

文章编号: 2095-2163(2024)03-0010-07

中图分类号: TP391

文献标志码: A

利用共有属性采样改进的 RippleNet 推荐模型

江山, 韩永华

(浙江理工大学 计算机科学与技术学院, 杭州 310018)

摘要: 针对 RippleNet 模型推荐结果中存在的用户兴趣属性失真问题, 提出一种利用共有属性采样改进的推荐模型 RippleNet-CA。首先, 该模型通过共有属性采样替代原模型的随机采样, 来重构知识图谱上的扩展偏好集合 (RippleSet), 基于增加 RippleSet 内部节点间的共有属性频数、增强多跳扩展偏好之间的相关性, 来提高用户特征包含的信息量。此外, 将用户特征和物品特征送入模型, 计算用户点击物品的概率, 通过排序点击概率得到 top - k 推荐。实验结果表明, 该模型在用户历史偏好多样的场景下, 推荐结果更加符合用户的偏好习惯, 推荐指标 AUC (Area Under Curve) 和多样性均获得了相应提升。

关键词: 推荐系统; 知识图谱; 共有属性采样; RippleNet 推荐模型

An improved RippleNet recommendation model using common attribute sampling

JIANG Shan, HAN Yonghua

(School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: To address the distortion problem of user interest attributes in the recommendation results of the RippleNet recommendation model, an improved RippleNet recommendation model RippleNet-CA using shared attribute sampling is proposed. First of all, the model reconstructs the extended preference set (RippleSet) on the knowledge graph by replacing the random sampling of the original model with shared attribute sampling and improves the information contained in user features based on increasing the frequency of shared attributes among nodes within the RippleSet and enhancing the correlation between user historical preferences and recommendation preferences. In addition, user features and item features are fed into the model to calculate the probability of users clicking on items, and top - k recommendations are obtained by ranking the click probabilities. The experimental results show that the model's recommendation results are more consistent with users' preference habits in scenarios with diverse user history preferences, and the recommendation metrics AUC (Area Under Curve) and diversity are improved accordingly.

Key words: recommendation system; knowledge graph; common attribute sampling; RippleNet recommendation model

0 引言

随着互联网信息技术的飞速发展, 互联网上的内容数量呈指数级增长, 随之而来的信息过载问题日益突出。传统推荐系统虽然降低了人们获取信息的成本, 提高了用户获取有用信息的效率, 一定程度上解决了信息过载问题, 但仍然存在冷启动、数据稀疏、可解释性^[1-2]差、推荐内容与用户兴趣^[3-6]不符等问题。知识图谱作为辅助信息, 能够解决推荐系统的数据稀疏和冷启动问题, 并且为推荐结果提供了可解释性。

RippleNet^[7]模型引入谷歌构建好的电影知识图谱, 将用户历史偏好物品与知识图谱实体对齐^[8], 在对齐后的知识图谱上, 以用户评分超过4分

的电影作为扩散起点集合, 以知识图谱中连接实体的关系边向外扩散2次, 将得到的实体集合作为用户兴趣的扩展^[9]。扩展偏好解决了用户电影交互数据集稀疏的问题。通过扩展偏好筛选出的推荐电影, 在知识图谱上的高相关概率扩散路径也可以解释推荐这部电影的原因, 增强了推荐结果的可解释性。其不足之处, 是在偏好扩散采样过程中, 固定采样邻居个数的采样方法, 在邻居较多和较少时, 可能采样到无关语义下的物品, 从而引入噪声。固定采样邻居个数随机采样得到的偏好, 不能准确区分用户对不同偏好物品之间相同属性的喜爱程度, 从而导致模型准确率下降。同时, 单次采样无法得到用户兴趣属性种类的真实分布, 导致推荐结果中用户

作者简介: 江山(1995-), 男, 硕士研究生, 主要研究方向: 推荐系统、数据挖掘。

通讯作者: 韩永华(1976-), 女, 博士, 讲师, 主要研究方向: 计算机视觉与模式识别。Email: han_yong_huahan@zstu.edu.cn

