DOI:10.20033/j.1003-7241.(2025)06-0067-06

信息熵加权下的图书馆数字资源完整性自动化评价方法

任阳红

（杨凌职业技术学院，陕西 杨凌 712100）

摘 要：随着数字化时代的到来，图书馆数字资源的管理和维护变得越来越重要，然而传统的评价方法往往无法准确地衡量图书馆数字资源的完整性。研究提出一种融合k-prototypes算法和信息熵加权的图书馆数字资源完整性自动化评价方法。该方法综合考虑了数据属性的类型差异、属性之间的关联性和重要性。经实验表明，研究算法聚类纯度超过0.9；图书馆目录数据集内部样本的离散程度最大，图书馆的数据覆盖率的平均分值为8.4，比数据准确性和数据更新性的平均分值分别高0.21和3.9，为图书馆数字资源的管理和维护提供了一种有效的工具和方法。

关键词：信息熵加权；图书馆数字资源；自动化评价；聚类分析

中图分类号：TP18 文献标识码：A 文章编号：1003-7241(2025)06-0067-06

Automatic Evaluation Method for the Integrity of Library Digital Resources Under Information Entropy Weighting

REN Yang-hong

(Yangling Vocational & Technical College, Yangling 712100, China)

Abstract: With the advent of the digital age, the management and maintenance of library digital resources are becoming more and more important. However, traditional evaluation methods are often unable to accurately measure the integrity of library digital resources. Therefore, an automated evaluation method for the integrity of library digital resources is proposed in this study, which integrates k-prototypes algorithm and information entropy weighting. The method comprehensively considers the type difference of data attributes, the relevance and importance among attributes. The experimental results show that the clustering purity of the proposed algorithm exceeds 0.9. The sample within the library catalog dataset has the largest degree of dispersion, and the average score of the library's data coverage is 8.4, which is 0.21 and 3.9 higher than the average score of data accuracy and data updating, respectively. To sum up, this study provides an effective tool and method for the management and maintenance of library digital resources.

Keywords: information entropy weighting; library digital resources; automatic evaluation; cluster analysis

0 引言

随着数字化时代的到来，图书馆数字资源的数量和种类不断增加，对图书馆数字资源的管理和维护提出了新的挑战[1]。其中，评估图书馆数字资源的完整性是一个重要的任务。完整性指的是图书馆数字资源数据集中缺失或不完整信息的程度。国内学者李琳采取熵加权聚类算法对图书馆数据进行分类，实验结果表明所提出的聚类方法可以提高处理书籍的精度和稳定度[2]。国外学者Ma团队将最佳-最坏方法和熵方法相结合，构建了一个图书馆应用性能评价标准体系，经实验表明该方法可以准确检测图书馆存在的问题[3]。了解图书馆数字资源的完整性可以帮助图书馆管理者更好地了解和维护数字资源，提供更好的服务[4]。然而，由于数字资源的多样性和复杂性，传统的评价方法往往无法准确地衡量图书馆数字资源的完整性[5]。传统方法往往只考虑数据属性的类型差异，而忽略了属性之间的关联性和重要性。因此，此次研究中，提出了一种融合k-prototypes算法和信息熵加权的图书馆数字资源完整性自动化评价方法。该方法综合考虑了数据属性的类型差异、属性之间的关联性和重要性。通过使用k-prototypes算法，能够有效地处理同时包含数值型和分类型属性的图书馆数字资源数据集。此次研究的创新点为通过引入信息熵加权的评价方式，并融合k-prototypes算法和信息熵加权的方法对图书馆数字资源完整性评价，以更加准确地衡量图书馆数字资源的完整性。此次研究的贡献在于提供一种有效的

*基金项目：陕西省重点研发计划项目(2024NC-YBXM-207)；杨凌职业技术学院2023年校内基金项目(SKYB-2364)；杨凌职业技术学院2024年校内教改项目(JG24092)

