doi: 10.13413/j.cnki.jdxblxb.2024009

基于知识嵌入技术的制度文件推荐算法

李 鑫[1], 王文迪[2], 张 伟[3], 冯 浩[4], 韩霄松[2]
(1. 吉林大学 原子与分子物理研究所, 长春 130012; 2. 吉林大学 计算机科学与技术学院, 长春 130012;
3. 吉林大学 后勤处, 长春 130012; 4. 吉林大学 政策与法律办公室, 长春 130012)

摘要: 针对制度文件推荐过程中传统算法存在准确率和推荐效率较低的问题, 提出一种基于知识嵌入的文本推荐算法. 通过将知识图谱中的知识转化为特征向量, 并结合神经网络模型, 有效提高了推荐系统在处理海量多样数据时的准确性和稳定性. 实验结果表明, 该算法在面对冷启动和用户兴趣多样化的情况下, 表现出了优于传统方法的推荐精度和稳定性, 为大规模推荐系统中的数据稀疏问题和个性化需求提供了更高效和可靠的解决方案, 有助于改善用户体验并提升推荐系统的整体性能.

关键词: 知识图谱; 词嵌入; 图嵌入; 神经网络
中图分类号: TP181 **文献标志码**: A **文章编号**: 1671-5489(2024)06-1377-07

Recommendation Algorithm for Institutional Documents Based on Knowledge Embedding Technology

LI Xin[1], WANG Wendi[2], ZHANG Wei[3], FENG Hao[4], HAN Xiaosong[2]
(1. *Institute of Atomic and Molecular Physics, Jilin University, Changchun* 130012, *China*;
2. *College of Computer Science and Technology, Jilin University, Changchun* 130012, *China*;
3. *Logistics Department, Jilin University, Changchun* 130012, *China*;
4. *Policy and Legal Office, Jilin University, Changchun* 130012, *China*)

Abstract: Aiming at the problem of low accuracy and low recommendation efficiency in the traditional algorithms during the recommendation process of institutional documents, we proposed a text recommendation algorithm based on knowledge embedding. By transforming knowledge in the knowledge graph into feature vectors and combining them with neural network models, the accuracy and stability of the recommendation system were effectively improved when dealing with massive and diverse data. Experimental results show that the proposed algorithm exhibits better recommendation accuracy and stability than the traditional methods in the face of cold start and diverse user interests. It provides a more efficient and reliable solution to the data sparsity problem and personalisation requirements in large-scale recommendation systems, which helps to improve the user experience and enhance the overall performance of the recommendation system.

Keywords: knowledge graph; word embedding; graph embedding; neural network

随着信息化水平的不断提高和人工智能技术的快速发展, 海量数据不断涌现, 如何高效地处理并

收稿日期: 2024-01-05.

第一作者简介: 李 鑫(1983—), 男, 汉族, 博士, 副研究员, 从事高教管理和大数据分析的研究, E-mail: jlulx@jlu.edu.cn.
通信作者简介: 韩霄松(1984—), 男, 汉族, 博士, 副教授, 从事机器学习和优化算法的研究, E-mail: hanxiaosong@jlu.edu.cn.

基金项目: 吉林省高教科研项目(批准号: JGJX2023D42).

分析这些数据，从中快速提取所需信息并挖掘隐藏的高价值信息，已成为当前的研究热点之一[1-2]. 社会运转中产生的数据量呈指数级增长，信息接收者获得的信息量远超出其处理信息的能力，大量冗余和无效信息使信息接收者难以准确获取有效信息，用户面对海量数据，通常难以在短时间内准确找到自己所需的信息，对用户体验产生了极大影响. 因此，一个高效的推荐系统能在信息过载的情况下帮助用户找到感兴趣的信息，极大提高信息获取的效率.

传统推荐模型还面临冷启动问题，即对于新用户或新项目，系统缺乏足够的历史数据，导致推荐效果较差. 例如，当一个新用户首次进入某个网站时，系统无法根据其偏好为其提供个性化推荐. 针对这种情况，利用知识嵌入技术可有效改善冷启动导致的困境. 王玉奎等[3]将知识图谱作为辅助信息融入推荐系统，以应对数据稀疏和冷启动问题；吴晓等[4]提出了一种结合注意力机制与GRU循环神经网络的序列推荐模型，以更好地利用时间信息；孙远灿等[5]针对评分矩阵过大的问题，提出了先对实体进行分类的方案；李君等[6]通过引入注意力机制解决输入向量稀疏的问题；张栩翔等[7]引入了图嵌入算法，以提高图计算的效率.

