《智能信息处理》课程考试

基于知识图谱的推荐算法研究

杨显鹏

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2020 年 12 月 10 日

基于知识图谱的推荐算法研究

杨显鹏

(大连海事大学 信息科学技术学院, 辽宁省大连市 中国 116000)

摘 要 随着互联网的发展，数据开始呈爆炸性的方式增长，随之而来的是海量的脏数据。为了不受这些脏数据的影响得到高质量的数据，推荐系统开始引起人们的广泛关注。传统的推荐算法比如协同过滤(CF) 太过于依赖用户和物品的交互信息，因此遗留的数据稀疏和冷启动问题成为学术界一直需要攻克的难题。最近，知识图谱由于其三元组的易理解性以及丰富的语义信息而被广泛应用在推荐系统中。由此提出一种基于知识图谱嵌入模型 TransE 和图神经网络的推荐算法，使用图神经网络可以提取用户和物品的高阶特征，知识图谱嵌入算法模型可以使提取出的特征信息更加丰富，在相关数据集上的实验结果表明该算法的推荐性能表现优异。

关键词 推荐系统; 知识图谱; 图神经网络; 图嵌入

中图法分类号 **** DOI 号 *投稿时不提供 DOI 号* 分类号

Research of Recommendation Algorithm Based on Knowledge Graph

Xianpeng Yang

[1)](College of Information Science & Technology , Dalian Maritime University, Dalian , China)

Abstract

With the development of the Internet，data began to grow along with huse amounts of dirty data in an explosive way. In order to get higher quality data without being affected by these dirty data，people begin to focus on Recommendation Systems. Traditional recommendation algorithms，such as Collaborative Filtering (CF), rely too much on the interactive information between users and items，so the problems of data sparsity and cold start have become difficult problems to be solved by the academia all the time. Recently, Knowledge Graph has been widely used in recommendation systems because of its trituple's intelligibility and rich semantic information. For these reasons，we propose a recommendation algorithm model，which is based on knowledge embedding model TransE and Graph neural network. Higher-order representation of users and items can be extracted by using the Graph neural network. Knowledge graph embedding algorithm model can extract the characteristic information in more abundant way. Experimental results on related datasets show that the recommended performance of the algorithm is excellent.

Key words recommendation system; knowledge graph; graph neural network; graph embedding

1 引言

推荐系统像很多其他基于海量数据的任务一样受益于深度神经网络的发展，而知识图谱作为典型的图结构数据包含着实体到实体之间的关系，这对用户的兴趣分析和建模具有一定的辅助作用。基于矩阵分解的协同过滤(Collaborative Filtering，CF)是商业领域最成功的方法之一，然而，基于 CF 的方法依赖于用户和项目之间过去的交互，这将导致冷启动问题(不推荐没有交互的项目) 。为缓解这一问题，研究人员通常会采取一些措施去整合辅助信息，比如社交网络、图片和评论等。

在众多种类的辅助信息中，知识图谱被广泛使用，其以机器可读的头-关系-尾(head-relation-tail) 三元组形式组成并包含丰富的结构信息。研究人员先后利用知识图谱在节点分类、句子补全和摘要生成等应用中取得了成功。此后出现了基于知识图谱感知的推荐模型，其中许多都受益于图神经网络(Graph neural network，GNN) 捕捉图中的高阶结构并细化嵌入用户和项目的特征。 如 RippleNet 传播用户在知识图谱中潜在的偏好并探索其更深层次的兴趣; 图 卷 积 网 络 (Knowledge Graph Convolutional Networks，KGCN) 利用卷积操作来产生高阶的连通性的物品特征; 图注意力网络(Knowledge graph attention network，KGAT) 使用注意力机制隐式地为图中不同的邻域节点指定权重系数。
由于推荐系统的高维和异质性，在推荐系统中使用知识图谱仍是一个挑战。一种可行的方法是通过知识图嵌入 (Knowledge Graph Embedding，KGE) 方法先行预处理知识图谱，该方法可将图谱中的实体和关系映射到低维向量表示。常用的 KGE 方法侧重于建模严格的语义相关性(如 TransE 模型假定头 + 关系 = 尾) ，因此这类方法非常适合于图文应用领域，如知识图谱补全和链接预测。