收稿日期：2023-12-25

工具和方法，帮助图书馆管理者评估和维护数字资源的完整性。

1 基于信息熵加权的图书馆资源自动化评价模型构建

1.1 融合k-prototypes算法的信息熵加权法改进研究

数字资源的完整性评价是图书馆和信息学领域中的一个重要研究领域，为评估数字资源的质量和可信度，确保用户获取高质量的信息，研究引入信息熵加权法用来评估资源完整性[6-7]。信息熵是信息理念中用于衡量随机变量中包含的信息量或者不确定性的程度。随机变量的信息熵的定义公式如式(1)所示。

$$H_R(x)=\frac{1}{1-a}\log\int f^a(x)dx, a>0, a\neq 1 \tag{1}$$

式中，x为随机变量，a为常数。若a=2时，则信息熵的表达式可简化为式(2)所示。

$$H_{R1}(x)=-\log\int f^2(x)dx \tag{2}$$

式中，常数a=2，因此式(2)为式(1)的二阶熵。接着为了计算数据中的概率密度，运用概率密度进行样本的密度分析，其表达式如式(3)所示。

$$\hat{f}(x)=\frac{1}{N}\sum_{i=1}^{N}W_{\sigma^2}(x,x_i) \tag{3}$$

式中，W_{σ^2}为窗函数，通过在每个样本点周围放置一个窗口，并将窗口内的样本点贡献到该点的概率密度估计中；σ^2为窗宽。窗函数的表达公式如式(4)所示。

$$W_{\sigma^2}(x,x_i)=\frac{1}{(2\pi)^{d/2}\sigma^d}\exp\left(-\frac{(x-x_i)^T(x-x_i)}{2\sigma^2}\right) \tag{4}$$

式中，d为样本维度；x_i为样本点；T为转置公式。

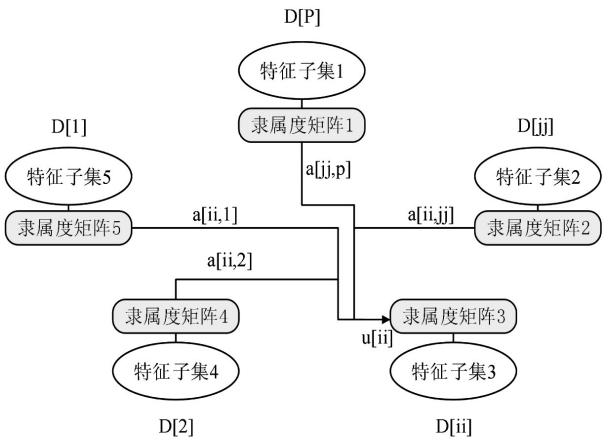


图1 k-prototypes算法的聚类结构获取示意图

然而在图书馆海量资源中，如何进行数据自动化处理评价是信息熵研究的难点。对于具有不同类别的数据集，如果某个类别的数据点数量远远超过其他类别，那么这个类别的数据点的信息熵将会被高度加权，从而可能导致其他类别的数据点被忽略或轻视。因此为更好地实现图书馆数据数字资源完整性的自动化评价，研究引入k-prototypes算法对传统的信息熵加权法进行改进，以实现图书馆数据完整性的自动化评价。图1为k-prototypes算法的聚类结构获取示意图。

信息熵加权是一种用于衡量信息量的不确定性程度的方法，可以用于对数字资源的完整性进行评价。在数字资源中，完整性评价是指对资源中存在的错误、缺失或损坏等问题进行检测和评估的过程。在信息熵加权下的完整性评价中，首先需要对数字资源进行分析，提取其中的特征信息。然后，根据不同的特征信息的重要性，给予其相应的权重。接下来，可以使用k-prototypes算法对数字资源进行聚类分析。k-prototypes算法是一种扩展的k-means算法，可以同时处理数值型和分类型数据[8]。通过将数字资源划分为不同的簇，可以更好地识别资源之间的相似性和差异性[9-10]。针对自动化问题，在数据预处理中，k-prototypes算法自动处理混合数据类型，无需手动将数据分为数值型和分类型。k-prototypes算法的目标函数如式(5)所示。

$$D(x,z)=\sum_{j=1}^{p}(x_j-z_j)^2+\gamma\sum_{j=p+1}^{q}\sigma(x_j,z_j) \tag{5}$$