收稿日期: 2023-03-12

兴趣属性多样性和推荐满意度下降。

刘伟江等学者^[10]考虑到 RippleNet 模型在偏好扩散过程中,随着跳数的增大,偏好传播中噪声引起的偏好失真导致最下层的物品与最上层的物品相关概率大幅下降的问题,提出了“同级参评用户与同级受评项目具有同级评分”的思想,通过采样减少了不必要的计算量,降低了无关语义路径^[11]引入的噪声。首先制定优先选择由高词频语义关系连接的知识图谱节点的偏好扩散规则,然后对偏好传播中的噪声进行修正,最后构建用户兴趣送入模型进行预测。这种方法将隐式反馈细化为3个级别,提高了 RippleNet 推荐的准确率。不足之处是,没有考虑到用户兴趣属性多样性的问题,也无法同时满足模型推荐的多样性、新颖性和用户满意度。

文献[12-15]通过对项目高阶邻居节点的聚合,有效地提升推荐性能,通过注意力机制识别项目周围实体的重要性。虽然相比 RippleNet 模型的随机选择邻居方式更具优势,但仍忽视了实体之间的联系。针对于此,文献[16]考虑到了实体之间的联系,重构用户兴趣特征,提出了基于知识图谱共同邻居排序采样的推荐模型(Knowledge Graph Convolutional Network - Public Neighbors sorting, KGCN-PN)。首先,通过共同邻居排序采样,保留了偏好传播过程中上下层之间共有属性最多的邻居节点,提高了采样后用户兴趣所包含的信息量;然后,使用卷积神经网络提取采样后的偏好数据集的特征,送入 KGCN 模型进行推荐;在数据稀疏的情况下提高了推荐结果的 AUC、ACC;不足之处是 KGCN 模型的推荐结果缺乏可解释性。

文献[17]将知识图谱推荐系统与多目标优化算法结合,同时从推荐的准确率、多样性和可解释性三个方面优化,为推荐系统评价标准提供了新的研究思路。目前,推荐算法性能大多从准确率进行评估。然而,不断提高推荐结果的准确性,可能导致多样性和可解释性下降,造成“信息茧房”^[18]效应。真正好的推荐系统应该同时兼顾多种评估指标。虽然知识图谱为推荐系统带来了多样性、可解释性等方面性能提升,但现有研究却很少从这些性能上进行评估。本文从提高用户兴趣属性多样性角度出发,通过设计偏好共有属性采样,对 RippleNet 模型进行了改进。

1 RippleNet-CA 模型

本章介绍了与本文工作相关的模型,其中包括2类模型。一类是向外传播(outward propagation)模

型,代表模型是 RippleNet。该模型模拟水波向外扩散,将知识图谱上扩散的实体视为用户潜在的偏好物品,通过计算层级之间的相关性,筛选出相关性高的用户潜在感兴趣^[19]的物品。另一类是向内聚合模型,代表模型是 KGCN 模型。KGCN 在学习知识图谱中实体特征时,聚合了该实体的邻居特征表示。通过增加迭代次数,邻居的定义可以扩展到多跳之外,从而实现了对用户潜在兴趣的挖掘。

上述2类模型各有优缺点,RippleNet 模型可以自动挖掘用户潜在偏好,可解释性好,但是当知识图谱规模扩大时,运行时间会呈指数级增长。KGCN 推荐准确率高,而且当知识图谱增大时运行时间变化不大,但由于使用图神经网络替代相关概率筛选偏好邻居节点,无法提取路径进行解释,可解释性差。由此可见,这2类模型都没有考虑到知识图谱偏好物品之间的共有属性。KGCN-PN 解决了 KGCN 忽视偏好间邻居关系的问题,而本文主要解决了 RippleNet 忽视偏好间共有属性的问题。

RippleNet-CA 模型框架如图1所示。由图1可知,RippleNet-CA 模型框架主要由采样模块和预测模块两部分组成。

采样模块使用 RippleSet 共有属性采样,替代偏好传播过程中原模型的随机采样。通过 RippleSet 共有属性采样重构偏好特征,使 RippleSet 内部节点间的相关性得到了增强,用户偏好特征包含的信息量得到了提高。采样模块的构思来自于偏好传播过程中干涉加强现象的启发。

