《智能信息处理》课程考试

基于知识图谱和 Bi-LSTM 的推荐算法

王壮

作业	分数[20]
得分	

2021 年 12 月 17 日

基于知识图谱和 Bi-LSTM 的推荐算法

王壮

(大连海事大学计算机科学与技术辽宁省大连市中国 116026)

摘 要：目前现有基于模型的推荐算法多是将评分数据输入到深度学习模型中进行训练，得出推荐结果。其缺陷在于无法 对预测结果进行可解释性分析。除此之外，无法有效地解决算法的冷启动问题。因此，本文提出一种基于知识图谱和 Bi-LSTM 的推荐算法，来有效解决算法的可解释性和冷启动问题。首先将获取到的数据集进行预处理，生成预编码向 量，根据数据集结点的连接性，构建专业领域知识图谱。其次利用知识图谱的元路径提取技术获取到多条用户-物品路 径信息，将其输入到 Bi-LSTM 中，在路径经过的各结点处加入一层注意力机制，目的是为了模型能够有效地获取到较远 结点的信息。最后将多条路径的训练结果输入到平均池化层中，用以区分不同路径的重要程度，利用交叉熵损失函数对 模型进行训练，从而得出预测结果。实验结果表明，与传统基于循环神经网络模型的推荐算法相比，该算法可有效地提 升算法的可解释性以及预测准确性，并缓解算法的冷启动问题。

关键词：知识图谱；双向循环神经网络；注意力机制；可解释性；冷启动

Recommendation Algorithm Based on Knowledge Graph and Bi-LSTM

Wangzhuang

(Computerscienceandtechnology，Dalianmaritimeuniversity,LiaoningDalian,116026,China)

Abstract: At present， most of the existing model-based recommendation algorithms input the score data into the deep learning model for training to get the recommendation results． Its defect is that it is unable to analyze the interpretability of the prediction results． In addition， the algorithm can not effectively solve the cold start problem． Therefore， this paper proposes a recommendation algorithm based on knowledge map and Bi-LSTM to effectively solve the problem of interpretability and cold start of the algorithm． Firstly，the data set is preprocessed to generate precoding vector． According to the connectivity of data aggregation points， the domain knowledge map is constructed． Secondly, the meta path extraction technology of knowledge map is used to obtain multiple user item path information，which is input into Bi-LSTM． A layer of attention mechanism is added to each node of the path， so that the model could effectively obtain the information of remote nodes． Finally，the training results of multiple paths are input into the average pooling layer to distinguish the importance of different paths． The cross-entropy loss function is used to train the model and the prediction results are obtained． The experimental results show that，compared with the traditional recommendation algorithm based on the cyclic neural network model，this algorithm can effectively improve the interpretability and prediction accuracy of the algorithm，and alleviate the cold start problem of the algorithm．

Key words: knowledge graph; bidirectional recurrent neural network; attention mechanism; interpretability; cold start

1 引 言

随着信息社会的飞速发展，人们可以方便地从各 种渠道获取到丰富的信息，但在面对庞杂的数据时， 很难在较短时间内筛选出有效信息，并做出相应判 断。在这种情况下，推荐算法的产生为人们解决了这 一难题 [1]。推荐算法可以利用互联网工具，在庞大 的数据中个性化地提取出有效信息，从而为人们提供 精准的服务，以此来满足用户的需求 [2]。推荐系统 可以在用户无明确需求的情况下，根据用户的历史行 为做出合理的推荐 [3]。并且在用户使用推荐系统的 同时，及时记录下用户的行为数据，为后续推荐提供 数据来源。

传统的协同过滤推荐系统 [4-6]，主要利用了用户和物品的属性信息，并且依赖用户的行为来进行物品推荐，在算法准确性上有显著的提高，但依然存在冷 启动和可解释性等问题 [7]。然而，这在提升算法的准确率和召回率方面尤为重要，因此，引入基于模型 的推荐算法。基于模型的推荐算法，多是将原始数据 信息输入到已构建好的深度学习模型当中，进行模型 训练，得出推荐结果。而由于深度学习模型原本就存在一定程度上的不可解释问题，因此引入知识图谱的概念[8],知识图谱作为先验知识，可以为推荐算法提 供语义特征。通过构建专业领域知识图谱，将带有语义的数据输入到模型当中，从而解决模型的可解释性 差问题。

