

文章编号:1671-5896(2026)04-0956-07

混合语义相似度提取下高维重叠 数据智能模糊检索算法

沈飞洋, 刘 涛

(杭州电子科技大学 法学院, 杭州 310018)

摘要: 针对在数据智能检索过程中, 单一语义相似度的局限性可能会影响数据相关性的判断, 导致检索结果的排序敏感度不高的问题, 提出了混合语义相似度提取下高维重叠数据智能模糊检索算法。建立高维重叠数据的信息粒度空间, 通过密度聚类与模糊度量, 识别出空间中的重叠数据所在的区域, 从而在不改变数据结构的前提下, 实现高维数据降维。混合均值函数、近似线性化统计、规则方法3种语义相似计算方法, 综合分析降维数据的相关性。在最近邻算法的遍历下, 按序提取出语义相似度较高的数据信息, 生成数据检索列表。算例测试结果表明, 该算法所表现出的 NDCG(Normalized Discounted Cumulative Gain) 指标值平均为 0.867, 检索结果具有较高的排序敏感度, 检索质量更高, 具有良好的实践应用前景。

关键词: 数据智能检索; 数据模糊检索; 高维重叠数据; 混合语义相似度; 数据检索; 智能检索

中图分类号: TP391

文献标志码: A

Intelligent Fuzzy Retrieval Algorithm for High Dimensional Overlapping Data under Hybrid Semantic Similarity Extraction

SHEN Feiyang, LIU Tao

(School of Law, Hangzhou Dianzi University, Hangzhou 310018, China)

Abstract: In the process of intelligent data retrieval, the limitation of a single semantic similarity may affect the judgment of data relevance, resulting in a low ranking sensitivity of the retrieval results. To alleviate this problem, an intelligent fuzzy retrieval algorithm is proposed for high-dimensional overlapping data under hybrid semantic similarity extraction. The information granularity space of high-dimensional overlapping data is established. Through density clustering and fuzzy measurement, the regions where the overlapping data is located in the space are identified, thereby achieving the dimension reduction of high-dimensional data without changing the data structure. Three semantic similarity calculation methods, namely the mixed mean function, approximate linearization statistics, and rule method, are used to comprehensively analyze the correlation of dimensionality reduction data. Under the traversal of the nearest neighbor algorithm, the data information with higher semantic similarity is extracted in sequence to generate the data retrieval list. The results of the case study show that the average NDCG(Normalized Discounted Cumulative Gain) index value exhibited by this algorithm is 0.867. The retrieval results have a high ranking sensitivity and higher retrieval quality, and have a good practical application prospect.

Key words: data intelligent retrieval; fuzzy data retrieval; high-dimensional overlapping data; mixed semantic

收稿日期: 2025-06-25

基金项目: 浙江省自然科学基金资助项目(POF1D6215623)

作者简介: 沈飞洋(1999—), 男, 杭州人, 杭州电子科技大学硕士研究生, 主要从事人工智能研究, (Tel)86-19858190382 (E-mail) Shenfeiyang0606@163.com; 刘涛(1982—), 男, 河南尉氏人, 杭州电子科技大学副教授, 硕士生导师, 主要从事社会人类学与人工智能研究, (Tel)86-19858190382 (E-mail) Shenfeiyang0606@163.com。

similarity; data retrieval; intelligent retrieval

0 引言

在现代信息系统中,高维重叠数据具有广泛分布。这种数据不仅会导致检索过程存在维度灾难的问题,还可能导致算法无法准确区别重叠区域的数据特征,增加了数据匹配的难度。其使传统的数据检索方法的检索效率低下,同时,还难以得出高质量的检索结果,已经无法满足现有数据检索工作的实际需求。为此,人们针对提升数据检索质量进行了深入探讨,并提出了不同的数据检索方法。