预测模块的描述如下:RippleNet 是一个结合知识图谱端到端的推荐模型,通过重构物品 v 和用户 u 的特征表示,计算用户点击物品的概率,并据此预测用户 u 对物品 v 的潜在兴趣。假设所有用户兴趣分布满足正态独立同分布,最大化观测到的知识图谱 G 和交互矩阵 Y 的模型参数的后验概率。模型的目标是通过知识图谱关注用户的对物品的偏好,其训练的预测函数为:

$$y_{uv} = F(u, v, \delta) \quad (1)$$

其中, y_{uv} 表示用户 u 点击物品 v 的概率, δ 表示函数 F 的参数。

使用小批量梯度下降算法优化损失函数为:

$$\begin{aligned} \min L = & -\log(p(Y| \Theta, G) \cdot p(G| \Theta) \cdot p(\Theta)) = \\ & \sum_{(u,v) \in Y} -(y_{uv} \log \sigma(\mathbf{u}^T \mathbf{v}) + (1 - y_{uv}) \log (-\sigma(\mathbf{u}^T \mathbf{v}))) + \\ & \frac{\lambda_2}{2} \sum_{r \in R} \|\mathbf{I}_r - \mathbf{E}^T \mathbf{R} \mathbf{E}\| + \frac{\lambda_1}{2} (\|\mathbf{V}\|_2^2 + \|\mathbf{E}\|_2^2 + \sum_{r \in R} \|\mathbf{R}\|_2^2) \end{aligned} \quad (2)$$

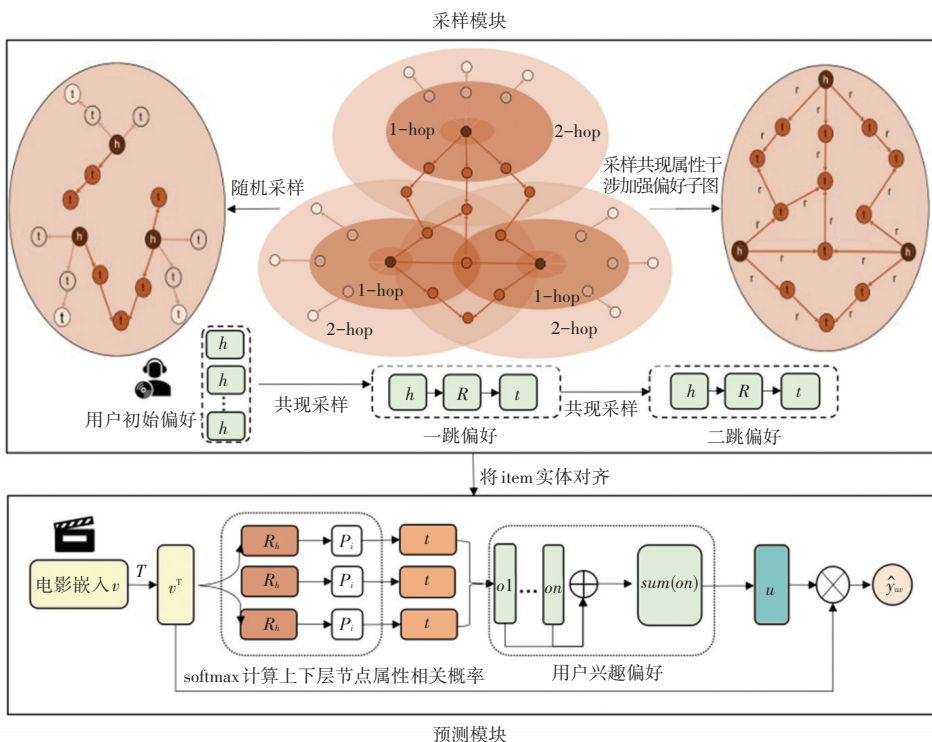


图1 RippleNet-CA 模型框架

Fig. 1 A RippleNet-CA model framework

1.1 偏好传播

1.1.1 偏好特征构造

由图1可知,波纹中心点就是种子节点。波纹集合是由与种子节点有 k 跳(hop)距离的三元组组成。一跳节点表示一个三元组关系可以连接的节点,知识图谱作为偏好传播的距离由跳数 k 决定,而第零阶段则代表用户的历史种子节点。用户 u 的 k 跳波纹集和(RippleSet)被定义为头节点 h 在 \mathcal{E}_u^{k-1} 中的三元组集合:

$$S_u^k = \{(h, r, t) \mid (h, r, t) \in \mathcal{E}_u^{k-1}\}, k = 1, 2, \dots, H \quad (3)$$

RippleNet的用户特征聚合了多跳RippleSet的节点特征和路径特征,并且能够给出相关性最高的路径作为可解释的推荐结果。 h 和 v_i 在关系空间 R_i 的相似度也叫做相关概率 p_i ,计算公式为:

$$p_i = \text{softmax}(v^T R_i h_i) = \frac{\exp(v^T R_i h_i)}{\sum \exp(v^T R h)} \quad (4)$$

