基于知识图谱共同邻居排序采样的推荐模型

李世宝[①] 张益维*[①] 刘建航[①] 崔学荣[①] 张玉成[②]

[①](中国石油大学(华东)海洋与空间信息学院 青岛 266580)

[②](中国科学院智能农业机械装备工程实验室 北京 100190)

摘 要：知识图谱作为辅助信息可以有效缓解传统推荐模型的冷启动问题。但在提取结构化信息时，现有模型都忽略了图谱中实体之间的邻居关系。针对这一问题，该文提出一种基于共同邻居排序采样的知识图谱卷积网络(KGCN-PN)推荐模型，该模型首先基于共同邻居数目对知识图谱中的每个实体邻域进行排序采样；其次利用图卷积神经网络沿着图谱中的关系路径将实体自身信息与接收域信息逐层融合；最后将用户特征向量与融合得到的实体特征向量送入预测函数中预测用户与实体项目交互的概率。实验结果表明该模型在数据稀疏场景下相较其他基线模型性能均获得了相应提升。

关键词：知识图谱；推荐系统；排序采样；图卷积神经网络

中图分类号：TP391.1, TP311 **文献标识码**：A **文章编号**：1009-5896(2021)12-3522-08

DOI: 10.11999/JEIT200735

Recommendation Model Based on Public Neighbor Sorting and Sampling of Knowledge Graph

LI Shibao[①] ZHANG Yiwei[①] LIU Jianhang[①] CUI Xuerong[①] ZHANG Yucheng[②]

[①](*College of Oceanography and Space Informatics*, *China University of Petroleum* (*East China*), *Qingdao* 266580, *China*)

[②](*CAS Engineering Laboratory for Intelligent Agricultural Machinery Equipment*, *Beijing* 100190, *China*)

Abstract: The knowledge graph as auxiliary information can effectively alleviate the cold start problem of traditional recommendation models. But when extracting structured information, the existing models ignore the neighbor relationship between entities in the graph. To solve this problem, a recommendation model based on KnowledgeGraph Convolutional Networke-Public Neighbor (KFCN-PN) sorting sampling is proposed. The model first sorts and samples each entity's neighborhood in the knowledge graph based on the number of public neighbors; Secondly, it uses graph convolutional neural networks to integrate the entity's own information and the receiving domain information along the graph's relationship path layer by layer; Finally, the user feature vector and the entity feature vector obtained by the fusion are sent to the prediction function to predict the probability of the user interacting with the entity item. The experimental results show that the performance of this model is improved compared with other baseline models in data sparse scenarios.

Key words: Knowledge graph; Recommendation system; Sorted sampling; Graph convolutional neural network

1 引言

随着云计算、Web开发框架等一系列技术的不断进步，推荐系统广泛应用于在线平台以满足用户的个性化需求从而缓解信息过载带来的影响[1]。传统的推荐模型基于协同过滤算法[2–4]，该算法存在冷启动问题，且很难将用户没有浏览过但可能感兴趣的商品推荐给用户。文献[5,6]通过引进用户-项目对，社交网络等辅助信息的方法来解决因数据稀疏造成的冷启动问题。知识图谱作为异构图(heterogenous graph)的一种，图中节点对应实体，边对应实体之间的关系。图谱中实体间丰富的

收稿日期：2020-08-21；改回日期：2021-01-14；网络出版：2021-01-19

*通信作者： 张益维 yiwei9084@gmail.com

基金项目：国家自然科学基金(61972417, 61872385, 91938204)，国家重点研发计划(2017YFC1405203)，中国科学院科技服务网络计划(KFJ-STS-ZDTP-074)，中央高校基本科研业务费专项资金(18CX02134A, 19CX05003A-4, 18CX02137A)

Foundation Items: The National Natural Science Foundation of China(61972417, 61872385, 91938204), The National Key Research and Development Project(2017YFC1405203), The CAS Science and Technology Service Network Initiative(KFJ-STS-ZDTP-074), The Fundamental Research Funds for the Central Universities (18CX02134A, 19CX05003A-4, 18CX02137A)