本文综合以上模型的设计思想，提出一种基于知识嵌入的推荐算法，将知识图谱中的知识转化为特征向量，通过计算机对这些向量进行处理与分析，以完成基于知识数据的推荐任务[8]. 同时，考虑到算法在实际应用中的有效性以及制度文件作为特定文档类型的特殊性，本文算法将以更简洁的步骤提供更精确的推荐结果. 对比实验结果表明，该算法在推荐任务中具有优势.

1 算法流程

图1为本文算法流程. 由图1可见，本文提出的基于知识嵌入的推荐算法流程主要分为3个步骤：知识提取、知识嵌入和模型训练. 其中模型训练采用黑盒模型和白盒模型. 黑盒模型通过神经网络进行训练，尽管其推荐结果准确，但缺乏解释性. 而白盒模型则采用基于向量相似度的评分和评价算法，以更好地解释推荐结果.

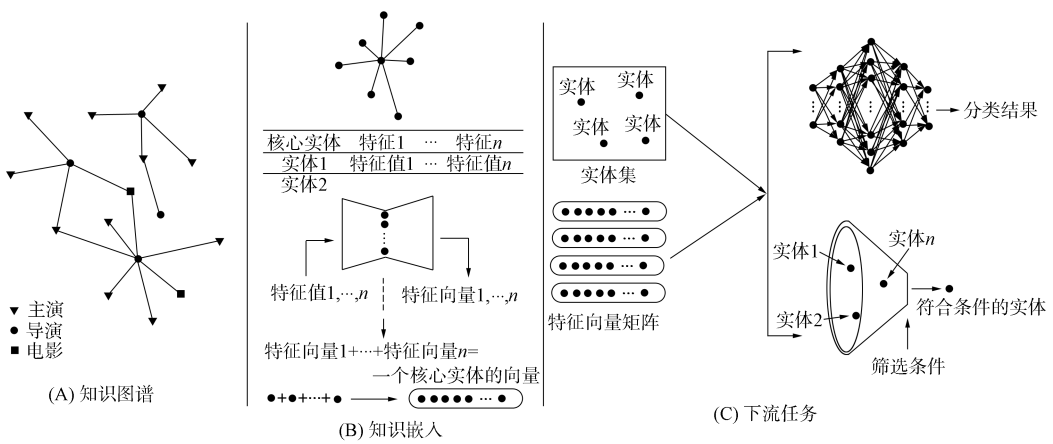


图1 本文算法流程

Fig. 1 Flow chart of proposed algorithm

1.1 知识图谱

受文献[3]模型启发，存储在知识图谱中的先验知识可有效应对冷启动问题，因此，本文算法第一步从知识图谱中获取先验知识.

知识图谱中存储的信息包括节点的属性信息和边关系的结构信息. 知识提取流程如图2所示. 首先，将所有属性视为核心属性，遍历所有与之相连的特征属性，即以所有节点为中心，遍历其相邻节点；其次，选择具有最多特征属性种类且分布最均匀的属性作为核心属性，对应的实体则作为核心实体；最后，将所有与核心实体相连的实体属性视为该核心实体的特征属性，并构建特征属性清单. 通过上述过程，边关系数据被转化为核心实体的特征属性信息，节点的属性信息则转化为特征属性的值. 这种方法集成了原始图谱中实体节点直接连接的结构信息，并将其转化为对核心实体的特征描

述，最终可以成功提取知识图谱中的部分先验知识.

