



计算机工程
Computer Engineering
ISSN 1000-3428, CN 31-1289/TP

《计算机工程》网络首发论文

题目: 融合知识图谱和协同过滤的推荐模型
作者: 康雁, 李涛, 李浩, 钟声, 张亚钊, 卜荣景
DOI: 10.19678/j.issn.1000-3428.0056234
网络首发日期: 2019-12-13
引用格式: 康雁, 李涛, 李浩, 钟声, 张亚钊, 卜荣景. 融合知识图谱和协同过滤的推荐模型. 计算机工程. <https://doi.org/10.19678/j.issn.1000-3428.0056234>



网络首发: 在编辑部工作流程中, 稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定, 且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式(包括网络呈现版式)排版后的稿件, 可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定; 学术研究成果具有创新性、科学性和先进性, 符合编辑部对刊文的录用要求, 不存在学术不端行为及其他侵权行为; 稿件内容应基本符合国家有关书刊编辑、出版的技术标准, 正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性, 录用定稿一经发布, 不得修改论文题目、作者、机构名称和学术内容, 只可基于编辑规范进行少量文字的修改。

出版确认: 纸质期刊编辑部通过与《中国学术期刊(光盘版)》电子杂志社有限公司签约, 在《中国学术期刊(网络版)》出版传播平台上创办与纸质期刊内容一致的网络版, 以单篇或整期出版形式, 在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊(网络版)》是国家新闻出版广电总局批准的网络连续型出版物(ISSN 2096-4188, CN 11-6037/Z), 所以签约期刊的网络版上网络首发论文视为正式出版。




融合知识图谱和协同过滤的推荐模型

康雁, 李涛, 李浩, 钟声, 张亚钊, 卜荣景

(云南大学, 软件学院, 云南 昆明 650500)

摘 要: 针对已有协同过滤推荐算法可解释性不高和基于内容推荐算法信息提取困难、推荐效率低等问题, 提出了一种融合知识图谱和协同过滤的高效推荐模型。其中, 一个模型首先获取知识图谱的推理路径并基于 TransE 算法将路径嵌入成向量, 接着利用 LSTM 和 soft attention 机制捕获路径推理的语义, 之后运用池化操作区分不同路径推理的重要性, 最终经过全连接层和 sigmoid 函数获得预测评分。另一个模型根据知识图谱表示学习的语义相似性, 利用协同过滤算法的思想获得预测评分。按预测评分的准确度将两个模型有效地融合, 最终获得可解释的混合推荐模型。模型在 MovieLens 数据集上进行了实验分析。与相关代表性算法相比, 实验结果具有较好的推荐解释性和更高的推荐准确率。

关键词: 知识图谱; 协同过滤; 深度学习; 混合推荐; 知识表示学习

开放科学(资源服务)标识码(OSID): 

A Recommendation Model Based on Fusion of Knowledge Graph and Collaborative Filtering

Kang Yan, Li Tao, Li Hao, Zhong Sheng, Zhang Yachuan, Bu Rongjing

(Department of Software, Yunnan University, Kunming, Yunnan 650500, China)

【Abstract】 Aiming at the problem of low interpretability of existing collaborative filtering algorithm, difficulty in extracting information and getting higher recommendation efficiency from existing content-based recommendation algorithm. An efficient hybrid recommendation model based on knowledge graph and collaborative filtering is proposed. One model firstly gets reasoning paths of the knowledge graph and embeds the paths into vectors based on TransE algorithm. Then it captures the semantics of paths reasoning by LSTM and soft attention mechanism. Then it uses pooling operation to distinguish the importance of different paths reasoning. Finally, it gets the prediction score through the full connection layer and sigmoid function. Another model uses the idea of collaborative filtering algorithm to obtain prediction scores based on the semantic similarity of knowledge graph representation learning. According to the accuracy of prediction score of two models, the two models are fused effectively, and finally an interpretable hybrid recommendation model is generated. The model is tested on MovieLens dataset. Compared with the related representative algorithms, the experimental results have good recommendation interpretability and higher recommendation accuracy.