语义关联可以帮助系统挖掘出不同实体之间的潜在关系，合理延伸用户的兴趣，从而提高推荐结果的精度和多样性[7]。

现有融合知识图谱的推荐模型大致可分为3类：基于嵌入的方法[8–11]、基于路径的方法[12–14]和嵌入加路径的混合方法[15,16]。(1)基于嵌入的方法中，为了提高模型的推荐质量，文献[8]依次使用翻译嵌入(Translating Embeddings, TransE)[17]算法、去噪编码器和卷积编码器分别提取项目的图谱结构特征、文本特征和视觉特征，最后将上述3个特征与项目自身偏移量求和作为项目最终的特征；为了增强模型对细粒度用户偏好的建模能力，文献[11]首先利用TransE等图谱嵌入算法将实体与关系映射到低维的特征向量中，然后将关系向量填充到键值对记忆网络(Key-Value Memory Networks, KV-MN)的密钥槽，其次使用循环神经网络(Recurrent Neural Network, RNN)的序列偏好向量作为查询条件读取特定用户的值向量，最后将序列偏好向量与经过编码附带注意力权重的值向量结合作为用户偏好的最终表示。常用的图谱嵌入方法侧重于对严格的语义关系进行建模，因此相较于推荐更适用于图谱补全和链路预测之类的任务场景；(2)基于路径的方法中，为了利用异构信息挖掘用户的高阶兴趣，文献[12]首先引入基于元路径的潜在特征来表示不同类型路径上用户和项目之间的连通性，然后定义全局和个性化两个层次的推荐模块，最后通过贝叶斯优化算法对模型进行训练；针对融合异构信息网络(Heterogeneous Information Network, HIN)的推荐模型无法学习元路径的显式表示以及未考虑元路径与用户-项目对之间的相互作用这两个问题，文献[13]结合元路径的上下文，设计了一种3向神经交互模型：首先基于优先级采样高质量的路径实例，然后使用共同注意力机制改进元路径上下文、用户以及上下文的表示。基于路径的方法在设计有效的基础路径时需要大量的领域知识，因此对于有着不同类型和关系的复杂知识图谱而言设计模板路径需要大量的专业人才，成本过高；(3)为了解决基于嵌入和基于路径两种方法的局限性，文献[7]通过沿着知识图谱中的关系路径迭代地扩展用户的潜在兴趣来刺激用户的偏好在实体集合上传播，最终由用户历史点击项目激活的多个实体集合叠加形成用户的偏好分布，该分布用于预测用户最终的点击概率。嵌入加路径的方法通过端到端的训练可以利用图谱中的高阶信息探索用户的潜在兴趣，克服了前两种方法的局限性。

上述3种方法中基于混合方法的推荐模型效果相对最佳，但是在对图谱中的实体邻居进行采样时，现有模型采用的机制均忽略了图谱中实体之间的邻居关系。为了更好地利用知识图谱中的结构信息，本文在构建知识图谱中每个实体的接收域时基于共同邻居数目对中心实体的邻居实体进行排序采样，提出了基于共同邻居排序的知识图谱卷积网络(Knowledge Graph Convolutional Networks-Public Neighbors sorting, KGCN-PN)模型，该模型首先基于共同邻居数目对知识图谱中的每个实体邻域进行排序采样。然后利用图卷积神经网络沿着图谱中的关系路径将实体自身信息与接收域信息逐层融合。其次将用户特征向量与融合得到的实体特征向量送入预测函数中预测用户与实体项目交互的概率。最后根据概率大小进行Top-K推荐。KGCN-PN在MovieLens-20M和Last.FM两个公开的数据集上与最近的基线模型相比，感受性曲线下的面积值(Area Under Curve, AUC)分别获得了1.9%和4.4%的提升。实验结果证明了KGCN-PN的有效性，同时在缺少用户-项目交互数据的冷启动场景中也保持了较强的推荐性能。

2　推荐问题建模

典型的推荐场景中有用户集合$\mathcal{U}=\{u_1,u_2,\cdots,u_M\}$和项目集合$\mathcal{V}=\{v_1,v_2,\cdots,v_N\}$。用户-项目交互矩阵$\boldsymbol{Y}\in\mathbb{R}^{M\times N}$由用户隐式反馈$y_{uv}$组成，其中$y_{uv}=1$表示用户$u$和项目$v$之间有过历史交互行为，例如浏览、点击或购买；否则$y_{uv}=0$。此外还有由3元组$\{h,r,t\}$构成的知识图谱$G$，其中$h\in\mathcal{E}$，$r\in\mathcal{R}$，$t\in\mathcal{E}$分别表示3元组的头部实体，关系和尾部实体，$\mathcal{E}$和$\mathcal{R}$分别代表知识图谱中的实体集合与关系集合。给定用户-项目交互矩阵$\boldsymbol{Y}$和知识图谱G，本文的任务是预测用户u是否会对之前未进行过交互的项目v产生潜在兴趣。具体来说，需要学习一个预测函数$\widehat{y}_{uv}=\mathcal{F}(u,v|\theta,\boldsymbol{Y},G)$，其中$\widehat{y}_{uv}$表示用户$u$和项目$v$的交互概率，$\theta$是函数$\mathcal{F}$的模型参数。

3　KGCN-PN模型架构

3.1　基于共同邻居排序的采样算法(PN)

尽管知识图谱中丰富的实体信息有助于提升推荐效果，但是过多的实体会导致模型在训练过程中因计算存储开销太大而出现无法收敛的情况。为了降低模型的计算存储开销，同时保持每批样本在训练过程中的时间复杂度相同，现有基于知识图谱的推荐模型均通过采样固定大小的邻居实体来不断扩大给定实体的邻域集合，同时将其邻域特征与自身特征合并来计算给定实体的特征表示。另外，知识图谱是由实体间不同的关系作为路径将所有实体相连构成的语义网络。为了挖掘用户潜在的偏好，增

