计算机科学
COMPUTER SCIENCE

融合知识图谱的负采样方法

陆海洋，柳先辉，侯文龙

引用本文

陆海洋，柳先辉，侯文龙. 融合知识图谱的负采样方法[J]. 计算机科学，2025，52(3)：161-168.

LU Haiyang, LIU Xianhui, HOU Wenlong. Negative Sampling Method for Fusing Knowledge Graph[J]. Computer Science, 2025, 52(3): 161-168.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240500015

融合知识图谱的负采样方法

陆海洋[1]　柳先辉[2]　侯文龙[1]
1 同济大学电子与信息工程学院　上海 201804
2 同济大学电子与信息工程学院 CAD 研究中心　上海 201804
(Ocean@tongji.edu.cn)

摘　要　为了解决信息过载的问题,推荐系统被广泛研究。由于很难获取大量高质量的显式反馈数据,隐式反馈数据成为训练推荐系统的主流选择。从未标记的数据中采样负例,即负采样,对于训练基于隐式反馈的推荐模型非常重要。现有推荐系统的负采样方法往往只关注如何选择包含更多用户偏好信息的强负样例,却没有考虑强负样例的假阴性问题。为了降低采样得到的负样例的假阴性概率并提高其信息量,提出了一种融合知识图谱的负采样方法。首先,根据用户-项目知识图谱构建负样例候选集;然后,通过基于贝叶斯分类的方式从候选集中筛选假阴性概率最小的负样例;最后,基于 Mixup 策略引入正混合技术构建强负样例。为了验证所提出方法的有效性,在两个公开数据集上进行了实验。结果表明,与现有方法相比,所提方法表现更优。

关键词:负采样;知识图谱;推荐系统;正混合
中图分类号　TP391

Negative Sampling Method for Fusing Knowledge Graph

LU Haiyang[1],LIU Xianhui[2] and HOU Wenlong[1]
1 College of Electronic and Information Engineering,Tongji University,Shanghai 201804,China
2 CAD Research Center,College of Electronic and Information Engineering,Tongji University,Shanghai 201804,China

Abstract　In order to solve the problem of information overload,recommender systems have been widely studied. Since it is difficult to obtain a large amount of high-quality explicit feedback data,implicit feedback data becomes the mainstream choice for training re-commender systems. Sampling negative instances from unlabeled data,i. e. negative sampling,is crucial for training recommendation models based on implicit feedback data. The previous negative sampling methods often focus on how to select hard negative instances that contain more user preference information,without considering the false negative problem. In order to reduce the false negative probability of negative instances obtained from sampling and make them more informative,a negative sampling method that integrates knowledge graph is proposed. Firstly,constructing a candidate instance set based on the user-item knowledge graph. Then,the negative instance with the lowest false negative probability is selected from the candidate set through a Bayesian classification approach. Finally,based on the Mixup strategy,positive mixing technology is introduced to construct the hard negative instance. To evaluate the effectiveness of the proposed method,validation was conducted on two public datasets. The results show that compared with previous methods,the method proposed in this paper performs better.

Keywords　Negative sampling,Knowledge graph,Recommender system,Positive mixing

1　引言

近年来,关于推荐系统的研究不断发展。推荐系统旨在对用户的兴趣进行建模,从而为用户推荐可能感兴趣的项目,包括电影、音乐和书籍等。推荐系统又可以分为 3 类,分别为基于协同过滤的推荐系统、基于内容的推荐系统以及混合推荐系统。基于协同过滤的推荐根据交互数据中用户或项目的相似性对用户偏好进行建模;而基于内容的推荐则利用项目的内容特征进行推荐;混合推荐则是对前两种方法的综合运用。

基于协同过滤的算法由于能够有效地捕捉用户的偏好以及可以方便快捷地应用于多种场景,因此得到了广泛的应用[1-3]。基于知识图谱的推荐[4-7]则引入了项目知识图谱作为对交互数据的补充,使用图神经网络来探索用户的偏好,缓解了基于协同过滤的算法中的冷启动和数据稀疏的问题,从而也得到了广泛的研究和应用。上述推荐方法往往需要从观察

到稿日期:2024-05-06　返修日期:2024-08-20
基金项目:国家重点研发计划(2022YFB3305700)
This work was supported by the National Key Research and Development Program of China(2022YFB3305700).
通信作者:柳先辉(xianhuiliu488@163.com)

到的用户-项目交互中对用户兴趣进行建模，从而进行推荐。但在许多情况下，获得大量高质量的显式反馈以用于推荐是一件困难的事，因此用户的隐式反馈（例如点击或购买行为）已成为训练这些推荐模型的默认选择[8]。

在基于隐式反馈的推荐模型的训练中，每个观察到的用户-项目交互通常被视为用户对这个项目感兴趣，这些项目构成了正样本。由于缺乏显式反馈数据，即无法明确用户不喜欢哪些项目，研究人员通常会从用户未交互的项目中随机选取一些项目作为用户不感兴趣的项目，即负样本。然后，基于正样本和负样本对隐式推荐模型进行优化，使正样本的得分高于负样本[9]。与许多半监督学习的问题类似，这些基于隐式反馈的推荐模型高度依赖于挖掘出好的负样本来提供高质量的用户偏好信息。因此研究人员对如何进行高质量的负采样进行了探究，这些方法主要可以分为两类，即静态负采样和强负采样。静态负采样中，每个未交互项目被采样为负例的概率不随训练发生变化。强负采样则主要利用推荐模型对不同未交互项目生成的分数来调整对不同未交互项目选择的概率。