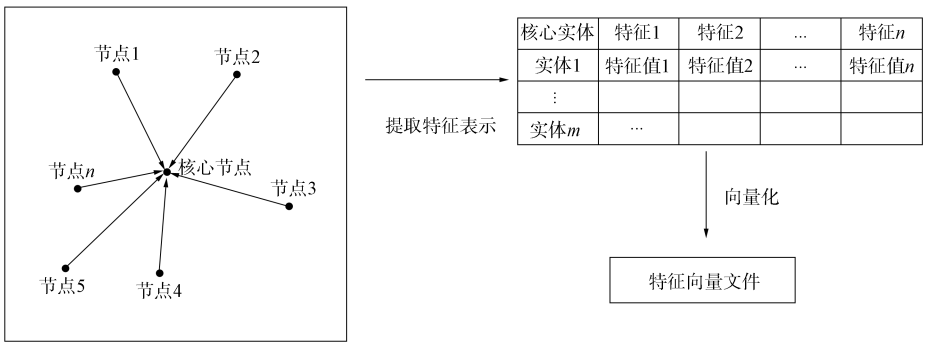


图 2　知识提取流程

Fig. 2　Knowledge extraction process

1.2　知识嵌入

知识嵌入的过程是将知识图谱中的节点和边的关系信息映射为特征向量，以便进行后续计算. 知识图谱由节点(实体)和边(关系)构成，其结构非常适合捕获实体之间的复杂关系. 在推荐系统中，知识图谱中的先验知识被转化为可用于计算的低维向量形式，使系统可以基于这些特征向量进行推荐.

本文采用 Word2Vec[9] 模型实现知识嵌入. 在 Word2Vec 模型中，CBOW(continuous bag of words)模型通过上下文预测中心词，而 Skip-Gram 模型则通过中心词预测上下文. 由于 Skip-Gram 模型在处理低频特征时性能较好，因此它在知识图谱的嵌入中更适合捕捉稀疏特征的语义关系. 在训练过程中，模型通过大量文本数据学习词语的共现模式，逐渐调整词向量，使语义相似的词在向量空间中彼此接近.

将每个核心实体特征值对应的向量进行拼接，以获得可以描述该实体特征的特征向量，该向量可用作推荐任务的数据基础. 但在实际操作中，一个实体的一些属性的属性值可能并不唯一，如某电影的类型可能既是恐怖类型又是喜剧类型. 对于这些特例需进行特殊处理，在嵌入模型训练前，先将每条拥有复合属性的实体数据扩写成多条数据，确定每条数据的每个属性只有一个属性值，然后再按 7∶3 提取出训练数据集和测试数据集，从而扩大了样本空间，可提高模型的准确率.

1.3　推荐任务

在获取了先验知识的嵌入表示后，需要以此为基础训练推荐模型完成不同的推荐任务. 传统的推荐算法主要包括基于内容的推荐算法、协同过滤推荐算法、混合推荐算法和基于矩阵分解的推荐算法等[10-11].

如果推荐任务的目标是分类并且不是很关注结果的可解释性，即可采用黑盒模型. 将先验知识作为输入，用户对文件的关注程度抽象成不同等级的输出，一个推荐任务即可抽象成多分类问题. 而黑盒模型的可解释性有限，为更好地说明推荐该文件的理由，可采取基于向量相似程度的评分算法和评价算法的白盒模型. 该模型可根据两个文件的相似程度打分，据此选取与用户历史关注程度最相似的文件作为模型输出，最后评价算法的准确度.

黑盒模型使用基于神经网络模型的推荐算法，首先将前文得到的所有特征向量数据与开源电影数据集(Movie_Lens)的用户评分数据相结合，按 7∶3 构建出神经网络模型训练需要的训练集和测试集. 模型以 ReLU 作为输入层和隐藏层的激励函数，以 Softmax 作为输出层的激励函数，损失函数设为 Binary_crossentropy，考虑到训练结果和训练的时间成本，训练步数选择 40 步，网络的深度选择 5 层. 同时，保留训练历史，用于结果分析. 白盒模型的评分算法如下：

$$\mathrm{Score}=\sum_{i=1}^{n}(|\boldsymbol{s}_i|+|\boldsymbol{s}_i-\boldsymbol{y}_i|)^2\Big/\sum_{i=1}^{n}\boldsymbol{s}_i^2, \tag{1}$$

其中 $\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_n$ 表示实体特征属性的特征向量，$\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_{n-1}$ 表示与其测量相似度的实体对应特征

属性的特征向量，Score 值越接近 1，两个实体越相似.

一般取两个实体所有属性的相似评分数值在前 1/5～3/5 的平均值作为两个实体的相似打分. 而对于某些包含多个属性值的特征属性，可选择打分最低的几个属性值作为代表，例如一个文件可以包含多个主题词，而读者在寻找一个文件的类似文件时可能对部分主题词更关注，对此任务设计算法时不需要关注所有的主题词.

2 电影数据集黑盒实验

黑盒实验流程为：获取需要的数据并进行数据预处理；选择核心实体并构建特征文件；使用词嵌入算法 Word2Vec 获取向量文件，将获得的知识嵌入数据和用户评分数据进行融合，构建神经网络需要的训练集和测试集；最后构建一个 5 层神经网络模型完成推荐任务.