强推荐结果的可解释性，模型需要将邻域集合沿着不同的路径扩展至多跳以外来捕捉用户的远距离兴趣。KGCN-PN通过扩展加聚合的方式将实体特征嵌入推荐模型中；同时它可以自动寻找从用户历史到潜在兴趣的可解释路径，期间不需要任何形式的人工设计。为了描述用户在知识图谱上逐层延伸的兴趣偏好，本文使用$N(h)=\{t|(h,r,t)\in G\}$表示与实体h直接相连的实体集合，采用递归的方式定义项目u在知识图谱上的k跳邻域集合

$$\begin{aligned}\varepsilon_u^k=&\{t|t\in N(t)\text{ and }(h,r,t)\in\mathcal{G}\text{ and }h\in\varepsilon_u^{k-1}\},\\&k=1,2,\cdots,\end{aligned}\tag{1}$$

其中，$\mathcal{E}_v^0=\{v\}$即为种子集合，H是预先定义的最大跳数。用户邻域集合可视为用户的历史兴趣在知识图谱上的自然延伸。而自然延伸的过程需要不断地对实体邻居进行采样，本文将每次采样得到的接收域定义为

$$S(e)=\{t|t\in N(e)\},\ |S(e)|=K\tag{2}$$

其中，K为采样大小。由此将式(1)中的$N(\cdot)$替换为$S(\cdot)$以获得更为紧凑的项目邻域集合，相应地式(1)中的用户邻域集合也会随之更新。值得注意的是，实体之间的共同邻居数量越多，说明实体之间的联系越紧密，即接收域经过编码得到的实体特征越丰富。然而传统方法在采样时忽略了待采样实体与中心实体之间的邻居关系，本文即根据上述事实提出了基于共同邻居的排序采样算法。

在KGCN-PN模型中，$S(e)$也称为实体e的单层接收域，图1为灰色实体的两层接收域示意图，其中黑色实体为采样过程中进入接收域的实体，白色实体为被忽略的实体，实体间箭头上的数字为共同邻居数目，为了描述简便，采样大小取3(模型实际训练过程中K值大应随训练集规模调整，4.5.2节给出了KGCN-PN在不同采样大小下的AUC值)。具体采样过程如下：

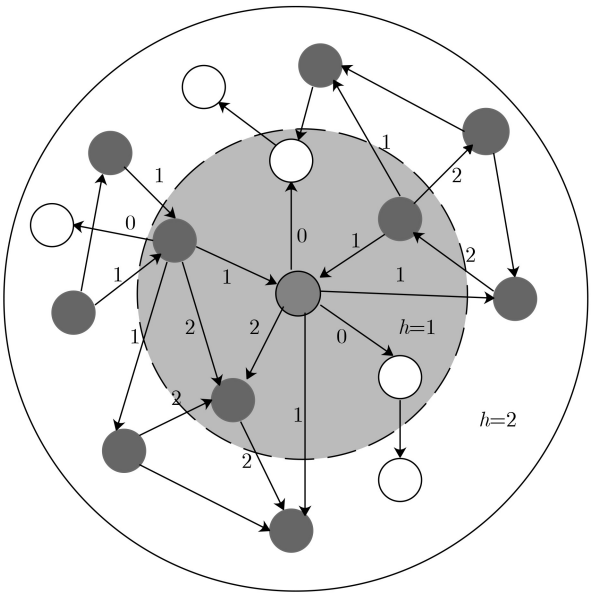


图 1 灰色实体的两层接收域

步骤 1　首先查找与中心实体e直接相连的实体集合$N(e)$，并统计$N(e)$中的实体个数$l(e)$。比较$l(e)$与采样值K的大小：如果$l(e)<K$，则在$N(e)$内选取K个实体(可重复)组成接收域$S(e)$；如果$l(e)\geq K$，则进入步骤2；

步骤 2　计算中心实体e的邻居列表$L(e)$：遍历集合$N(e)$取出每个元组内的第1个元素放入$L(e)$中；

步骤 3　计算共同邻居列表$T(e)$：对$L(e)$中的每个实体i执行步骤2中的操作得到$l(e)$个不同的邻居列表，分别与$L(e)$求交集，将每个交集中的元素个数放入$T(e)$中

$$T(e)=\bigcup_{i\in L(e)}L(i)\cap L(e)\tag{3}$$

步骤 4　对$T(e)$中的元素进行归并排序；

步骤 5　取出排序后的序列中后K个元素进行遍历，查找每个元素在$T(e)$中的索引位置并放入采样列表$I(e)$，遍历结束后删除采样列表中重复的索引，最后在$N(e)$中查找采样列表中索引对应的实体即可得到中心实体的单层接收域$S(e)$。

表1中的算法1给出了上述采样流程的伪代码。

下面证明本文采样方法的有效性：设实体e单层接收域的总信息量为$H(e)$，代表每个采样实体

表 1　基于共同邻居(PN)排序的实体采样算法(算法1)

输入：中心实体的直接邻居集合$N(e)$
输出：中心实体的单层接收域$S(e)$
(1) Function Public-Neighbors-Sampling(e)
(2) if $l(e)<K$
(3) $S(e)\leftarrow$ choose K entities from $N(e)$
(4) else
(5) for $i=0,1,\cdots,l(e)-1$ do
(6) $L(e)$.append($N(e)[i][0]$)
(7) $T(e)\leftarrow U_{i\in L(e)}L(i)\cap L(e)$
(8) $P(e)\leftarrow$do Merge Sort on $T(e)$
(9) for $e\in P(e)[l(e)-K:l(e)]$ do
(10) $I(e)$.append(the indices of e in $T(e)$)
(11) Remove duplicate index in $I(e)$
(12) for $i\in I(e)$ do
(13) $S(e)$.append($N(e)[i][0]$)
(14) return $S(e)$