现有的负采样方法虽然已经取得了初步的成功，但仍然面临着一些挑战。

1）候选负样例整体信息量不高。强负采样方法往往会为每个用户构建一个负样例候选集，然后对候选负样例计算得分以挑选强负样例，从而降低时间复杂度。但是，这些方法往往都是通过随机从用户所有未交互的项目中挑选项目来构建负样例候选集，没有较好地利用用户-项目交互矩阵中的信息，导致挑选出来的候选负样例整体信息量不高。

2）采样得到的负样例无法做到既“强”又“真”。现有的大部分强负采样方法都通过不同的方式从候选集中挑选出了得分较高的强负样例，但没有考虑假阴性问题，因为假负样例也会有较高的得分，导致使用这些负采样方法的模型的稳健性较差，而且降低了模型的表现。近期有研究虽然考虑了负样例的假阴性问题[10]，但是为了降低采样得到的负样例的假阴性概率而牺牲了一定的负样例的信息量，即“强度”，降低了模型性能的上限。

为了解决上述问题，本文提出了一种融合知识图谱的负采样方法。针对候选负样例整体信息量不高的问题，本文基于用户-项目交互矩阵构建用户-项目知识图谱，从知识图谱的特定区域内挑选负样例构建负样例候选集，从而提升候选负样例的整体信息量。针对强负采样方法中无法兼顾降低采样得到的负样例的假阴性概率又提高其信息量的问题，使用基于贝叶斯分类思想的方法，计算负样例候选集中不同负样例为假负样例的先验概率和后验概率，挑选出假阴性概率最小的负样例。最后，基于 Mixup[11] 的思想，将正样例的信息混合进挑选出来的负样例中来提升负样例的质量，使最终得到的负样例既是低假阴性概率的，又有着较多的信息量，从而提升负采样的效果。

综上所述，本文的主要贡献如下：

1）提出了一种根据知识图谱构建负样例候选集的方法，能够提高候选负样例的整体的信息量；

2）提出了一种简单有效的融合知识图谱的负采样方法，其能够较好地兼顾解决负样例的假阴性问题和负样例的低质量问题，提升推荐模型的训练效果；

3）在两个公开的数据集上进行了实验，结果表明，相比于以往的工作，本文提出的负采样方法能够更好地提升模型的推荐效果。

2 相关工作

2.1 推荐系统

随着互联网的不断发展，网络资源也呈现出指数型增长的趋势，而推荐系统则可以有效地缓解信息过载的问题，为用户推荐合适的信息。推荐系统从本质上来说就是对人的偏好进行模拟，通过推荐算法对需要进行推荐的数据进行处理和分析，然后将相应的推荐结果返回给用户。推荐算法是推荐系统的核心。根据推荐算法的不同，推荐系统又可以分为 3 类，分别为基于协同过滤的推荐系统、基于内容的推荐系统以及混合推荐系统。基于协同过滤方法的推荐根据交互数据中用户或项目的相似性对用户偏好进行建模；而基于内容的推荐则利用项目的内容特征进行推荐；混合推荐是对前两种方法的综合运用。

基于协同过滤的推荐系统因有着强大的性能以及广泛的应用场景，得到了广泛的研究[12]。这种方法主要通过分析用户-项目交互矩阵来建模用户的偏好，从而进行推荐。文献[13]通过分析交互矩阵来确定不同用户间以及不同用户与其交互过的项目之间的差异，从而根据差异有针对性地为用户推荐合适的项目。而文献[14]则将神经网络引入协同过滤算法，通过计算用户和项目之间的非线性关系进行更精准的推荐。基于知识图谱的推荐系统则是对基于协同过滤的推荐系统的一种改进，主要缓解了其数据稀疏和冷启动的问题[15]。文献[16]通过衡量知识图谱上不同用户节点之间的加权元路径的相似程度来衡量不同用户的相似性，以进行推荐。文献[17]则通过图神经网络（Graph Neural Networks，GNN）在知识图谱上迭代地传播用户的偏好信息来更新用户和项目的嵌入向量，通过计算用户向量和项目向量的内积来表示项目的得分，并将得分高的项目推荐给用户。

2.2 用于推荐系统的负采样

现有的推荐系统基本都根据隐式反馈数据进行训练，通过给正样例赋予比负样例更高的得分来模拟用户的偏好。在没有显式负样本的情况下，有效挖掘负样例，即负采样，对这些系统的性能有着重大的影响[18]。推荐系统中的负采样主要是从用户未交互过的项目中挑选一些项目作为负样例。负采样方法主要可以分为两类，分别为静态负采样和强负采样。

2.2.1 静态负采样

如果从用户未交互的项目中选择项目作为负样例，那么，通过给不同的项目设置不同的权重，便能根据负例分布进行采样。当每个项目被采样为负例的概率不随训练发生变化时，这种采样方法就被称为静态负采样。

在静态负采样中，最常用的是随机负采样。随机负采样也被称为均匀负采样，即随机地从用户未交互过的项目中

选择一个作为负例。文献[19-20]提出的方法都采用了随机负采样。在不考虑负采样的研究中，研究者们一般使用随机负采样作为采样方法，以便公平地和基线(baseline)进行比较。另一种启发式的静态负采样则为基于流行度对未交互过的项目进行带权采样。流行度可以通过交互频次来反映，$P_n(v)\propto deg(v)^{\alpha}$，即项目 v 被选为负例的概率和 v 的流行度的 α 次方有关系。文献[21-23]都采用了基于流行度的负采样方法，将项目在训练集中的流行程度作为候选负例的权重。其中，文献[21]基于项目流行度对缺失数据进行加权，提高了传统矩阵分解机的性能；文献[22]则通过提高对高流行度的项目的采样概率来缓解推荐系统中的冷启动问题。