【Key words】 knowledge graph; collaborative filtering; deep learning; hybrid recommendation; knowledge representation learning
DOI:10.19678/j.issn.1000-3428.0056234.

0 概述

随着大数据时代的到来, 各互联网公司尤其是各大国内外的电商网站对利用数据的推荐算法研究越来越重

视。目前, 对个性化推荐的研究已有很多成果。推荐算法主要有协同过滤推荐、基于内容的推荐和混合推荐等^[1]。

协同过滤推荐算法^[2,3,4,5]尝试利用人群意愿进行推

基金物品: 国家自然科学基金(No.61762092, No.61762089); 云南省软件工程重点实验室开放基金项目(No. 2017SE204)

作者简介: 康雁 (1972-), 女, 博士/副教授, 硕士生导师, 主要研究方向为深度学习和自然语言处理。李涛 (1995-), 男, 通讯作者, 硕士研究生, 主要研究方向为推荐系统; 李浩 (1970-), 男, 教授, 硕士生导师, 主要研究方向为分布式计算和网络计算; 钟声 (1995-), 男, 硕士研究生, 主要研究方向为软件工程和推荐系统; 张亚钊 (1996-), 女, 硕士研究生, 主要研究方向为推荐系统; 卜荣景 (1996-), 女, 硕士研究生, 主要研究方向为推荐系统。

E-mail: 365318663@qq.com

荐^[6]。协同过滤虽然在推荐准确性上取得了显著的效果,但与许多基于内容的推荐算法相比解释推荐结果不直观^[7]。基于内容的推荐^[8,9]试图用各种可用内容信息对用户和物品进行特征建模。物品内容对用户来说通常是容易理解的,所以在基于内容的推荐中,通常可以直观地向用户解释为什么该物品被推荐。但基于内容的推荐在不同的推荐背景下收集所需要的内容信息是一项耗时的任务^[7]。而构建知识图谱只需要利用实体与实体之间的关系,这将很大程度上减少基于内容的推荐提取内容信息的工作量。知识图谱作为一种新兴的辅助数据也引起了学术界和工业界的广泛关注^[10]。

近年来,由于强大的表示学习能力,深度学习在图像处理、自然语言处理和语音识别等领域取得了巨大的突破和成就,也为推荐系统领域的研究带来新的机遇。但这些基于深度学习的推荐^[11,12,13]大多数是基于矩阵分解的思想,只考虑了用户与物品的评分数据,这个原因抑制了模型的推荐效果。^[14]

混合推荐算法^[15,16]一般能够同时综合多种推荐算法的优良性,具有更好的推荐效果。将知识图谱推荐、深度学习和混合推荐算法联系起来,可以使整个推荐方法在具有可解释的情况下也可能取得更高推荐准确率。

综合上述背景,本文利用相关技术,提出了一种结合知识图谱、深度学习和协同过滤的混合推荐模型—HCKDC (Hybrid Recommendation Model Combining Knowledge Graph, Deep Learning and Collaborative Filtering)。其中,一个模型将知识图谱推理结合深度学习,生成第一种预测评分;另一个模型将协同过滤思想结合知识图谱表示学习的语义相似性,计算得到第二种预测评分。预测评分的准确性决定了其对最终模型的影响程度。最终按模型的重要程度将两个模型有效地融合,得到可解释的推荐结果。在 MovieLens-100K 数据集上实验结果表明,和近期先进的可解释推荐模型 RKGE (Recurrent Knowledge Graph Embedding)^[17]和 RippleNet^[18]相比,该模型在具有较好解释性情形下能取得更高的推荐准确率,具有先进性。

