融合知识图谱与注意力机制的产学合作推荐研究

程楠楠[1] 彭吉琼[1] 吴 璇[1]
CHENG Nannan PENG Jiqiong WU Xuan

摘 要 针对产学合作数据稀疏导致推荐精准性和解释性不高问题，提出了一种基于融合知识图谱与领域感知注意力机制的产学合作推荐模 IUCR-SKG，旨在为企业用户推荐合适的科研团队，促进产学合作效率。首先构建了产学合作知识图谱和合作关系图，然后合并异构图并使用 RotatE 技术优化嵌入表示，接着采用基于注意力机制的领域融合方法扩展用户实体，最后使用多层感知机循环叠加结构预测企业用户和科研团队发生合作的概率。以“人工智能 AI”领域为例，所提出的模型性能相较于基准模型均有显著提升，且注意力机制将邻域信息融合，扩展了企业用户的合作关系，使得推荐效果相较其他注意力机制的嵌入算法效果也得到一定的改善。

关键词 知识图谱；注意力机制；产学合作推荐

doi：10.3969/j.issn.1672-9528.2024.02.014

0 引言

在现代技术经济中，对于企业，特别是技术密集型的中小企业来说，紧靠内部的知识创造提升创新能力从而提升企业竞争力是相对不够的，还需要依靠外部合作，特别是与高校的技术合作 [1]。然而在实际推进产学合作的进程中，面对技术信息的庞杂，如何进行技术机会的识别和合作伙伴的选取是亟待解决的问题 [2]。

当下知识信息爆炸的时代，技术信息，特别是高校的学术资源也呈现爆发式增长，面对海量且庞杂的数据，推荐系统是最高效的资源利用手段之一。但高校的学术资源数量大、类型多，有一定的异质性，而企业的创新需求又面临多样性、模糊性等特点，因此传统的推荐算法会面临产学合作关系数据稀疏问题 [3]。知识图谱辅助信息推荐算法能有效缓解该类问题，提升推荐结果的准确性和解释性 [4]。为了促进产学合作效率、提高产学合作推荐的准确性，本文深入分析知识图谱在产学合作推荐中的技术优势和应用场景，旨在为企业用户提供围绕技术主题场景的合作伙伴推荐方案。

1 相关工作

关于产学合作推荐研究可以分为基于合作的主题内容推荐、基于合作关系网络的推荐两大类。

基于主题内容的推荐主要是借助文本分析技术等表征产学合作的技术主题，计算和用户需求的匹配程度，并将其高分结果推荐给目标用户。Takahiro 等人 [5] 提取学者出版物和学者主页作为学者档案，为用户推荐合适的合作者，在学者主题知识提取的过程中，出版物的时间表征学者某段时间的研究兴趣。Zhang 等人 [6] 通过引入出版物的时间对通过主题模型得到的学者主题向量进行指数加权，通过数据驱动的监督学习模型提取数据中固有特征得到学者的特征表示。闫晓慧等人 [7] 融合专利和论文信息的内容挖掘对企校创新合作推荐研究。

基于合作关系网络的推荐将合作关系映射到网络图中，节点代表作者，边代表作者间的合作关系，合作推荐就是对尚未产生联系的节点之间未来产生链接的可能性进行预测，根据预测结果进行推荐。张金柱等人 [8] 基于深度学习的网络表示学习方法，学习研究者在所处网络里的语境信息，形成每个研究者的稠密、低维向量表示，最后通过向量相似度指标计算研究者间的语义相似度，实现合作预测和推荐。陈文杰 [9] 构建了一个基于超图结构的科研合作超网络，基于共同邻居和资源分配来构建超图的结构相似性指标，以实现科研合作推荐。