与中心实体共同邻居的总和；$\varphi_e(v)$表示基于共同邻居排序采样得到的实体v与中心实体e的共同邻居个数；$\mathcal{R}_e(v)$表示随机采样得到的实体v与中心实体e的共同邻居个数。则上述两种算法采样得到的接收域信息差为

$$\sum_{v\in S(e)}\log_2\varphi_e(v)-\sum_{v\in S(e)}\log_2\mathcal{R}_e(v) \tag{4}$$

式(4)展开为

$$\begin{aligned}&\log_2\varphi_e(v_1)+\log_2\varphi_e(v_2)+\cdots+\log_2\varphi_e(v_k)\\&\quad-\log_2\mathcal{R}_e(v_1)-\log_2\mathcal{R}_e(v_2)-\cdots-\log_2\mathcal{R}_e(v_k)\end{aligned} \tag{5}$$

整理得到

$$\begin{aligned}&\log_2\varphi_e(v_1)-\log_2\mathcal{R}_e(v_1)+\log_2\varphi_e(v_2)-\log_2\mathcal{R}_e(v_2)\\&\quad+\cdots+\log_2\varphi_e(v_k)-\log_2\mathcal{R}_e(v_k)\end{aligned} \tag{6}$$

由分析可知，式(6)只有在两种采样方式采样得到的实体均相同时结果为0，其余情况结果均大于0，换言之基于共同邻居排序采样得到的实体接收域相较于随机采样信息量更大。

3.2 知识图谱上的图卷积过程(KGCN)

在获得知识图谱中每个实体的单层接收域后，本文需要将知识图谱转化为针对特定用户的权重图。现实生活中，不同的用户有着不同的兴趣偏好：针对同一件商品，用户a可能看重商品的外观，用户b更注重商品的质量。本文使用用户-关系评分函数$f_u(r)$来描述关系r对于用户u的重要程度：$f_u(r)=\boldsymbol{u}\times\boldsymbol{r}$，其中$\boldsymbol{u}$和$\boldsymbol{r}$分别为用户和关系的特征向量。

本文通过计算项目v的邻域线性组合来描述v的邻域拓扑结构

$$v^u_{S(v)}=\sum_{e\in S(v)}\tilde{f}_u(r_v,e)\cdot\boldsymbol{E} \tag{7}$$

其中，$\boldsymbol{E}$是实体e的特征向量，$\tilde{f}_u(r_{v,e})$是经过正则化的用户-关系分数

$$\tilde{f}(r_v,e)=\frac{\exp(f_u(r_v,e))}{\sum\limits_{e\in S(v)}\exp(f_u(r_v,e))} \tag{8}$$

如图2所示，在每个图卷积层中，实体的最终特征是由其自身和邻域实体共同决定的，经过多层传播，最终的输出结果v^u整合了实体自身的初始特征和L层的接收域特征。

表2中的算法2给出了图卷积层沿着知识图谱中的关系路径依次提取实体邻域特征的伪代码，其中L表示接收域的最大层数，特征向量的后缀$[l]$表示当前接收域的层数。对于给定的用户-项目对(u,v)，首先使用逐层迭代的方式计算v的接收域$\mathcal{M}$。然后循环L次“扩展加聚合”的操作：在每次循环中，计算当前层的接收域中每个实体的邻域特征，这些特征与前一层的特征$e^u[l-1]$进行聚合得到的新特征将用于下一轮循环。卷积层的最后一步是将v的特征向量$\boldsymbol{v}$和其邻域特征$v^u_{S(v)}$通过聚合器整合到向量$\boldsymbol{v}^u$中

$$\mathrm{agg}_{\mathrm{sum}}=\sigma(\boldsymbol{W}\cdot(\boldsymbol{v}+\boldsymbol{v}^u_{S(v)})+\boldsymbol{b}) \tag{9}$$

最终的L层实体特征v^u与用户特征向量$\boldsymbol{u}$一同送入预测函数$\mathcal{F}$中，计算得到用户$\boldsymbol{u}$与项目v发生交互的概率$\hat{y}_{uv}=\mathcal{F}(u,\boldsymbol{v}^u)$。

模型完整的损失函数为

$$\begin{aligned}\mathcal{L}=\sum_{u\in U}\Bigg(&\sum_{v:y_{uv}=1}\mathcal{J}(y_{uv},\hat{y}_{uv})\\&-\sum_{i=1}^{T^u}E_{v_i\sim P(v_i)}\mathcal{J}(y_{uv_i},\hat{y}_{uv_i})\Bigg)+\lambda\|\mathcal{F}\|_2^2\end{aligned} \tag{10}$$

其中，$\mathcal{J}(y_{uv},\hat{y}_{uv})$为真实概率和预测概率之间的交叉熵损失，$P$为负采样分布，$T^u$是用户$u$的采样个数，最后一部分是L2正则项。本文中$T^u=|\{v:y_{uv}=1\}|$，$P$遵循均匀分布。