目前，基于知识图谱的推荐算法主要是利用知识 图谱表示技术或知识图谱元路径提取技术 [9-10]。基于知识图谱表示技术的推荐算法，主要是利用原始数 据集构建知识图谱，利用建模的方式将结点和关系在向量空间中进行表示，通过计算从而生成低维稠密向量。Zhang 等人 [11] 利用 embedding 将交互信息、文本、图片进行编码，提取出多类型的语义特征，与推荐 物品相融合，提出 CKE 框架，充分学习物品的各属性 特征，从而提高算法的准确性。Wang 等人 [12] 提出了 CDL 模型，利用知识图谱表示学习得到物品结构化 信息，并利用去噪编码器网络学习编码层文本表示向 量和视觉表示向量，并将这些表示向量与物品的潜在 因子向量相结合，利用矩阵分解算法完成推荐。但其 问题在于关系类型较少且实体间关系较为稠密，导致 推荐的准确性不佳。Wang 等人 [13] 提出了 DKN 模型，利用知识图谱表示技术以及卷积神经网络对句子进行学习，并加入注意力机制，实现新闻推荐。但这些模型并未有效解决算法的可解释性以及冷启动等 问题。

基于知识图谱元路径提取技术的推荐算法，主要是利用已构建好的知识图谱不同结点之间的连通性，来提取出不同的路径信息，并结合协同过滤模型进行 推荐。Qian 等人 [14] 提出了一种全新的路径相似度计算方法 PathSim，通过计算不同路径的相似程度得出 相似物品，从而进行推荐。Shi 等人 [15] 提出了一种基于元路径的策略，利用带权元路径来对路径相似度进 行计算，构造带有权重的正则化项损失函数，来进行模型训练。Wang 等人 [16] 提出了 RippleNet 模型，将 知识图谱中的路径以及结点信息作为额外的知识加 入到推荐算法中，将用户对某一物品的喜好理解为水 波传播，以某一用户作为中心,不断向外发散，对各层 发散结果进行相似度计算，并融合各层结果，得出最 终预测值，实现物品推荐，但无法有效地解决算法的 冷启动问题。因此，本文提出一种基于知识图谱和 Bi-LSTM (双向循环神经网络)的推荐算法 KG-BiLSTM，将知 识图谱嵌入技术和知识图谱元路径提取技术相结合，在提升算法准确率的同时，解决算法的可解释性和冷 启动问题。在 MovieLens-1M 和 Yelp 数据集上进行 大量实验，实验表明该算法与主流推荐算法相比，在精确率、召回率和 MRR 指标上有较大提升。该算法 主要工作有如下几个方面:

1) 以有向图的形式构建专业领域知识图谱，将 用户属性信息与物品属性信息作为不同类型的结点 加入到知识图谱当中。并利用知识图谱元路径提取技术，提取出知

识图谱中不同用户-物品的路径信息， 存储于路径信息文件当中。

2）将不同路径的相应结点信息进行编码，生成嵌入向量，将各路径的各结点嵌入向量输入到 BiLSTM 当中，根据前后结点的信息，生成 Bi-LSTM 隐 藏层向量，使整条路径具有语义含义。

3）为充分捕捉到较远距离结点的信息，引入注 意力机制，将每条路径中的各结点隐藏层向量输入到 注意力机制当中，从而进一步提升预测的准确性。

4）将各路径信息输入到平均池化层中，用于区分不同路径的重要程度。对不同路径的预测信息进 行融合，利用交叉熵损失函数进行模型训练，不断更 新模型参数，得出最终预测结果。

5）将本文提出的 KG-BiLSTM 推荐算法与主流基于知识图谱的推荐算法在 MovieLens-1M 和 Yelp 数 据集上进行对比实验，并根据原始数据集，设置用户评分小于 4 的集合为冷启动数据集。根据 Precision、Recall 和 MRR 指标以及可解释性分析，验证该算法 在原始数据集和冷启动数据集上均可有效提升推荐 算法的准确性，并对推荐结果做出合理解释。