基于主题内容推荐的方法虽然解释性强且实现简单，但是仅使用文本的语义相似度去衡量用户的合作倾向不免偏颇，而且主题关键字确定严重依赖用户明确的需求和背景知识。基于合作关系网络的推荐会面临关系数据稀疏和冷启动问题。而知识图谱技术由于拥有丰富的实体、关系和属性知识，兼具内容推荐和关系网络推荐的优势，可有效缓解上述

1. 江西科技学院 江西南昌 330098
[基金项目] 江西科技学院自然科学项目（ZR2104）；江西省教育厅科学技术研究项目（GJJ2202609）

问题。李锴君等人[10]提出基于学术知识图谱和主题特征嵌入的论文推荐方法。知识图谱融入推荐算法目前主要有三种方式[11]：基于嵌入的方法、基于路径的方法和基于图神经网络的方法。

2 研究设计

针对产学合作数据稀疏进而推荐过程未充分捕捉高校产学合作属性信息和合作网络关系，最后导致推荐精准性和解释性不高的问题，本文提出了一种基于融合知识图谱与领域感知注意力机制的产学合作推荐模（industry-university collaborator recommendation based on subject knowledge graph，IUCR-SKG）。如图 1 所示，模型框架由 3 部分组成：（1）产学合作知识联合异构图构建：构建知识图谱，并将其和用户 - 项目二部图组成异构图；（2）特征表示：将知识图谱融合的异构图进行向量化建模，并采用基于注意力机制的领域融合方法扩展用户表示；（3）推荐预测：使用多层感知机循环叠加结构学习企业用户 - 合作伙伴的交互关系，计算得分并排序，最后给出企业用户偏好的合作伙伴列表 $\{c_1,c_2,\cdots,c_{10}\}$。

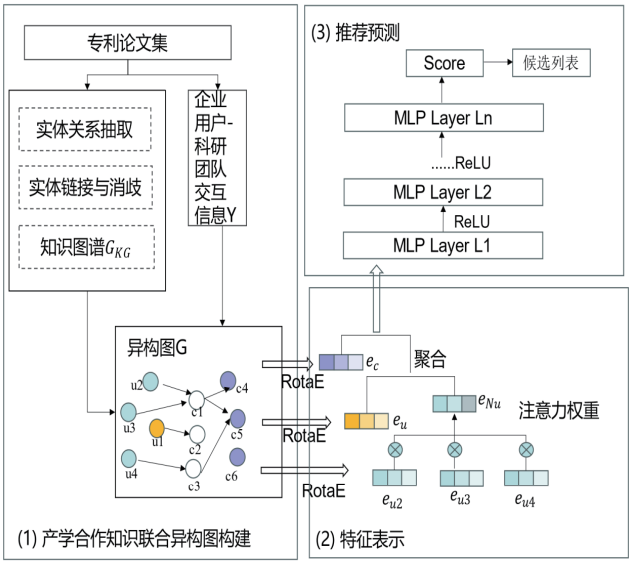

图 1 企业科研合作者推荐模型 IUCR-SKG 框架图

2.1 符号及问题定义

企业用户集为 $\boldsymbol{U}=\{u_1,u_2,u_3,\cdots\}$，项目集是科研合作团队 $\boldsymbol{C}=\{c_1,c_2,c_3,\cdots\}$，团队名取其高校或者科研院所机构在学术资源中排名第一的姓名。企业用户与产学合作者的交互矩阵 $\boldsymbol{Y}=\{y_{uc}\,|\,u\in U,c\in C\}$，当 y_{uc}=1 时，表示企业用户与科研合作者之间存在交互合作；当 y_{uc}=0 时，表示没有交互。根据交互矩阵 $\boldsymbol{Y}$ 表示企业用户 - 产学合作者的交互图 $G_{UC}=\{(u,c)\,|\,y_{uc}=1\}$。