式中，x和z分别为数据点和簇组；γ为权重系数；σ为数据点到簇组距离的模态。

1.2 图书馆数字完整性自动化评价模型构建

在构建图书馆数字完整性自动化评价模型中需要对图书馆的数字资源进行确定，并对图书馆的数字资源进行评价指标的确定[11-12]。首先运用改进后的信息熵加权法对图书馆数字资源的相似度进行提取，图书馆资源的信息有序度表达式如式(6)所示。

$$\begin{bmatrix}X\\p\end{bmatrix}=\begin{bmatrix}X_1\ X_2\ldots X_n\\p_1\ p_2\ldots p_n\end{bmatrix} \tag{6}$$

式中，X图书馆离散型数字资源的随机变量；P为X的有序度。

若信息熵越低，则图书馆的数字资源越有序，其相似度也越高。而图书馆数字资源的某一信息熵可以表示为式(7)。

$$H(X)=E[I(X_i)] \tag{7}$$

式中，X_i为某一信息资源；$I(X_i)$为X_i的有序度的对数负值；E为图书馆资源有序化程度。对其进行转化后可以得到式(8)。

$$H(X_1)=\sum_{i=1}^{n}p_i\log_2\frac{1}{p_t} \tag{8}$$

式中，p_i为图书馆资源中i分类的有序度。

	C1	C1	C1	C1	C1	C1	C1	C1	C1	C1	C1
U1			$R_{1,3}$								
U2		$R_{2,2}$		$R_{2,4}$				$R_{2,8}$			
U3							$R_{3,7}$				$R_{3,11}$
U4	$R_{4,1}$										
U5							$R_{5,7}$				$R_{5,11}$
U6				$R_{6,4}$				$R_{6,8}$			

图2 图书馆数字资源的评分矩阵

对于图书馆数字资源信息有序度的描述，为资源信息是否统一提供了判断标准。接着对图书馆数字资源的用户相似度进行分析，形成对图书馆数字资源的评分公式，评分差值如式(9)所示。

$$D=\left|R_A-R_B\right|=(\left|d_1\right|,\left|d_2\right|,...\left|d_N\right|) \tag{9}$$

式中，R_A和R_B为用户A和用户B对图书馆资源进行评价的分数；d_1、d_2、d_3为两个用户对每一资源打分的差值比较。用户的评分矩阵图如图2所示。

图2中，U为用户；C为图书馆的数字资源；$R_{i,j}$为资源评分结果。图书馆数字资源的评分矩阵是一个由用户和资源组成的二维矩阵，其中每个元素表示用户对资源的评分。评分矩阵的行表示用户，列表示资源，矩阵中的值表示用户对对应资源的评分。最后通过用户评分可以得出图书馆资源完整性的评价和资源协同阈值，其阈值表达式如式(10)所示。

$$O=\overline{R_A}+\frac{Sim(A,B)*(R_B-\overline{R_B})}{\sum Sim(A,B)} \tag{10}$$

式中，$\overline{R_A}$为用户A对所有评价的图书馆数字资源给出的评分结果；Sim为用户A和用户B的评分结果相似度，接着对图书馆数字资源的完整性进行模型的搭建。

融入k-prototypes算法的信息熵加权数字资源完整性的自动化评价模型结构可以包括以下几个部分，首先是数据预处理，其次为特征选择，根据数字资源完整性的评价指标和需求，选择合适的特征。接着是信息熵计算，根据选取的特征，计算每个特征的信息熵。然后为加权处理，根据计算得到的特征的信息熵，对其进行加权处理，得到每个特征的权重。最后对模型进行评估，评估构建的完整性评价模型的性能和准确性。并根据评估结果进行模型的分析和改进，进一步优化模型的准确性和效果。以上是融入k-prototypes算法的信息熵加权下数字资源完整性的自动化评价模型的结构。

