

知识图谱嵌入的高阶协同过滤推荐系统

徐 兵¹ 弋沛玉¹ 王金策² 彭 舰¹

1 四川大学计算机学院 成都 610065

2 山西能源学院计算机与信息工程系 山西 晋中 030600

(1064937023@qq.com)

摘 要 针对推荐系统存在的数据稀疏问题,传统的协同过滤方法无法捕捉辅助信息之间的相关性,从而降低了推荐的准确度,文中提出 KGE-CF 模型,引入了知识图谱作为辅助信息,利用知识图谱中多源结构性的数据来缓解数据稀疏问题。KGE-CF 结合多层感知机捕获高阶非线性特征的能力,能够学习出用户与项目更深层次的交互信息,从而提升推荐质量。首先,KGE-CF 模型将用户的历史交互项目与知识图谱中的实体进行映射,并且利用知识图谱的翻译模型进行训练,得到实体嵌入向量与关系向量,并依据“兴趣迁移”思想进一步学习出更为丰富的用户向量;然后,模型将学习得到的用户向量与项目向量拼接,作为多层感知机的输入,捕捉用户与项目之间的高阶特征信息;最后,通过一个 sigmoid 函数得到用户对候选项目的偏好程度。通过在真实数据集上的实验,证明了提出的 KGE-CF 模型在点击率预测和 top- k 两种推荐场景下均优于其他方法。

关键词: 知识图谱;多层感知机;推荐系统;协同过滤;数据稀疏性

中图法分类号 TP3

High-order Collaborative Filtering Recommendation System Based on Knowledge Graph Embedding

XU Bing¹, YI Pei-yu¹, WANG Jin-ce² and PENG Jian¹

1 College of Computer Science, Sichuan University, Chengdu 610065, China

2 Department of Computer and Information Engineering, Shanxi Institute of Energy, Jinzhong, Shanxi 030600, China

Abstract For the problem of data sparsity in recommendation systems, traditional collaborative filtering methods fail to capture the correlations between auxiliary information, which decreases the accuracy of recommendation. In this paper, we propose the KGE-CF model by introducing the knowledge graph as auxiliary information, and utilizing the multi-source structured data in knowledge graph to effectively alleviate data sparsity. KGE-CF integrates multi-layer perceptron to capture the high-order nonlinear features to learn deeper interaction information between users and items, which can improve the quality of recommendation. Specifically, we first map the user's historical interaction items with the corresponding entities in the knowledge graph, and use the translation model of the knowledge graph for training, through that we can get the entity embedding vector and the relation embedding vector, moreover the model can learn a richer user vector by interest disseminating. Then, after concatenating the obtained user vector and the item vector, we send it into the multi-layer perceptron to capture the high-order feature information between the user and the item. Finally, a sigmoid function is used to obtain the user's preference probability for candidate items. The experimental results on the real-world datasets prove that the proposed KGE-CF model in this paper has achieved the best recommendation performance than state-of-art methods.

Keywords Knowledge graph, Multi-layer perceptron, Recommendation system, Collaborative filtering, Data sparsity

1 引言

推荐系统(Recommendation System)是一种主动提供个性化信息的服务,其依据用户与项目的交互历史,去推测用户的喜好,进而呈现用户感兴趣的项目,满足用户信息过滤的需求,提升用户获取信息的效率。传统的推荐技术主要包括基于内容的推荐^[1]、协同过滤推荐^[2]及混合推荐^[3]。然而,在推

荐系统中通常存在数据稀疏性问题,即在应用协同过滤(Collaborative Filtering, CF)推荐算法时因缺少用户-项目交互数据,无法通过分析用户与项目的特征来提取潜在的偏好向量。

目前,随着深度学习技术在图像处理、机器翻译、智能对话等各领域得到普遍应用,推荐领域深度学习的应用同样取得了引人注目的成就,深度模型学习辅助信息的特征能有效地捕捉到用户和项目的潜在表示,进一步缓解了数据稀疏性

基金项目:四川省重点研发计划(2020YFG0089);四川大学-自贡市校地科技合作专项基金(2018CDZG-15);山西省青年科技研究基金项目(201801D221176)

This work was supported by the Key R&D Program of Sichuan Province, China(2020YFG0089), University-Local Science and Technology Cooperation Foundation of Sichuan University-Zigong(2018CDZG-15) and Natural Science Foundation for Young Scientists of Shanxi Province, China(201801D221176).