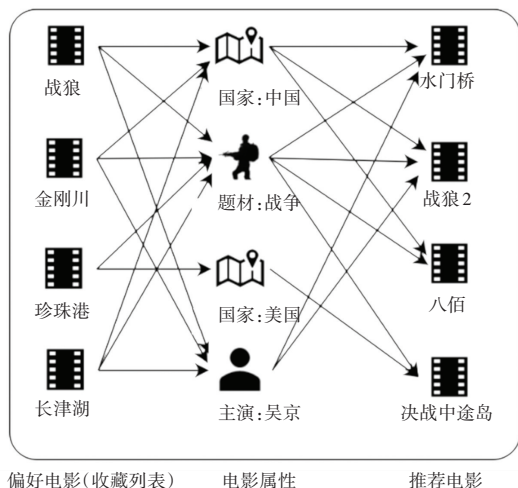
用户兴趣特征 u 的计算公式为:

$$u = \sum p_i t_i \quad (5)$$

最后,结合用户特征和物品特征计算预测的点击概率:

$$\hat{y}_{uv} = \text{sigmoid}(u^T v) \quad (6)$$

用户兴趣电影共有属性如图2所示。图2中,用户兴趣在构建的电影知识图谱传播过程中,奇数跳尾节点表示用户感兴趣电影的属性实体,偶数跳尾节点表示偏好属性对应的电影实体。



偏好电影(收藏列表) 电影属性 推荐电影

图2 用户兴趣电影共有属性

Fig. 2 User preference for co-occurring attributes

由图2可见,一跳波纹集合用户A的偏好的属性实体为:中国,战争,吴京等;二跳波纹集和用户A的扩展偏好电影实体为:水门桥,战狼2,八佰,决战中途岛。每部电影属性^[20]偏好在提取出的三元组中的 (r, t) 组合特征共有频次,代表了用户对该偏

好物品的偏爱程度。

用户兴趣经过两跳以上的传播,由于相关概率 p_i 小于 1,三跳及以后的偏好与种子相关性急剧下降且 RippleSet 呈爆炸性增长,极大地增加了计算量。因此,相关研究一般默认超参数跳数 $H = 2$ 。虽然文献[16]通过增强用户高阶兴趣,表示在 $H = 3$ 、 $H = 4$ 时获得了最优效果,但仍然存在计算量巨大的问题。

1.1.2 干涉加强

用户兴趣在知识图谱传播过程中存在像水波扩散一样的现象。随着波纹扩散,波纹逐渐衰减^[21],但当存在多个波源时,波纹间会相互叠加形成干涉加强^[6]区域。即如图 1 中采样模块所示,用户的偏好物品扩散出的波纹集会产生多个波纹,波纹间通过偏好物品间的共有属性产生联系,相互干涉叠加形成干涉加强区域,即图 1 中深色节点集合。

为此,本文提出一种设想:如果能够尽可能多地干涉加强区域(即采样出图 1 中采样模块右侧的子图)采样出来作为用户兴趣特征,则该模型将能更快速和充分地学习到用户兴趣,使推荐在用户兴趣属性层面更具准确性和多样性。

1.2 RippleNet 采样算法改进

1.2.1 随机采样

RippleNet 的采样方式借鉴了 GraphSage^[22]、FastGCN^[23]、AS-GCN^[24] 等的采样方式,对每个节点进行固定数量邻居的随机采样。与基于 GCN 改进模型的采样方式不同的是,GCN 模型学习的是图的整体嵌入,采样要考虑图的原始结构分布信息,需要学习包含图结构分布的图嵌入,计算开销巨大。而 RippleNet 仅考虑知识图谱作为辅助信息的偏好物品概率分布,并未考虑图的原始结构分布信息,因此无需学习图结构分布的图嵌入。RippleNet 采用三路张量分解来表示三元组结构信息,获得三元组的嵌入,而非整体知识图谱的图嵌入。作为辅助信息的知识图谱,应尽可能多地包含共有属性信息,否则作为辅助信息的信息增益将会减少。

RippleNet 采用随机采样的方法,不能很好地刻画用户兴趣物品属性在知识图谱中的分布情况,采样偏好过少会导致数据稀疏、冷启动,尤其是当固定采样个数较小的时候。随机选取节点的方式虽然能够大幅减少计算量,但同时也会使用户兴趣数据集具有不确定性,从而导致用户兴趣的不稳定,进而直接影响推荐结果准确性。