静态负采样方法不随训练发生变化，无法动态地调整候选负样例的分布来适应模型，也就难以挖掘更有利的负样本。

2.2.2 强负采样

强负采样与静态负采样不同，其通过为得分较高的项目分配更高的采样概率来调整候选负样例的分布，以找到强负例(hard negative)来提升模型的训练效果。

文献[24-27]都采用了强负例采样来提升推荐的效果。SRNS[28]首先观察到真负样例和假负样例都被分配了高分，而假负样例的分数方差相对较低。因此，SRNS 偏向选择方差较高、得分较高的负样例。MixGCF[29]通过图卷积模型来利用项目的邻居信息合成负样例，提高负样例的质量。DENS[8]通过解纠缠方法，解开项目的相关和不相关因素，并通过因素感知采样策略识别最佳负样本。除此之外，生成对抗网络(Generative Adversarial Networks，GAN)的思想也被运用到了强负采样中。与 GAN 类似，对抗式负采样给定两个推荐模型，一个作为判别器，一个作为生成器，基于对抗的思想进行训练。例如，IRGAN[30]利用生成对抗网络来玩极小极大游戏，其中生成器充当推荐，采样器用来愚弄鉴别器。AdvIR[31]结合了对抗性负采样和对抗性训练来生成信息丰富的负样例。

前面提到的强负采样算法虽然可以提升采样质量，获得强负例以提升模型推荐效果，但也存在着关键问题：借助模型评分来选择负例会加重采样时的假阴性问题。因为单从评分上看，伪负例和强负例都会得到较高分数，按照模型评分得到的分布进行采样会提升伪负例被采样到的概率。

3 相关定义

3.1 问题定义

给定项目集合 $V=\{v_1,v_2,v_3,\cdots,v_m\}$ 以及用户集合 $U=\{u_1,u_2,u_3,\cdots,u_n\}$，其中 m 为项目的数量，n 为用户的数量。用 $O=\{(u,v^+)\mid u\in U,v^+\in V\}$ 代表观察到的交互集合，即隐式反馈，其中每对 (u,v^+) 代表用户 u 与项目 v^+ 之间发生过交互。基于隐式反馈的推荐模型旨在从观察到的交互中建模用户的兴趣，交互过的用户-项目对通常会作为模型训练的正对，而用户未交互过的项目就是候选负样例。而负采样，就是根据每对正对 (u,v^+)，从此前用户 u 未交互过的项目中挑选一个合适的项目作为负样例 v^-；然后，通过 BPR 损失函数对推荐模型进行优化，为正对 (u,v^+) 赋予比负对 (u,v^-) 更高的分数。BPR 损失函数如下所示：

$$L_{\mathrm{BPR}}=\sum_{(u,v^+,v^-)}-\ln\sigma(\boldsymbol{e}_u\cdot\boldsymbol{e}_{v^+}-\boldsymbol{e}_u\cdot\boldsymbol{e}_{v^-}) \tag{1}$$

其中，$\boldsymbol{e}_u$ 代表用户 u 的嵌入向量，$\boldsymbol{e}_{v^+}$ 代表正样例 v^+ 的嵌入向量，$\boldsymbol{e}_{v^-}$ 代表负样例 v^- 的嵌入向量，向量间的内积被用来衡量正负对的得分；$\sigma(\cdot)$ 代表 sigmoid 激活函数。

3.2 用户-项目知识图谱的定义

给定项目集合 $V=\{v_1,v_2,v_3,\cdots,v_m\}$ 以及用户集合 $U=\{u_1,u_2,u_3,\cdots,u_n\}$，根据用户的隐式反馈构建用户-物品交互矩阵 $\boldsymbol{Y}=\boldsymbol{R}^{n\times m}$，矩阵中的每个 y_{uv} 由式(2)定义。

$$y_{uv}=\begin{cases}1, & \text{观察到 } u \text{ 与 } v \text{ 产生交互}\\ 0, & \text{未观察到 } u \text{ 与 } v \text{ 产生交互}\end{cases} \tag{2}$$

结合交互矩阵 $\boldsymbol{Y}$，构建用户-项目知识图谱。此处实际上是将交互矩阵转化为一个无向图 $G(E,R)$。其中，E 代表图中的节点，主要有两类，分别为用户节点以及项目节点；R 表示图中的关系，主要指 $\boldsymbol{Y}$ 中发生的用户-项目交互行为。图 1 给出了用户-项目知识图谱 G 的构建过程。

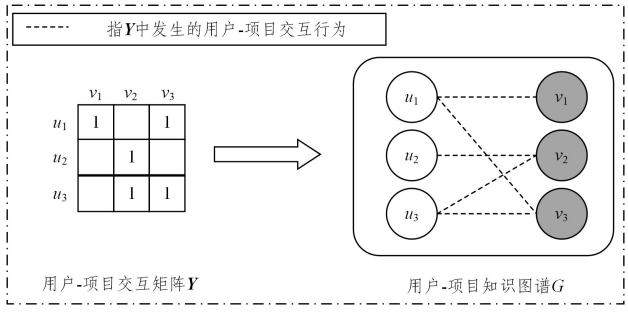

图 1 用户-项目知识图谱的构建过程

Fig. 1 User-item knowledge graph construction process

4 融合知识图谱的负采样方法

4.1 基于用户-项目知识图谱构建候选集

以往的基于知识图谱的推荐系统中，一般都采用迭代图神经网络的思想，在用户-项目知识图谱中逐层传播信息，来更新用户和项目的嵌入向量，从而探索用户的偏好。但是，在传播的时候，传播层的数量对推荐系统的性能有着巨大的影响，当传播层的数量增大到一定数值之后，继续增加传播层，会导致推荐系统的性能大幅下降，文献[4-5,32]都指出了这一问题。