3.3 算法复杂度分析

算法1计算项目接收域的时间复杂度为

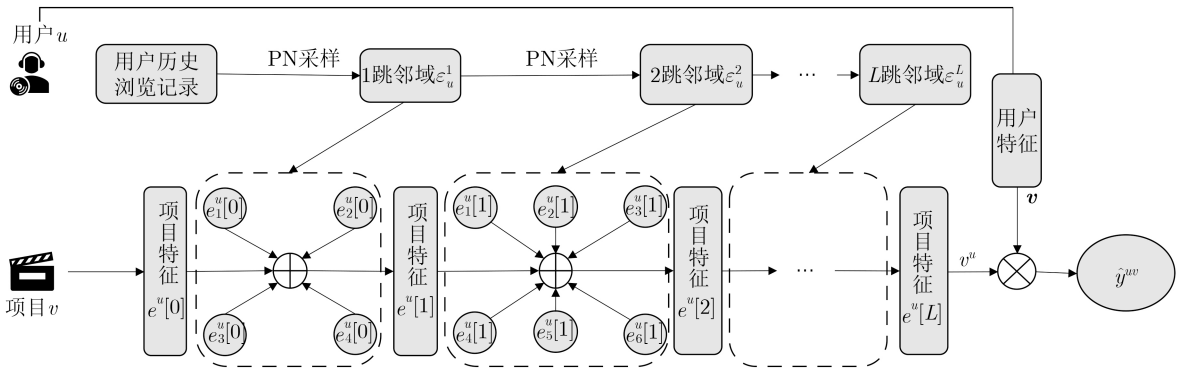


图 2 KGCN-PN模型架构

表 2 图卷积算法(算法2)

输入：交互矩阵$\boldsymbol{Y}$；知识图谱$G(\mathcal{E},\mathcal{R})$；接收域矩阵$\boldsymbol{S}$；

训练参数：$\{u\}_{u\in\mathcal{U}},\{e\}_{e\in\mathcal{E}},\{r\}_{r\in\mathcal{R}},\{W_i,b_i\}_{i=1}^{L}$；

输出：预测函数$\widehat{y}_{uv}=\mathcal{F}(u,v|\theta,\boldsymbol{Y},G)$

(1) while model does not converge do
(2) for (u,v) in Y do
(3) $\mathcal{M}[0]\leftarrow v$
(4) for $l=1,2,\cdots,L$ do
(5) $\mathcal{M}[l]\leftarrow\mathcal{M}[l-1]$
(6) for $e\in\mathcal{M}[l-1]$ do
(7) $\mathcal{M}[l]\leftarrow\mathcal{M}[l]\cup$Public-Neighbors-Sampling(e)
(8) return $\{\mathcal{M}[i]\}_{i=0}^{L}$
(9) $e^u[0]\leftarrow e,\forall e\in\mathcal{M}[0]$
(10) for $l=1,2,\cdots,L$ do
(11) for $e\in\mathcal{M}[l]$ do
(12) $e^u_{S(e)}[l-1]\leftarrow\sum_{e'\in S(e)}\tilde{f}_u\left(r_{v,e'}\right)e'^u[l-1]$
(13) $e^u[l]\leftarrow\text{agg}\left(e^u_{S(e)}[l-1],e^u[l-1]\right)$
(14) $v^u\leftarrow e^u[L]$
(15) Calculate the probability of interaction: $\widehat{y}_{uv}$
(16) Update (θ,W,b) in the direction of gradient descent
(17) return $\mathcal{F}$

$O(n^2+n\lg n)$，其中n为输入参数$N(e)$中实体的个数。算法2的每轮迭代中，每个循环的复杂度依次为Y,L,K^L,Kd和d^2，其中Y为用户-项目交互次数，L为最大跳数，K为采样大小，d为特征维度，因此算法2中图卷积的时间复杂度为$O(YLK^L(Kd+d^2))=O(YLK^{L+1}d+YLK^Ld^2)$。

4 实验分析

本节将对KGCN-PN模型在电影和音乐两个推荐场景中的表现进行评估。

4.1 数据集

MovieLens-20M是电影推荐中广泛使用的基准数据集，由MovieLens网站上约2×10^7条评价历史构成；Last.FM包含了Last.fm音乐平台上2000多名用户的收听信息。表3列出了两个数据集的统计内容。

表 3 实验数据集统计情况

	MovieLens-20M	Last.FM
用户数	138159	1872
项目数	16954	3846
交互次数	13501622	42346
实体数	102569	9366
关系种类	32	60
3元组数	499474	15518

4.2 测试配置

为了验证KGCN-PN模型的效果，本文将与下述的4个基线模型作对比：

奇异值分解(Singular Value Decomposition, SVD)[2]是典型的基于协同过滤的推荐模型，其中用户与项目交互的概率由各自的特征向量相乘计算得到：$y_{uv}=\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v}$。

协同知识嵌入(Collaborative Knowledge Embedding, CKE)[8]是基于嵌入方法的代表模型，基本思想是将知识图谱作为辅助信息引入模型的同时结合协同过滤方法完成推荐任务。