产学合作知识图谱 $G_{KG}=\{(h,r,t)\,|\,h,\mathrm{t}\in E,r\in R\}$，其中 h、t 分别表示头实体和尾实体，r 表示实体之间的关系。E 表示知识图谱实体集合。由于 G_{UC} 中的科研项目团队 C 是产学合作知识图谱 G_{KG} 的实体 E 的子集，因此可将两图构建企业用户 - 科研团队 - 实体异构图 $G=\{(h,r,t)\,|\,h,t\in E',r\in R'\}$，其中 $E'=E\cup U,R'=R\cup R_{UC}$，此时将知识图谱中的辅助信息融合进项目集。

IUCR-SKG 模型的任务是：输入企业用户项目矩阵 $\boldsymbol{Y}$ 和知识图谱 G_{KG}，输出一个可学习的偏好预测函数 $F=(u,c\,|\,Y,G_{KG},\theta)$ 来预测企业用户 u 与科研合作团队 c 发生合作的可能性，其中 θ 表示模型的参数集。

2.2 产学合作知识图谱

产学合作知识图谱实体主要为 5 类，分别是科研团队（取该组织的第一作者姓名为团队名称）、组织（科研团队所在的组织，可以是大学或者研究所）、主题（科研团队研究的主题领域）、合作企业（共同发表专利和论文的某个企业公司）、合作组织（共同发表专利和论文的大学或研究所）。这 5 类实体广泛存在于各大科研数据库的免费公开信息中，在具体操作中，主要提前对文本数据的主题进行处理建模。本文使用潜在狄利克雷分布模型（latent Dirichlet allocation, LDA），对论文专利的题名、关键词和摘要数据进行预处理训练提取科研团队的主题分布。因此各实体之间的关系如下：科研团队工作于某个组织机构，科研团队研究主题领域，科研团队与合作组织是学学交互，与合作企业之间是产学合作。

2.3 特征表示

2.3.1 特征表示

企业用户 - 科研团队 - 实体异构图包含多种类型的节点和关联关系，因此每种类型节点有不同的特征空间。本文采用 RotatE 模型[12]来为节点和关系生成低维嵌入表示，该方法能够建模和推断各种关系模式，包括对称 / 反对称、反演和合成。不同于传统的嵌入表示，它采用一种自对抗负采样方法代替传统负采样方法，根据特定的概率分布和当前已嵌入的模型来采样负样本，因此其对负样本的区分能力更强，模型的泛化能力也更好。RotatE 利用了欧拉公式，将关系视为从头实体向尾实体的旋转。对于公式（1）的 RotatE 的距离如公式（2）所示。

$$t_i=h_i r_i\ ,\ where\ h_i,r_i,t_i\in R \tag{1}$$

$$d_r(h,t)=\|\,h\circ r-t\,\| \tag{2}$$

式中：$\circ$ 运算为 Harmad product，具体的运算法则为：将给定向量 / 矩阵的同行同列对应的元素乘在一起，形成一个新的向量 / 矩阵。当三元组越接近事实，头尾实体距离越接近时，

得分函数值就越低。

2.3.2 基于领域感知的注意力机制

企业科研合作者推荐模型 IUCR-SKG 的核心理念是利用知识图谱 KG 中实体之间的相关关联关系来学习实体表示，最大化挖掘知识信息。因此，对上文构建的每个子企业用户 - 科研团队 - 实体异构图 G，使用注意力机制将邻居节点的领域信息融合到中心节点的嵌入表示中。Self-attention 注意力机制不仅关注当前位置，还能通过位置编码获取上下文语义，其计算公式为：

$$Attention(Q,K,V)=Soft\max(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 是三个相似度匹配矩阵，是 Self-attention 内部三个相同尺寸的权重矩阵与嵌入向量表示相乘得来的。d_k 表示当前向量 $\boldsymbol{k}$ 的维度，做降维使用。