2 图书馆资源评价算法实现和模型训练结果评估

2.1 k-prototypes算法在数字资源评价的性能评估

为对图书馆数字信息进行进一步的分析，实验选取某高校图书馆数据库进行验证，并从高校图书馆数字资

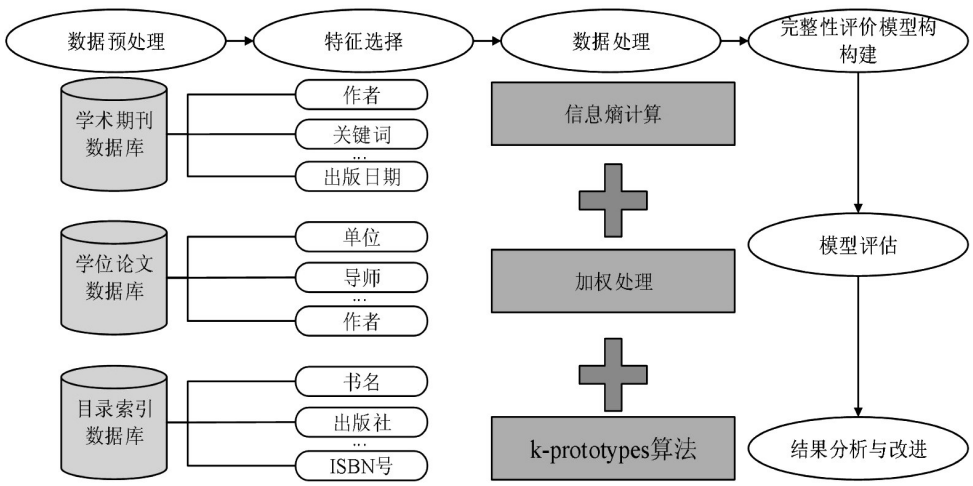


图3 图书馆数字资源的完整性评价模型

源的时效性、全面性、权威性、实用性、特色性、规范性、连续性出发建立数字资源完整性的评价指标建立，指标表如表1所示。

表1 数字资源评价指标的信息熵、效用值和权重

数字资源内容	信息熵	效用值	权重
时效性	0.771 5	0.228 5	0.115 3
全面性	0.823 3	0.176 7	0.245 6
权威性	0.768 9	0.231 1	0.103 2
实用性	0.795 1	0.204 9	0.161 4
特色性	0.624 1	0.375 9	0.120 7
规范性	0.789 1	0.210 9	0.150 7
连续性	0.683 4	0.316 6	0.103 1

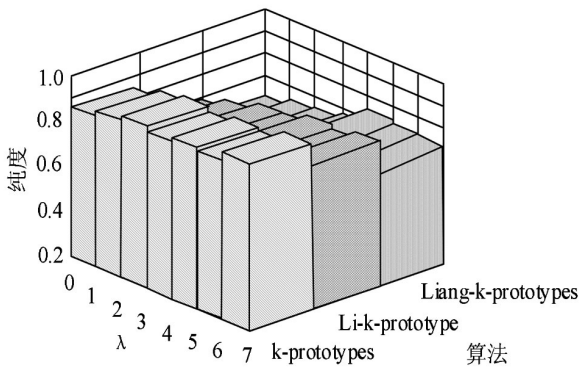


图4 数据集在不同λ值中的聚类纯度比较

如表1所示，其中数字资源的全面性的权重最高为0.245 6，因为数字资源的全面性最能代表数字资源的完整性。首先，全面的数字资源可以提供更多的信息资源，满足用户对多领域、多类型信息的需求。其次为数字资源的适用性和规范性，其权重分别为0.161 4和0.150 7。接着为测试研究算法在图书馆数字资源完整性评价的优势，选取某高校图书馆的学术期刊数据集(academic journal dataset，AJ)，学位论文数据集(dissertation dataset，DD)，图书馆目录数据集(library catalog dataset，LC)进行聚类纯度的检验，与Liang-k-prototypes和Li-k-prototype算法进行比较。比较结果如图4所示。

由图4可知，随着λ的增加，算法的聚类纯度都呈现出上升的趋势，其中研究的k-prototypes算法聚类纯度在λ大于等于3时，纯度超过0.9，而Li-k-prototype算法的聚类纯度始终低于0.8，Liang-k-prot-otypes算法的聚类纯度最低，始终保持在0.6以下。综上可知，研究算法有利于对图书馆不同数据集进行聚类分析。接着在不同实验次数下，应用融合k-prototypes算法的信息熵加权法对收集的图书馆数字资源完整性对应的各指标进行评分，评分结果如图5所示。