本文在此基础上提出一种基于 GNN 和知识图谱嵌入模型 TransE 的推荐算法，先把相应的知识图谱信息通过 KGE 算法映射到高维的向量空间，再将图谱输入到相应的 GNN 之中; 实验表明，更高维的语义信息可以提升神经网络的学习能力，使最后的推荐性能有所提升。

2 相关概念介绍

2.1 推荐系统

对于一个典型的推荐系统，用户 u 和物品 v 的集合通常表示为$\boldsymbol{U}=\{u_1,u_2,u_3\cdots\}$和$\mathrm{V}=\{v_1,v_2,v_3\cdots\}$。根据用户和物品的历史行为数据，可得到二者间的交互矩阵为$\boldsymbol{Y}=\{y_{uv}\mid u\in\boldsymbol{U},v\in\boldsymbol{V}\}$。如果一对用户和物品间存在关联（如打分、点击等）那么$y_{uv}=1$反之$y_{uv}=0$。

对于用来提升推荐性能的知识图谱，用三元组集合$\boldsymbol{G}=\{(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})\mid h,t\in\boldsymbol{\varepsilon},r\in\boldsymbol{R}\}$来表示，其中每个三元组$(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$包含了知识图谱的头节点 h、尾节点 t 和二者之间的关系 r。$\boldsymbol{\varepsilon}=\{e_1,e_2,\ldots\}$ 表示所有的实体(包括头节点和尾节点)集合，$\boldsymbol{R}=\{r_1,r_2,r_3\cdots\}$为知识图谱的关系集合。
一般来说，知识图谱中的实体有一些代表着物品 v，而这些存在于知识图谱中的物品通常多个实体有关联，所以把物品 v 有关的实体集合表示为 N(v) 推荐系统可以在这个集合的基础上寻找到用户的潜在兴趣实体。最终的预测函数表示为

$$\hat{y}_{uv}=f(u,v;\Theta,\boldsymbol{G})$$

式中$\hat{y}_{uv}$为用户 u 对物品 v 感兴趣的概率；Θ 为整个函数f的所有参数；G 为知识图谱的三元组集合。

2.2 知识图谱嵌入模型 TransE

为使推荐模型和知识图谱更好地融合，使用知识图谱嵌入模型先行处理知识图谱数据，得到语义信息更为丰富的向量; 然后再将得到的相关嵌入向量输入到后续的 GNN 模型中。

传统的知识图谱一般使用本体语言表示，深度学习给予了一个更为明确的思路: 用向量的方式来表示知识图谱。这种形式在需要进行的任务中，如预测、推理等，具有更强的可扩展性与可表达性。嵌入模型目标是把一对对三元组编码为低维的向量形式。知识图谱嵌入模型的目的是向低维向量空间中嵌入多关系数据的实体和关系，同时还能保留数据中的结构信息。

表示学习的目的是将需要表达的对象(知识 图谱等) 表达为机器可以理解的实值向量的形式。对于知识图谱，表示学习目标是图谱中的实体和关系，然后构建模型将实体和关系映射到低维向量空间中进行后续的推理或预测任务，TransE 模型是表示学习的一个经典方法。

TransE 模型属于翻译模型：其将实体和关系表示为同一空间中的向量，对于给定的三元组$(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$ 模型将其中的关系$\boldsymbol{r}$看成头节点$\boldsymbol{h}$到尾节点$\boldsymbol{t}$的平移向量，即$h+r\approx t$; 这种思想来自于词向量空间的平移不变性，TransE 模型如图 1 所示。

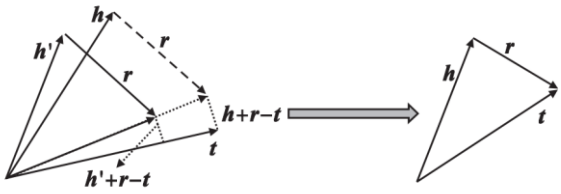

图 1 TransE 模型原理

在训练模型过程中，模型会不断调整其参数，使得知识图谱中的$\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}$的距离尽可能的小。模型的优化目标为

$$L=\sum_{(\boldsymbol{h},\boldsymbol{r},t)\in S}\sum_{(h',r,t')\in S}[\gamma+d(\boldsymbol{h}+\boldsymbol{r},\boldsymbol{t})-d(\boldsymbol{h}'+\boldsymbol{r},\boldsymbol{t}')]_+$$