通信作者:彭舰(jianpeng@scu.edu.cn)

问题,并结合协同过滤方法获取用户项目的最终向量,从而为用户进行有效推荐。He 等^[4]提出的 NCF 模型证明了深度学习与传统推荐算法融合能有效提升推荐的准确性。然而,深度学习具有“数据饥饿”的特点,需要大量数据才能进行有效学习,提升推荐效果,又因深度学习缺乏对数据的真正理解,只是机械地对已知数据进行拟合,在考虑不同类型的辅助信息时,因数据的不相关性,在提取特征时产生的噪声会造成推荐偏差。

随着知识图谱的出现,实体与实体之间的语义关系的探索逐渐引起了学者们的关注。研究表明,知识图谱的表示学习可以将实体与关系嵌入到低维向量空间中,低维稠密的向量不仅可以用来反映知识图谱的结构特征,还可以表示实体与实体间的基本语义关系。因此,将知识图谱作为辅助信息引入推荐能有效缓解数据的稀疏性。Wang 等^[5]提出的 RippleNet 创新性地用“兴趣迁移”思想,将知识图谱作为辅助信息引入推荐系统,有力地解决了数据稀疏性问题,然而其缺乏对用户/项目向量更充分地交叉。

针对上述不足,本文将具有多源结构性数据的知识图谱作为一种辅助信息引入推荐系统中,结合深度学习提出了 KGE-CF(Knowledge Graph Embedding&Collabprative Filtering)模型。结构性的数据关系能表示数据间的相互关系,多源性的数据为深度学习提供了更多的数据来源。该模型将知识图谱与深度学习结合,并融入协同过滤,学习统一、稠密且低维的用户和项目的潜在向量表示,能更好地捕获用户与项目之间的非线性关系特征,提升预测准确率。本文的主要贡献如下:

(1)通过深度学习对知识图谱嵌入后的用户与项目向量进行交互关联,捕捉嵌入的全局高阶交互潜在特征,提升推荐结果的准确性。

(2)通过在 3 个真实数据集上的实验结果表明,本文模型在点击率预估场景和 top- k 推荐场景下的性能均优于其他算法。

(3)通过对不同稀疏度的数据集实验表明,借助知识图谱丰富的语义信息,能够缓解数据稀疏问题,保持推荐的高效。

本文第 2 节介绍了相关研究工作;第 3 节描述了模型构建过程;第 4 节介绍了具体的实验过程、实验结果分析;最后总结全文。

2 相关工作

本节介绍了与本文工作相关的模型,包括基于深度学习的推荐系统和基于知识图谱的推荐系统。

2.1 基于深度学习的推荐系统

近年来,深度学习利用其自动提取特征的能力来缓解数据稀疏,在推荐系统中的应用已证明了其有效性。文献[6]提出的 wide&deep 模型,将线性与非线性特征通道相结合,以平衡记忆能力和泛化能力,在缓解数据稀疏的同时,提高了推荐的准确性和扩展性。文献[7]提出的 DeepFM 模型,结合因子分解机和多层感知机,利用因子分解机对低阶交互进行建模,使用多层感知机对高阶特征进行交互建模,从而学习低阶与高阶特征交互,增强了数据的表达能力,丰富了用户-项目的交互信息。文献[4]提出的 NCF 模型,利用多层感知机建模用户和项目之间的非线性关系,结合低阶信息与高阶信息

来近似用户-项目的交互函数。文献[8]提出的 NFM 模型使用一个池化层来计算因子分解机的一阶特征交互项,然后将得到的向量输入到一个多层感知器。此外,还有其他神经网络用于推荐,例如文献[9]通过利用自编码器来预测用户-项目矩阵的缺失评分,从而使推荐达到更高的精度,文献[10]基于循环神经网络将协同过滤转化为序列预测问题,从而提高短期兴趣的预测精度。本文使用多层感知机结合矩阵分解的通用框架来构建本文模型。

2.2 基于知识图谱的推荐系统

为解决数据稀疏问题,将知识图谱作为辅助信息(side information)引入推荐系统,通过知识图谱中实体之间丰富的关系能丰富项目和项目的联系、增强推荐算法的挖掘能力,从而有效地弥补用户-项目间交互信息的稀疏性。基于知识图谱的推荐系统通过基于嵌入的方法、基于路径的方法及两者结合的方法这 3 种方式来应用知识图谱。基于嵌入的方法有:文献[11]提出的 CKE 模型,利用 TransR^[12]方法对知识图谱的结构信息进行处理,得到实体的结构化信息向量,结合获得的文本信息向量和视图信息向量,形成项目的潜在表示;文献[13]提出的 DKN 模型,专注于新闻推荐,利用 TransD 方法学习知识图谱中的实体向量,并学习实体一跳范围的上下文实体向量,使用多通道对不同空间的向量进行拼接,解决了新闻推荐的三大难题;文献[15]提出的 MKR 模型,采用多任务学习框架,将推荐系统与知识图谱特征学习任务交替优化训练,利用知识图谱嵌入协助推荐任务,使推荐更具灵活性和适应性。基于路径的方法有:文献[16]提出的 RKGE 模型,使用循环网络对实体间的所有路径的语义进行自动建模,同时采用池化学习实体间路径的重要性,并将其融合到推荐中,从而发现用户对项目的偏好;文献[17]提出的 KPRN 模型,通过对实体的顺序依赖性和连接用户-项对的路径的复杂关系建模,并通过路径推理用户偏好,为推荐提供解释性。两者相结合的方法有:文献[5]提出的 RippleNet 模型,将知识图谱嵌入通过偏好迁移的方式融入推荐中,并自动发现用户交互历史与候选项目间的连接路径;文献[18]提出的 KTUP 模型,根据用户偏好某个项目的原因建模,将推荐任务与知识补全相结合,以提升推荐结果的准确性和解释性。