产学合作是一个复杂且持续的场景，科研团队的合作经验相对其科研能力在产学合作伙伴的选择时更为重要 [13]。因此，企业用户 u 对某个科研伙伴 c 的偏好程度可以用科研伙伴 c 与用户 u 周围邻居之间的配对相似度表示。运用公式（4）进行领域感知注意力机制计算，得到偏好系数。

$$\Pi_h(h,r,t)=G(\mathrm{e}_t,\{e_t\,|\,t\in N_h\})=g(\{f(e_t,e_{t'})\,|\,t\in N_h\}) \tag{4}$$

式中：e_t 表示头实体，N_h 表示头实体的邻居集，$e_{t'}$ 表示当前头实体的邻居尾实体的嵌入表示，$G(\cdots)$ 是注意力函数，$g(\cdots)$ 是注意力池化函数，f 是成对注意函数，计算公式为：

$$f(e_t,e_{t'})=\mathrm{Re}LU(\cos(e_t\circ e_r,e_{t'})) \tag{5}$$

式中：e_r 表示关系实体的嵌入表示，使用余弦相似度计算后通过 ReLU 进行归一化处理。按系数计算后可得到邻居节点的嵌入表示为：

$$e_{N_h}=\sum_{(h,r,t)\in N_h}\pi_h(h,r,t)e_t \tag{6}$$

接下来，使用双交互聚合器 Fagg 进行计算，它可以更好地捕捉节点之间的关系。计算公式为：

$$Fagg=\mathrm{Re}LU(w_1\cdot(e_h+e_{N_h}))+\mathrm{Re}LU(w_2\cdot(e_h\circ e_{N_h})) \tag{7}$$

式中：W_1、W_2 是可训练的权重矩阵，e_h 是头实体的嵌入表示，e_{Nh} 是邻居节点，是 Hadamard 乘积计算。

进一步循环堆叠多个领域传播聚合层，以计算实体 l 阶后的向量，本文将 l 层的实体表示为：

$$e_h^l=Fagg(e_h^{l-1},e_{N_h}^{l-1}) \tag{8}$$

2.4 推荐预测

在通过聚合方法有效对企业用户实体与领域实融合建模后，接下来本文开始设置循环的叠加结构，也就是对融合后的企业用户向量和科研合作者向量进行拼接后，再多次进行非线性变换，输出结果为推荐得分，公式为：

$$\begin{aligned}&Z_1=\begin{bmatrix}\tilde{e}_u\\ e_c\end{bmatrix}\\&Z_2=\alpha_2(W_2{}^TZ_1+b_2)\\&\cdots\cdots\\&Z_l=\alpha_l(W_l{}^TZ_{l-1}+b_l)\end{aligned} \tag{9}$$

式中：α_l 表示激活函数，W 和 b 表示可训练的参数和偏移。最终的得分计算公式为：

$$Score=\delta(h^TZ_l) \tag{10}$$

式中：δ 是 sigmoid 归一化函数。

3 实验与结果分析

3.1 数据采集和数据预处理

本文以“人工智能 OR AI”关键词为例，在 CNKI、SooPAT 专利数据库中使用高级搜索获取近五年的核心期刊及以上论文和专利数据，主要获取字段为论文 / 专利名称、作者团队 / 发明人团队、单位 / 申请人机构、发表日期、关键词 / 主分类号、摘要等，完成后按照名称 + 作者团队去重处理，并将专利的主分类号按照其对应表翻译成关键词和论文的关键词合并，合并后使用文本主题 LDA 取 TOP1 主题词作为子主题领域。按照合作定义，本实验将单位 / 申请人机构超过 2 个的定义为合作，把有企业参与的合作类型定位为产学合作。作者团队 / 发明人团队数据全部处理为所在结构的第一作者的团队。结合《高等学校科技统计资料汇编》，给出各个高校所在的省份地区、科研院所和企业对其名称使用正则提取省份地区，剩下的手动补全，相关实体数量数据如表 1 所示。