2.1 数据集

为验证本文算法，选择经典的数据集 MovieLens，该数据集包含 9 000 多条用户评分信息和电影属性信息，包括电影名称、上映时间、所属国家、用户评分等，这些信息非常适合捕获用户的多样化兴趣，同时有助于处理冷启动问题. 为增强实验的通用性，还通过数据爬虫从互联网电影资料库(IMDB)获取了电影的更多相关信息，包括主演、导演、编剧等信息用于补充数据.

2.2 基准算法

实验设置 11 个基准算法对比分析本文提出的算法性能. 这些基准算法分别代表了不同的推荐方法，包括基于注意力机制、矩阵分解和神经网络的模型. 设置多个基准算法的目的是为更全面验证本文算法在精度和稳定性上的优势. 为与本文算法区分及方便，基准算法使用 LabelEncoder 对字符串形式的属性值进行编码.

1) AFM(attentional factorization machines)[12]：一种基于注意力机制的推荐算法，其结合了因子分解机和注意力机制的优点.

2) AutoInt(automatic feature interaction learning via self-attentive neural networks)[13]：一种结合了自动特征交互技术和神经网络模型以提高推荐效果的推荐算法.

3) CCPM(convolutional click prediction model)[14]：一个基于卷积神经网络的点击率预测模型，其利用卷积层捕获特征之间的局部和高阶非线性关系，从而提升模型的预测能力.

4) DCN(deep & cross network)[15]：一种深度学习的推荐算法，其结合了深度神经网络和交叉网络的优点，用于处理推荐系统中的特征交叉问题，能更好地捕获特征之间的高阶交互，从而在推荐系统中取得良好的效果.

5) DeepCrossing[16]：一种用于推荐系统的深度学习算法，其核心思想是通过深度神经网络自动捕获特征之间的复杂非线性关系，以更好地理解用户和物品之间的复杂关系，提高推荐效果.

6) DeepFM(deep factorization machines)[17]：一种融合了深度学习和因子分解机的推荐算法，其核心思想是将输入特征同时输入到 FM 模型和神经网络模型中，通过 FM 模型学习特征之间的二阶交互关系，通过神经网络模型学习高阶特征之间的交互关系.

7) FM(factorization machines)[18]：一种用于推荐系统的机器学习算法，它可以处理稀疏数据并且对特征之间的交互进行建模，因此在推荐任务中性能优异.

8) FFM(field-aware factorization machines)[19]：一种用于推荐系统的机器学习算法，它是因子分解机的扩展，旨在更好地处理字段感知信息，其优点是能更好地捕捉特征之间的字段感知交互，因此在处理推荐系统等需要考虑不同特征领域交互的问题上性能优异.

9) FGCNN(feature generation by convolutional neural network)[20]：一种用于推荐系统的算法，其结合了特征因子化和图卷积神经网络的思想，可以有效地处理推荐系统中的稀疏和高维特征.

10) NFM(neural factorization machines)[21]：一种基于神经网络的推荐算法，其结合了 FM 和神经网络的优势. NFM 的目标是解决传统推荐算法在处理大规模稀疏数据时的效率问题，并提高推荐的准确性.

11) PNN(product-based neural networks)[22]：一个用于推荐系统的深度学习模型，其采用了一种称为“product layer”的结构捕捉用户和物品之间的交互信息，PNN 模型在处理稀疏和高维特征的推荐问题上性能优异，能捕捉到特征之间的非线性关系.

2.3　实验结果与分析

表 1 列出了不同算法的对比实验结果. 由表 1 可见，与其他 11 种基准算法相比，本文算法在 610 名用户的推荐任务中表现出较高的稳定性和准确性，在精度和鲁棒性方面具有显著优势. 尽管该算法的训练时间较长，但考虑到计算硬件性能的持续提升，这一时间开销可以接受.