3 KGE-CF 模型

本文的目标是学习用户与候选项目的交互概率,用于预测用户 u 对其未交互的项目 j 的点击概率,以表示用户 u 对项目 j 的兴趣偏好程度。

3.1 问题定义

用户集合 $U=\{u_1, u_2, \dots, u_M\}$ 和项目集合 $I=\{i_1, i_2, \dots, i_N\}$, M 和 N 分别表示用户数量和项目数量。使用向量 p_u 和 p_j 分别作为输入用户 u 和候选项目 j 的潜在表示,通过全连接层 $f(\cdot)$ 学习用户-项目间的交互,最后将 sigmoid 函数 $\sigma(\cdot)$ 作为激活函数,输出用户对项目的偏好程度,预测函数的表达式如下:

$$\hat{y}_{uj} = \sigma(f(p_u, q_j)) \quad (1)$$

在项目端的知识图谱中,实体和实体之间丰富的语义关联为挖掘用户兴趣提供了潜在视角。本文通过实体嵌入、兴趣迁移、特征提取 3 个步骤,将用户在知识图谱中的相关实体融入推荐,以缓解数据稀疏的问题,丰富用户的表示。

(1)实体嵌入,项目和实体为现实世界中一个对象的两种表达,将用户交互历史的项目映射为知识图谱中的头实体,从而在知识图谱中构成三元组集合 (h,r,t) 。

(2)兴趣迁移,嵌入的实体 $h,t \in \mathbb{R}^d$ 和关系 $r \in \mathbb{R}^d$,通过关系矩阵 $R \in \mathbb{R}^{d \times d}$,将实体从实体空间投影到对应的关系空间,从而进行兴趣迁移,利用实体的多跳邻居来学习用户嵌入表示。

(3)特征提取,学习到的用户向量与项目向量利用多层感知机来捕获隐含特征的能力,预测用户对项目的点击概率,形成推荐。

3.2 模型架构

本文依照多层感知机构建推荐算法的通用方法^[6,8],将知识图谱融入协同过滤中,提出了 KGE-CF(知识图谱嵌入协同过滤)模型,如图 3 所示,该模型由 3 个关键层组成:1)Embedding 层,将输入的高维离散特征转换为固定长度的连续特征,进行知识迁移嵌入;2)MLP 层,对得到的用户向量与项目向量进行交互学习,捕获用户-项目之间隐含的高阶非线性特征;3)Output 层,将 sigmoid 函数作为激活函数,最终输出给定用户对目标项目的偏好程度,表达用户兴趣大小。

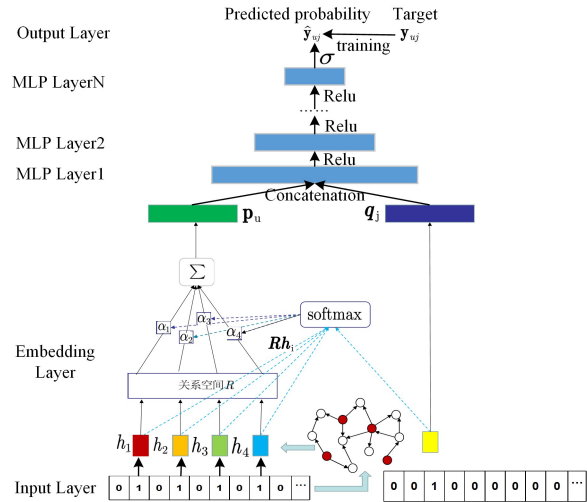


图 1 KGE-CF 的模型结构
Fig. 1 Model structure of KGE-CF

3.2.1 输入层(input layer)

输入层的输入包括两部分:用户-项目交互历史 $\mathcal{H}_u$ 、候选项目 item j 。 $\mathcal{H}_u$ 通过用户与项目的隐式反馈(点击、购买、观看等),从侧面反映用户对项目的兴趣,如式(2)所示:

$$\mathcal{H}_u = \begin{cases} 1, & \text{用户 } u \text{ 与项目 } i \text{ 存在交互} \\ 0, & \text{其他} \end{cases} \quad (2)$$

其中, $\mathcal{H}_u \in \mathbb{R}^{M \times N}$, M 为用户的数量, N 为项目的数量。

3.2.2 嵌入层(embedding layer)

将输入层编码后的输出作为嵌入层的输入,将用户历史交互的项目融入知识图谱作为实体,通过知识图谱翻译模型,得到用户兴趣迁移相关实体,以解决数据稀疏性问题。用户点击历史映射为知识图谱中的头实体 h ,将其作为知识图谱的种子实体。利用知识图谱的关系 r ,在关系空间 R 中发现相关尾实体 t ,则其构成兴趣迁移的三元组 $\mathcal{F}_u$,具体定义如式(3)所示:

$$\mathcal{F}_u = \{(h, r, t) | (h, r, t) \in G \cap h \in \mathcal{H}_u\} \quad (3)$$