2 相关工作

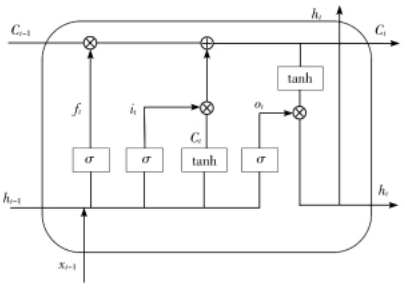
2.1 知识图谱

知识图谱最早的应用是提升搜索引擎的能力，提 高用户的搜索质量，其概念最早是由 Google 于 2012 年 5 月 17 日提出。随后在智能问答、自然语言处理、 推荐算法、数据分析等场合被广泛应用。

知识图谱 [17] 实质上是一个语义网络，由图的形 式来形象地描述不同事物之间的关系，通常由节点和 边组成。知识图谱中的节点用于表示实体和概念。实体是出现在知识图谱中的各类具体结点，概念是具有相同属性的某一类结点的集合。知识图谱中的边 表示不同结点之间的关系和属性。知识由三元组 h，r，t）进行表示，其中 h 为头实体，r 为关系，t 为尾 实体,一个完整的知识图谱需要由若干个不同的三元 组构成。在知识图谱表示学习中，通常是将知识图谱 嵌入到模型中，通过模型训练，使 h + r≈t，从而得到 知识的低维向量表示。而知识图谱元路径提取通常 是提取出整条路径的结点信息，将路径结点信息输入 到模型中进行训练,生成整条路径的低维向量，从而 赋予路径特定语义 [18]。

2.2 循环神经网络 LSTM

长短期记忆网络 LSTM 是一种循环神经网络的 变形 [19]。主要是为了解决传统循环神经网络中存在的长依赖问题，适用于处理间隔时间较长的事件。 LSTM 内部结构如图 1 所示。该结构由 3 个门控单元和 1 个细胞状态组成。其中细胞状态为 ct，类似一个链式结构，用于整合和记录信息。而门控单元用于选择信息是否允许通过以及通过量，以此来控制细胞状态。门控单元分别为输入门 it、遗忘门 ft 和输出门 ot。 输入门决定需要更新的信息，遗忘门决定需要覆盖哪 些信息，而输出门决定哪些信息被输出。由于 LSTM 拥有处理较长时间间隔信息的特性，因此被应用于知识图谱当中，可根据路径的选择来记录结点信息，用于后续模型输入。



图一 LSTM 内部结构图

2.3 注意力机制

注意力机制最初被应用于机器翻译领域，目前在人工智能领域，注意力机制可融合多种神经网络模型，在自然语言处理、推荐系统、统计学习等方面有较为广泛的应用。

注意力机制[20]类似人类的视觉机制，在对某事物进行观察时，通常倾向于关注较为重要的信息， 而忽视相关度较低的信息，

从而有助于做出最终决策。基于此原理，注意力机制可根据输入信息，对其重要性权值进行计算，将得到的重要性权值与输入信息相乘,再将乘以不同权值的输入信息相加，得出最终的 注意力向量。其模型如图 2 所示.

