

基于动态图神经网络的医学实体链接方法研究

李博诚 李乃适 周婧雅 王 怡

(中国医学科学院北京协和医院病案科 & WHO 国际分类家族合作中心 北京 100730)

〔摘要〕 **目的/意义** 克服传统深度学习方法缺点，进一步提高医学实体链接性能。**方法/过程** 构建序列信息图和主体信息图以获取多维度潜在文本信息，采用动态图神经网络进行全局关联，充分利用模型内部信息，并基于注意力机制生成最终序列关联。基于真实临床疾病诊断数据构建实证数据集，并进行消融实验和对比实验。**结果/结论** 该方法精确率为 90.34%，召回率为 89.72%，*F1* 值为 90.03%。消融实验和对比实验结果显示，改进策略有效且必要。

〔关键词〕 实体链接；图神经网络；电子病历；注意力机制

〔中图分类号〕 R-058 **〔文献标识码〕** A **〔DOI〕** 10.3969/j.issn.1673-6036.2025.07.009

Study on Medical Entity Linking Method Based on Dynamic Graph Neural Network

LI Bocheng, LI Naishi, ZHOU Jingya, WANG Yi

Department of Medical Record, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & WHO Family of International Classification Collaborating Center in China, Beijing 100730, China

〔Abstract〕 **Purpose/Significance** To overcome the shortcomings of traditional deep learning method, and to further improve the performance of medical entity linking. **Method/Process** Sequence infographics and subject infographics are constructed to obtain multi-dimensional potential text information. Dynamic graph neural networks are adopted for global association, the internal information of the model is fully utilized, and the final sequence association is generated based on the attention mechanism. An empirical dataset is constructed based on real clinical disease diagnosis data, ablation experiments and comparison experiments are conducted. **Result/Conclusion** The precision rate of the method is 90.34%, the recall rate is 89.72%, and the *F1* value is 90.03%. The results of the ablation and comparison experiments show that the improvement strategy is effective and necessary.

〔Keywords〕 entity linking; graph neural network; electronic medical record (EMR); attention mechanism

〔修回日期〕 2025-06-10

〔作者简介〕 李博诚，博士，主管技师，发表论文 24 篇；通信作者：王怡，研究员。

〔基金项目〕 中央高水平医院临床科研业务费（项目编号：2025-PUMCH-C-006）；中央高水平医院临床科研业务费（项目编号：2022-PUMCH-B-091）。

1 引言

随着信息技术的发展和电子病历系统的普及，医院每天会产生大量医疗文本。由于医生的使用习惯不同，加之部分医学名词受外文翻译影响，同一医学概念可能存在不同的表达方式^[1]，这极大影响了医疗数据的加工与利用。实体链接（又称实体消歧）指判

断不同知识体系中的实体是否指向现实世界中的同一对象的过程^[2]，其核心任务是将非规范化的实体表述映射到标准化实体库中的对应条目^[3]。高质量的医学实体链接对提升临床决策支持系统准确性、促进电子病历结构化与互操作性意义重大。本研究基于电子病历数据，提出基于动态图神经网络的实体链接方法，以期进一步提升医学实体链接性能。

2 相关研究

2.1 基于语义信息的链接算法

基于语义信息的链接算法将实体链接问题转化为多个文本分类或聚类问题，典型算法包括支持向量机、FastNewman 等^[4]。此类方法虽然能取得一定效果，但随着深度学习及相关技术的发展，其局限性逐渐显现：不仅依赖人工特征工程，还易出现特征稀疏问题。针对上述问题，Yamada I 等^[5]基于大规模语料进行预训练，生成实体的向量表示，通过双向编码器表征（bidirectional encoder representations from transformers, BERT）编码局部和全部实体信息进一步增强实体链接的准确率；Hamel O 等^[6]提出名为 LinKED - S2S 的实体链接方法，在 DBpedia 等通用数据集测试取得良好效果。

2.2 基于结构信息的关联算法

图神经网络（graph neural networks, GCN）是基于结构信息关联算法的主要代表。Peng H 等^[7]将文本信息转化为文本图，通过 GCN 实现文本分类；Yao L 等^[8]利用文档中词语共现关系丰富文本语义，并利用文本图卷积神经网络提高分类性能；Huang L 等^[9]建立具有全局共享参数的文本图存储全局信息。对于实体链接任务，李丽双等^[10]针对中文医学实体链接问题提出并行图神经网络；Globerson A 等^[11]将循环信念传播用于集体推理；Wang Z 等^[12]基于知识三元组，利用 GCN 学习图节点向量表示，进行实体相似性度量。