式中$[x]_+$表示函数取值大于零时取值不变，小于令时则取零，这种函数一般称之为合页损失函数;γ为一个正确三元组与错误三元组之前的间隔修正，γ越大，两个三元组之前被修正的间隔就越大，则对向量的修正就越严格(一般都设置为 1)；d 为$\boldsymbol{h}+\boldsymbol{r}$和$\boldsymbol{t}$两个向量之间的距离，一般使用的是 L1 或 L2 范数；S 为用来训练的三元组的集合。模型的目标是让正确三元组之间的距离变小、错误三元组的距离变大; 所以如果函数取值大于零，则表示需要对模型的参数进行调整，训练流程如表 1 所示。

表 1 TransE 算法整体流程

输入：训练集 $\boldsymbol{S}=\{(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})\}$，实体集合 $\boldsymbol{E}$，关系集合 $\boldsymbol{R}$，修正间隔 γ，嵌入维度 k

输出：实体 $\boldsymbol{E}$ 和关系 $\boldsymbol{R}$ 集合的嵌入向量

初始化 $r\leftarrow$联合分布$\left(-\frac{6}{\sqrt{k}},\frac{6}{\sqrt{k}}\right)$

$e\leftarrow$联合分布$\left(-\frac{6}{\sqrt{k}},\frac{6}{\sqrt{k}}\right)$

开始循环

$S_{batch}\leftarrow$采样(S,b) //对 $\boldsymbol{S}$ 进行采样

$T_{batch}\leftarrow\phi$//初始化 T 的三元组集合

for $(h,r,t)\in S_{batch}$ **do**

$(h',r,t')\leftarrow$采样$(S_{(h,r,t)})$ //进行负采样

$T_{batch}\leftarrow T_{batch}\cup\{(h,r,t),(h',r,t')\}$

end for

更新参数：

$$\sum_{((h,r,t),(h',r',t'))\in T_{batch}}\nabla[\gamma+d(h+r,t)-d(h'+r,t')]_+$$

结束循环

模型在知识图谱上的采样步骤如图 2 所示。假设用户点击的物品为 v 1，以 v 1 为中心向外扩散 一个步长，将相关的实体放入一个集合 N(v)中，然后根据集合中每个实体的嵌入特征将集合的所 有特征聚合成一个向量，再以此特征为中心继续 重复之前的步骤向外扩散。

采样后网络会根据图卷积网络(Graph Convolution Networks，GCN)层进行特征提取，GCN 模型如图 3 所示。对每个物品 v，先根据上面的采样结果对邻域节点计算相关系数$\pi_{r_{v,e}}^{u}$代表着对用户 u 来说物品 v 和实体 e 的关联程度)，u 代表 不同用户，$r_{v,e}$表示物品 v 和邻域节点 e 间的关系。相关系数计算公式为

$$\pi_{r_{v,e}}^{u}=w_r(\text{concat}([\boldsymbol{u},\boldsymbol{r},\boldsymbol{v}]))+\boldsymbol{b}_r$$

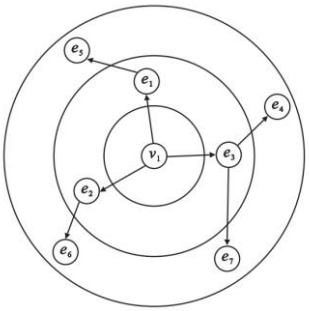

图 2 采样步骤

式中的w_r和$\boldsymbol{b}_r$皆为可训练参数。

得到相关系数后将其输入到一个 Softmax 层，可以得到新的系数$\pi_{r_{v,e}}^{u}$，然后对邻域实体的嵌入向量进行聚合并生成最终的面向用户的邻居信息 n，将 n 与原始的 v 向量通过聚合函数 agg (•)进行关联计算，可以得到模型最终需要的实体特征v',相关公式为