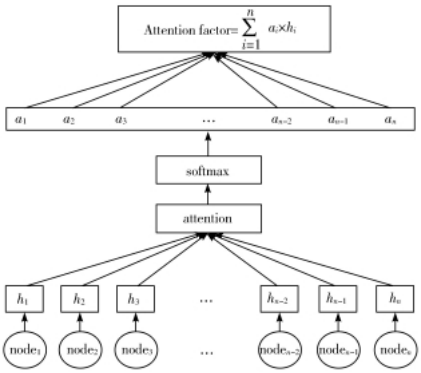


图 2 注意力机制模型图

利用该模型可实现将不同结点的特征信息进行有效融合，从而防止因路径过长导致较远结点特征无法完整提取的问题。

3 基于知识图谱和 Bi-LSTM 的推荐算法

本章将详细描述 Bi-LSTM 推荐算法的主要内容。该模型主要分为 5 层: 预处理层、嵌入层、Bi-LSTM 层、注意力层、预测层(包括池化层和全连接层) 。基于知识图谱和 Bi-LSTM 推荐算法的模型如图 3 所示。该算法首先将获取到的数据集进行分类预处理，对于每个用户和物品结点赋予唯一的 id 值，并构建一个有向图，将用户结点、物品结点、用户各属性结点以及 物品各属性结点添加到有向图中，并用边对各属性和关系相连，接着构建专业领域知识图谱。对已构建好 的知识图谱，提取每个用户-物品对的多条路径信息，设置路径条数最大值，将提取出的路径信息统一存储到文件中，作为后续模型的输入。其次，对各用户、物品及其属性信息进行编码，生成嵌入向量。然后，构建双向循环神经网络，将各路径的嵌入向量依次输入到神经网络当中，进行模型训练。同时引入注意力机制，充分获取前端较远结点的特征信息。最后，将各 条路径的隐藏层向量输入到平均池化层中，根据路径的重要性程度进行向量整合，并利用交叉熵损失函数完成模型训练

3.1 问题描述

推荐算法通常是根据用户对物品的已评分信息 预测未被评分物品的分值，算法的原始数据主要分为 用户集合 U = { U1，U2，…，Ui} 和物品集合 V = { V1，V2，…，Vj} ，其中 i 为用户总数，j 为物品总数。将用户与物品的交互信息存储于矩阵 R i × j 中，若 rij = 1 则表示用户 i 对物品 j 产生过交互，若 rij = 0 则表示用 户 i 对物品 j 未产生过交互。

将用户、物品及其属性信息映射到知识图谱 KG 当中，用有向图的形式来对知识图谱进行存储，头实体和尾实体对应有向图中的结点，关系对应有向图的边。由于知识图谱是由多个三元组组成，因此每个用户-物品交互信息可表示为(U，interact，V) 三元组的 形式，interact 表示用户对物品产生过交互。并且物 品-属性信息同样也可表示为三元组形式，将物品属性信息以结点的形式添加到知识图谱中。

对已构建好的知识图谱提取不同用户-物品对的多条路径信息，将路径信息进行编码，输入到模型中训练出相应的用户和物品预测向量，对预测向量相乘 从而得到每个用户-物品的预测值，生成 top-N 推荐列表 [21]，将预测推荐列表与用户已交互的列表进行比 对,以此来计算出模型的精确率和召回率。

3.2 KG-BiLSTM 算法

3.2.1 预处理层

对已构建好的知识图谱提取不同用户-物品对的路径信息。在该推荐算法当中(Ui，Vj) 表示一个实体对，其中 U 代表用户集合，V 代表物品集合。i 代表某 个用户，j 代表某个物品。若存在用户 i 对物品 j 产生 过交互，则利用不同长度的路径来连接该实体对。即 存在 P(Ui，Vj) = { P1，P2，…，Pm } ，m 的值代表用户物品对路径条数，对于其中的任意一条路径存在 pm

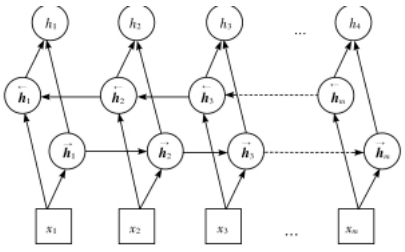
= e0 r →1 e1 r →2 e2 → … r →n en。其中 e0 = Ui、en = Vj ，pm 代表用户-物品对的第 m 条偏好路径，n 表示经 过的结点个数。r1，r2，…，rn 表示连接偏好路径相对 应的各个关系。其中所经过的 e1、e2 表示经过的物品属性结点。每条路径有其不同的语义表示。将生成 的各条路径送入嵌入层，进行编码。

3.2.2 嵌入层

引入嵌入层的目的是在于将高维稀疏的特征向量转换成低维向量，便于模型输入。对于预处理层中 涉及的各个结点 ei，首先生成预编码向量。针对已构 建好的路径，提取路径中涉及的各个结点，将其输入 嵌入层中，进行模型训练。利用嵌入层生成每个结点 相对应的嵌入向量 en = { en1，en2，en3，…，enr } ，r 表示 嵌入向量的维数。从而将每个结点表示成低维向量的形式。每个嵌入向量中的各个分量代表结点的特征，将含有结点特征的嵌入向量输入隐藏层中，使路 径包含一定的语义信息。