由图 2 可知,RippleNet 模型在 H 跳邻居上随机

地采样固定个数的邻居作为 H 阶偏好,这种采样方式会丢失用户对不同偏好的喜爱程度信息。比如用户 A 对中国电影偏好远比美国电影要多,结果通过单次固定个数的随机采样,大概率会丢失用户对不同类型电影属性的比例信息,导致模型误判美国或其它国家电影为用户最喜欢的电影类型,从而影响推荐结果的准确性。再比如,用户 A 有 100 部喜欢的电影,其中不同国家电影比例为中国:美国:俄国=18:1:1。按照原有的 RippleNet 模型采样算法,随机采样 32 部电影,单次采样的结果中 3 个国家电影的比例很难与采样前保持一致,无论是对美国和俄国的电影进行更多次或更少次的采样都会引入噪声。事实上,电影的属性不止一种,用户感兴趣的属性也不止一种,RippleNet 模型也无法让推荐的电影每种属性都能与用户的偏好相符。

1.2.2 构造属性-类别权重矩阵

定义属性集合 $C = \{c_1, c_2, \dots, c_s\}$, 用来表示 (r, t) 组合特征的不同类别;用户偏好物品集合 $U = \{u_1, u_2, \dots, u_m\}$, 根据用户-评分数据和属性-类别数据,统计得到属性-类别次数矩阵^[25-26]。偏好属性-类别权重矩阵见表 1。表 1 中, n_{ij} 为用户 u_i 所喜欢物品属性在知识图谱中 (r, t) 共现的次数。规定每个用户偏好物品对属性次数的总贡献值为 1,因此对于有多个类别属性的用户偏好物品,应为该类别偏好属性在偏好物品总类别中出现的频率。如:某用户偏好物品有 3 个偏好属性,则每个偏好属性的次数为 1/3。

表 1 偏好属性-类别权重矩阵

Table 1 Preferred attribute-category weight matrix

	c_1	c_2	\dots	c_s
u_1	n_{11}	n_{12}	\dots	n_{1s}
u_2	n_{21}	n_{22}	\dots	n_{2s}
\dots	\dots	\dots	\dots	\dots
u_m	n_{m1}	n_{m2}	\dots	n_{ms}

进一步可计算出用户偏好物品 u 对偏好属性类别 c 的权重为:

$$p_{uc} = \frac{n_{uc}}{\sum_{c=c_1}^{c_s} n_{uc}} \quad (7)$$

其中,分子 n_{uc} 表示用户偏好物品 u 的属性 c 在知识图谱中共现的次数,分母表示用户偏好物品 u 喜欢的所有物品不同属性的总频数之和。

1.2.3 共有属性采样算法

针对 RippleNet 模型在扩散偏好过程中采用随

机采样邻居节点带来的问题,本文提出基于RippleSet共有属性的RippleNet的推荐算法,重构用户兴趣 u 。算法设定用户偏好误差门限 T ,可用于调节用户偏好误差容忍度,即容许与用户历史偏好有一定的差距。 $T \in [0, 1]$,设置为0时,表示用户偏好权重至少与历史偏好相等;大于0时,表示用户偏好权重比历史偏好权重减少;值越大,减少的程度越大。共有属性采样算法描述见算法1。

算法1 RippleSet共有属性采样

输入 交互矩阵 \mathbf{Y} 、知识图谱 G 、固定采样个数 k 、偏好误差门限 T

输出 RippleSet

初始化:采样概率列表 $\mathbf{P} \leftarrow []$ 采样索引列表 $\mathbf{I} \leftarrow []$

1. 为每个用户计算波纹集合 $\{S_u^k\}_{k=1}^H$
 2. for (h, r, t) in G do
 3. 统计不同 h 共有的 (r, t) 类别数 m 和频数 n_{uc}
 4. 计算 m 类 (h, r, t) 概率数组 \mathbf{P} , $\mathbf{P}_{uc} = n_{uc}/\text{sum}(n_{uc})$
 5. if $m = k$ then
 6. 查找 m 类 (h, r, t) 对应索引得到索引列表 \mathbf{I}
 7. 根据概率数组 \mathbf{P} ,采样个数 k 和索引数组 \mathbf{I} 进行共有属性采样
 8. if $m < k$ then
 9. 按 n_{uc} 排序 (r, t)
 10. 查找前 m 类 (h, r, t) 对应索引得到索引列表 \mathbf{I}
 11. 根据概率数组 \mathbf{P} ,采样个数 k 和索引数组 \mathbf{I} 进行共有属性采样
 12. if $m > k$ then
 13. 按 n_{uc} 排序 (r, t)
 14. 计算前 $k \times (1 + T)$ 类 (h, r, t) 概率数组 \mathbf{P} , $\mathbf{P}_{uc} = n_{uc}/\text{sum}(n_{uc})$
 15. 查找前 $k \times (1 + T)$ 类 (h, r, t) 对应索引得到索引列表 \mathbf{I}
 16. 根据概率数组 \mathbf{P} ,采样个数 k 和索引数组 \mathbf{I} 进行共有属性采样
 17. end for
 18. return RippleSet
- 文献[16]通过信息熵公式,证明了共有属性采