个性化实体推荐(Personalized Entity Recommendation, PER)[12]是基于路径方法的代表模型，它将知识图谱视为异构的信息网络并从中提取路径特征来表示用户与项目之间的潜在联系。

RippleNet[7]是基于混合方法的代表模型，类似于记忆网络，该模型沿着知识图谱中的关系探索用户的潜在兴趣。

4.3 实验设置

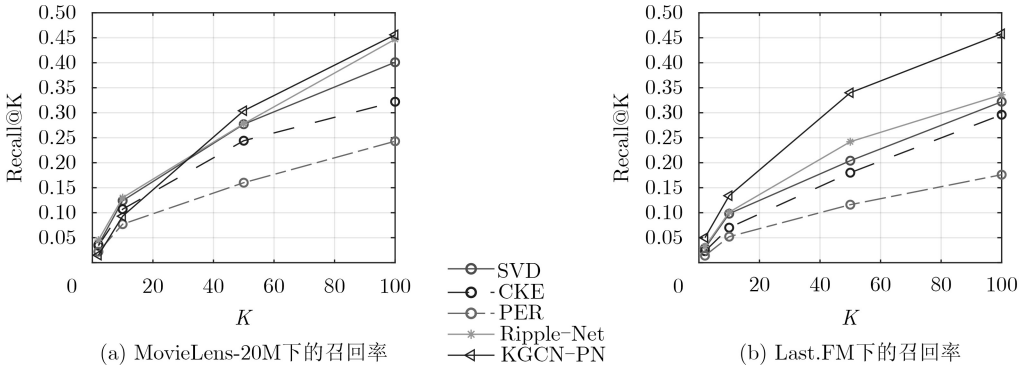
KGCN-PN中，$\mathcal{F}$为内积，最后一层接收域的激活函数σ为tanh，其余层的激活函数为ReLu。每个数据集中训练集、验证集和测试集的比例为6:2:2。本文在以下两个实验场景中评估模型性能：(1)点击率预测场景。本文使用训练好的模型对测试集中每个用户-项目对的交互情况进行预测，评价指标为模型的AUC值。(2)Top-K推荐场景。本文使用训练好的模型为测试集中的每个用户挑选K个预测点击率最高的项目，并选择Recall@K, Precision@K, ndcg@K 3个指标对推荐集分别进行评估。模型的所有训练参数由Adam算法进行优化。

4.4 实验结果分析

表4与图3—图5分别列出了点击率预测和Top-K推荐的实验结果。从中可以观察到KGCN-PN在音乐数据集上的提升要高于电影，这表明KGCN-PN能够有效缓解冷启动问题，因为Last.FM中的

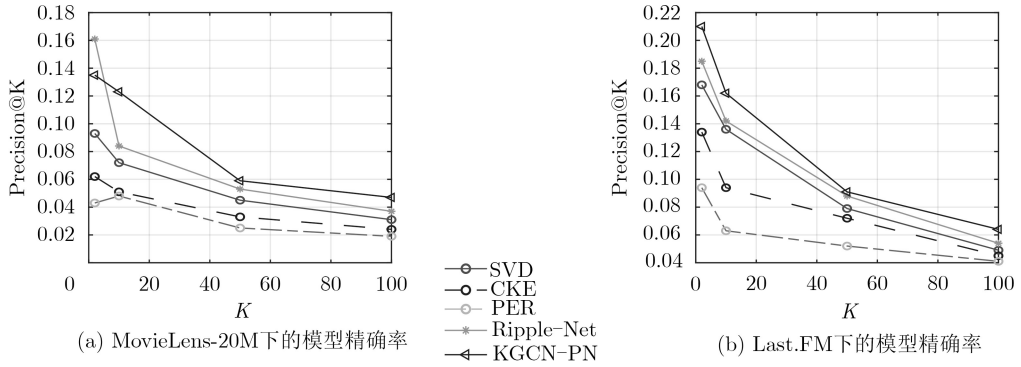
表 4 不同模型在点击率预测场景下的AUC值

模型	Movie	Music
SVD	0.963	0.769
PER	0.832	0.633
CKE	0.924	0.744
Ripple-Net	0.960	0.770
KGCN-PN	**0.979**	**0.804**



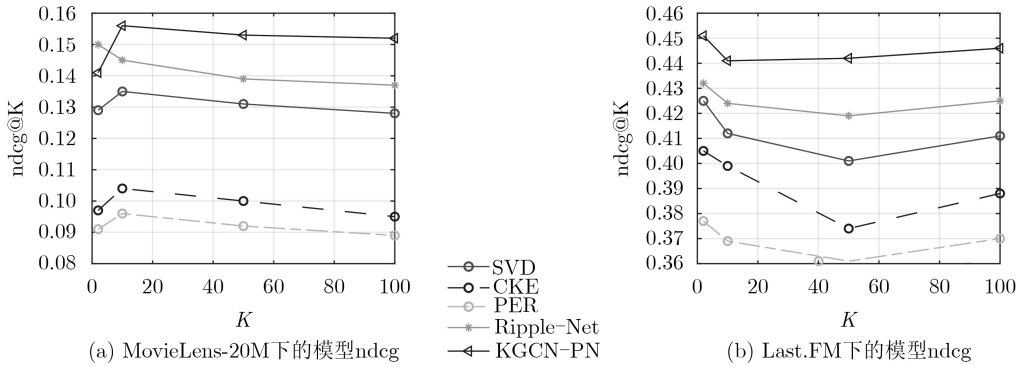
(a) MovieLens-20M下的召回率　　(b) Last.FM下的召回率