从负采样的角度来看，这是因为在较少传播层数(又称传播跳数)的情况下，传播信息可以较好地探索用户的偏好。这些离用户比较近的项目包含了较多用户偏好信息，因此可以提高推荐系统的性能。但是当跳数增大到一定程度之后，这些离用户较远的项目包含的用户的偏好信息就较少。如果使用这些项目来更新嵌入向量，会引入一定的噪声，从而导致模型性能的下降。这也可以用社会影响理论[33]来解释，即社交网络中的用户会相互影响，导致产生相似的偏好。类似地，与同一项目交互过的用户意味着其很大可能属于同一社区，并有着相似的偏好。以用户-项目知识图谱中的用户节点 u_0 为例，其三跳邻居 $v_1(u_0\rightarrow v_0\rightarrow u_1\rightarrow v_1)$ 包含较多的 u_0 的偏好信息。这是因为 u_0 和 u_1 都与 v_0 有过交互，他们很可能有着

相似的偏好。因此，u_1 交互过的物品 v_1 也很有可能与 u_0 产生交互，包含较多的 u_0 的偏好信息。

基于上述发现，可以根据用户-项目知识图谱来构建每个用户的负样例候选集，为后续挑选出高质量的负样例做准备。与直接通过交互矩阵构建负样例候选集相比，通过转化为知识图谱的方式不仅更加清晰直观，而且能够提升构建负样例候选集的效率。知识图谱的图形化表示使得候选负样例的挑选过程更加明晰，方便理解挑选出来的项目与用户的关系，并提供一定的可解释性，更加能够通过高效的节点游走算法来加快负样例候选集的构建过程。

具体来说，以用户节点 u 为中心，选择该用户节点的 2-hop 到 K-hop 内的邻居集合中的项目作为负样例候选集 M_u 的来源，即通过均匀采样从用户节点的 2-hop 到 K-hop 内的邻居集合中采样 n 个项目构建 M_u，n 为 M_u 的大小。通常来说，K 一般取 3 或 5，因为用户节点的一跳邻居为交互过的项目，显然为正样例，不能采为负样例。二跳邻居和四跳邻居则为其他的用户节点，也不能作为负样例。与将所有用户未交互过的项目作为负样例候选集的来源相比，这种方式挑选到的候选负样例信息量更高，包含了更多的用户偏好信息，也能进一步提高后续步骤得到的负样例的质量。图 2 给出了为用户 u 构建 M_u 的过程（K 取 3）。

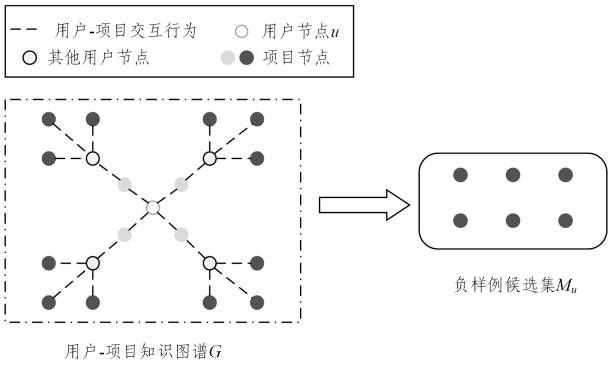

图 2 负样例候选集的构建过程

Fig. 2 Candidate instance set construction process

4.2 基于贝叶斯分类的负样例挑选

以往的强负采样方法都会遇到假阴性问题，因为伪负样例和强负样例都会被推荐器给予高分。如果错误地将伪负样例用作训练，那么模型将会学习到错误的用户偏好信息，导致性能下降。因此，需要基于构建好的负样例候选集 M_u，挑选出一个采样风险最小的负样例，即假阴性概率最小的负样例，而不是直接选择信息量最大即得分最高的负样例。

文献[10]指出负例的预测分数越高，它是假负例的概率密度就越高，它是真负例的概率密度就越低；而且随着训练的继续，这种现象更加明显。这表明，推荐模型对假负样例的评分高于真负样例。

因此可以采用文献[10]提出的方法，给定一个正对 (u, k)，从负样例候选集 M_u 中挑选出一个采样风险最小的负样例 j。具体来说，对于一个候选负样例 i，其为真负样例的后验概率可以采用式(3)所示的贝叶斯公式计算。

$$\begin{aligned} P(tn|x_i) &\propto P(x_i|tn)P_{tn}(i) \\ P(x_i|tn)P_{fn}(i) &= 2[1-F(x_i)]f(x_i)P_{tn}(i) \end{aligned} \tag{3}$$

其中，x_i 指项目 i 的得分，可以通过向量内积得到；$P(x_i|tn)$ 指 i 为真负样例的类条件密度，$P_{tn}(i)$ 是 i 为真负样例的先验概率；$f(x_i)$ 为所有未交互过项目的得分分布函数，$F(x_i)=\int_{-\infty}^{x_i} f(t)\mathrm{d}t$ 则为相应的累积分布函数。