上述研究扩展了实体链接的技术路径，但现有方法仍存在两个主要局限：一是目前基于 GCN 的方法多依赖词共现和序列特征构建文本图，未能充分

挖掘深层次的语义信息；二是在模型推理过程中，未能有效利用已确定的实体链接关系信息。

3 多维动态图神经网络实体链接模型构建

本研究提出的多维动态图神经网络实体链接模型结构，见图 1，主要包括多维文本图构建、基于动态图神经网络的全局关联、基于注意力（attention）机制的序列关联生成 3 个主要部分。多维文本图构建用于将原始临床诊断信息转化为文本图^[13]；基于动态神经网络的全局关联利用动态图神经网络对转化后的文本图进行全局关联；最终在第 3 部分利用注意力机制完成序列关联生成与结果输出，同时将模型生成的实时关联结果信息输入动态神经网络中。

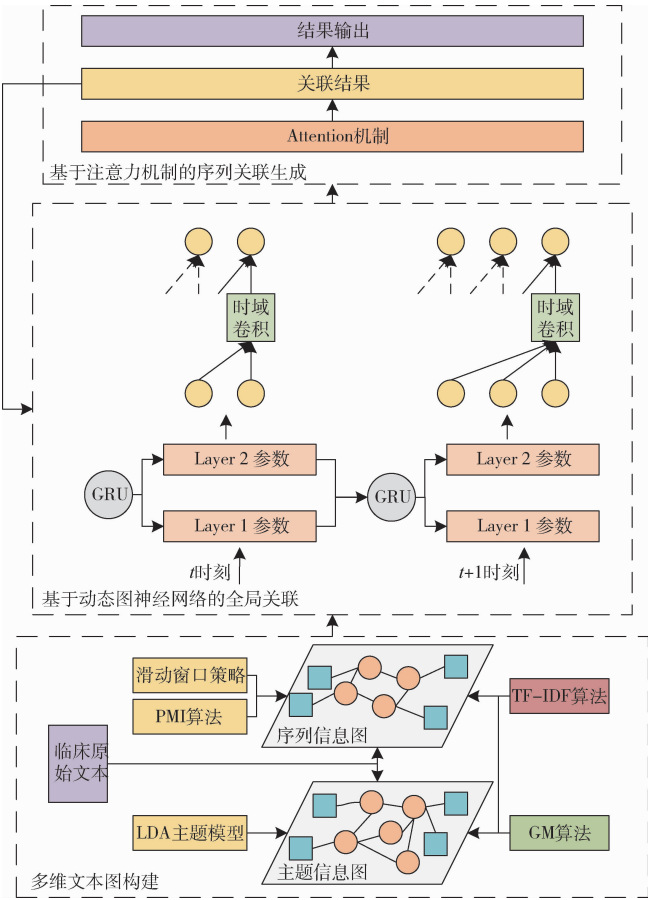


图 1 多维动态图神经网络实体链接模型结构

3.1 多维文本图构建

设待处理语料库 C ，包含 m 个文档，文档集合

为 $D = \{D_1, D_2, \dots, D_m\}$ ，包含 n 个不重复的词，每个词表示为 W_i ，词语集合 $W = \{W_1, W_2, \dots, W_n\}$ 。对语料库 C 构建图 $G = (V, E, A)$ 。 $V(|V| = q)$ 表示图 G 中的节点集，节点数量为 $q = m + n$ ， E 表示图 G 中边的集合， $A \in R^{q \times q}$ 表示图 G 的邻接矩阵，见图 2。

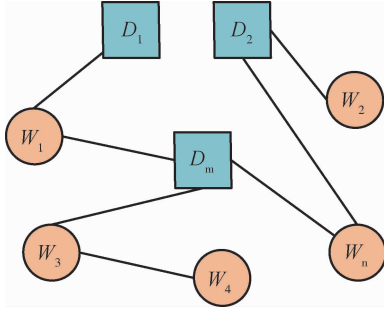


图 2 文本图示意

3.1.1 序列信息图构建 序列信息图侧重文本中的序列上下文关系。在自然语言中，文字的排列顺序在一定程度上影响其表达的含义。使用滑动窗口 (sliding window, SW) 策略^[14] 构建文本序列信息图，利用点互信息法^[15] 计算两个词节点之间的连线权重 $EW_{\text{seq}}(w_i, w_j)$ 。其中， $p(w_i, w_j)$ 表示词对 w_i 和 w_j 在同一 SW 出现的概率， $N(w_i, w_j)$ 表示同时包含词对 w_i 和 w_j 的 SW 数， $N_{\text{windows}}(w_i, w_j)$ 表示 SW 的总数， $p(w_i)$ 表示 w_i 出现在 SW 中的概率， $N_{\text{windows}}(w_i)$ 表示包含 w_i 的 SW 数。