3.2.3 Bi-LSTM 层

将每个用户-物品对的各条路径信息，按其结点 的先后顺序依次输入到双向循环神经网络当中，考虑到前端结点与末尾结点距离较远，本文采用 Bi-LSTM 模型，可以充分融合前端结点的特征信息，从而保证特征提取的完整性。其中双向循环神经网络 BiLSTM 如图 4 所示。



图四 双向循环神经网络 Bi-LSTM 模型图

3.4 本文算法流程

本文算法流程如算法 1 所示。

算法 1 KG-BiLSTM 算法

输入：用户-物品交互矩阵 R、物品属性信息、用 户、物品和物品属性的预编码向量。

输出：根据预测值生成用户 U 对物品 V 的 top-N 推荐列表。

1）构建知识图谱 KG。将用户-物品交互矩阵R和物品属性信息中涉及的物品及其属性以结点的形 式添加到有向图中并用边对结点相连。

2）提取知识图谱 KG 中不同用户-物品对的路径 信息。

3）将已提取到的路径信息输入到嵌入层中，利用 embedding 生成嵌入向量。

4）将各路径中涉及的结点嵌入向量输入到 BiLSTM 层中，利用式(1)～式(6) 计算出各结点的 BiLSTM 层隐藏向量。

5）将每条路径上的各结点隐藏向量，输入到注意力层中，用于融合较远结点的属性信息，利用式 (7) ～ 式(8)计算得出整条路径的注意力层向量。

6）整合各条路径的注意力层向量，将该向量输入到平均池化层中，利用式(9)对不同路径的重要性程度进行区分，根据重要性权值对多条路径向量进行融合。

7）将生成的平均池化层向量输入到全连接层中，用于对向量进行降维，利用式(10)计算得出用户物品对的最终预测向量。

8）利用式(11)计算出模型的损失函数值。

9）利用 SGD 优化器对参数进行梯度更新，通过不断训练，更新模型参数值，以达到使损失函数最小化的目的。

4 结束语

本文针对推荐算法中存在的可解释性和冷启动 等问题提出了 KG-BiLSTM 算法，该算法将知识图谱 实体嵌入和路径提取相结合，并充分考虑路径中各个结点的特征信息，针对用户已交互的物品，根据其各 结点的类型，学习不同路径的特征，并最终进行向量 融合，实现 Top-N 预测。该模型在 2 个真实数据集上进行了实验分析，其结果均优于目前主流算法。在未来，还将考虑融入用户结点的属性信息，从而进一步 提高算法的预测精确率。

[1] SHAPI R A B. R ecommender Systems Handbook [M] . Springer, 2011 .

[2] JANNACH D，ZANKER M，FELFERNIG A，et al． Recommender Systems: An Introduction [M] ． Cambridge University Press，2010．

[3] XIA F， LIU H F， LEE I， et al ． Scientific article recommendation: Exploiting common author relations and historical preferences ［J］． IEEE Transactions on Big Data，2016，2(2) : 101-112．

[4] 杨武，唐瑞，卢玲． 基于内容的推荐与协同过滤融合的新闻推荐方法 ［J］． 计算机应用，2016，36(2) : 414-418．

[5] LINDEN G, SMITH B, YORK J .Item-to-item collaborative filtering ［J］． Internet Computing，2003，7(1) : 76-80．

[6] 李玉省． 个性化推荐系统关键技术研究［D］． 北京: 北京邮电大学，2016．

[7] PRIYANKA R D, VIJENDRA S． Systematic evaluation of social recommendation systems: Challenges and future [J] ． International Journal of Advanced Computer Science and Applications.2016.7(4),DOI:10.14569/IJACSA .2016 .070420．

[8] 徐增林, 盛泳潘, 贺丽荣, 等 ． 知识图谱技术综述 [J] ． 电子科技大学学报，2016，45(4) : 589-606．