表 1　不同算法对比实验结果

Table 1　Comparison experimental results of different algorithms

算法	平均值	中值	方差	算法	平均值	中值	方差
本文	0.929	0.931	0.042	DeepFM	0.667	0.666	0.182
AFM	0.652	0.652	0.165	FFM	0.567	0.560	0.148
AutoInt	0.665	0.666	0.221	FGCNN	0.655	0.651	0.162
CCPM	0.662	0.666	0.183	FM	0.624	0.617	0.168
DCN	0.658	0.652	0.156	NFM	0.656	0.666	0.165
DeepCrossing	0.646	0.642	0.175	PNN	0.861	0.857	0.034

3　文件数据集的白盒实验

3.1　数据集

实验使用的数据集来源于吉林大学近年来发布的各类制度文件，共 500 多篇. 每篇文件记录了多个特征属性，包括发布时间、发布人、发布部门、发文编号等. 其中发布时间标识了每个文件的发布日期，以便于分析文件的时间分布和发布趋势；发布人记录了负责发布的具体工作人员，以帮助追踪文件的权威性；发布部门涉及不同的职能单位，如教务处、人力资源处等，有助于分析文件的内容来源和职能分布；发文编号则为每个文件分配了唯一编号. 此外，文件还包含各级标题以及时效性等信息，进一步增强了数据集的分析价值和广泛适用性.

3.2　实验流程

从知识图谱获取先验知识后，还需进行数据扩充. 首先，对原文件集的 PDF 数据进行预处理，得到程序可直接处理的文本数据；其次，通过 Jieba 分词和 LDA 主题词分析模型分析出每个文件的主题，将其作为新属性加入属性清单中的属性集中，作为该文件的特征属性之一；再次，使用 Word2Vec 模型分析出文件集所有文件特征属性的词向量作为该文件的特征属性向量；最后，将两个文件的特征属性向量通过评分算法打分，以衡量两个文件的相似度.

3.3　评价算法

考虑到效率，实验选择通过分词数据的相似程度衡量两个文件的内容相似性：设 X 是一篇文件文本内容的所有分词数量，Y 是另一篇文件文本内容的所有分词数量，Z 是两篇文件分词重合的数量，则两篇文件的评价算法公式可表示为

$$\text{Score_true} = 2 \times Z/(X+Y). \tag{2}$$

对每篇文件，将分别使用本文算法、Word2Vec 自带的相似算法(Models)和直接使用字符匹配(Words)3 种方法选择的相似文件通过该评价算法进行打分，并求取平均值，作为该方法对该文件的效果评价.

3.4　实验结果

相似文件选取过程中 3 种不同方法的性能对比如图 3 所示. 图 3(A)显示了在所有文件中每种方法获得最佳结果的比例，图 3(B)则对比了 3 种方法的效果稳定性. 由图 3 可见，本文算法在 80%的情况下性能优于其他两种方法，不仅在文件评分的平均值上更高，而且在效果上表现出更强的稳定性. 表明本文算法在相似文件选取任务中具有较好的性能和鲁棒性.

图 4 为以部分医学研究生管理相关文件为例作为历史数据(A)，通过本文算法(B)、Word2Vec 自

带的相似算法(C)和直接使用字符匹配(D)三种方法的推荐结果. 由图4可见，利用本文算法与利用Word2Vec自带的算法推荐的文件部分相似，与直接字符匹配的推荐结果差异较大. 根据评价函数，本文算法的推荐结果在文件内容上与历史文件最相近.