王娅云等^[1]提出了基于数据分类的数据检索方法。通过提取数据特征,利用定义的损失函数优化聚类中心的设计,实现跨域数据的分类处理。通过数据特征向量相似度的计算,输出得到数据检索结果。该方法能将检索时间降低至 0.23 ms,检索效率较高。但数据分类难以处理空间重叠的数据,无法实现精确区分,分类性能不理想,进而导致检索结果容易出现排序混杂的问题,排序敏感度不高。陈庆华^[2]提出了基于哈希学习算法的数据检索方法。通过编码神经网络的实体抽取,建立数据实体的知识图谱。在哈希学习函数的处理下,通过二进制哈希码的相似度匹配,生成数据检索结果。其适用于非结构化的数据类型,避免了数据冲突问题。但在实际应用中,数据的二进制哈希码可能存在长度不足的问题,其将增加数据检索过程的碰撞概率,导致相似序列的哈希码冲突,排序敏感度下降。陈雪等^[3]提出了基于映射关系的数据检索方法。构建数据结构间的实体关系,定义数据检索内容的五元组。通过计算数据实体的映射距离,分析数据的条件概率分布,进而利用 Tag 方法实现数据检索。该方法能消除数据模态的影响,实现多模态数据之间的检索。但其难以处理动态变化的数据集,映射关系的更新不够及时,导致排序结果不敏感。李中等^[4]提出了基于深度卷积神经网络的数据检索方法。计算数据空间的全局特征量,在深度卷积神经网络的特征提取下,分析数据哈希码的汉明距离参数。并以此作为指标,检索出目标数据信息。其所需的计算量较小,能节省数据检索过程的通信开销。但如果实际构建的深度卷积神经网络泛化性过强,可能导致细节数据实特征的丢失,使检索结果排序敏感度下降。

针对上述问题,笔者提出了一种混合语义相似度提取下高维重叠数据智能模糊检索算法。其通过对数据语义相似度的优化计算,通过遍历并排序,得到排序敏感度更高的检索结果。

1 高维重叠数据智能模糊检索算法设计

1.1 高维重叠数据的模糊度量与降维

为解决数据的高维灾难问题,笔者采用距离度量学习方法对其进行非线性降维,并针对高维数据类别边界模糊的重叠问题进行模糊度量,明确数据边界,便于后续语义相似度的计算。

建立高维重叠数据的信息粒度空间,如下^[5]:

$$S = g(x)/|U|, \quad (1)$$

其中 S 表示数据信息粒度空间, x 表示该空间中的高温重叠数据, $g(x)$ 表示数据信息粒度的基数, $|U|$ 表示数据粒度结构复杂度。

在该空间中,采用切比雪夫距离,进行任意维度上的数据距离度量。如下:

$$d(x_i, x_j) = \max_{i,j} (g |x_i - x_j|), \quad (2)$$

其中 $d(x_i, x_j)$ 表示高维重叠数据的距离度量, x_i 和 x_j 分别表示信息粒度空间中第 i, j 个数据。

根据数据的距离度量,通过密度聚类,识别出空间中的重叠数据。考虑高维重叠数据的边界集可能存在为空的情况,则有^[6-7]:

$$\begin{cases} c = w_1 \sum_{x_i \in C_i} x_i / C_i + w_2 \sum_{x_j \in C_j} x_j / C_j, \\ w_1 + w_2 = 1, \quad w_1 > w_2, \end{cases} \quad (3)$$

其中 c 表示类簇中心, C_i 和 C_j 分别表示高维重叠数据 x_i, x_j 所属的聚类簇, w_1 表示下近似区域权重, w_2 表示边界权重。

根据聚类结果, 识别出高密度的数据区域, 得到数据的重叠区域^[8]。针对该区域内的高维数据进行模糊度量, 如图1所示。

如图1所示, 数据 x_3 位于重叠区域, 可以属于两个不同的类簇, 而 x_1 和 x_2 则属于一个单独的类簇。针对该数据分布情况, 笔者将图1中点的虚线和横短的虚线视为无需参与计算的距离度量, 以此建立重叠数据边界模糊度量的计算准则, 如下:

$$\beta = \max\left(0, x_k - \gamma \frac{N}{M} C(x_k)\right), \quad (4)$$

其中 β 表示数据的边界模糊度量, x_k 表示位于重叠区域的数据, γ 表示粗糙系数, N 表示交叉数据量, M 表示总数据量。

通过模糊度量, 处理重叠区域样本, 避免传统数据降维处理中将重叠区域强行分离导致原始数据结构破坏的问题^[9]。基于此, 构建高维重叠数据降维处理的降维矩阵, 如下:

$$W = \sum_{x_k}^N \|\lambda \cdot \varphi(x)\|_2^\beta, \quad (5)$$

其中 W 表示降维矩阵, λ 表示降维处理的非线性权重参数向量, $\varphi(x)$ 表示非线性映射函数。

通过上述步骤, 在不改变数据结构的情况下, 完成高维重叠数据的降维处理。

1.2 数据混合语义相似度计算

面向降维后的数据, 采用混合语义相似度计算的方法, 分析数据与检索词间的相关性关系。将简单的关键词匹配检索过程改进为深层语义理解的检索过程, 提升检索质量。

确定降维处理后, 数据语义的维度数如下^[10-11]:

$$Z = \cos\langle r_s, r_x \rangle, \quad (6)$$

其中 Z 表示降维数据的语义维度数, r 表示数据 x 的语义词汇, r_x 表示相应的投影向量, r_s 表示该数据中的停用词投影向量。

考虑数据语义维度数的不同, 可能使维度数相差较大的数据语义相似度计算结果出现较大偏差^[12]。为此, 结合均值函数、近似线性化统计等方法进行语义相似度的混合计算, 改善单一计算方法的局限性, 得到综合衡量的数据语义相似度。

均值函数下的语义相似度如下^[13-14]:

$$L_1 = \frac{1}{Z} \sum_{i=1}^M f(x, P(Z), a), \quad (7)$$

其中 L_1 表示均值函数计算得到的语义相似度, f 表示均值函数, $P(Z)$ 表示数据语义维度数的条件概率分布, a 表示目标检索词。

近似线性化统计下的语义相似度如下:

$$L_2 = (2r_x - r_a^2)^\vartheta, \quad (8)$$

其中 L_2 表示近似线性化统计计算得到的语义相似度, r_a 表示目标检索词语义的投影向量, ϑ 表示线性化系数。

规则方法下的语义相似度如下^[15]:

$$L_3 = (1 - \varphi)R(2\alpha - Z(x)) + \varphi R(2\alpha - Z(a)), \quad (9)$$

其中 φ 表示可信度, R 表示语义维度数的归一化操作, α 表示规则阈值。

基于上述, 采用整体混合的方法, 计算出混合语义相似度, 如下:

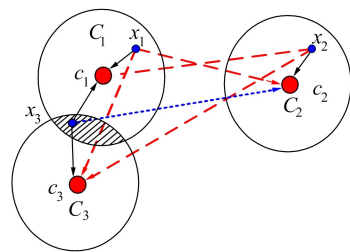


图1 高维重叠数据的模糊度量

Fig. 1 Fuzzy measurement of high-dimensional overlapping data

$$L(x,a)=1-\frac{L_1\sum_{i=1}^ML_3^2}{\varepsilon(L_2-1)},$$

(10)

其中 $L(x,a)$ 表示数据的混合语义相似度, ε 表示秩统计参数。

通过上述步骤, 完成数据混合语义相似度的计算。

1.3 相似度提取下的数据智能模糊检索

在目标数据库中, 根据检索词计算所得的混合语义相似度, 采用最近邻搜索的方法遍历所有数据, 提取出语义相似度最高的数据样本, 并进行排序, 输出得到数据的智能检索结果。

定义最近邻搜索的索引向量, 如下^[16]:

$$\boldsymbol{I}=\arg \min \left\|x^m-L(x, a)\right\|_2^2-e(x),$$

(11)

其中 $\boldsymbol{I}$ 表示索引向量, m 表示子向量编码器, $e(x)$ 表示向量量化误差。

利用该索引向量, 遍历目标数据库, 选择出满足一定语义相似度阈值的数据样本。阈值的确定由实际检索需求制定, 可以为检索数据量, 语义相似度大小等^[17]。并在其过程中, 为确保最近邻搜索方向为最快遍历方向, 引入了一个梯度向量作为索引向量遍历方向的指导。如下:

$$\nabla \boldsymbol{\Phi}=\left\{\frac{\partial I}{\partial \theta_1}, \frac{\partial I}{\partial \theta_2}, \cdots, \frac{\partial I}{\partial \theta_M}\right\},$$