其中, G 表示知识图谱。

通过兴趣迁移 $\mathcal{F}_u$,由交互历史得到兴趣发现的相关实体集合,如式(4)所示:

$$\mathcal{T} = \{t | (h, r, t) \in \mathcal{F}_u \cap h \in \mathcal{H}_u\} \quad (4)$$

由头实体 h 发现尾实体 t 的过程中,对于实体间的每一种关系都存在概率 α_i 对应相关的尾实体 t 。该概率表示,在兴趣迁移三元组 $\mathcal{F}_u$ 中,用户点击的历史项目 i 映射为 h 后在所对应的关系空间 R 中,与候选项目 j 的相似度。其计算式如式(5)所示:

$$\alpha_i = f(h, r_i, q_j) = q_j^T h r_i \quad (5)$$

通过 f 计算出权重,再通过 softmax 进行归一化,如式(6)所示:

$$\alpha_{ij} = \frac{\exp(f(h, r_i, q_j))}{\sum_{(h, r, t) \in \mathcal{F}_u} \exp(f(h, r_i, q_j))} \quad (6)$$

其中, $h_i \in \mathbb{R}^d$, $R_i \in \mathbb{R}^{d \times d}$, $q_j \in \mathbb{R}^d$ 分别为头实体 h_i 、关系 r_i 和项目 j 的嵌入向量, d 为嵌入维度。

通过以上计算,得到用户的兴趣迁移向量 p_u ,如式(7)所示:

$$p_u = \sum \alpha_i t_i \quad (7)$$

3.2.3 多层感知机层(MLP layer)

将嵌入层输出的用户向量 p_u 与项目向量 q_j 拼接后的向量 v 作为多层感知机层的输入,通过感知机学习用户和项目之间的非线性高阶特征,本文采用三层的多层感知机,并在每层之间选取线性整流函数(ReLU)作为激活函数来避免梯度消失,多层感知机的表达式如式(8)所示:

$$\begin{aligned} v &= \text{concat}(p_u, q_j) = \begin{bmatrix} p_u \\ q_j \end{bmatrix}, \\ z_1 &= \text{relu}(W_1^T v + b_1) \\ z_2 &= \text{relu}(W_2^T z_1 + b_2) \\ &\vdots \\ z_L &= \text{relu}(W_L^T z_{L-1} + b_L) \end{aligned} \quad (8)$$

其中, W 和 b 是需要学习的参数,在模型训练过程中会不断更新。

3.2.4 输出层(output layer)

KGE-CF 模型的最终目标是得到最接近真实值的预测值 $\hat{y}_{uj}$,使推荐效果最优。在输出层,本文将多层感知机的最后一层作为输出,激活函数为 sigmoid,将输出的预测值 $\hat{y}_{uj}$ 控制在 $[0,1]$ 范围内,预测值表示将项目推荐给用户概率的大小,表达式如式(9)所示:

$$\hat{y}_{uj} = \sigma(\text{relu}(W_o^T z_L)) \quad (9)$$

其中, $\sigma(\cdot)$ 表示 sigmoid 函数,本文选择交叉熵作为损失函数,其表达式如式(10)所示:

$$L(\theta) = -\sum [y_{uj} \log \hat{y}_{uj} + (1 - y_{uj}) \log (1 - \hat{y}_{uj})] \quad (10)$$

3.3 KGE-CF 训练

算法 1 描述了 KGE-DL 的训练过程。首先基于 $\mathcal{H}_u$ 和 G 构造的 $\mathcal{F}_u$ 得到 p_u ,然后通过 Adam 优化算法训练获取最优模型。

算法 1 KGE-CF 的学习

输入:交互矩阵 $\mathcal{H}_u$,知识图谱 G ,神经网络层数 L ,嵌入空间维度 L
输出:KGE-CF
1. 初始化所有参数

- 2. 为每个用户计算兴趣集合 $\mathcal{F}_u$
- 3. Repeat:
- 4. 从 $\mathcal{H}_{in}$ 中采样一个批量
- 5. 从 G 中采样一个批量
- 6. 将用户向量与项目向量拼接成新向量,输入 MLP 层进行训练。
- 7. 在 MLP 的最后一层将所获得的向量通过 sigmoid 激活函数转为取值在 $[0,1]$ 的预测值,通过预测值和真实值的交叉熵损失迭代更新模型参数,直到满足收敛条件
- 8. end repeat

4 实验

4.1 数据集

本文使用如下 3 个公开的数据集来评估本文模型和对比模型。

(1)MovieLens 1M:电影评分数据集,包含来自 6 040 个用户对 3 706 部电影的 1 000 209 个评分,此数据集的稀疏度为 95.53%。

(2)Book-Crossings:图书评分数据集,包含 105 284 个用户,340 557 本书,1 149 780 个评分,数据稀疏度为 99.99%。

(3)Last. FM:包含来自在线音乐网站的 1 892 个用户对 14 632 个歌手的 92 834 次的点击数,数据稀疏度为 99.72%。

(4)KB4Rec^[19]:其将推荐系统中的项目与 Freebase 知识图谱中的实体进行关联,经过过滤处理后得到的知识图谱推荐项目关联数据,具体统计信息如表 1 所列。