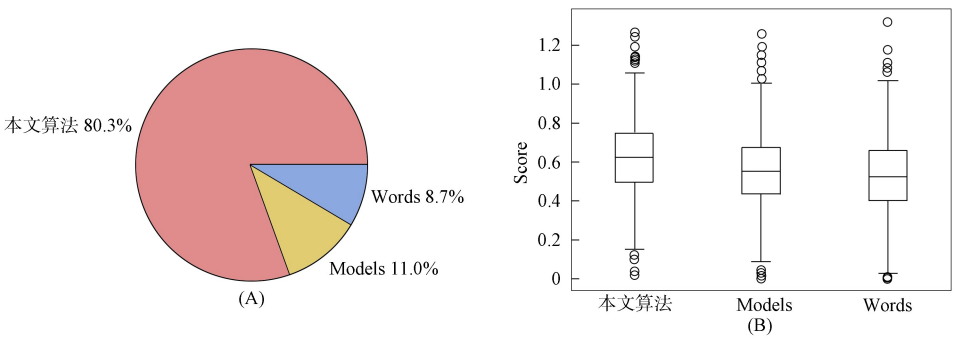


图3 3种不同方法对所有文件的推荐效果

Fig. 3 Recommendation effects of three different methods on all files

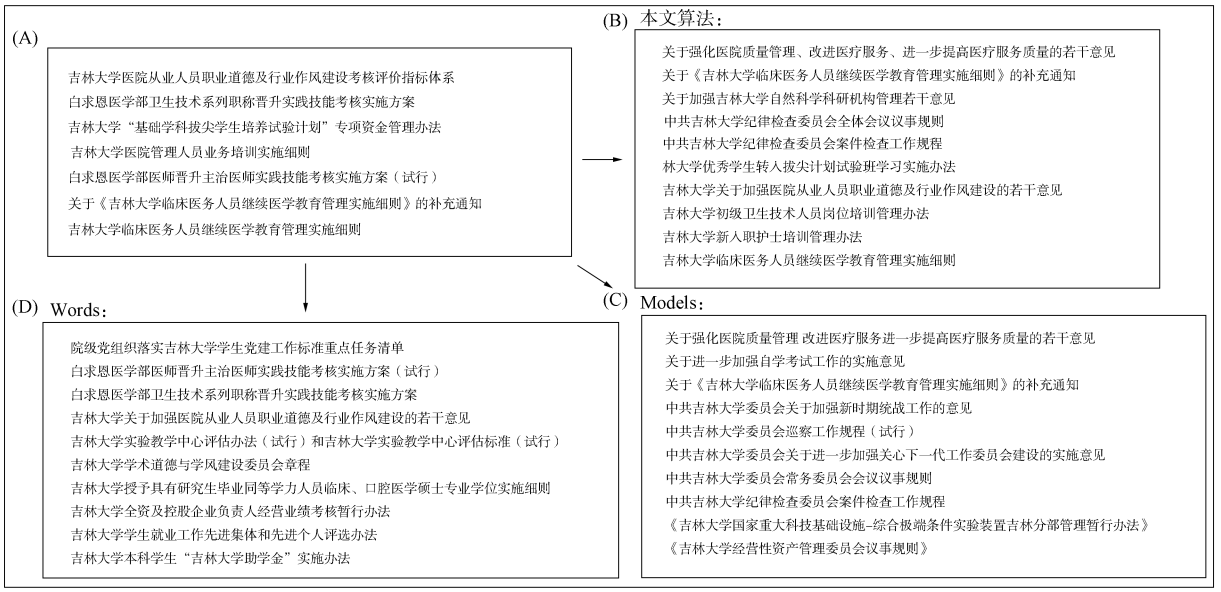


图4 使用3种方法的样例

Fig. 4 Examples of using three methods

综上所述，针对文件推荐缺乏先验数据进行训练而导致推荐准确率和效率较低的问题，本文采用相近的电影数据库作为训练集，提出了一种基于知识嵌入的推荐算法，融合了知识图谱与推荐模型的优势，通过将知识图谱中的实体和关系信息抽象为包含特征及关系结构的向量数据，应用于神经网络模型的训练和相似度计算，从而实现了高效的推荐. 实验结果表明，本文方法在制度文件推荐系统中表现出较高的准确率和稳定性，性能优越.

参 考 文 献

[1] 车万翔，窦志成，冯岩松，等. 大模型时代的自然语言处理：挑战、机遇与发展[J]. 中国科学：信息科学，2023，53(9)：1645-1687.（CHE W X，DOU Z C，FENG Y S，et al. Natural Language Processing in the Era of Large Models：Challenges，Opportunities，and Developments[J]. Chinese Science：Information Science，2023，53(9)：1645-1687.）

[2] 黄立威，江碧涛，吕守业，等. 基于深度学习的推荐系统研究综述[J]. 计算机学报，2018，41(7)：1619-1647.（HUANG L W，JIANG B T，LÜ S Y，et al. Survey on Deep Learning Based Recommender Systems[J]. Chinese Journal of Computers Application，2018，41(7)：1619-1647.）

[3] 王玉奎，郭秀娟. 融入知识图谱的电影推荐算法[J]. 福建电脑，2023，39(9)：28-30.（WANG Y K，GUO X J.