(12)

其中 $\nabla \boldsymbol{\Phi}$ 表示梯度向量, θ 表示梯度向量的下降方向。

在最近邻搜索的语义相似度提取下, 将提取出的数据样本按照语义相似度的大小进行排序, 生成数据智能检索列表。

通过上述步骤, 完成混合语义相似度提取下高维重叠数据智能模糊检索算法的研究设计。

2 算例测试

2.1 实验数据

为验证所提算法的检索性能, 对多领域的数据智能检索进行了实验测试。

实验测试的领域包括学术、影视剧以及商品检索等。其中, 学术检索领域的实验数据来源于知网上的 5 000 篇期刊文献以及公开的 arXiv 数据集。在 arXiv 数据集的存储库中选择 2 204 作为实验数据, 包括14 274 篇中文科技文献。影视剧领域的实验数据来源于 Netflix 网站以及 TMDb(The Movie Database) 影视数据集等。商品领域的实验数据来源于淘宝、Amazon、Online Auctions 等网站。在数据爬取工具的支持下, 得到的实验数据集如表 1 所示。

表 1 实验数据集概况

Tab. 1 Overview of experimental dataset

类别	数据量	类别	数据量	类别	数据量
学术	19 274	影视剧	18 544	商品	9 485

将上述数据以 JSON(JavaScript Object Notation) 格式存储到实验数据库中。为各领域的数据, 定义检索词的属性, 如表 2 所示。

表 2 数据检索关键词属性

Tab. 2 Keyword attributes for data retrieval

类别	关键词属性
学术	文献标题、作者姓名、期刊名称、文献级别、文献领域、主题、关键词等
影视剧	影视剧名称、参演人员、导演姓名、上映时间、剧情关键词、主题、情感关键词、拍摄技术等
商品	商品名称、商品品类、规格参数、颜色与外观、商品用途、品牌名称、使用人群等

面向上述的实验数据, 对数据检索进行测试实验。

2.2 数据智能检索算例

利用上述构建的实验数据集, 面向笔者算法进行数据智能检索测试。以实验数据集中的学术领域

数据为例，得到以下实验结果。

针对原始数据集存在的高维重叠问题，采用模糊度量的方法进行了降维处理。其过程中，原始实验数据集所呈现出的类簇重叠状态如图 2 所示。

从图 2 可看出，在聚类分析下，原始实验数据集存在重叠区域，部分数据可以归属于不同的类簇。通过对该类簇的边界进行模糊度量，在降维矩阵下，实现原始数据的降维处理。通过对降维数据的可视化分析可知，笔者算法的数据降维并未改变原始数据集的数据结构，降维性能较优。

经过降维处理后，采用混合语义相似度的方法，对数据与检索词间的语义相似度进行了计算，得到性能对比结果如图 3 所示。

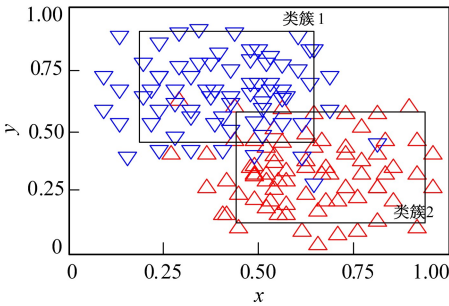


图 2 数据类簇重叠状态

Fig. 2 Overlapping state of data clusters

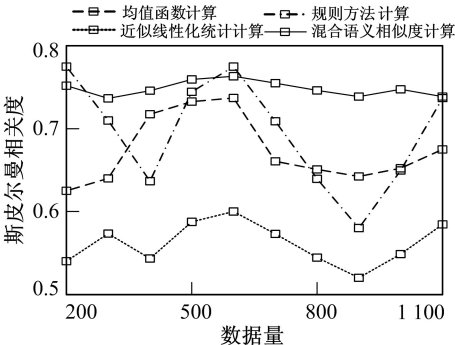


图 3 混合语义相似度计算性能对比

Fig. 3 Performance comparison of hybrid Semantic similarity calculation

从图 3 可看出，与单一的语义相似度计算方法相比，混合相似度的计算方法能大幅度改善计算性能，计算结果表现出的斯皮尔曼相关度更高。说明笔者所设计的计算方法能有效改善单一计算方法的局限性，综合各方法的准确度与稳定度，实现了性能更优的语义相似度计算。