表 1 实验相关数据统计

统计对象	实体	数量
数据集	企业用户	127 444
	科研团队	150 525
	合作交互	40 472
	产学合作关系占比	0.14
知识图谱	科研团队	150 525
	组织	45 157
	主题	202 525
	合作企业	37 450
	合作组织	30 104

另外，数据处理中为了样本的平衡性，将每条记录数据转换成企业用户 - 合作伙伴对作为正样本，然后随机抽取未与企业用户合作过且与正样本数量一致的合作伙伴作为负样本。利用 Neo4j 构建知识图谱，如图 2 所示，并将科研合作伙伴与知识图谱三元组中的实体建立对齐关系。

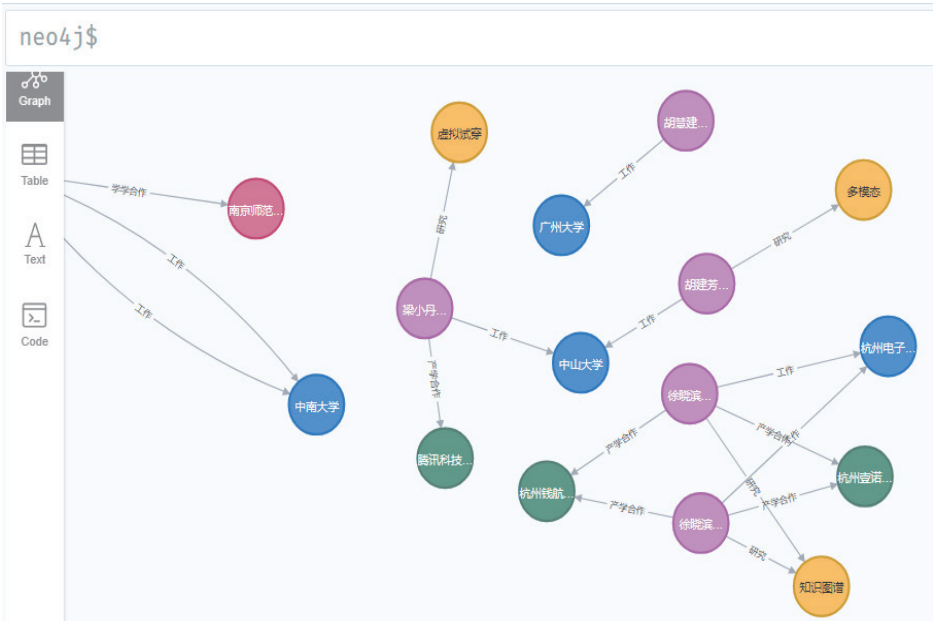

图 2 知识图谱部分截图

3.2 实验环境

本文环境配置如下：Windows 11 操作系统，显卡 NVIDIA CeForce RTX 3060 Ti，编程语言 Python 3.8。使用 Xavier 方法对可训练参数进行初始化。

3.3 对比模型

本实验选取以下模型进行实验模型的性能对比分析，其中一类为传统推荐模型，另一类为融合知识图谱的推荐模型。

BPRFM[14]：基于协同过滤的推荐，主要通过矩阵分解方式得到用户和项目的表示。

CBF[15]：基于内容的推荐，实验主要使用专利论文主题的相似度。

CKE[16]：一种基于嵌入的知识感知推荐方法，主要利用 TransR 学习项目的知识信息和结构表示。

DKN[17]：融入知识图谱信息的推荐模型，与本文不同，该方法对实体向量和领域均使用 CNN 进行特征提取表示。

3.4 评价指标

本文为 TOP-N 推荐场景，采用归一化折损累计增益 NDCG@K、Recall@K、HR@K 进行评价对比。其中，NDCG@K 主要评估模型的排序性能，测试集中与企业用户发生合作交互的正样例在推荐列表的位置越靠前，则该值越高，也就是符合企业用户偏好的，根据其位置累计增益，反之则增益为 0，最终根据排序位置计算折损。计算公式为：