样方法对采样结果包含信息量的提高:设 o_i 的总信息量为 $H(o_i)$,代表每个用户不同偏好物品的扩展三元组中 (r, t) 频数总和; $P_e(v)$ 表示基于用户不同偏好物品的扩展三元组中共有 (r, t) 排序采样得到的 (r, t) 共有频数; $R_e(v)$ 表示随机采样得到的 (r, t) 共有频数; $P_e(v)$ 大于 $R_e(v)$,则上述2种算法采样得到的 o_i 信息差为:

$$\sum_{o \in U} \log_2 P_e(v) - \sum_{o \in U} \log_2 R_e(v) \quad (8)$$

式(8)展开为:

$$\log_2 P_e(v_1) + \log_2 P_e(v_2) + \dots + \log_2 P_e(v_k) - \log_2 R_e(v_1) - \log_2 R_e(v_2) - \dots - \log_2 R_e(v_k) \quad (9)$$

整理得到:

$$\log_2 P_e(v_1) - \log_2 R_e(v_1) + \log_2 P_e(v_2) - \log_2 R_e(v_2) + \dots + \log_2 P_e(v_k) - \log_2 R_e(v_k) \quad (10)$$

分析可知,式(8)的计算结果大于或等于0。因为式(10)只有在2种采样方式得到的实体均同时结果为0,其余情况结果均大于0。换言之,基于共有属性偏好^[25]排序采样得到的 o_i 相较于随机采样信息量更大,随机采样在一定程度上丢失了偏好特征信息。

由图2可知,《战狼》和《金刚川》在考虑电影题材和主演时是高度相似的,但在考虑年代或导演时相似度会小很多,因此共有关系越多,2部电影之间的相似度越高。从图1的采样模块可以看出,经过偏好共有属性采样得到的用户兴趣子集中,共有关系频数增多,上下层间关联度 p_i 增大。又因为 p_i 正比于 u , u 正比于 \hat{y}_{uv} ,所以该采样方法会使用户点击物品的 \hat{y}_{uv} 增大,这在一定程度上验证了本文提出的设想。

1.3 实例分析

本节针对用户兴趣传播中一跳RippleSet属性分布进行实例分析。以Movielens1M数据集中用户 u_{20} 在电影知识图谱上对齐的偏好电影为例,用户 u_{20} 喜欢电影编号为2307、1693、2286的3部电影。用户 u_{20} 知识图谱一跳路径三元组见表2。表2中,这3部电影分别有15、11、11条一跳路径三元组。这3部电影的共有偏好属性权重的比例为language2458 : genre2624 : genre2495 : genre2535 : genre3731 = 3 : 3 : 2 : 2 : 2。其余三元组路径,因为没有共有属性,其二跳三元组路径与用户兴趣物品的相关性大幅下降。

表2 用户 u_{20} 知识图谱一跳路径三元组Table 2 User u_{20} knowledge graph one-hop path triples

head	relation	tail	head	relation	tail	head	relation	tail
2307	language	2458	2286	language	2458	1693	language	2458
2307	rating	2477	2286	genre	2495	1693	genre	2535
2307	genre	2495	2286	genre	2624	1693	genre	2624
2307	genre	2535	2286	star	3458	1693	genre	3731
2307	genre	2624	2286	rating	3721	1693	genre	3991
2307	country	2755	2286	genre	6365	1693	rating	5040
2307	genre	3731	2286	country	8627	1693	writer	22467
2307	genre	6365	2286	writer	12409	1693	director	44366
2307	star	6647	2286	director	12409	1693	writer	79811
2307	director	33776	2286	director	51141	1693	star	114980
2307	writer	34635	2286	writer	99426	1693	director	125298
2307	writer	43078						
2307	director	47425						
2307	writer	56000						
2307	director	111972						