根据优化的语义相似度计算结果，在最近邻搜索算法的遍历提取下，排序所提取出的数据信息，生成最终的检索列表，如图 4 所示。

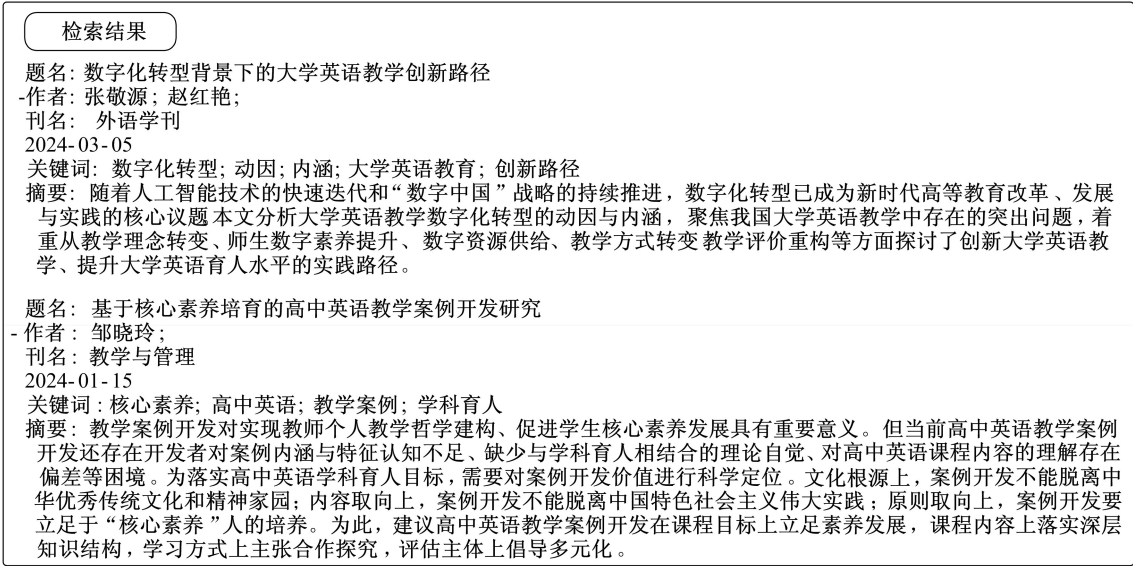


图 4 数据智能检索结果(部分)

Fig. 4 Intelligent data retrieval results (partial)

根据算例测试所输入的检索关键词“英语教学”，笔者算法筛选出了满足阈值的数据信息。其中，排名前二的两条检索结果如图 4 所示。基于用户的实际检索需求，该检索结果为用户所需内容，命中率

较高。结果验证了笔者算法在数据检索中的应用可行性。

2.3 检索性能评估

为进一步探讨所提算法的检索性能, 采用 NDCG 作为指标进行了多组实验的性能评估。该指标反映了检索结果的相关性等级以及数据排序位置的有效性, 能量化评估检索结果的排序敏感度。同时, 实验采用了基于哈希学习算法、映射关系的数据检索方法作为对比方法, 从而更加直观地分析所提算法的检索性能。经过测试, 得到各方法检索的指标评估结果如图 5 所示。

从图 5 可看出, 与其他两种检索方法相比, 所提算法表现出的检索结果 NDCG 指标水平明显更高, 平均能达到 0.867。说明所提算法对于相关数据的排序与其实际的语义相似度更为接近, 更符合目标检索关键词的内容, 更加有效地区别了数据的相关性程度, 具有优良的排序敏感度。其主要得益于所提算法混合了多种语义相似度计算方法, 成功改善了单一计算方法的局限性, 从而提升了数据智能检索的性能, 得到更加优质的检索结果。