$$\tilde{\pi}_{r_{v,e}}^{u}=\frac{\exp(\pi_{r_{v,e}}^{u})}{\sum_{e'\in N(v)}\exp(\pi_{r_{v,e}}^{u})}$$

$$\boldsymbol{n}=\sum_{e\in N(v)}\tilde{\pi}_{r_{v,e}}^{u}\boldsymbol{e}$$

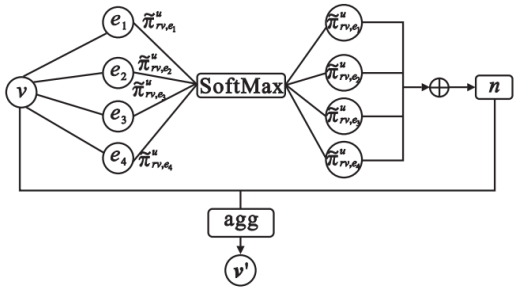

图 3 GCN 层结构

通过这种独特的关系注意力机制，能够将知识图谱和给出的用户、物品和关系信息进行融合并挖掘出用户更深层次的潜在兴趣。

结合前面介绍的 TransE 模型，本算法的整体流程如图 4 所示。输入一个特定的用户-物品(u，v) ，模型先根据 KGE 算法得到 v 的邻域节点向量，然后再输入到 GNN 中得到用户表征 u 和物品表征 v'，最终计算出相应的点击概率$\hat{y}_{uv}$。

3 学习算法

对于数据集中的每一对 u、v，模型最终都会得到用户 u 和物品 v 的特征向量 U 和 V’，然后计算出 u 对 v 的点击概率$\hat{y}_{uv}$，公式为

$$\hat{y}_{uv} = \sigma'(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}')$$

式中σ'为 sigmoid 函数。

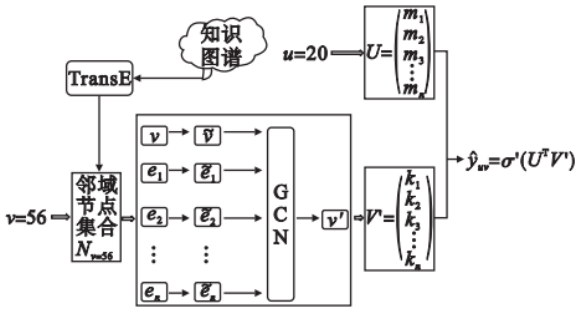

图 4 算法整体流程图

对模型进行优化时使用的是交叉熵损失函数，同时还使用了负采样策略解决训练数据的正 负样本失衡的情况。整体目标函数计算公式为

$$L = \sum_{u\in U}\left(\sum_{v:y_{uv}=1} £(y_{uv},\hat{y}_{uv}) - \sum_{i=1}^{Nu} E_{V_i\sim p(v_i)}£(y_{uv},\hat{y}_{uv})\right)$$

式中£为交叉熵损失函数；p 为对样本负样本的联合分布。公示的第二项为 L2 正则化。

4 实验结果

本实验采用的是电影推荐领域最常用的数据集 MovieLen，该数据集包含一百万个用户对电影的评分数据(评分在 1 ～ 5 之间) 、2445 部电影以及 6036 个用户。知识图谱来自于微软提供的开源的 Satori 数据库，通过相应的数据预处理，得到了适用于算法模型的 120 万条三元组数据、18 万个实体及 12 种关系。

作为一个经典的点击率预测问题，实验中使用准确率(Accuracy，ACC) 和曲线下面积(Area Under Curve，AUC) 两个评价指标。ACC 代表模型推荐的准确率，值越高说明模型性能越好。为克服样本不均衡问题，二分类问题常常把 AUC 也作为分类器的评价指标，其值越接近 1 代表分类器越优秀。