由图5可知，在收集的样本中，图书馆数字资源的特色性平均评分最高，约为9.5，其次为数据资源的权威性，平均评分值约为9.4，而该图书馆数字资源的全面性的平均评分为9.1，实用性评价为9.1，时效性的平均分值为8.5，规范性和连续性的平均分值分别为8.7和8.8。由此可知该图书馆的特色性和权威性较强，拥有丰富的藏书资源，包括各类图书、期刊、学位论文、会议论文等，而在数字资源的全面性方面有待提高。

2.2 图书馆数字资源完整性自动化模型的实验评估

接着针对图书馆数字资源的AJ、DD、LC三种数据集，通过研究的图书馆数字资源完整性自动化模型对三种数据集进行离群点检测。首先进行数据预处理，将混合数据(包含数值型和分类型数据)转换成k-prototypes算法能够处理的形式。接着利用k-prototypes算法对数据进行聚类，得到k个聚类簇。AJ数据集检测结果如图6所示。

图6(a)为图书馆数据AJ数据集的离群点检测结果，其中有14个离群点；图6(b)为图书馆数据DD数据集的离群点检测结果，其中有12个离群点；图6(c)为图书馆数据LC数据集的离群点检测结果，其中有17个离群点。其中LC数据集的离群点平均分值为22.2，比AJ数据集和DD

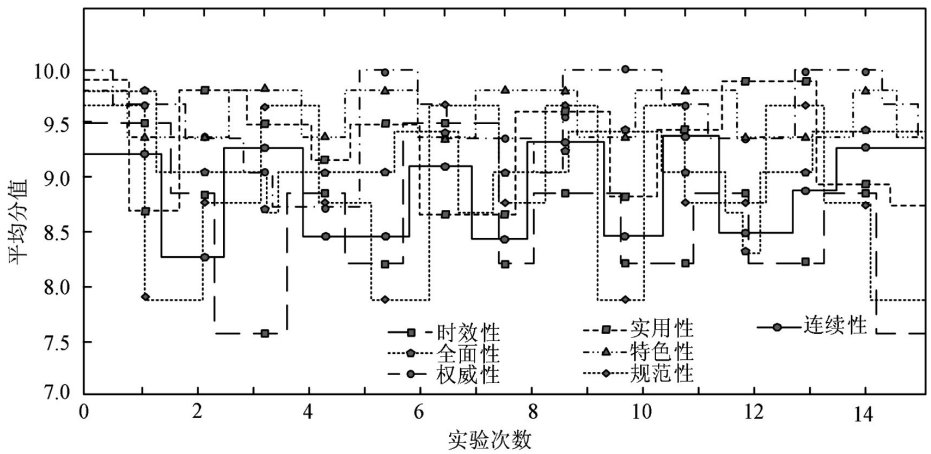
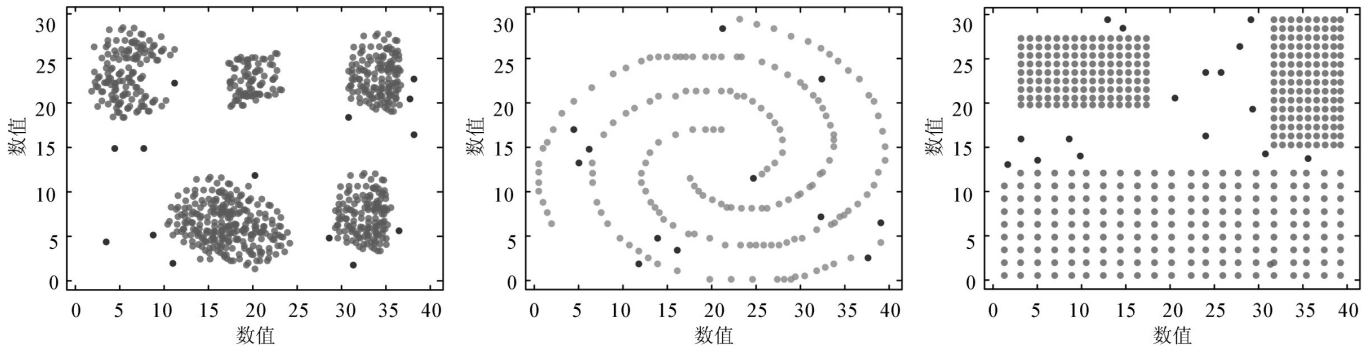