$$EW_{\text{seq}}(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \quad (1)$$

$$p(w_i, w_j) = \frac{N(w_i, w_j)}{N_{\text{windows}}(w_i, w_j)} \quad (2)$$

$$p(w_i) = \frac{N(w_i)}{N_{\text{windows}}(w_i)} \quad (3)$$

3.1.2 主题信息图构建 主题信息图 G^{top} 反映文本中的主题关联度。利用隐含狄利克雷分布 (latent Dirichlet allocation, LDA)^[16-17] 获取语料库的潜在主题分布 $T = \{t_1, t_2, \dots, t_k\}$ ，通过词语的相关性度量表示边连权重 $EW_{\text{top}}(w_i, w_j)$ ，通过困惑度 (perplexity)^[18] 测度最优主题数 k 。其中， $w\lambda_i$ 表示词语 w_i 的相关性度量， U 表示 w_i 和 w_j 出现在同一 Topic 的次数； W 为语料库中所有词语的集合， M 为语料库中文档的数量， W_d 为文档 d 中的词， N_d 表示每个文

档中的词数， $p(W_d)$ 表示文档中词 w 出现的概率。

$$EW_{\text{top}}(w_i, w_j) = w\lambda_i + w\lambda_j - U + 1 \quad (4)$$

$$\text{perplexity}(W) = \exp \frac{\sum_{d=1}^M \log p(W_d)}{\sum_{d=1}^M N_d} \quad (5)$$

3.1.3 边连与节点权重设置 对于文档-词语边，通过词频-逆文档频率 (term frequency-inverse document frequency, TF-IDF)^[19] 衡量词与文档之间的边连权重。其中， F_{TF} 表示词频， $n_{i,j}$ 为词语 w_i 在文档 D_j 中出现的次数， $\sum_k n_{k,j}$ 表示文档 D_j 中所有词汇出现次数的总和。 F_{IDF} 为逆文档频率， $|D|$ 为语料库中的文档总数， N_{w_i} 表示包含词语 w_i 的文档数目。

$$F_{\text{TF}_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (6)$$

$$F_{\text{IDF}_{i,j}} = \log \frac{|D|}{1 + N_{w_i}} \quad (7)$$

$$F_{\text{TF-IDF}_{i,j}} = F_{\text{TF}_{i,j}} \times F_{\text{IDF}_{i,j}} \quad (8)$$

$$N_{w_i} = |\{j: w_i \in d_j\}| \quad (9)$$

在既往研究基础上，利用引力模型 $GM^{[20]}$ 表征各图中的节点权重，节点 v_i 的权重 VW_i 计算方式如下。其中， vk_i 表示与节点 v_i 相连的边数，即节点 v_i 的度 (degree)， d_n 表示节点 v_i 和 v_j 之间的最短距离， VW 正比于节点重要性，即 VW 的值越大节点的重要程度越高。

$$VW_i = \sum_{i \neq j} \frac{vk_i \times vk_j}{d_n^2} \quad (10)$$

3.2 基于动态图神经网络的全局关联

为弥补传统实体链接模型逐个链接、未能充分利用实体间关系的不足，引入反馈动态机制。即通过空间和时域双层图卷积结构，将已链接信息添加到后续任务，充分表征实体间潜在关系，捕获图的动态演变过程。本研究选取 BERT 进行语义表示，通过余弦相似度算法完成实体间局部关联，为后续 GCN 中的空间卷积提供初始数据。

3.2.1 空间卷积层 引入动态反馈机制，将已确定链接关系的实体补充到文本图中，进一步扩充文本信息。设图 $G = \{G_1, G_2, \dots, G_T\}$ ，其中 $G_t = (V_t, E_t)$ 为 t 时刻 G 的状态， E_t 为 t 时刻边的集合。通过 GCN

