证，结果表明随着训练集和测试集的比例提高，随机森林回归算法的准确率也随之提升，准确率达到了 81%，证明本研究提出的绘画作品价值评估指标体系具备可用性。

参考文献(References)

[1] 刘翔宇. 中国当代艺术品交易机制研究[D]. 济南：山东大学，2012.

[2] 冯坤，杨强，常馨怡，等. 基于在线评论和随机占优准则的生鲜电商顾客满意度测评[J]. 中国管理科学，2021，29(2)：205-216.

[3] 解欣. 艺术品价值评估和价格影响因素研究[D]. 长沙：湖南大学，2017.

[4] 庄穆妮,李勇,谭旭,等. 基于 BERT-LDA 模型的新冠肺炎疫情网络舆情演化仿真[J]. 系统仿真学报,2021,33(1):24-36.
[5] 刘晋霞,侯倩倩. 热点主题特征维度的识别指标体系构建及实证研究:以我国制氢领域为例[J]. 情报杂志,2022,41(9):150-158.
[6] TAN X,ZHUANG M N,LU X,et al. An analysis of the emotional evolution of large-scale Internet public opinion events based on the BERT-LDA hybrid model[J]. IEEE access,2021,9:15860-15871.
[7] ALAMMARY A S. BERT models for Arabic text classification: a systematic review[J]. Applied sciences,2022,12(11):5720.
[8] 徐红,张斯婷,李凌方. 基于 LDA 模型与共词分析法的农村阅读推广主题发现与热点分析[J]. 情报科学,2022,40(10):67-73.
[9] 吕鲲,项旻昊,靖继鹏. 基于 LDA2Vec 和 DTM 模型的颠覆性技术主题识别研究:以能源科技领域为例[J]. 图书情报工作,2023,67(12):89-102.
[10] 俞琰,赵乃瑄. 基于辅助集的专利主题分析领域停用词选取[J]. 数据分析与知识发现,2018,2(11):95-103.
[11] ALSHANIK F,APON A,HERZOG A,et al. Accelerating text mining using domain-specific stop word lists[C]//IEEE. Proceedings of the 2020 IEEE International Conference on Big Data. Piscataway:IEEE,2020:2639-2648.
[12] 郝秀慧,方贤进,杨高明. 基于 TFIDF+LSA 算法的新闻文本聚类与可视化[J]. 计算机技术与发展,2022,32(7):34-38,45.
[13] 关鹏,王曰芬. 科技情报分析中 LDA 主题模型最优主题数确定方法研究[J]. 现代图书情报技术,2016(9):42-50.

作者简介:

李天义(2000-),男,硕士生。研究领域:艺术品价值评估。
刘勤明(1984-),男,博士,教授。研究领域:人工智能,维护调度。

(上接第 57 页)

从 ResNet 系列网络的表现来看,随着网络层数的增加,使用分割前后图像训练的模型之间的识别准确率差异逐渐减小。这可能是因为随着网络层数的不断加深,模型参数量越来越大,而分割图像后减少了图像中包含的信息量,因此使用分割后图像进行训练的模型识别准确率的提升逐步变小,而关于 Vision-Transformer 模型使用分割后的图像训练反而降低其识别准确率的原因,可能与此类似。

4 结论(Conclusion)

本文提出了一种先分割、后识别的西瓜叶片病害识别算法,经过一系列的调优和测试后,使用分割后的叶片图像进行病虫害分类的识别准确率达到 92.9%,能够满足瓜农防治病虫害的基本需求。但是,本研究还存在一些不足,需要进一步改进与优化。

首先,实验中使用的数据集图像虽然是在自然环境下拍摄的,但未采用不同的相机拍摄,也没有设置不同的分辨率,图像大小固定不变,导致图像的多样性不足,鲁棒性有待加强。其次,每个样本中仅包含一张病虫害的叶片和只包含一种病虫害。因此,在复杂环境下的病虫害识别还需进一步验证。

针对以上不足,提出四项改进措施:(1)使用不同相机,并设置不同分辨率,以增加图像的多样性;(2)收集包含多种病虫害的叶片样本,以便在复杂环境下验证算法的识别能力;(3)增加在不同光线和角度下拍摄的图像,以扩充数据集,提高算法的泛化能力;(4)研究更多类型的神经网络,通过调参和优化,进一步提高识别准确率,并实现实时识别。

参考文献(References)

[1] 苏婷婷,牟少敏,董萌萍,等. 深度迁移学习在花生叶部病害图像识别中的应用[J]. 山东农业大学学报(自然科学版),2019,50(5):865-869.
[2] 王娜,王克如,谢瑞芝,等. 基于 Fisher 判别分析的玉米叶部病害图像识别[J]. 中国农业科学,2009,42(11):3836-3842.
[3] 张建华,孔繁涛,吴建寨,等. 基于改进 VGG 卷积神经网络的棉花病害识别模型[J]. 中国农业大学学报,2018,23(11):161-171.
[4] HU X G,YANG H G. DRU-net:a novel U-net for biomedical image segmentation[J]. IET image processing,2020,14:192-200.
[5] 蒋雪源,陈青梅,黄初华. 基于动态遍历的分层特征网络视觉定位[J]. 计算机工程,2021,47(9):197-202.
[6] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[DB/OL]. (2017-12-06)[2023-04-13]. https://arxiv.org/abs/1706.03762v5.
[7] 徐光宪,冯春,马飞. 基于 UNet 的医学图像分割综述[J]. 计算机科学与探索,2023,17(8):1776-1792.
[8] LI Y Y,WANG Z Y,YIN L,et al. X-Net:a dual encoding-decoding method in medical image segmentation[J]. The visual computer,2023,39:2223-2233.
[9] 汪健,梁兴建,雷刚. 基于深度残差网络与迁移学习的水稻虫害图像识别[J]. 中国农机化学报,2023,44(9):198-204.
[10] TOUVRON H,BOJANOWSKI P,CARON M,et al. ResMLP:feedforward networks for image classification with data-efficient training[J]. IEEE transactions on pattern analysis and machine intelligence,2023,45(4):5314-5321.

作者简介:

向宇杰(2002-),男,本科生。研究领域:人工智能。
向元平(1981-),女,硕士,讲师。研究领域:图像处理,模式识别。