2 实验验证

2.1 数据集

本文使用 MovieLens1M 作为实验数据集。该数据集中包含了 6 036 位用户记录与 3883 部电影,共有 1 000 209 条评分(评分在 1~5 之间),经过预处理过滤掉无效评分数据,还剩 753 772 条评分。753 772 条评分中,376 886 条评分大于 4、376 886 条评分小于 4,评分大于 4 的组成用户历史偏好记录。本文的知识图谱是在 Wang 等学者^[7]基于 Microsoft Satori 知识库构建的置信度大于 0.9 电影知识图谱基础上形成的。首先,通过 Microsoft Satori 知识图谱和电影数据集中的用户历史偏好记录中电影实体对齐,匹配出用户偏好电影的知识图谱子集;然后,在电影知识图谱上扩展用户的交互记录,解决了用户偏好数据稀疏的问题。

MovieLens1M 数据集的数据稀疏主要是指用户和电影的交互不充分,平均每部电影的评分数和平均每个用户评分过的电影数远小于用户数和电影数。计算 MovieLens1M 的数据集稀疏度为 96%,即无评分数据占据整个样本空间 96%,需要从 4% 的评分数据中学习用户偏好。数据集稀疏度的计算见式(11):

$$\text{数据集稀疏度} = \frac{\text{总评分数}}{\text{用户数} \times \text{电影数}} \quad (11)$$

2.2 评价指标

本文用 AUC (Area Under Curve) 及用户偏好属性多样性 $Diversity$ ^[26] 作为评估指标。

根据用户偏好物品属性-类别权重矩阵,每个用户对应一个属性分布向量,若各个偏好属性类别分布均匀,则信息熵最大,表明用户偏好属性多样性需求越大。用户偏好属性多样性定义如式(12)所示:

$$Diversity(u) = - \sum_{c=1}^n p_{uc} \log_n p_{uc} \quad (12)$$

2.3 实验结果分析

2.3.1 最大跳数的影响

为了验证 RippleNet-CA 的改进对 RippleNet 模型深层兴趣挖掘能力的提升,不同最大跳数 H 对模型性能的影响,本文进行了对比实验,实验结果见表 3。

表3 不同 H 对模型 AUC 的影响Table 3 Effect of different H on model AUC

模型	H			
	1	2	3	4
RippleNet	0.916	0.921	0.915	0.918
RippleNet-CA	0.918	0.919	0.932	0.930
KGCN	0.972	0.976	0.974	0.514

实验表明,随着 H 的增大,RippleNet-CA 模型性能提升明显。主要原因在于通过共现属性采样,

用户高阶偏好物品间的共有属性变得更集中,相关性更强。当 $H = 4$ 时模型没有像 KGCN 一样崩溃。

2.3.2 用户偏好属性多样性

通过控制采样用户偏好误差门限 T 的大小,可以调节共现属性采样的力度大小。误差门限 $T = [0.2, 0.4, 0.6, 0.8]$, 该参数对用户偏好属性多样性 (*Diversity*) 的影响见表 4。

表 4 不同 T 对模型 *Diversity* 的影响

Table 4 Effect of different T on model *Diversity*

模型	T			
	0.2	0.4	0.6	0.8
RippleNet-CA	0.480	0.559	0.662	0.635

实验表明,当用户偏好误差门限为 0.6 时,采样得到的结果属性多样性最高,极端情况下,模型相当于直接按共现频次排序采样。

3 结束语

本文提出基于知识图谱中共有属性采样的推荐模型 RippleNet-CA。首先通过共有属性采样算法重构知识图谱中每个实体的扩散偏好,然后使用偏好知识图谱的嵌入特征并结合用户特征进行项目推荐。实验结果表明在电影推荐场景中,RippleNet-CA 在推荐多样性和 *AUC* 性能上优于原模型。该模型不仅适用于电影推荐,也适用于其他构建了相应知识图谱的领域。通过本次研究实验发现,改进的 RippleNet-CA 模型相比于 RippleNet 模型,推荐结果电影属性更加多样,更符合用户历史偏好属性,对用户高阶潜在兴趣的挖掘能力更强。

参考文献

[1] 刘语晗. 基于知识图谱的可解释推荐算法研究[D]. 北京:北京邮电大学,2021.

[2] 朱海萍,赵成成,刘启东,等. 基于互惠性约束的可解释就业推荐方法[J]. 计算机研究与发展,2021,58(12):2660-2672.