$$NDCG@K=\frac{1}{I}\sum_{i=1}^{k}\frac{2^{r_i}-1}{\log_2(i+1)} \tag{11}$$

式中：r_i 是企业用户历史合作伙伴与待推荐伙伴之间的相似度，即产学合作伙伴在推荐列表位置 i 的相似度，I 表示按照最佳排序方案下所计算的折损累计增益。

Recall@K 召回率指标，表示如下，该值越高，表示推荐性能越好。

$$Recall@K=\frac{\sum_{u\in U}|P(u)\cap U(u)|}{\sum_{u\in U}|P(u)|} \tag{12}$$

式中：$P(u)$ 是测试集中用户 u 交互合作的正样例科研团队列表，$U(u)$ 是实验生成的用户 u 的推荐列表。HR@K 衡量前 K 个推荐的产学合作伙伴中，正确预测企业用户感兴趣的科研团队在其中所占的比例，主要评估排序的质量。

4 实验结果与分析

4.1 基准模型对比分析

在 K=10 时，本文模型与各基准模型的对比结果如表 2 所示。实验结果表明，CF 算法效果较差，在数据稀疏的场景下，该类算法推荐的偏差较大。CBF 相对 CF 有提升，用户对于科研伙伴的兴趣，首先要考虑科研伙伴的主题相似性。基于嵌入的 CKE 模型通过将矩阵分解和知识图谱简单结合在一定程度上提高了模型性能。使用了注意力机制的 DKN 比未使用注意力机制融合知识的模型效果更好。本文的 IUCR-SKG 模型能取得相对较优的效果，主要是其更好地捕捉除了知识信息外的结构信息，还使用注意力机制将邻域信息融合，扩展企业用户的合作关系，获得更加丰富的节点信息。

表 2 实验结果对比（K=10）

推荐模型	NDCG@K	Recall@K	HR@K
BPRFM	0.066 1	0.199 1	0.291 3
CBF	0.075 7	0.186 2	0.313 8
CKE	0.080 1	0.227 5	0.358 5
DKN	0.082 1	0.231 8	0.371 0
IUCR-SKG	**0.086 7**	**0.258 7**	**0.393 1**

4.2 消融实验

为了探究模型框架中各个模块的设置对最终推荐效果的影响，本文对以下三种变体进行消融实验，以验证模型的有效性。

（1）IUCR-SKG-1：去除 RotatE 模型的嵌入层设计。

（2）IUCR-SKG-2：不使用领域信息。

（3）IUCR-SKG-3：不使用领域感知的注意力机制，改用加权平均值的方式做权重。

从表 3 可以看出各个设置项对于模型效果的必要性，IUCR-SKG 模型通过 RotatE 先对原有的交互信息进行语义扩充，再根据领域感知注意力机制对目标节点进行邻居信息聚合，得到更加丰富的新实体语义信息，进而帮助稀疏关系和冷启动用户构建产学合作偏好，从而达到较好的推荐性能。

表 3 实验结果对比（K=10）

推荐模型	NDCG@K	Recall@K	HR@K
IUCR-SKG-1	0.085 9	0.248 6	0.378 9
IUCR-SKG-2	0.082 1	0.248 7	0.354 6
IUCR-SKG-3	0.083 7	0.248 3	0.378 4
IUCR-SKG	**0.086 7**	**0.258 7**	**0.393 1**

5 结论

针对产学合作关系数据稀疏问题，本文提出了一种提高产学合作科研伙伴推荐效果的模型框架，该方法融合了来自产学合作知识图谱的实体关系结构特征进行推荐预测。同时，使用 RotatE 技术对嵌入模型进行优化，采用了领域感知注意力机制对给定节点的邻居进行有效聚合，充分利用节点信息。在“人工智能 AI”领域的产学合作推荐数据上，推荐算法 IUCR-SKG 相较于传统的协同过滤、CB、CKE 均有显著提升，而且使用注意力机制将邻域信息融合，扩展企业用户的合作关系，获得更加丰富的节点信息，使得推荐效果相较于 DKN 也有一定的提升。

参考文献：

[1] RAN C J, SONG K, YANG L. An improved solution for partner selection of industry-university cooperation [J]. Technology analysis & strategic management, 2020,4:1-15.