实验过程中先将数据集的评分进行二进制编码，阈值设置为 4，即评分低于 4 的编码为 0，其余的编码为 1。同时根据模型在验证集上的表现不断调整模型的超参数，在向量的嵌入维度方面进行相关的对比实验，其它参数不变的情况下把维度 d 设置在 2 ～ 64 维之间(以 2 的幂指数增长)。在数据集上的 AUC 变化如图 5 所示。

由图 5 可以看出，随着 d 在 2 ～ 8 维之间的尺度逐渐增大，AUC 也逐渐变大，因为嵌入的尺寸更大，编码的信息就更为丰富，但在 8 维之后性能开始下降，这可能是因为过拟合所致，因此实验中最终把嵌入维度设置为 8。

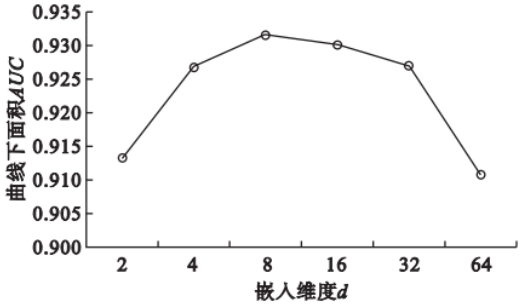

图 5 嵌入维度 d 与 AUC 的关系曲线

本文在数据集上进行了对比实验，结果表明

本文改进后的算法对比业界中表现良好的模型 FM、CKE、 RippleNet、KGCN 取得了较好的结果。基于协同过滤的 FM 算法由于没有使用异构的知识图谱信息表现最差; 基于正则化的 CKE 忽视了图谱的高阶连接性; 虽然 RippleNet 和 KGCN 表现同样优异，但本文额外多出的 KGE 模 块使得本文的算法相较于上述各算法在 ACC 的 表现方面分别取得了 2.3% 、2.4% 、1.2% 、1.2% 的提升，在 AUC 层面也分别获得了 2。9% 、2.3% 、1.0%、1.6%的增益，具体实验结果对比如表 2 所示。

表 2 实验结果对比

模型	*ACC*	*AUC*
FM	0.9101(－2.3%)	0.8328(－2.9%)
CKE	0.9095(－2.4%)	0.8376(－2.3%)
KGCN	0.9222(－1.2%)	0.8489(－1.0%)
RippleNet	0.9208(－1.2%)	0.8435(－1.6%)
本文模型	0.9316	0.8570

表中的百分数是其它模型对比本文模型结果获得的百分比增益。

5 结论

提出一种基于知识图谱嵌入模型 TransE 和图神经网络的推荐算法，通过知识图谱模型可以挖掘出用户更深层次的潜在兴趣，实验证明了该模型的优越性。该方法也可用于需要知识图谱结构性信息的领域，如社交网络或文本处理等。

参 考 文 献

[1] 祁燕,岳添骏,杨大为.基于用户打分和评论的推荐算法研究[J].沈阳理工大学学报,2018,37(02):11-17.

[2] 张吉祥,张祥森,武长旭,赵增顺.知识图谱构建技术综述[J/OL].计算机工程:1-16[2021-12-17].https://doi.org/10.19678/j.issn.1000-3428.0061803.

[3] Hongwei Wang (2019). Knowledge Graph Convolutional Networks for Recommender Systems.WWW.

[4] Breitfuss Arno,Errou Karen,Kurteva Anelia,Fensel Anna. Representing emotions with knowledge graphs for movie recommendations[J]. Future Generation Computer Systems,2021(prepublish):

[5] Chen Jiaying,Yu Jiong,Lu Wenjie,Qian Yurong,Li Ping. IR-Rec: An interpretive rules-guided recommendation over knowledge graph[J]. Information Sciences,2021,56

[6] Wang YueQun,Dong LiYan,Li YongLi,Zhang Hao. Multitask feature learning approach for knowledge graph enhanced recommendations with RippleNet.[J]. PloS one,2021,16(5):

[7] Enrico Palumbo,Diego Monti,Giuseppe Rizzo,Raphaël Troncy,Elena Baralis. entity2rec: Property-specific knowledge graph embeddings for item recommendation[J]. Expert Systems With Applications,2020,151(prepublish):