图 3 Top-K推荐中不同模型的召回率



(a) MovieLens-20M下的模型精确率　　(b) Last.FM下的模型精确率

图 4 Top-K推荐中不同模型的精确率



(a) MovieLens-20M下的模型ndcg　　(b) Last.FM下的模型ndcg

图 5 Top-K推荐中不同模型的ndcg

数据相比于MovieLens-20M更加稀疏；未引入知识图谱的SVD模型的性能实际上优于引入知识图谱的CKE和PER模型，这说明后两个模型没有充分利用知识图谱中的实体特征和人工设计的路径信息；Ripple-Net相比于其他基线模型有更优的性能，因为它也利用了图谱的多跳邻居结构，这说明捕捉知识图谱中实体的邻域信息对推荐任务有重要意义；KGCN-PN的AUC值相较于基线模型中性能最高的Ripple-Net,在Movie和Music数据集上分别提升了1.9%和4.4%。

如图3所示，在Movie数据集下当推荐的项目数K超过35时，KGCN-PN的召回率开始超过其他基线模型；在数据相对稀疏的Music数据集下KGCN-PN的召回率一直优于其他基线模型。图4中KGCN-PN的精确率在Movie和Music数据集上分别平均超过了Ripple-Net 15.72%和12.37%。图5中KGCN-PN在两个数据集上的ndcg均高于基线模型，同时折线的走势平缓，说明KGCN-PN的鲁棒性和稳定性较强。

4.4.1 冷启动场景中的实验结果

推荐系统中引入知识图谱的一个主要目标是缓解数据稀疏问题。本文在验证集与测试集大小固定的情况下，分别使用MovieLens-20M完整训练集的20%～100%5个不同大小的训练集进行训练，图6列出了5种情况下的AUC曲线。当训练集大小为完整训练集的20%时，相较于完整训练集4个基线模

型的AUC值分别下降了8.1%, 2.6%, 3.0%和3.9%, KGCN-PN仅下降了1.7%。这表明在用户-项目交互数据稀疏的情况下，KGCN-PN仍能保持良好的预测性能。

4.4.2 超参数敏感性分析

本文首先分析KGCN-PN对接收域层数L的敏感性：其他参数固定的情况下，依次更改L进行实验，实验结果如表5所示。L取1或2时模型效果最好，取4时效果最差，这是因为计算中心实体的邻域特征时过大的L会引入不必要的实体特征从而导致模型过度平滑。

本文其次分析了采样大小对模型性能的影响，实验结果如表6所示。起初随着K的增大，模型性能会提升，因为适当的采样大小能够编码更多的用户和实体信息；但是过大的采样数会导致模型过拟合。

5 结束语

本文提出基于知识图谱中共同邻居排序采样的推荐模型。KGCN-PN首先通过共同邻居排序算法获取知识图谱中每个实体的有效邻域，然后使用图卷积神经网络提取实体的邻域特征并结合用户特征进行项目推荐。实验结果表明在电影和音乐推荐场景中，KGCN-PN的性能优于其他基线模型。该模型不仅适用于电影和音乐推荐，也适用于其他构建了相应知识图谱的领域。

本文通过实验发现图卷积层数太深不仅会使模型收敛速度变缓，还会导致实体在训练过程中丢失自身特征，因此在保持GCN层数不变的前提下，如何更有效地学习实体特征将是下一步工作的重点。

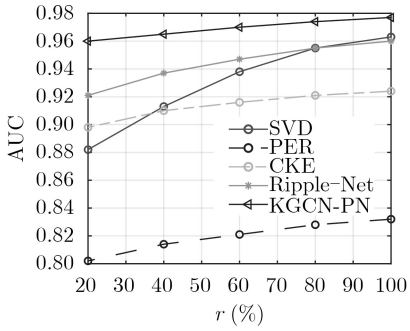


图 6 模型在不同规模训练集下的AUC曲线

表 5 KGCN-PN在不同接收域层数下的AUC值

	K			
	1	2	3	4
MovieLens-20M	0.976	**0.977**	0.969	0.501
Last.FM	**0.803**	0.787	0.527	0.512

表 6 KGCN-PN在不同采样大小下的AUC值

	K					
	2	4	8	16	32	64
MovieLens-20M	**0.977**	0.977	0.976	0.976	0.973	0.970
Last.FM	0.792	0.799	0.791	**0.804**	0.802	0.702

参 考 文 献

[1] ZHENG Guanjie, ZHANG Fuzheng, ZHENG Zihan, *et al*. DRN: A deep reinforcement learning framework for news recommendation[C]. The 2018 World Wide Web Conference, Lyon, France, 2018: 167–176. doi: 10.1145/3178876.3185994.