[2] CHUNG J, KO N, YOON J. Inventor group identification approach for selecting university-industry collaboration partners[J].Technological forecasting and social change, 2021, 171:120988.1-120988.11.

[3] 唐浩 . 基于领域知识图谱的学术资源推荐算法研究 [D]. 宁波 : 宁波大学 ,2020.

[4] 陈珊珊 , 姚苏滨 . 基于知识图谱与邻域感知注意力机制的推荐算法研究 [J/OL]. 计算机科学 ,1-16[2023-09-18].http://kns.cnki.net/kcms/detail/50.1075.TP.20231113.1006.008.html.

[5] TAKAHIRO K, DAIKI T. Proposal of a hybrid recommendation algorithm to support the discovery for mashup applications[J]. Journal of information technology & software engineering, 2021,11:1-5.

[6] ZHANG Q, MAO R, LI R. Spatial–temporal restricted supervised learning for collaboration recommendation[J]. Scientometrics, 2019,119(3): 1497-1517.

[7] 闫晓慧 , 马博闻 , 邓三鸿 , 等 . 融合专利与论文信息的内容挖掘和引用基础的企校创新合作推荐研究 [J]. 现代情报 , 2023, 43(3):13-25.

[8] 张金柱 , 于文倩 , 刘菁婕 , 等 . 基于网络表示学习的科研合作预测研究 [J]. 情报学报 , 2018,37(2):132-139.

[9] 陈文杰 . 基于超图的科研合作推荐研究 [J]. 数据分析与知识发现 , 2023, 7(4):68-76.

[10] 李锴君 , 牛振东 , 时恺泽 , 等 . 基于学术知识图谱及主题特征嵌入的论文推荐方法 [J]. 数据分析与知识发现 , 2023, 7(5): 48-59.

[11] YANG Y, HUANG C, XIA L, et al. Knowledge graph contrastive learning for recommendation[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2022: 1434-1443.

[12] SUN Z, DENG Z H, NIE J Y, et al. Rotate: knowledge graph embedding by relational rotation in complex space[EB/OL].(2019-02-26)[2023-09-27]. http://arXiv preprint arXiv:1902.10197.

[13] 程楠楠 , 金欢 . 本科高校产学合作效能测度及影响因素研究 [J]. 中国高校科技 ,2023(7):10-15.

[14]RENDLE S, FREUDENTHALER C, GANTNER Z, et al. BPR: bayesian personalized ranking from implicit feedback[EB/OL].(2012-05-09)[2023-10-11].http:// arXiv preprint arXiv:1205.2618.

[15] BAI X, WANG M, LEE I, et al. Scientific paper recommendation: a survey[J].IEEE access,2019,7:9324-9339.

[16] ZHANG F, YUAN N J, LIAN D. Collaborative knowledge base embedding for recommender systems[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY: ACM. 2016:353-362.

[17] WANG H, ZHANG F, XIE X, et al. DKN: deep knowledge-aware network for news recommendation[C]//Proceedings of the 2018 World Wide Web Conference on World Wide Web. Lyon, France. 2018:1835-1844.

【作者简介】

程楠楠（1987—），女，江苏南通人，高级工程师，博士，研究方向：知识推荐算法、机器学习、教育大数据。

彭吉琼（1988—），女，江西南昌人，讲师，硕士，研究方向：教育大数据、产学合作研究。

吴璇（1995—），女，江西南昌人，助教，硕士，研究方向：大数据分析与应用。

（收稿日期：2023-11-30）