表 1 KB4Rec 数据统计
Table 1 KB4RecData statistics

Dataset	# Items	# Linked-Items	# users	# Interactions
MovieLens	18 345	18 057	138 493	19 984 024
Amazon-book	367 982	34 476	603 668	8 898 041
LFM-1b(tracks)	615 823	337 349	79 133	15 765 756

稀疏度为在用户-项目矩阵中,无评分数据的元素占整个矩阵空间的比率,表达式如下:

稀疏度=1-
$$\frac{\text{评分数据}}{\text{用户数} \times \text{项目数}} \tag{11}$$

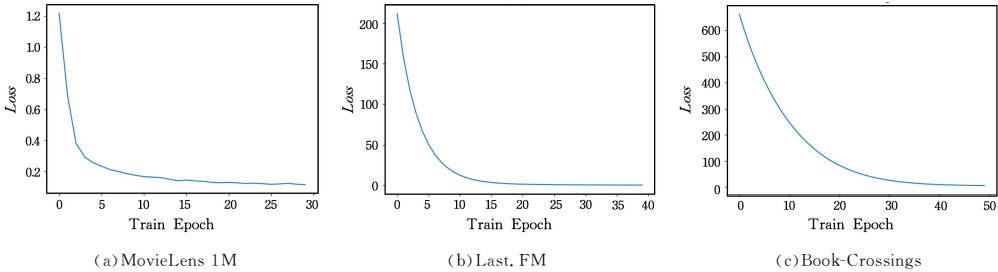


图 2 KGE-CF 模型的收敛情况
Fig. 2 Loss convergence of KGE-CF model

4.4 对比模型

本实验采用以下较为经典的推荐算法作为对比算法。

(1)BPRMF 算法^[20]:一种使用贝叶斯个性化排名(BPR)损失来优化基于隐式反馈的矩阵分解模型。

(2)NCF 算法^[4]:使用多层感知机来替代矩阵分解中的简单内积操作,使用户-项目的向量做更充分的交叉,引入更多的非线性特征,以增强模型的表达能力。

(3)CKE 算法^[11]:将项目的知识结构特征、文本特征和图

4.2 评价指标

本文针对点击率预测(CTR)推荐场景,采用 AUC(ROC 曲线下面积)来评估本文方法的有效性。AUC 反映了模型的排名能力,即模型预测时将正例(positives)排在负例(negatives)前面的概率,其定义如式(12)所示:

$$AUC = \frac{\sum_{ins_i \in positiveclass} rank_{ins_i} - \frac{A \times (B+1)}{2}}{A \times B} \tag{12}$$

其中, $ins_i \in positive$ 表示正例, $rank_{ins_i}$ 表示第 i 条样本依据概率从小到大的排名序号, A 和 B 分别是正例的个数和负例的个数。

本文针对 top- k 推荐场景,采用式(13)一式(15)中描述的准确率 Precision、召回率 Recall 和 F1-Score 指标来测量 top- k 推荐的有效性。准确率指在推荐系统中,正确推荐的项目数与推荐项目总数的比率;召回率指在推荐系统中,正确推荐的项目数与总项目数的比率;F1-Score 得分是准确率和召回率的调和平均值。

$$Precision = \frac{\sum_u |R_u \cap T_u|}{\sum_u |R_u|} \tag{13}$$

$$Recall = \frac{\sum_u |R_u \cap T_u|}{\sum_u |T_u|} \tag{14}$$

$$F-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{15}$$

4.3 模型参数设置

对于选取的数据集,在实验过程中本文按照传统的训练集、验证集和测试集划分比率 6:2:2 随机选取。在训练过程中,初始权重选择满足均值为 0、标准差为 0.01 的高斯分布随机数,通过验证集上的最优 AUC 值来手动调整学习率 learning rate、批次大小 batch_size、迭代次数 epoch、嵌入维数等相关超参数,使用自适应估计(Adam)优化器学习获取模型。

KGE-CF 模型在 3 个数据集上的损失收敛如图 2 所示。

像特征进行拼接,作为项目的最终表示,从而引入协同过滤推荐。

(4)RippleNet 算法^[6]:是一种利用偏好传播的思想,在知识图谱中传播用户的潜在偏好,以丰富用户的特征表示。

4.5 实验结果与分析

(1)CTR 场景下的性能分析

本文在 MovieLens 1M,Book-Crossings,Last. FM 这 3 个数据集上进行实验,各种方法在点击率预测上的 AUC 性能如表 2 所列。

表 2 各种推荐方法在 3 个数据集上的 AUC 表现

Table 2 AUC performance of each recommended method on three datasets

Model	MovieLens 1M	Book-Crossings	Last, FM
BPRMF	0.771	0.653	0.669
NCF	0.898	0.674	0.764
CKE	0.796	0.667	0.678
RippleNet	0.919	0.721	0.776
KGE-CF	0.941	0.739	0.797