由于直接计算分布函数 $f(\cdot)$ 较为困难，因此用条件概率的分数形式来消除密度函数。一个候选负样例 i 为真负样例的概率可以用式(4)计算。

$$\begin{aligned} unbias(i) &\triangleq P(tn|x_i) \\ &= \frac{P(x_i|tn)P_{tn}(i)}{P(x_i|tn)P_{tn}(i)+P(x_i|fn)P_{fn}(i)} \\ &= \frac{f(x_i)[1-F(x_i)]P_{tn}(i)}{f(x_i)[1-F(x_i)]P_{tn}(i)+F(x_i)f(x_i)P_{fn}(i)} \\ &= \frac{[1-F(x_i)][1-P_{fn}(i)]}{1-F(x_i)-P_{fn}(i)+2F(x_i)P_{fn}(i)} \end{aligned} \tag{4}$$

其中，$F(x_i)$ 可以用式(5)计算，即所有未交互过项目集合 L_u 中得分小于 i 的项目的比例。

$$F(x_i)=\frac{\#\{x_l<x_i \mid l\in L_u\}}{\#\{L_u\}} \tag{5}$$

其中，$P_{fn}(i)$ 为候选负样例 i 为假负样例的先验概率，可以通过流行度来计算。因为一个项目被交互过的次数越多，代表其越流行，那么它大概率会被推送给某个用户，用户可能与其发生交互。因此一个项目的流行度越高，其为假负样例的先验概率就越高，具体可以用式(6)计算。

$$P_{fn}(i)=\frac{pop_i}{N} \tag{6}$$

其中，N 代表用户-项目交互总次数；pop_i 代表 i 的流行度，即被交互次数。

基于上述内容，给定正样例 k，计算采样候选负样例 i 为负样例的风险，如式(7)所示。

$$R(i|k)=P(fn|i)\cdot info(i)-P(tn|i)\cdot\lambda\cdot info(i) \tag{7}$$

其中，$info(i)=1-\sigma(x_k-x_i)$ 代表采样 i 为负样例对更新模型参数所做的贡献，即损失梯度大小；λ 代表效果尺度。

最后，可以通过式(8)，从 M_u 中挑选一个采样风险最小的负样例 j。

$$\begin{aligned} j &= \arg\min_{i\in M_u} R(i|k) \\ &= \arg\min_{i\in M_u}[1-unbias(i)]\cdot info(i)-\lambda\cdot unbias(i)\cdot info(i) \\ &= \arg\min_{i\in M_u} info(i)\cdot[1-(1+\lambda)\cdot unbias(i)] \end{aligned} \tag{8}$$

4.3 基于 Mixup 思想的负样例增强

基于上述负样例挑选方法，可以从负样例候选集 M_u 中挑选出采样风险最小的负样例 j，但这个负样例 j 却不一定是信息量最大的一个，也就是说，上述负样例挑选方法为了降低采样风险而牺牲了一定的负样例的信息量。

在 CV 领域，Mixup[11] 的思想已经被普遍用来进行数据增强，通过简单的线性插值方法，可以实现较好的效果。文献[34]基于 Mixup 的思想，通过往负样例中注入正样例的信息来增强负样例，从而提升模型训练的效果。该项工作将正混合

技术用于对比学习领域，并且取得了非常好的效果。

受这些工作的启发，将正混合技术用于推荐系统领域，通过往挑选得到的负样例中注入正样例的信息来增强负样例，从而弥补基于贝叶斯分类的负样例挑选中对负样例信息量的牺牲。该项技术只针对交互矩阵或其转化成的知识图谱中的部分项目，并不是整体用于矩阵或知识图谱。

具体来说，给定三元组(u,k,j)，其中u代表用户，k代表正样例，j代表经过负样例挑选得到的采样风险最小的负样例，使用式(9)对j进行增强。

$$\boldsymbol{e}_{j'}=\partial\boldsymbol{e}_k+(1-\partial)\boldsymbol{e}_j \tag{9}$$

其中，$\boldsymbol{e}_j$和$\boldsymbol{e}_k$分别代表j和k的嵌入向量，∂代表预定义的混合系数，$\boldsymbol{e}_{j'}$代表经过数据增强后的负样例的嵌入向量。

最终得到三元组(u,k,j')，其中j'代表增强后的负样例，(u,k)和(u,j')则为模型训练所用到的正对和负对。上述方法只针对挑选得到的负样例进行增强，而不是整体运用于交互矩阵或知识图谱。

4.4 融合知识图谱的负采样方法等价算法

根据上述负采样过程，可以得到融合知识图谱的负采样方法的等价算法，如算法1所示。

算法1 融合知识图谱的负采样方法

输入：训练集$R=\{(u,k)\}$；M_u的来源集合N_u；推荐模型；M_u的大小n；超参数∂，K以及λ

输出：训练好的推荐模型

```
for epoch=1,2,…,T do:
    采样一个 mini-batch R_batch∈R
    for each(u,k)∈R_batch do
        通过均匀采样构建 M_u⊆N_u
        通过基于贝叶斯分类的负样例挑选方法挑选出 j∈M_u
        通过正混合方法得到 j'
        通过(u,k,j')计算 BPR 损失函数
    end
通过 BPR 损失函数更新模型
end
```

其中，N_u指以用户节点u为中心的2-hop到K-hop内的邻居集合(去除用户节点)。首先，基于用户-项目交互矩阵构建用户-项目知识图谱；然后，根据K的大小为每个用户构建N_u。接下来就可以用算法1进行负采样，从而训练推荐模型。

时间复杂度分析：首先是构建M_u，时间复杂度为$O(n)$。对每个M_u中的项目，计算式(5)的时间复杂度为$O(|L_u|)$，计算式(6)的时间复杂度为O(1)，计算式(8)的时间复杂度为O(1)，最后是正混合的时间复杂度为O(1)。总体来说，所提出的负采样方法是一种线性时间复杂度的方法，能够较为高效地采集假阴性概率较低、信息量较高的负样例，提升模型的推荐效果。