[3] 张雪茹,官磊. 基于知识图谱的用户兴趣推荐算法[J]. 计算机应用,2021,41(S2):59-65.

[4] 熊中敏,舒贵文,郭怀宇. 融合用户兴趣的图神经网络推荐模型[J]. 计算机科学,2022,49(6):165-171.

[5] 邓明通,刘学军,李斌. 基于用户兴趣和动态兴趣的多样性推荐方法[J]. 小型微型计算机系统,2018,39(9):2029-2034.

[6] 段文善. 融合知识图谱的多层兴趣建模及其在推荐系统中的应用[D]. 太原:太原理工大学,2021.

[7] WANG Hongwei, ZHANG Fuzheng, WANG Jialin, et al. RippleNet: Propagating user preferences on the knowledge graph for recommender systems [C]//Processing of the 27th ACM International Conference on Information and Knowledge

Management. New York; ACM, 2018: 417-426.

[8] 钱玲飞,崔晓蕾. 基于数据增强的领域知识图谱构建方法研究[J]. 现代情报,2022,42(3):31-39.

[9] 张彬,徐建民,吴姣. 大数据环境下基于知识图谱的用户兴趣扩展模型研究[J]. 现代情报,2021,41(8):36-44.

[10] 刘伟江,郝一哲. 融合知识图谱的 RippleNet 推荐模型优化研究[J]. 现代情报,2021,41(9):64-73,83.

[11] 许荣海,王昌栋. 基于异构信息网络元路径张量分解的深度推荐系统[J]. 信息安全学报,2021,6(5):77-87.

[12] WANG Hongwei, ZHAO Miao, XIE Xing, et al. Knowledge graph convolutional networks for recommender systems [C]//Proceedings of the World Wide Web Conference. Lyon, France; dblp, 2019: 3307-3313.

[13] WANG Xiang, HE Xiangnan, CAO Yixin, et al. KGAT: Knowledge graph attention network for recommendation [C]//Proceedings of the 25th ACM SIGKDD International Conference. New York; ACM Press, 2019: 950-958.

[14] WANG Xiang, HUANG Tinglin, WANG Dingxian, et al. Learning intents behind interactions with knowledge graph for recommendation [C]//Proceedings of the Web Conference. Slovenia; dblp, 2021: 878-888.

[15] 梁顺攀,涂浩,王荣生,等. 融合重要性采样和池化聚合的知识图谱推荐算法[J]. 小型微型计算机系统,2021,42(5):967-971.

[16] 李世宝,张益维,刘建航,等. 基于知识图谱共同邻居排序采样的推荐模型[J]. 电子与信息学报,2021,43(12):3522-3529.

[17] XIE Lijie, HU Zhaoming, CAI Xingjuan, et al. Explainable recommendation based on knowledge graph and multiobjective optimization[J]. Complex & Intelligent Systems, 2021, 7(3): 1241-1252.

[18] 刘海鸥,李凯,何旭涛,等. 面向信息茧房的用户画像多样化标签推荐[J]. 图书馆,2022(3):83-89.

[19] 张波,赵鹏,张金金,等. 基于用户潜在兴趣的知识感知传播推荐算法[J]. 计算机应用研究,2022,39(9):2615-2620.

[20] 顾亦然,张远之,杨海根. 基于电影属性和交互信息的电影推荐算法[J]. 南京理工大学学报,2022,46(2):177-184.

[21] 曹渝昆,方一新,苗泽宇,等. 基于衰减传播的知识图谱增强推荐系统[J]. 计算机工程与应用,2023,59(10):180-186.

[22] HAMILTON W L, YING R, LESVOKEC J. Inductive representation learning on large graphs [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach; Curran Associates Inc., 2017:1025-1035.

[23] CHEN Jie, MA Tengfei, XIAO Cao. FastGCN: Fast learning with graph convolutional networks via importance sampling [C]//Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver, Canada; Vancouver Convention Center, 2018:1-15.

[24] HUANG Wenbing, ZHANG Tong, RONG Yu, et al. Adaptive sampling towards fast graph representation learning [C]//Advances in Neural Information Processing System. Montreal, Canada; MIT Press, 2018:4558-4567.

[25] 金志刚,朱琦,刘晓辉. 基于属性偏好和邻居信任度的协同过滤推荐算法[J]. 南开大学学报(自然科学版),2021,54(3):25-30.

[26] 刘莉. 基于用户多样性偏好的 top-N 推荐算法[J]. 计算机工程与应用,2021,57(17):116-121.