从表 2 可以看出,本文提出的 KGE-CF 在 3 个数据集上的 AUC 均获得了较好的性能,证明了本文方法的有效性。

具体而言,相比本文的基础模型 RippleNet,其在 MovieLens 1M, Book-Crossings, Last, FM 上 AUC 分别提升了 2.2%, 1.8%, 2.1%。5 个模型在 3 个数据集上的性能从高到低分别是电影、音乐和书籍。这是因为在 3 个数据集中,用户和项目之间的平均连接数不同,电影相比其余两个数据集有更丰富的交互,能供学习潜在嵌入特征的信息更多。

(2)top-*k* 场景下的性能分析

模型在 Precision@K, Recall@K 和 F1@K 这 3 个数据集上的性能对比如图 3—图 5 所示。

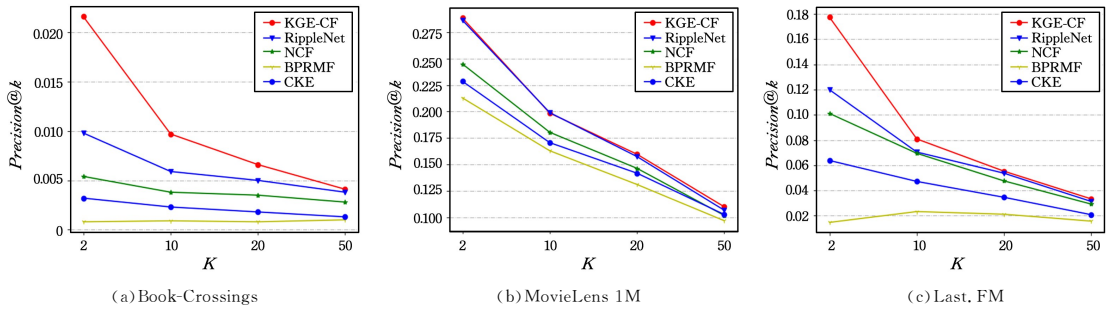


图 3 在 top-*k* 中推荐 Precision@K 的结果

Fig. 3 Results of Precision@K in top-*k* recommendation

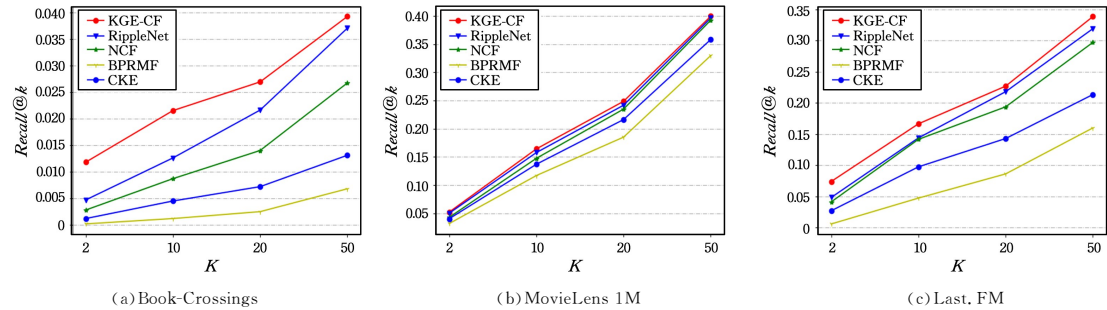


图 4 在 top-*k* 中推荐 Recall@K 的结果

Fig. 4 Results of Recall@K in top-*k* recommendation

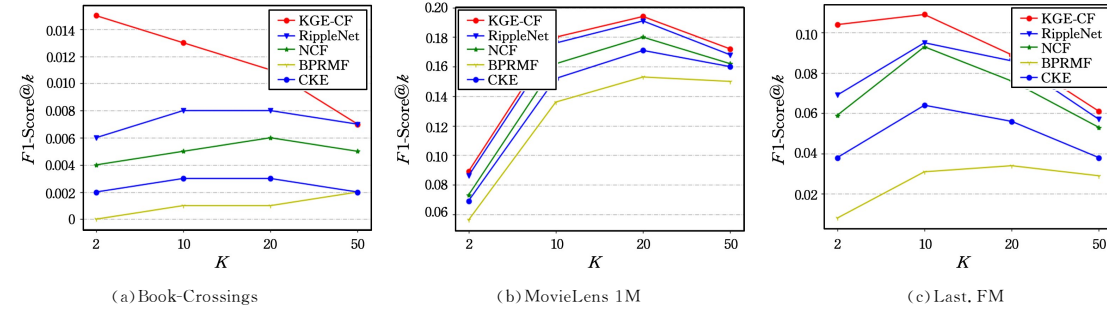


图 5 在 top-*k* 中推荐 F1-Score@K 的结果

Fig. 5 Results of F1-Score@K in top-*k* recommendation

从图 3—图 5 中可以看出,KGE-CF 在 MovieLens 1M, Book-Crossings 和 Last, FM 3 个数据集上均取得了不错的表现。BPRMF 算法作为一种传统推荐算法,在 3 个数据集上的准确率、召回率及 F1-score 都是最低的。NCF 算法相比 BPRMF 算法,通过深度学习技术挖掘出用户和项目的潜在特征,一定程度地缓解了数据稀疏性,提升了推荐性能。CKE 算法仅优于 BPRMF 算法,因为其只利用了知识图谱的结构嵌入,未引入文本和图像特征。RippleNet 算法通过将知识图谱的实体与关系特征共同嵌入,缓解了数据稀疏性,获得了更细粒度的特征,使推荐性能获得提升。