5 实验

5.1 数据集

为了验证所提出方法的有效性，在两个真实的数据集上进行了实验。这两个数据集分别为电影数据集MovieLens-100k以及MovieLens-1M，其中用户-项目交互的具体数据如表1所列。

表1 数据集的相关信息

Table 1 Statistics of datasets

Dataset	User	Item	Interaction	Density/%
MovieLens-100k	943	1682	100000	6.30
MovieLens-1M	6040	3952	1000209	4.47

5.2 实验指标以及对比方法

为了对实验结果进行较好的评估，采用以下评价指标。

1)精确率(Precision，P)：表示在模型给出的TOP-K物品列表中，用户真正感兴趣的物品占物品列表中物品的比例。

2)召回率(Recall，R)：表示用户真正感兴趣的物品出现在模型给出的TOP-K物品列表中的比例。

3)归一化折损累计增益(Normalized Discounted Cumulative Gain，NDCG)：一种衡量排序质量的指标。该项指标为模型给出的TOP-K物品列表中的每一项物品赋予一个分数，这个分数由物品的排序位置以及该物品是否是用户真正感兴趣的所决定。排序越靠前，是用户感兴趣的，则该物品的分数越高。这个指标能够较为综合地反映模型推荐效果的好坏。

以上3个指标，都是值越高，模型的推荐效果越好。

同时，为了检验所提方法是否有效，设置了以下对比负采样方法与其进行对比。

1)RNS[5]：进行随机负采样，认为用户未交互过的每个项目被采样到的概率是相等的。

2)PNS[21]：基于流行度的负采样，根据项目流行度对项目进行采样。流行度可以通过交互频次来反映，$P_n(v)\propto deg(v)^{\alpha}$，即物品$v$被选为负例的概率与$v$的流行度的$\alpha$次方有关系。

3)AOBPR[23]：对全局排名较高的负样例进行过采样。采样概率与$\exp(-rank(j|u)/\lambda)$呈正相关，其中$rank(j|u)$代表项目j在用户u所有未交互过的项目中的得分排名。

4)DNS[24]：每次挑选得分最高的未交互项目作为负样例，每个未交互过的项目的被采样概率与其得分成正相关。

5)SRNS[28]：首先观察到真负样例和假负样例都被分配了高分，而假负样例的分数方差相对较低，因此偏向选择方差较高、得分较高的项目作为负样例。

6)BNS[10]：通过分析假负样例和真负样例的得分分布情况，设计了一种贝叶斯分类方法，每次挑选采样风险最小的候选负样例作为负样例。

5.3 实验细节

采用经典的基于隐式反馈的协同过滤推荐模型——矩阵分解机MF[35]作为推荐模型。两个数据集的嵌入维度都设置为$d=32$，学习率$\alpha=0.01$，正则化参数$reg=0.01$，batch size $b=1$，训练epoch数$T=100$。对于其他负采样方法，都设置同样的模型参数，以便公平地比较性能。

对于MovieLens-100k数据集，所提方法的超参数分别为$\lambda=5$，$n=5$，$\partial=0.8$，$k=3$。对于MovieLens-1M数据集，所提方法的超参数分别为$\lambda=5$，$n=5$，$\partial=0.6$，$k=3$。

5.4 实验结果及分析

不同负采样方法的推荐结果如表2所列，最优结果以粗体表示。

表 2 不同负采样方法的推荐结果

Table 2 Recommended results for different negative sampling methods

Dataset	Method	TOP-5			TOP-10			TOP-20		
		Precision	*Recall*	*NDCG*	*Precision*	*Recall*	*NDCG*	*Precision*	*Recall*	*NDCG*
MovieLens-100k	RNS	0.3879	0.1297	0.4132	0.3358	0.2152	0.3961	0.2711	0.3287	0.3956
	PNS	0.2651	0.0871	0.2703	0.2341	0.1479	0.2643	0.1958	0.2377	0.2719
	AOBRR	0.3954	0.1365	0.4172	0.3296	0.2143	0.3931	0.2691	0.3352	0.3971
	DNS	0.4042	0.1402	0.4297	0.3337	0.2197	0.4029	0.2719	0.3401	0.4051
	SRNS	0.3947	0.1335	0.4168	0.3379	0.2161	0.3987	0.2723	0.3365	0.4007
	BNS	0.4195	0.1458	0.4547	0.3459	0.2283	0.4209	0.2757	0.3459	0.4168
	Proposed	**0.4316**	**0.1481**	**0.4605**	**0.3586**	**0.2309**	**0.4308**	**0.2853**	**0.3503**	**0.4261**
MovieLens-1M	RNS	0.3835	0.0804	0.4139	0.3507	0.1325	0.3634	0.2762	0.2149	0.3469
	PNS	0.3466	0.0689	0.3744	0.3105	0.1150	0.3261	0.2479	0.1891	0.3095
	AOBPR	0.3934	0.0891	0.4239	0.3523	0.1458	0.3732	0.2822	0.2340	0.3621
	DNS	0.4057	0.0924	0.4369	0.3609	0.1523	0.3868	0.2918	0.2451	0.3726
	SRNS	0.3947	0.0876	0.4331	0.3511	0.1513	0.3936	0.2743	0.2336	0.3874
	BNS	0.4192	0.0985	0.4427	0.3613	0.1595	0.4076	0.3001	0.2507	0.3896
	Proposed	**0.4262**	**0.1013**	**0.4484**	**0.3657**	**0.1641**	**0.4128**	**0.3038**	**0.2575**	**0.3953**