A Film Recommendation Algorithm Integrating Knowledge Graph [J]. Fujian Computer, 2023, 39(9): 28-30.)
[4] 吴晓，成卫青. 结合注意力机制与循环神经网络的序列推荐模型 [J]. 软件导刊，2023，22(10)：95-104.（WU X，CHENG W Q. A Sequence Recommendation Model Combining Attention Mechanism and Recurrent Neural Networks [J]. Software Guide，2023，22(10)：95-104.）
[5] 孙远灿，程凯，陈文静，等. 图书分类推荐算法研究 [J]. 福建电脑，2023，39(9)：64-66.（SUN Y C，CHENG K，CHEN W J，et al. Research on Book Classification Recommendation Algorithms [J]. Fujian Computer，2023，39(9)：64-66.）
[6] 李君，倪晓军. 融合注意力机制的知识图谱推荐模型 [J]. 软件导刊，2023，22(3)：118-124.（LI J，NI X J. A Knowledge Graph Recommendation Model Integrating Attention Mechanism [J]. Software Guide，2023，22(3)：118-124.）
[7] 张栩翔，汤玉祺，赵文，等. 基于知识图谱和图嵌入的个性化学习资源推荐 [J]. 计算机系统应用，2023，32(5)：180-187.（ZHANG X X，TANG Y Q，ZHAO W，et al. Personalized Learning Resource Recommendation Based on Knowledge Graph and Graph Embedding [J]. Computer System Application，2023，32(5)：180-187.）
[8] WANG L，ZHANG W，HE X L，et al. Supervised Reinforcement Learning with Recurrent Neural Network for Dynamic Treatment Recommendation [EB/OL]. (2018-07-04) [2023-02-15]. https://arxiv.org/abs/1807.01473.
[9] JOULIN A，GRAVE E，BOJANOWSKI P，et al. Bag of Tricks for Efficient Text Classification [EB/OL]. (2016-07-06)[2023-03-10]. https://arxiv.org/abs/1607.01759.
[10] ZHANG W，LIN Y G，LIU Y，et al. Self-supervised Reinforcement Learning with Dual-reward for Knowledge-Aware Recommendation [J]. Applied Soft Computing，2022，131：109745-1-109745-12.
[11] LIU X Y，TAN H B，CHEN Q H，et al. RAGAT：Relation Aware Graph Attention Network for Knowledge Graph Completion [J]. IEEE Access，2021，9：20840-20849.
[12] XIAO J，YE H，HE X G，et al. Attentional Factorization Machines：Learning the Weight of Feature Interactions via Attention Networks [EB/OL]. (2017-08-15)[2023-01-10]. https://arxiv.org/abs/1708.04617.
[13] SONG W P，SHI C C，XIAO Z P，et al. AutoInt：Automatic Feature Interaction Learning via Self-attentive Neural Networks [EB/OL]. (2018-10-29)[2023-03-01]. https://arxiv.org/abs/1810.11921.
[14] LIU Q，FENG Y，WU S，et al. A Convolutional Click Prediction Model [C]//ACM International on Conference on Information & Knowledge Management. New York：ACM，2015：1743-1746.
[15] WANG R X，FU B，FU G，et al. Deep & Cross Network for Ad Click Predictions [EB/OL]. (2017-08-17) [2023-02-01]. https://arxiv.org/abs/1708.05123.
[16] SHAN Y，HOENS T R，JIAO J，et al. Deep Crossing：Web-Scale Modeling without Manually Crafted Combinatorial Features [C]//The 22nd ACM SIGKDD International Conference. New York：ACM，2016：255-262.
[17] GUO H F，TANG R M，YE Y M，et al. DeepFM：A Factorization-Machine Based Neural Network for CTR Prediction [EB/OL]. (2017-03-13)[2023-01-20]. https://arxiv.org/abs/1703.04247.
[18] RENDLE S. Factorization Machines [C]//2010 IEEE International Conference on Data Mining. Piscataway，NJ：IEEE，2011：995-1000.
[19] JUAN Y C，ZHUANG Y，CHIN W S，et al. Field-Aware Factorization Machines for CTR Prediction [C]//Conference on Recommender Systems. New York：ACM，2016：43-50.
[20] LIU B，TANG R M，CHEN Y Z，et al. Feature Generation by Convolutional Neural Network for Click-through Rate Prediction [C]//The Word Wide Web Conference. New York：ACM，2019：1119-1129.
[21] HE X G，CHUA T S. Neural Factorization Machines for Sparse Predictive Analytics [J]. ACM Sigir Forum，2017，51：355-364.
[22] QU Y R，CAI H，REN K，et al. Product-Based Neural Networks for User Response Prediction [C]//2016 IEEE 16th International Conference on Data Mining (ICDM). Piscataway，NJ：IEEE，2016：1149-1154.

（责任编辑：韩　啸）