KGE-CF 算法优于对比算法中表现最好的 RippleNet,其结合了知识图谱嵌入和深度学习,解决了数据稀疏性,提取了用户-项目之间的高阶关系特征。

(3)嵌入维数影响分析

为探索 KGE-CF 在不同嵌入维度时的表现,本文在 3 个数据集上通过不断增大嵌入维数 Dimension 来观察其对 AUC 的影响,如图 6 所示。由图 6 可见,在一定范围内,随着嵌入维度的增加,能够有效提高 KGE-CF 的性能,但当超过一定范围时,随着嵌入维数的增加,KGE-CF 的性能略微下降。这说明从知识图谱中编码有限的知识时,模型能学到更

多特征,从而性能得到提升,然而当编码的知识超过一定阈值时,模型会过拟合,导致性能下降。此外,还可以发现,在相对密集的 MovieLens-1M 数据集上,随着维数继续增大,AUC

指标下降得更为明显,在相对稀疏的数据集 Book-Corssings 和 Last.FM 上,下降稍显缓和,说明在稀疏数据上使用知识图谱嵌入技术更能减小对超参数嵌入维数的影响。

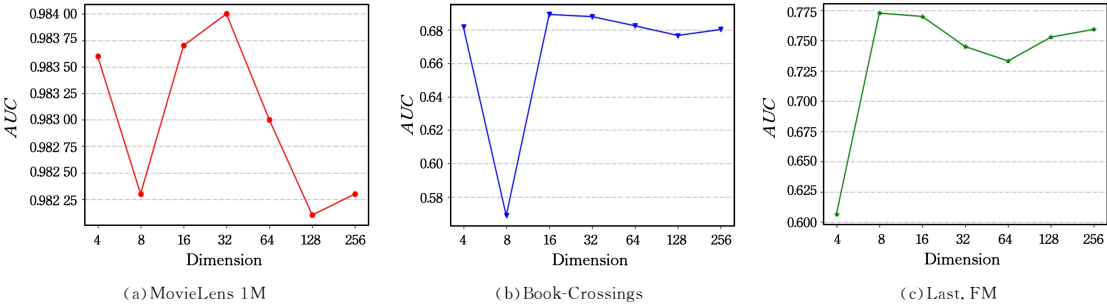


图 6 本文算法 3 个数据集上在不同维度下的 AUC
Fig. 6 Result of AUC for different dimensions on all three datasets

(4)算法处理稀疏数据能力分析
为了验证 KGE-CF 模型与对比模型在不同稀疏度时推荐性能的变化情况,本节在 MovieLens 1M 数据集上进行实验。

通过以下步骤从原始 ML-1M 数据集中随机选取 3 个数据稀疏集,在保持用户数和项目数不变的情况下,第一个数据稀疏集的评分大小为原数据集评分的 70%,第二个数据稀疏集的评分大小为第一个数据稀疏集评分的 50%,第三个数据稀疏集为第二个数据稀疏集评分大小的 20%

步骤 1 从原始用户-项目矩阵中随机选择一条评分记录,通过判断删除该记录后,该数据集的用户数和条目数是否不变。如果不变,删除此条评分;否则重复步骤 1。

步骤 2 在步骤 1 的多次重复之后,创建原始数据集的第一个稀疏性版本,以“-A”为后缀命名。

步骤 3 基于以-A 为后缀的数据集重复步骤 1 和步骤 2,以同样的方式创建以-B 后缀的数据集的第二个稀疏性版本。

同理,可得到以-C 为后缀的稀疏数据集,3 个不同稀疏度的数据集的统计如表 3 所列。

而本文提出的 KGE-CF 模型的 AUC 仅降低了 1.5%,这表明:

(1)在数据稀疏度提升时,4 种算法的推荐性能均有降低,但 KGE-CF 算法的推荐性能仍优于对比算法。

(2)KGE-CF 算法较对比算法更适用于数据稀疏度较高的情况,仍然保持了较高的推荐准确性。

结束语 本文针对推荐中存在的数据稀疏性问题,提出了 KGE-CF 模型,该模型将知识图谱与神经网络结合应用于协同过滤推荐算法,首先基于知识图谱的嵌入方法,扩展用户-项目交互数据的项目规模,缓解交互数据的稀疏性,然后通过神经网络对高阶特征的学习能力,学习用户向量和项目向量的深层次表示,提升用户-项目隐式关系。实验结果表明,基于知识图谱与深度学习的 KGE-CF 协同过滤推荐算法能够有效提升推荐质量。另外,本文还在 3 个不同稀疏度的数据集上进行了实验,实验结果表明提出的 KGE-CF 模型优于本文的对比模型,将知识图谱与深度学习关联结合到推荐系统中有助于缓解推荐系统中存在的数据稀疏性问题,进一步提高了推荐系统的推荐质量。

