融入知识图谱的电影推荐算法

王玉奎 郭秀娟

(吉林建筑大学电气与计算机学院 长春 130118)

摘 要 传统的电影推荐系统可为用户推荐感兴趣的电影，但存在数据稀疏和冷启动问题。本文提出了ALS-G算法，将知识图谱作为辅助信息融入推荐系统中。首先利用图卷积神经网络GCN对电影属性知识图谱进行卷积获得电影的向量，将该电影向量与用户向量作为输入，然后利用ALS推荐算法预测该用户对该电影的喜爱程度。在MovieLens-1M数据集上的验证结果表明，本文的方法不仅能够缓解数据稀疏和冷启动问题，而且还提高了推荐系统的准确性。

关键词 知识图谱；电影推荐系统；ALS-G 算法

中图法分类号 TP183 DOI:10.16707/j.cnki.fjpc.2023.09.006

Algorithm for Movie Recommendation Based on Knowledge Graph

WANG Yukui, GUO Xiujuan

(Department of Electrical and Computer Science, Jilin Jianzhu University, Changchun, China, 130118)

Abstract Raditional movie recommendation systems can recommend movies of interest to users, but they have issues with sparse data and cold start. This paper proposes the ALS-G algorithm, which integrates the Knowledge graph as auxiliary information into the recommendation system. First, the graph Convolutional neural network GCN is used to convolve the film attribute Knowledge graph to obtain the film vector, and the film vector and user vector are used as inputs. Then, the ALS recommendation algorithm is used to predict the user's preference for the film. The validation results on the MovieLens-1M dataset show that our method not only alleviates data sparsity and cold start issues, but also improves the accuracy of the recommendation system.

Keywords Knowledge Graph; Movie Recommendation System; ALS-G Algorithm

1 引言

随着信息技术的飞速发展和视频播放设备的更新迭代[1]，电影大数据的信息量逐渐庞大。用户量剧增带来了数据的爆炸式增长，产生了信息过载现象。信息接收者接收到的信息远远超出了其处理信息的能力，大量无效冗余信息数据对信息接收者产生选择干扰，使其无法准确获得有效信息。推荐系统作为一种信息过滤的系统，有效帮助用户在信息过载的情况下获取感兴趣的信息。

传统的推荐算法有协同过滤推荐、基于内容的推荐和混合算法推荐。但这些推荐系统存在着数据稀疏和冷启动问题。为了缓解以上两个问题，在推荐模型中添加相关的知识图谱辅助信息能够取得更好的效果[2]。

自然语言是人类知识最主要的表达载体，但是文本字符串对机器并不友好，机器在理解人类语言方面仍然面临困难。知识图谱采用图的方式来描述和表达知识，是结构化的语义知识库，用符号的形式来描述现实世界中的概念及其相互关系。通常用头实体、关系、尾实体的三元组来描述知识图谱[3]。知识图谱相比于文本字符串，能够使机器理解起来更加容易。

将知识图谱作为辅助信息整合到推荐系统中，不仅有效缓解了数据稀疏和冷启动问题，而且能够

王玉奎，男，1999年生，主要研究领域为机器学习、推荐系统。E-mail:1561862010@qq.com。郭秀娟（**通信作者**），女，1961年生，主要研究领域为数据挖掘、图形图像处理、数值计算。E-mail:779523836@ qq.com。

提高推荐系统的准确性。本文在传统推荐算法的基础上，提出 ALS-G 算法。相比于 ALS（Alternating Least Squares）算法，本算法引入了电影属性信息知识图谱，缓解了推荐系统的数据稀疏和冷启动问题。利用图卷积网络 GCN（Graph Convolutional Network）对电影属性知识图谱卷积获得电影的向量，丰富了电影 Embedding，提高了推荐算法的准确度。但它不足的是，增加了图卷积模块，算法的时间和空间成本增加。