[2] 司亚利, 张付志, 刘文远. 基于签到活跃度和时空概率模型的自适应兴趣点推荐方法[J]. 电子与信息学报, 2020, 42(3): 678–686. doi: 10.11999/JEIT190287.
SI Yali, ZHANG Fuzhi, and LIU Wenyuan. An adaptive point-of-interest recommendation method based on check-in activity and temporal-spatial probabilistic models[J]. *Journal of Electronics & Information Technology*, 2020, 42(3): 678–686. doi: 10.11999/JEIT190287.

[3] KOREN Y. Factorization meets the neighborhood: A multifaceted collaborative filtering model[C]. The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, USA, 2008: 426–434. doi: 10.1145/1401890.1401944.

[4] 伊华伟, 张付志, 巢进波. 基于模糊核聚类和支持向量机的鲁棒协同推荐算法[J]. 电子与信息学报, 2017, 39(8): 1942–1949. doi: 10.11999/JEIT161154.
YI Huawei, ZHANG Fuzhi, and CHAO Jinbo. Robust collaborative recommendation algorithm based on fuzzy kernel clustering and support vector machine[J]. *Journal of Electronics & Information Technology*, 2017, 39(8): 1942–1949. doi: 10.11999/JEIT161154.

[5] WANG Hongwei, ZHANG Fuzheng, HOU Min, *et al*. SHINE: Signed heterogeneous information network embedding for sentiment link prediction[C]. The Eleventh ACM International Conference on Web Search and Data Mining, Marina Del Rey, USA, 2018: 592–600. doi: 10.1145/3159652.3159666.

[6] CHENG H T, KOC L, HARMSEN J, *et al*. Wide & deep learning for recommender systems[C]. The 1st Workshop on Deep Learning for Recommender Systems, Boston, USA, 2016: 7–10. doi: 10.1145/2988450.2988454.

[7] WANG Hongwei, ZHANG Fuzheng, WANG Jialin, *et al*. RippleNet: Propagating user preferences on the knowledge graph for recommender systems[C]. The 27th ACM International Conference on Information and Knowledge Management, Torino, Italy, 2018: 417–426. doi: 10.1145/3269206.3271739.

[8] ZHANG Fuzheng, YUAN N J, LIAN Defu, *et al*. Collaborative knowledge base embedding for recommender systems[C]. The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, California, San Francisco, USA, 2016: 353–362. doi: 10.1145/2939672.2939673.

[9] WANG Hongwei, ZHANG Fuzheng, XIE Xing, *et al*. DKN: Deep knowledge-aware network for news recommendation[C]. The 2018 World Wide Web Conference, Lyon, France, 2018: 1835–1844. doi: 10.1145/3178876.3186175.

[10] WANG Hongwei, ZHANG Fuzheng, ZHAO Miao, *et al*. Multi-task feature learning for knowledge graph enhanced recommendation[C]. The World Wide Web Conference, San Francisco, USA, 2019: 2000–2010. doi: 10.1145/3308558.3313411.

[11] HUANG Jin, ZHAO W X, DOU Hongjian, *et al*. Improving sequential recommendation with knowledge-enhanced memory networks[C]. The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, USA, 2018: 505–514. doi: 10.1145/3209978.3210017.

[12] YU Xiao, REN Xiang, SUN Yizhou, *et al*. Personalized entity recommendation: A heterogeneous information network approach[C]. The 7th ACM International Conference on Web Search and Data Mining, New York, USA, 2014: 283–292. doi: 10.1145/2556195.2556259.

[13] HU Binbin, SHI Chuan, ZHAO W X, *et al*. Leveraging meta-path based context for top- n recommendation with a neural co-attention model[C]. The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, United Kingdom, 2018: 1531–1540. doi: 10.1145/3219819.3219965.

[14] ZHAO Huan, YAO Quanming, LI Jianda, *et al*. Meta-graph based recommendation fusion over heterogeneous information networks[C]. The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, Canada, 2017: 635–644. doi: 10.1145/3097983.3098063.

[15] WANG Xiao, WANG Ruijia, SHI Chuan, *et al*. Multi-component graph convolutional collaborative filtering[J]. *The AAAI Conference on Artificial Intelligence*, 2020, 34(4): 6267–6274. doi: 10.1609/aaai.v34i04.6094.

[16] WANG Xiang, HE Xiangnan, CAO Yixin, *et al*. KGAT: Knowledge graph attention network for recommendation[C]. The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, USA, 2019: 950–958. doi: 10.1145/3292500.3330989.

[17] BORDES A, USUNIER N, GARCIA-DURÁN A, *et al*. Translating embeddings for modeling multi-relational data[C]. The 26th International Conference on Neural Information Processing Systems, Lake Tahoe, USA, 2013: 2787–2795.

李世宝：男，1978年生，硕士，副教授，硕士生导师，研究方向为移动计算、无线通信.

张益维：男，1995年生，硕士生，研究方向为知识图谱推荐技术.

刘建航：男，1978年生，博士，副教授，研究方向为车联网.

崔学荣：男，1979年生，博士，教授，研究方向为智能感知.

张玉成：男，1980年生，博士，副研究员，研究方向为智能信息处理.

责任编辑：余　蓉