本文通过利用项目端的知识图谱展开了研究,但对于用户端的知识图谱依然存在的丰富的关联信息,因此,如何将用户端的知识图谱数据融入推荐系统,从而进一步提升推荐性能,是一个可改进的方向。同时,本文模型在运行效率上有所欠缺,使得本文的研究在真实场景中的运行效率仍然具有一定局限性。未来随着知识图谱嵌入技术的发展,将更大规模的图谱数据引入推荐系统的实验,能进一步帮助提升模型的效果。

表 3 不同稀疏度的数据集统计
Table 3 Datasets statistics with different level of sparsity

Datasets	# Users	# Items	# Ratings	# Sparsity/%
ML1M-A	6040	3706	739 843	96.694 8
ML1M-B	6040	3706	447 590	98.0004
ML1M-C	6040	3706	200 279	99.1053

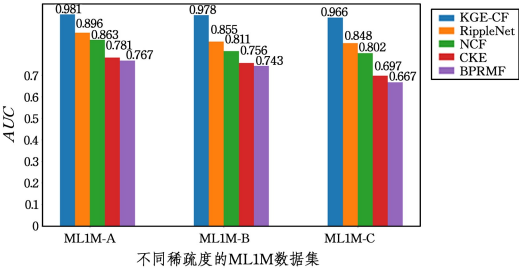


图 7 ML1M 数据集在不同稀疏度时的 AUC
Fig. 7 AUC on the ML1M datasets with different level of sparsity

从图 7 可以观察到,对于稀疏数据集 ML1M-A,在稀疏数据集 ML1M-C 时,4 个对比模型 RippleNet,NCF,CKE 和 BPRMF 的 AUC 分别降低了 4.8%,6.1%,8.1%和 9.9%。

参 考 文 献

[1] MARTÍNEZ L, PÉREZ L G, BARRANCO M. A multigranular linguistic content-based recommendation model[J]. International Journal of Intelligent Systems, 2007, 22(5): 419-434.

[2] SU X, KHOSHGOFTAAR T M. A survey of collaborative filtering techniques [J/OL]. Advances in Artificial Intelligence, 2009. <https://doi.org/10.1155/20091421425>.

[3] NILASHI M, BIN IBRAHIM O, ITHNIN N. Hybrid recommendation approaches for multi-criteria collaborative filtering [J]. Expert Systems with Applications, 2014, 41(8): 3879-3900.

[4] HE X, LIAO L, ZHANG H, et al. Neural collaborative filtering [C] // Proceedings of the 26th International Conference on

World Wide Web. 2017;173-182.

[5] WANG H,ZHANG F,WANG J,et al. Ripplenet:Propagating user preferences on the knowledge graph for recommender systems[C]// Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018; 417-426.

[6] CHENG H T,KOC L,HARMSSEN J,et al. Wide & deep learning for recommender systems [C] // Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. 2016: 7-10.

[7] GUO H,TANG R, YE Y, et al. DeepFM: a factorization-machine based neural network for CTR prediction[J]. arXiv:1703.04247,2017.

[8] HE X,CHUA T S. Neural factorization machines for sparse predictive analytics [C] // Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval. 2017;355-364.

[9] SEDHAIN S,MENON A K,SANNER S, et al. AutoRec: Autoencoders meet collaborative filtering[C]// Proceedings of the 24th international conference on World Wide Web. 2015: 111-112.

[10] DEVOOGHT R,BERSINI H. Long and short-term recommendations with recurrent neural networks[C]// Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization. 2017;13-21.

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

[12] LIN Y,LIU Z,SUN M,et al. Learning entity and relation embeddings for knowledge graph completion[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2015: 2181-2187.

[13] 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. 2018;1835-1844.

[14] JI G, HE S,XU L, et al. Knowledge graph embedding via dynamic mapping matrix[C]// Proceedings of the 53rd Nnual

Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (volume 1:Long papers). 2015;687-696.

[15] WANG H,ZHANG F,ZHAO M,et al. Multi-task feature learning for knowledge graph enhanced recommendation[C]// The World Wide Web Conference. 2019;2000-2010.

[16] SUN Z,YANG J,ZHANG J,et al. Recurrent knowledge graph embedding for effective recommendation[C] // Proceedings of the 12th ACM Conference on Recommender Systems. 2018;297-305.

[17] WANG X,WANG D,XU C,et al. Explainable reasoning over knowledge graphs for recommendation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(1): 5329-5336.

[18] CAO Y,WANG X,HE X,et al. Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences[C] // The World Wide Web Conference. 2019;151-161.

[19] ZHAO W X,HE G,YANG K,et al. Kb4rec: A data set for linking knowledge bases with recommender systems[J]. Data Intelligence,2019,1(2): 121-136.

[20] RENDLE S,FREUDENTHALER C,GANTNER Z,et al. BPR: Bayesian personalized ranking from implicit feedback[J]. arXiv: 1205. 2618.2012.



XU Bing, born in 1989, M. S. candidate. His main research interests include data mining and recommendation system.



PENG Jian, born in 1970, Ph.D, professor, Ph. D supervisor, is a outstanding member of China Computer Federation. He main research interests include big data and wireless sensor network.