从表 2 中的结果可以看出，本文所提出的负采样方法在所有的评价指标上都取得了最好的结果。具体地，静态负采样 PNS 和 RNS 这两者的性能是最差的，这说明静态负采样不能够根据模型训练的情况来动态调整候选负样例的采样概率，导致不能挖掘出信息量更多的负样例，降低了模型的性能。与 DNS 方法相比，AOBPR 方法优先选择全局排名最高的项目作为负样例，而不是根据用户及其正样本的情况动态地选择负样例，这导致其性能不如 DNS。原因在于，所谓的强负样例并不是全局排名越高越好，而是需要根据用户及其正样例来决定。因为只有确定了用户及其正样例，才能真正确定这个用户的偏好情况，从而选择强负样例，也就是包含更多用户偏好信息的负样例。SRNS 方法和 BNS 方法都为了解决假阴性问题而对强负采样方法做出了改进。SRNS 方法根据实验所得的发现优先选择方差较高、得分较高的项目作为负样例，但是其用于选择高质量负例的线性平均操作限制了其有效性，导致性能不如 BNS 方法。作为对 BNS 方法的改进，本文提出的方法通过构建高质量的负样例候选集，以及最后的正混合技术，进一步提高了挑选出来的负样例的质量，较好地弥补了 BNS 方法中为了降低采样风险而牺牲了一定负样例的质量的问题，取得了最好的结果。

5.5 消融实验结果及分析

本文提出的方法相对于 BNS 方法来说，主要引入了两个组件来在降低采样风险的同时，进一步提高负样例的质量。为了研究这两个组件的影响，在 MovieLens-100k 数据集上进行了消融实验。首先通过随机从用户所有未交互过的项目中挑选项目构建 M_u，而不是从用户的 2-hop 到 K-hop 内的邻居集合中挑选，其余方法中的内容不变，主要对应表 3 中的 without1 结果。其次在从 M_u 中挑选出 j 之后，不采用正混合技术，直接将 j 用于模型的训练，其余方法中的内容不变，主要对应表 3 中的 without2 结果。总体的消融实验结果如表 3 所列。

表 3 消融实验结果

Table 3 Results of ablation experiment

Method	TOP-5			TOP-10			TOP-20		
	Precision	*Recall*	*NDCG*	*Precision*	*Recall*	*NDCG*	*Precision*	*Recall*	*NDCG*
BNS	0.4195	0.1458	0.4547	0.3459	0.2283	0.4209	0.2757	0.3459	0.4168
without1	0.4209	0.1464	0.4563	0.3470	0.2293	0.4229	0.2771	0.3472	0.4181
without2	0.4301	0.1472	0.4591	0.3573	0.2300	0.4295	0.2837	0.3488	0.4249
proposed	0.4316	0.1481	0.4605	0.3586	0.2309	0.4308	0.2853	0.3503	0.4261

可以看出，在进行消融之后，推荐结果均出现了不同程度的下降，但都优于 BNS 方法，从而证明本文所提出的方法确实做出了有效的改进。从 without1 结果可以看出，基于用户-项目知识图谱构建的负样例候选集能够提高候选负样例整体的信息量，从而提高后续采样的质量。从 wihout2 结果可以看出，使用正混合技术能够增强挑选出来的负样例，提升推荐的效果。

5.6 不同 K 的实验结果及分析

如图 3 所示，本文对比了在 MovieLens-1M 数据集上的不同跳数 $K(K\in\{3,5,7\})$ 的 TOP-20 推荐结果。从结果可以看出，随着跳数的增加，模型的推荐效果呈下降趋势。这是因为随着跳数的增加，离用户节点较远的项目也有可能被采样为负样例，但是这些项目包含较少的用户偏好信息，因此导致了模型性能的下降。

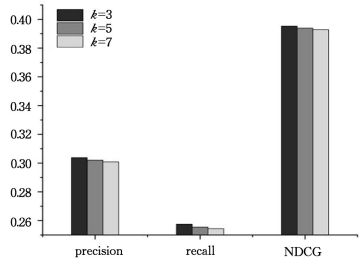

图 3 MovieLens-1M 数据集上不同跳数的实验结果

Fig. 3 Results of different hops on MovieLens-1M dataset

结束语 本文提出了一种融合知识图谱的负采样方法，

基于用户-项目知识图谱构建负样例候选集，然后通过基于贝叶斯分类的方法从候选集中挑选出采样风险最小的候选负样例，最后通过正混合技术增强挑选出来的负样例，最终将其用于模型的训练。该负采样方法能够较好地兼顾解决假阴性问题以及采样的负样例信息量不高的问题，从而提升模型的推荐效果。

本文虽然对应用于推荐系统中的负采样方法做出了一定的探索，但仍存在一些不足之处。首先，对于候选负样例为假负样例的先验概率建模，主要基于启发式的方法，仍有进一步优化的空间。其次，负采样面临着样本的信息量和无偏性的权衡，本文只是对此做出了一定的探索，并未达到这两者的最佳权衡。本文提出的方法中的正混合思想主要来源于对比学习，在未来的工作中，也可以考虑将更多对比学习中先进的研究结果用于负采样方法中，探索负样本信息量和无偏性之间更好的权衡策略，从而提升负采样的效果。除此之外，计算机视觉领域也有很多好的负采样策略，可以进行借鉴学习，将其拓展运用到推荐系统领域。

参 考 文 献

[1] GOLDBERG D, NICHOLS D, OKI B M, et al. Using collaborative filtering to weave an information tapestry[J]. Communications of the ACM, 1992, 35(12): 61-70.

[2] KABBUR S, NING X, KARYPIS G. Fism: factored item similarity models for top-n recommender systems[C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013: 659-667.