学习不同时刻的节点表示,第 l 层的节点嵌入更新 H^l , 以及 v_i, v_j 邻接矩阵权重 A_{ij} 的计算方式如下。其中, $\tilde{A} = A + I_n$ 表示合并后的矩阵, I_n 为单位矩阵, $\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$ 表示一个对称标准化拉普拉斯算子, $W^l \in R^{d_l \times d_{l+1}}$ 是第 l 层的可训练权重矩阵, σ 为非线性激活函数。

$$H^l = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{l-1} W^l) \quad (11)$$

$$A_{ij} = VW_i \times EW_{ij} \times VW_j \quad (12)$$

将不同时间步中 GCN 的每层参数视为一个序列,通过门控循环单元结构 (gated recurrent unit, GRU) 的更新机制学习 GCN 中的权重变化。其中 w_t^l 为 t 时刻 GCN 第 l 层的参数, f 为更新函数,利用 GRU 中的更新门 Z 和重置门 R 更新参数。其中 Z_t^l 为 t 时刻第 l 层 GRU 的更新门,用于确定更新信息的内容。

$$w_t^l = (X_t^{l-1}, W_{t-1}^l) \quad (13)$$

$$f(X_t^{l-1}, W_{t-1}^l) = (1 - Z_t^l) W_{t-1}^l + Z_t^l \tilde{W}_t^l \quad (14)$$

$$Z_t^l = \sigma(U_z^l X_t^{l-1} + V_z^l W_{t-1}^l + B_z^l) \quad (15)$$

随后利用重置门 R_t^l 对 t 时刻需要遗忘的信息进行选择。

$$R_t^l = \sigma(U_r^l X_t^{l-1} + V_r^l W_{t-1}^l + B_r^l) \quad (16)$$

最终通过更新门与重置门对 GCN 中的参数进行更新。其中 U 为模型输入, W 为权值矩阵, B 为偏置向量。

$$\tilde{W}_t^l = \tanh(U_w^l X_t^{l-1} + V_w^l (R_t^l W_{t-1}^l) + B_w^l) \quad (17)$$

$$W_t^l = (1 - Z_t^l) W_{t-1}^l + Z_t^l \tilde{W}_t^l \quad (18)$$

3.2.2 时域卷积层 通过时域卷积聚合模型信息,利用扩张卷积扩大时域卷积的作用域,增强模型对长距离依赖关系的提取能力。

$$F(x) = \sum_{i=0}^{k-1} f(i) \cdot X_{x-d-i}^l \quad (19)$$

其中 X_{x-d-i}^l 为第 l 层 GCN 的输出。采用 $ReLU$ 作为第 1 层 GCN 的激活函数, $softmax$ 作为第 2 层 GCN 的激活函数,在两层 GCN 中增加 dropout 层,防止过拟合现象。通过 GCN,每个节点均聚合了其邻域节点中的全部信息,进一步丰富了语义。

$$Z = \text{softmax}(\hat{A} \text{ReLU}(\hat{A} X W_0) W_1) \quad (20)$$

3.3 基于注意力机制的序列关联生成

利用注意力机制在已链接实体集合中优先选取

对模型最终效果影响最明显的实体。具体支持得分的计算方式如下。其中 $\hat{e}_i$ 表示已链接的实体, $H^{e_{l+1}}$ 为待链接的实体, $X_i^{e_i}$ 代表实体 e_i 通过动态 GCN 处理后的实体表示。

$$\mu(\hat{e}_i) = \max_{e_{l+1} \in e_{l+1}} H^{e_{l+1}} \cdot D \cdot X_i^{e_i} \quad (21)$$

$$\alpha(\hat{e}_i) = \frac{\exp[\mu(\hat{e}_i)]}{\sum_{e_j} \exp[\mu(\hat{e}_j)]} \quad (22)$$

随后计算未连接实体集合的相似度并更新全局得分。

$$\theta(e_{l+1}^j, S_l) = \sum_{e_i} \alpha(\hat{e}_i) \cdot X_i^{e_i} \cdot R \cdot H^{e_{l+1}^j} \quad (23)$$

4 实证研究

基于北京协和医院提供的真实临床疾病诊断数据构建实证数据集。专业编码人员依据《国际疾病分类第十次修订本 (ICD-10) 北京临床版》匹配原始临床诊断与 ICD 诊断词条,质控人员进行质控,共得到 105 928 条诊断数据。依据《涉及人的生命科学和医学研究伦理审查办法》第三十二条,北京协和医院伦理审查委员会免除本研究伦理审查,伦理证明编号为 1-25ZM0038。