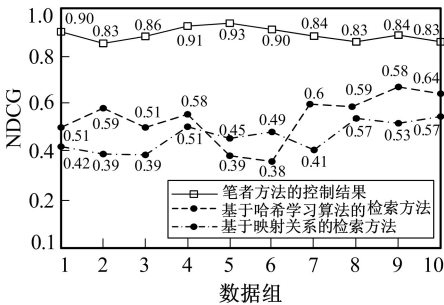


图 5 各方法检索性能对比

Fig. 5 Comparison of retrieval performance of various methods

3 结 语

笔者设计的混合语义相似度提取下高维重叠数据智能模糊检索算法, 能在不改变原始数据重叠结构的基础上, 有效完成数据的降维处理。结合混合语义相似度的计算, 成功避免了单一计算结果的限制, 能对数据的语义相似度进行更加准确的衡量, 为检索过程提供了可靠基础。结合最近邻搜索算法的遍历搜索, 得出了排序敏感度较高的检索结果。笔者算法在多个领域均能有效适用, 支持复杂的高维数据分析场景, 解决高维数据的检索难题。

参考文献:

[1] 王娅云, 李增伟, 魏永瑜, 等. 电力中台海量数据跨域快速检索方法设计 [J]. 信息技术, 2025(3): 151-156, 162.
WANG Y Y, LI Z W, WEI Y Y, et al. Design of Cross-Domain Fast Retrieval Method for Massive Data of Power Middle Office [J]. Information Technology, 2025(3): 151-156, 162.

[2] 陈庆华. 基于哈希学习算法的数字化医院非结构化数据智能检索 [J]. 中国卫生信息管理杂志, 2025, 22(1): 136-142.
CHEN Q H. Intelligent Retrieval of Unstructured Data in Digital Hospitals Based on Hash Learning Algorithms [J]. Chinese Journal of Health Informatics and Management, 2025, 22(1): 136-142.

[3] 陈雪, 陈艺丹, 何其森, 等. 基于映射关系的多模态电网设备数据实时检索 [J]. 电子设计工程, 2025, 33(5): 26-30.
CHEN X, CHEN Y D, HE Q M, et al. Real Time Retrieval of Multimodal Power Grid Equipment Data Based on Mapping Relationships [J]. Electronic Design Engineering, 2025, 33(5): 26-30.

[4] 李中, 郭云丽. 舰船通信关联信息目标数据检索方法研究 [J]. 舰船科学技术, 2024, 46(18): 159-162.
LI Z, GUO Y L. Research on Target Data Retrieval Methods for Ship Communication Association Information [J]. Ship Science and Technology, 2024, 46(18): 159-162.

[5] 甘臣权, 杨宏鹏, 祝清意, 等. 基于访问控制的可验证医疗区块链数据搜索机制 [J]. 重庆邮电大学学报(自然科学版), 2023, 35(5): 873-887.
GAN C Q, YANG H P, ZHU Q Y, et al. A Verifiable Medical Blockchain Data Search Mechanism with Access Control [J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2023, 35(5): 873-887.

[6] 梁英杰, 张明元, 姜锋, 等. 融合 Informer 和深度哈希的时序数据相似性搜索研究 [J]. 控制工程, 2023, 30(7): 1317-1323.
LIANG Y J, ZHANG M Y, JIANG F, et al. Research on Similarity Search of Time Series Data Based on Informer and Deep