2 算法概述

2.1 ALS

ALS 算法属于基于模型的推荐算法[4]。它通过对数据的训练得到模型，用于预测用户对物品的评分。在现实中，用户-电影评分矩阵通常非常庞大，且多为空值，难以直接计算。ALS 是将评分矩阵分解成多个矩阵的乘积的形式达到降低维度的目的。公式如下：

$$R_{m\times n} = U_{m\times k} I_{k\times n} \tag{1}$$

其中，评分矩阵$R_{m\times n}$可以通过用户矩阵$U_{m\times k}$和项目矩阵$I_{k\times n}$乘积的形式对空值进行填充，m 表示用户的数量，n 表示电影数量，k 表示特征的维度。通过补全$R_{m\times n}$达到预测的目的[5]。

2.2 ALS-G

ALS-G 算法框架可分为 3 层：嵌入层、卷积层、预测层。ALS-G 框架图如图 1 所示。

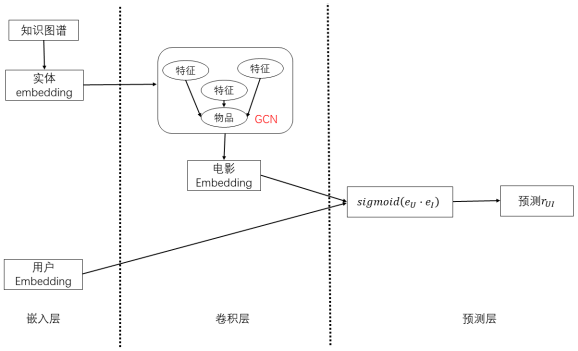

图 1 ALS-G 框架图

嵌入层主要分为两部分：第一部分是将所有用户进行 embedding 表示e_U，第二部分是将电影属性知识图谱中的所有实体进行 embedding 表示e_e。

卷积层是利用图卷积神经网络进行卷积，将知识图谱中的实体 Embeddinge_e作为卷积层的输入，基于 GCN 的消息传递机制，丰富电影 embedding 的表征[6]。GCN 的公式如下：

$$H^{l+1} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^l w^l\right) \tag{2}$$

其中，H^l是第l层的输入特征，H^{l+1}是指输出特征，w^l指线性变换矩阵σ是指非线性激活函数。在本算法中使用ReLU激活函数。其公式为：

$$ReLU(x) = max(0, x) \tag{3}$$

$\tilde{A}$表示有自连的邻接矩阵，定义如下：

$$\tilde{A} = A + I \tag{4}$$

其中，A表示邻接矩阵，I代表单位矩阵，代表自己和自己相连。$\tilde{D}$是自连矩阵的度矩阵，定义如下：

$$\tilde{D}_{ii} = \sum_j \tilde{A}_{ij} \tag{5}$$

经过卷积层后，得到电影的 embedding 表示e_I，通过向量内积形式来计算用户对目标图书的偏好预测分数：在数据处理时，将用户对电影的评分分为 0 或 1，最后在内积的基础上利用$sigmoid$函数来进行二分类预测[7]。定义如下：

$$\hat{y} = sigmoid(e_U \cdot e_I) \tag{6}$$

使用交叉熵损失函数(BCE)，训练时完整的损失函数为：

$$BCEloss(y,\ \hat{y}) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y}) \tag{7}$$

其中，$\hat{y}$表示 ALS-G 算法的预测值，y表示真实值。

ALS-G 算法伪代码如下所示：

Algorithm:

输入:用户项目评分矩阵 R;知识图谱 G 。

输出:目标用户对目标项目预测分值$\hat{y}$

步骤一：获取用户 embeddinge_U

步骤二：随机初始化知识图谱中实体 embeddinge_e

步骤三：利用公式(2)对 e_e做图卷积获得电影 embeddinge_I

步骤三：利用公式(6) $sigmoid(e_U \cdot e_I)$目标用户对目标项目预测分值$\hat{y}$

3 实验

3.1 数据集

实验使用的数据集为 Movielens-1M 中的用户物品评分三元组[8]，用户数量为 6036，电影数量为 2347，评价数量为 753772。对用户物品的评分三元

组进行文件预处理。将评分为 4 以上的（包含 4）的标注为 1（表示喜欢），将评分为 4 以下的标注为 0（表示不喜欢）。此时获得新的三元组数据（用户 ID，电影 ID，是否喜欢)。

3.2 实验环境

实验环境及实验配置如下：操作系统 Windows10，64 位操作系统，CPU 配置为 Inter(R)Core(TM)i10700kCPU@2.90GHz，8 GB 内存，采用 python 3.8 编程语言，深度学习框架为 Pytorch。CUDA Version 为 11.4。

研究将所有的数据按照 7: 3 的比例进行划分，分别用作模型的训练集与测试集。图卷积层的输出维度为 256。激活函数为 ReLU，使用 AdamW 优化器，初始学习率$\alpha = 0.001$.