按 7:2:1 划分训练集、测试集和验证集。以编码人员人工匹配结果为金标准,用准确率 (precision, P)、召回率 (recall, R) 和 $F1$ - Measure ($F1$) 评估实验结果,采用 5 折交叉验证法。实验环境与参数设置^[21-22],见表 1。

表 1 实验环境设置

实验环境	参数
GPU	GeForce RTX 1050Ti
CPU	E5-2678V3
开发环境	Python3.8.3、TensorFlow2.4.1
Epoch	50
LSTM 学习率	0.002
L2 正则化参数	0.000 1
Dropout	0.5
LDA - Alpha	50/K
LDA - eta	0.02

4.1 消融实验结果与分析

通过控制变量设计，验证构建多维文本图以及引入动态反馈机制的必要性与有效性，见表 2。其中 w/o 表示无， G^{seq} 如 (w/o) 表示模型中仅缺少主题信息图。

表 2 消融实验结果 (%)

模型	P	R	F1
G^{seq} (w/o)	88.28	87.78	88.03
G^{top} (w/o)	89.25	89.03	89.14
节点权重 (w/o)	89.84	89.27	89.55
动态反馈机制 (w/o)	88.69	88.54	88.61
本文方法	90.34	89.72	90.03

对比结果显示本文方法取得了最优结果。同时，各类文本图对模型的贡献也存在一定差异，从侧面验证了融合多维度文本信息的必要性，其中序列信息图对结果的影响最大。对比是否加入动态反馈机制的结果，该机制相较于传统 GCN 可以进一步丰富有效文本信息，提升综合性能。

4.2 对比实验结果与分析

本文方法与基线方法的对比结果，见表 3。本文方法的综合表现优于对比方法。

表 3 模型性能与对比实验结果 (%)

模型	P	R	F1
BiLSTM	82.48	82.88	82.68
SciBERT - BiGRU - CNN	85.56	84.96	85.26
TextGCN	86.74	87.13	86.93
Zhou X 等 ^[23]	87.16	86.32	86.74
本文方法	90.34	89.72	90.03

对于 BiLSTM 和 SciBERT - BiGRU - CNN 两个主要基于序列信息的模型，SciBERT - BiGRU - CNN 能通过预训练的 SciBERT 增强文本表示，利用 BiGRU 获取上下文语义信息，同时通过 CNN 获取文本局部关键特征，因而相较于 BiLSTM 表现更好。TextGCN 同样将文本视为图结构数据，相较于传统

深度学习模型，召回率明显提高，但该模型仅构建词共现、序列特征的文本图，语义提取能力相对有限。本研究通过多维文本图丰富语义信息，提升模型表现。Zhou X 等^[23]的研究用排序算法优化模型，却忽略了模型处理过程中被关联知识的利用，限制了最终综合表现。如在错误样例中，乳腺癌表达存在乳房和乳腺混用现象，部分医生描述肿瘤位置时习惯用括号补充。如“乳房恶性肿瘤（外上象限）”，部分医生习惯用“乳腺外上恶性肿瘤”表达，本研究模型能识别更多样表达方式。

5 结语

本研究针对现阶段 GCN 文本信息利用不足、现有研究未利用模型运行中已被链接实体信息的问题，在传统共现信息基础上补充序列和主题信息文本图，构建包含动态机制的 GCN 生成最终实体链接信息。通过对比与消融实验，验证了改进策略的有效性与必要性。本研究在一定程度上丰富了现阶段医学实体链接方法，但也存在不足。未来将进一步研究多维度语义融合方法，通过空域卷积模型降低模型运算成本，并引入大语言模型提升性能和效果。

作者贡献：李博诚负责研究设计、实验实施、论文撰写；李乃适负责完善研究方案、论文修订；周婧雅负责完善研究方案、构建数据集；王怡负责论文选题、审核与修订。

利益声明：所有作者均声明不存在利益冲突。

参考文献

- 彭琳, 宋珺, 熊玲珠, 等. 医学领域知识融合研究进展 [J]. 计算机工程与应用, 2024, 60 (9): 48 - 64.
- 张玥, 李韧, 杨建喜, 等. 基于深度神经网络的实体链接研究综述 [J]. 中文信息学报, 2024, 38 (8): 1 - 14.
- 李晓瑛, 李军莲, 邓盼盼, 等. 医学知识组织系统构建研究与应用实践 [J]. 数字图书馆论坛, 2020 (7): 30 - 35.
- 向宇, 郭云龙, 徐潇, 等. 多策略中文微博实体词消歧及实体链接 [J]. 计算机应用与软件, 2016, 33 (8): 12 - 17, 61.
- YAMADA I, WASHIO K, SHINDO H, et al. Global entity