- Hash [J]. Control Engineering of China, 2023, 30(7): 1317-1323.
- [7] 杨婷, 莫若玉, 张秀娟, 等. 轻量级缓存策略的关系型数据库全文搜索加强与扩展 [J]. 计算机应用, 2023, 43(8): 2431-2438.
- YANG T, MO R Y, ZHANG X J, et al. Enhancement and Expansion of Full-Text Search in Relational Databases Based on Lightweight Caching Strategy [J]. Journal of Computer Applications, 2023, 43(8): 2431-2438.
- [8] 赵建伟, 赵光, 王震, 等. 基于时域-频域哈希编码的电力海量空间数据智能化检索系统 [J]. 自动化与仪器仪表, 2025(4): 301-303.
- ZHAO J W, ZHAO G, WANG Z, et al. Intelligent Retrieval System for Massive Spatial Data of Electric Power Based on Time Domain Frequency Domain Hash Coding [J]. Automation & Instrumentation, 2025(4): 301-303.
- [9] 邵伟, 朱高宇, 于雷, 等. 高维数据的降维与检索算法 [J]. 山东大学学报(理学版), 2024, 59(7): 27-43.
- SHAO W, ZHU G Y, YU L, et al. Dimensionality Reduction and Retrieval Algorithms for High Dimensional Data [J]. Journal of Shandong University (Natural Science), 2024, 59(7): 27-43.
- [10] 刘亮龙, 牛秦洲. 基于模糊 C 均值聚类的异构数据信息检索仿真 [J]. 计算机仿真, 2024, 41(10): 482-486.
- LIU L L, NIU Q Z. Simulation of Heterogeneous Data Information Retrieval Based on Fuzzy C -Means Clustering [J]. Computer Simulation, 2024, 41(10): 482-486.
- [11] 赵松燕, 曲朝阳, 郭晓利, 等. 基于 MapReduce 的输电监测数据智能检索模型 [J]. 电力系统保护与控制, 2023, 51(22): 177-187.
- ZHAO S Y, QU C Y, GUO X L, et al. Intelligent Retrieval Model of Power Transmission Monitoring Data Based on MapReduce [J]. Power System Protection and Control, 2023, 51(22): 177-187.
- [12] 王辉, 张晓明, 鲍丽芳, 等. 基于云存储的电子档案数据跨域安全检索算法研究 [J]. 集成电路与嵌入式系统, 2025, 25(3): 66-72.
- WANG H, ZHANG X M, BAO L F, et al. Research on Cross-Domain Secure Retrieval Algorithm of Electronic Archives Data Based on Cloud Storage [J]. Integrated Circuits and Embedded Systems, 2025, 25(3): 66-72.
- [13] 刘洋, 毛克明. 基于自适应反馈机制的小差异化图像纹理特征信息数据检索 [J]. 江苏大学学报(自然科学版), 2025, 46(1): 73-81.
- LIU Y, MAO K M. Retrieval of Texture Feature Information Data for Small Differentiated Images Based on Adaptive Feedback Mechanism [J]. Journal of Jiangsu University (Natural Science Edition), 2025, 46(1): 73-81.
- [14] 焦世超, 关日鹏, 况立群, 等. 深度语义关联学习的基于图像视觉数据跨域检索 [J]. 计算机工程, 2024, 50(5): 190-199.
- JIAO S C, GUAN R P, KUANG L Q, et al. Image-Based Cross-Domain Visual-Data Retrieval with Deep Semantic Correlation Learning [J]. Computer Engineering, 2024, 50(5): 190-199.
- [15] 韩云娜. 基于多相似度模糊 C 均值聚类的不均衡流数据检索方法 [J]. 吉林大学学报(信息科学版), 2024, 42(4): 726-732.
- HAN Y N. Data Retrieval Method of Unbalanced Streaming Based on Multi-Similarity Fuzzy C -Means Clustering [J]. Journal of Jilin University (Information Science Edition), 2024, 42(4): 726-732.
- [16] 阚钊玉. 高可靠性移动通信网络数据库跨域安全检索研究 [J]. 微型电脑应用, 2024, 40(8): 46-50.
- KAN D Y. Research on Cross Domain Security Retrieval of Highly Reliable Mobile Communication Network Database [J]. Microcomputer Applications, 2024, 40(8): 46-50.
- [17] 张旭东, 谢民, 黄建平, 等. 基于数据挖掘的电力自动化系统运行数据中台资源检索技术研究 [J]. 安徽师范大学学报(自然科学版), 2023, 46(2): 119-125.
- ZHANG X D, XIE M, HUANG J P, et al. Research on Station Resource Retrieval Technology in Operation Data of Power Automation System Based on Data Mining [J]. Journal of Anhui Normal University (Natural Science), 2023, 46(2): 119-125.

(责任编辑: 刘东亮)