[3] KOREN Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008: 426-434.

[4] HE X, DENG K, WANG X, et al. Lightgcn: Simplifying and powering graph convolution network for recommendation[C]// Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 639-648.

[5] WANG X, HE X, WANG M, et al. Neural graph collaborative filtering[C]// Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019: 165-174.

[6] WANG X, HE X, CAO Y, et al. Kgat: Knowledge graph attention network for recommendation[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 950-958.

[7] GUO X W, XIA H B, LIU Y. Hybrid Recommendation Model of Knowledge Graph and Graph Convolutional Network[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(6): 1343-1353.

[8] LAI R, CHEN L, ZHAO Y, et al. Disentangled negative sampling for collaborative filtering[C]// Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. 2023: 96-104.

[9] RENDLE S, FREUDENTHALER C, GANTNER Z, et al. BPR: Bayesian personalized ranking from implicit feedback[J]. arXiv: 1205.2618, 2012.

[10] LIU B, WANG B. Bayesian Negative Sampling for Recommendation[C]// 2023 IEEE 39th International Conference on Data Engineering. 2023: 749-761.

[11] ZHANG H, CISSE M, DAUPHIN Y N, et al. mixup: Beyond empirical risk minimization[J]. arXiv: 1710.09412, 2017.

[12] SUN Z, GUO Q, YANG J, et al. Research commentary on recommendations with side information: A survey and research directions[J]. Electronic Commerce Research and Applications, 2019, 37: 100879.

[13] SARWAR B, KARYPIS G, KONSTAN J, et al. Item-based collaborative filtering recommendation algorithms[C]// Proceedings of the 10th International Conference on World Wide Web. 2001: 285-295.

[14] YOO H, CHUNG K. Deep learning-based evolutionary recommendation model for heterogeneous big data integration[J]. Transactions on Internet and Information Systems, 2020, 14(9): 3730-3744.

[15] RAO Z Y, ZHANG Y, LIU J T, et al. Recommendation methods and systems using knowledge graph[J]. Acta Automatica Sinica, 2021, 47(9): 2061-2077.

[16] SHI C, ZHANG Z, LUO P, et al. Semantic path based personalized recommendation on weighted heterogeneous information networks[C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. 2015: 453-462.

[17] 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.

[18] MA H D, FANG Y Q. Dynamic Negative Sampling for Graph Convolution Network Based Collaborative Filtering Recommendation Model[J]. Computer Science, 2023, 50(S2): 489-495.

[19] DIAZ-AVILES E, DRUMOND L, SCHMIDT-THIEME L, et al. Real-time top-n recommendation in social streams[C]// Proceedings of the Sixth ACM Conference on Recommender Systems. 2012: 59-66.

[20] CUI P, LIU S, ZHU W. General knowledge embedded image representation learning[J]. IEEE Transactions on Multimedia, 2017, 20(1): 198-207.

[21] HE X, ZHANG H, KAN M Y, et al. Fast matrix factorization for online recommendation with implicit feedback[C]// Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2016: 549-558.

[22] TOGASHI R, OTANI M, SATOH S. Alleviating cold-start problems in recommendation through pseudo-labelling over knowledge graph[C]// Proceedings of the 14th ACM International Conference on Web Search and Data Mining. 2021: 931-939.

[23] RENDLE S, FREUDENTHALER C. Improving pairwise learning for item recommendation from implicit feedback[C]// Proceedings of the 7th ACM International Conference on Web Search and Data Mining. 2014: 273-282.

[24] ZHANG W, CHEN T, WANG J, et al. Optimizing top-n collaborative filtering via dynamic negative item sampling[C]// Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2013: 785-788.

[25] ZHAO T, MCAULEY J, KING I. Improving latent factor models via personalized feature projection for one class recommendation[C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. 2015: 821-830.

[26] YU L, ZHOU G, ZHANG C, et al. RankMBPR: Rank-aware mutual bayesian personalized ranking for item recommendation [C]// Web-Age Information Management: 17th International Conference. 2016: 244-256.

[27] ZHAO Y, GUO G B, JIANG L Y. Adversarial Sampling for Social Recommender[J]. Journal of Cyber Security, 2021, 6(5): 88-98.

[28] DING J, QUAN Y, YAO Q, et al. Simplify and robustify negative sampling for implicit collaborative filtering[J]. Advances in Neural Information Processing Systems, 2020, 33: 1094-1105.

[29] HUANG T, DONG Y, DING M, et al. Mixgcf: An improved training method for graph neural network-based recommender systems[C]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 665-674.

[30] WANG J, YU L, ZHANG W, et al. Irgan: A minimax game for unifying generative and discriminative information retrieval models[C]// Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017: 515-524.

[31] PARK D H, CHANG Y. Adversarial sampling and training for semi-supervised information retrieval[C]// The World Wide Web Conference. 2019: 1443-1453.

[32] ZHAO Y H, LIU L, WANG H L, et al. Survey of Knowledge Graph Recommendation System Research[J]. Journal of Frontiers of Computer Science and Technology, 2023, 17(4): 771-791.

[33] ANAGNOSTOPOULOS A, KUMAR R, MAHDIAN M. Influence and correlation in social networks[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008: 7-15.

[34] KALANTIDIS Y, SARIYILDIZ M B, PION N, et al. Hard negative mixing for contrastive learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 21798-21809.

[35] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30-37.

LU Haiyang, born in 1999, postgraduate. His main research interests include knowledge graph and recommender systems.

LIU Xianhui, born in 1979, Ph.D, associate professor. His main research interests include machine learning, data mining and big data, and networked manufacturing.

(责任编辑:柯颖)