- disambiguation with BERT [C]. Seattle: The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022.
- 6 HAMEL O, FAREH M. Deep sequence to sequence semantic embedding with attention for entity linking in context of incomplete linked data [J]. Engineering applications of artificial intelligence, 2024, 134 (8): 108689.
- 7 PENG H, LI J, HE Y, et al. Large - scale hierarchical text classification with recursively regularized deep graph - cnn [C]. Lyon: The 2018 World Wide Web Conference, 2018.
- 8 YAO L, MAO C, LUO Y. Graph convolutional networks for text classification [C]. Hawaii: The AAAI Conference on Artificial Intelligence, 2019.
- 9 HUANG L, MA D, LI S, et al. Text level graph neural network for text classification [EB/OL]. [2025 - 01 - 24]. <https://arxiv.org/abs/1910.02356>.
- 10 李丽双, 王泽昊, 秦雪洋, 等. 基于平行交互注意力网络的中文电子病历实体及关系联合抽取 [J]. 中文信息学报, 2024, 38 (6): 108 - 118.
- 11 GLOBERSON A, LAZIC N, CHAKRABARTI S, et al. Collective entity resolution with multi - focal attention [C]. Berlin: The 54th Annual Meeting of the Association for Computational Linguistics, 2016.
- 12 WANG Z, LV Q, LAN X, et al. Cross - lingual knowledge graph alignment via graph convolutional networks [C]. Brussels: The 2018 Conference on Empirical Methods in Natural Language Processing, 2018.
- 13 宋豪俊, 李燕, 刘悦悦, 等. 融合子图结构的医学知识推理方法综述 [J]. 医学信息学杂志, 2025, 46 (1): 63 - 68, 92.
- 14 史文艺, 朱欣娟. 融合 BERT 和自注意力机制的张量图卷积网络文本分类 [J]. 计算机系统应用, 2025, 34 (3): 152 - 160.
- 15 WILLIAMS C K I. On suspicious coincidences and pointwise mutual information [J]. Neural computation, 2022, 34 (10): 2037 - 2046.
- 16 BLEI D M, NG A, JORDAN M I. Latent Dirichlet allocation [J]. The journal of machine learning research, 2003, 3 (1): 993 - 1022.
- 17 GU R, LIN L. Application of latent Dirichlet allocation and autoencoder to real estate datasets [J]. The journal of supercomputing, 2025, 81 (1): 1 - 24.
- 18 HUANG H, ZHU D, WANG X. Evaluating scientific impact of publications: combining citation polarity and purpose [J]. Scientometrics, 2021, 127 (9), 5257 - 5281.
- 19 张敏, 李唯, 范青. 基于语义信息的术语加权算法提升科技文献检索的准确性 [J/OL]. 图书馆杂志, 1 - 18 [2025 - 02 - 17]. <http://kns.cnki.net/kcms/detail/31.1108.g2.20241008.1725.011.html>.
- 20 YANG S, LIU Y. Short text classification method by fusing corpus features and graph attention network [J]. Journal of computer applications, 2022, 42 (5): 1324.
- 21 高永杰, 党建武, 张希权, 等. 基于互信息自适应的多模态实体对齐方法 [J]. 计算机应用研究, 2025, 42 (1): 106 - 110.
- 22 车超, 刘迪. 基于双向对齐与属性信息的跨语言实体对齐 [J]. 计算机工程, 2022, 48 (3): 74 - 80.
- 23 ZHOU X, MIAO Y, WANG W, et al. A recurrent model for collective entity linking with adaptive features [C]. Louisiana: The AAAI Conference on Artificial Intelligence, 2020.

《医学信息学杂志》开通微信公众号

《医学信息学杂志》微信公众号现已开通, 作者可通过该平台查阅稿件状态; 读者可阅览当期最新内容、过刊等; 同时提供国内外最新医学信息研究动态、发展前沿等, 搭建编者、作者、读者之间沟通、交流的平台。可在微信添加中找到公众号, 输入“医学信息学杂志”进行确认, 也可扫描右侧二维码添加, 敬请关注!



《医学信息学杂志》编辑部