3.3 推荐指标

实验选用两种经典推荐算法指标来进行对比评价：准确率（Accuracy）、精确率（Precision）。准确率（Accuracy）是预测准确的样本在所有样本中的比例；精确率（Precision）是预测准确的正例样本在所有预测为正例样本的比例[9]。公式如下：

$$Accuracy = \frac{TP+TN}{P+N} \tag{8}$$

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

其中，P表示正例数，N表示负例数，TP表示将正例预测为正例个数，TN表示将负例预测为负例个数，FP表示将负例预测为正例的个数。

3.4 实验结果及分析

为了验证电影知识图谱属性信息的作用，将本文提出的算法与 ALS 算法进行对照试验。平均损失为交叉熵损失的总和与参与训练数量的比，结果越小越好。两种算法的平均损失与训练批次的关系如图 2 所示。

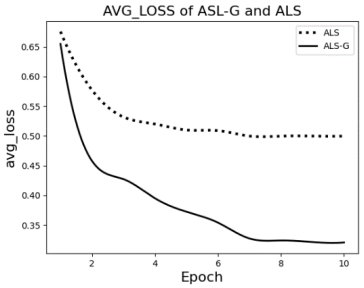

图 2 平均损失函数

由图 2 可知，虚曲线为 ALS 算法的平均损失函数，实曲线为 ALS-G 算法的平均损失函数。ALS 算法在第三轮迭代开始收敛，ALS-G 算法在第八轮开始收敛。加入电影知识图谱作为辅助信息的 ALS-G 算法比 ALS 算法平均损失函数的数值更小。

表 1 推荐指标

算法	Accuracy	Precision
ALS-G	0.8001	0.8441
ALS	0.8208	0.8560

由表 1 分析，ALS-G 算法相比于 ALS 算法在 Accuracy 和 Precision 都有所提升。其中 Accuracy 指标提升了 2.07%，Precision 指标提升了 1.19%。显然，将知识图谱作为辅助信息融合到电影推荐算法中提高了电影推荐算法的准确性。

4 结束语

本文提出的 ALS-G 算法在 ALS 推荐算法的基础上，利用 GCN 对电影属性知识图谱进行卷积来获得电影 Embedding，丰富了电影 Embedding 的表征。实验证明，将电影属性知识图谱作为辅助信息融合到推荐算法中相比于传统的 ALS 推荐算法拥有更好的准确度[10]。

参 考 文 献

[1] 项亮.推荐系统实战.北京:人民邮电出版社, 2012

[2] 李春英,武毓琦,汤志康,等.融合知识图谱的学习者个性化学习资源推荐.小型微型计算机系统，http://kns.cnki.net/kcms/detail/21.1106.tp.20230412.1830.002.html.,2023,08,04

[3] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE transactions on knowledge and data engineering, 2005, 17(6):734-749

[4] 陈帜.张文德,刘田.基于图卷积神经网络的图书推荐方法研究.情报探索,2022,300(10):1-5

[5] 徐雪东.刘晓东.基于时间加权 ALS 模型协同过滤推荐算法.电子设计工程,2022,30(14):39-43

[6] 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.南京理工大学学报,2022,46(02):177-184

[7] 冯舒.基于图卷积神经网络的协同过滤推荐模型研究[硕士学位论文].青岛理工大学,青岛,2022

[8] 张泽文.基于知识图谱的电影推荐系统的研究与实现[硕士学位论文].南昌大学,南昌,2021

[9] 於方仁.动手学推荐系统.北京:清华大学学出版社,2022

[10] 史宇涛.基于图网络结构的推荐系统模型研究[硕士学位论文].南京邮电大学,南京,2022