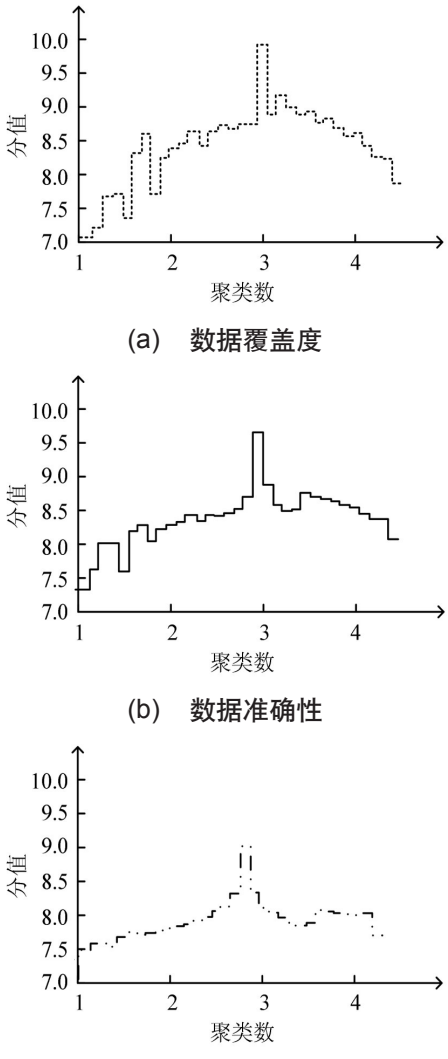
图5 图书馆数字资源完整性指标分结果



(a) AJ数据集检测结果 (b) DD数据集检测结果 (c) DD数据集检测结果

图6 图书馆数据集离群点检测

数据集的离群点检测的平均分值分别高7.2和6.8，说明该图书馆目录数据集内部样本的离散程度更大。最后应用研究模型，根据提供的样本数据集从数据覆盖度、数据准确性和数据更新性三个评价指标对该图书馆的数字资源完整性进行评分，首先是数据覆盖度，其评分结果如图7所示。



(a) 数据覆盖度

(b) 数据准确性

(c) 数据更新性

图7 图书馆数据资源完整性评价结果

由图7(a)可知该图书馆的数据覆盖率在聚类数为3时达到最高评分，数据覆盖率的平均分值为8.4。由图7(b)可知，数据准确率在聚类数为3时同样达到最高值为9.6，由图7(c)可知数据更新率最高分数低于9分，而该图书馆的数据覆盖率平均得分比数据准确性和数据更新性的平均分值分别高0.21和3.9。由此可知该图书馆的数据覆盖率较广，数据准确性较广，但数据的更新速度有待进一步提高。综上可知，融合k-prototypes算法的信息熵加权法可以有效对图书馆数字资源完整性进行自动化评分，为图书馆提高数字资源全面性提供现实依据并能够推动用户对知识和信息的需求。

3 结束语

随着数字化时代的到来，图书馆数字资源的数量和多样性不断增长。对于图书馆管理者来说，如何评估和确保数字资源的完整性成为一个重要的挑战。因此研究基于信息熵加权法并融合k-prototypes算法构建评价模型。结果可知，数字资源的全面性的权重最高为0.245 6。研究的k-prototypes算法聚类纯度最高平均纯度超过9.0。在数字资源评价中，图书馆数字资源的特色性平均评分最高，约为9.5，而图书馆目录数据集内部样本的离散程度最大。在数字资源完整性评价中，数据覆盖率度的平均分值最高，为8.4，比数据准确性和数据更新性的平均分值分别高0.21和3.9。综上所述，此次研究的算法模型能够有效实现对图书馆数字资源完整性的评估。此次研究的改进之处在于，未来应扩大对不同层次不同类型的图书馆数字资源完整性研究。

参考文献：

[1] 朱学芳，邢绍艳.基于用户需求的高校图书馆数字资源服务质量评价研究[J].情报科学，2022，40(3)：10-25.

[2] 李琳.应用于图书馆书籍分类的熵加权聚类算法[J].现代电子技术，2020，43(1)：3-16.

(下转第116页)