1 相关研究

推荐算法主要有协同过滤推荐、基于内容的推荐和混合推荐等。

1.1 协同过滤推荐

最早的协同过滤推荐算法是 Resnick 等人在[2]中提出的基于用户的协同过滤。之后 Sarwar 等人在[3]中介绍了基于物品的协同过滤算法。当协同过滤与 Koren 在文献[4]引入的潜在因子模型 (LFM) 相结合时,协同过滤取得了进一步的成功。另外也有一些对协同过滤算法的优化,如论文[5]提出的将潜在因子模型和邻域模型融合形成的多方面的协作过滤模型。协同过滤一定程度上可以根据其算法设计的原理进行解释。例如,基于用户的协同过滤推荐可以解释为:推荐的物品是与该用户类似的用户喜欢的物品。而基于物品的协同过滤可以解释为:推荐的物品是与用户喜欢的物品类似的物品。但其与许多基于内容的推荐算法相比解释推荐结果不直观。

1.2 基于内容的推荐

基于内容的推荐试图用各种可用内容信息对用户和物品进行特征建模。基于内容的推荐有 Ricci 等人^[8]和 Pazzani 等人^[9]提出的基于内容的协同过滤系统。虽然基于内容的推荐具有较好的可解释性,但在不同的推荐背景下收集所需要的内容信息是一项耗时的任务。

1.3 基于知识图谱的推荐

知识图谱包含用户和物品的丰富信息,这些特性也为推荐物品的生成提供更直观的、更具针对性的解释。且构建知识图谱只需要利用实体与实体之间的关系,这将很大程度上减少基于内容的推荐提取内容信息的工作量。

将知识图谱引入推荐系统的有根据连接两个实体的路径代表不同语义的扩展潜在因子模型^[19],如论文[20]中的元路径方法。这种方法有助于根据物品相似性推断用户偏好,从而生成有效的推荐。然而,元路径方法严重依赖于手工制作元路径的特性,需要额外领域的知识。而且手工制作的元路径特性往往不完整,很难覆盖所有可能的实体关系,从而阻碍了推荐质量的提高。和元路径相比,知识图谱表示学习能够自动学习获得知识图谱中实体的语义嵌入,拥有比元路径更好的效果^[17]。基于知识图谱表示学习的研究到现在也有很大的进展^[21]。其中备受关注的有 Translating 系列表示学习算法。Bordes 等人在[22]提

出了 TransE 算法, TransE 的基本方法是给定三元组 (h, r, t) , 用关系 r 向量作为头实体 h 向量和尾实体 t 向量之间的平移。

1.4 基于深度学习的推荐

Salakhutdinov 等人在[11]中第一次将深度学习应用在推荐系统, 并提出了一种基于受限玻尔兹曼机的协同过滤模型。Huang 等人在[12]中提出了深度结构化语义模型。2017 年, 何向南[13]等人进一步结合多层感知机模型和广义矩阵分解模型提出了神经协同过滤的算法, 进一步提升了推荐模型的性能。但这些算法模型大多数是基于矩阵分解的思想, 只考虑了用户与物品的评分数据, 这个原因抑制了模型的推荐效果。

1.5 混合推荐

混合推荐算法一般能够同时综合多种推荐算法的优良性, 具有更好的推荐效果。如吴玺煜等人在论文[15]

中对协同过滤推荐算法和基于知识图谱表示学习语义邻近推荐算法进行融合, 提升了推荐准确率。Wei Zhao 等人在论文[16]中用了四种方式将矩阵分解和 RNN 模型融合, 也提升了推荐准确率。

2 模型

我们的 HCKDC 模型由知识图谱与深度学习结合模型 RCKD (Recommendation Model Combining Knowledge Graph and Deep Learning) 和知识图谱与协同过滤结合模型 RCKC (Recommendation Model Combining Knowledge Graph and Collaborative Filtering) 组成。RCKD 能够模拟知识图谱中的推理过程, 解决信息提取困难和可解释性不高的问题。RCKC 具有较好可解释性并能提高推荐质量。整个模型框架如图 1 所示。

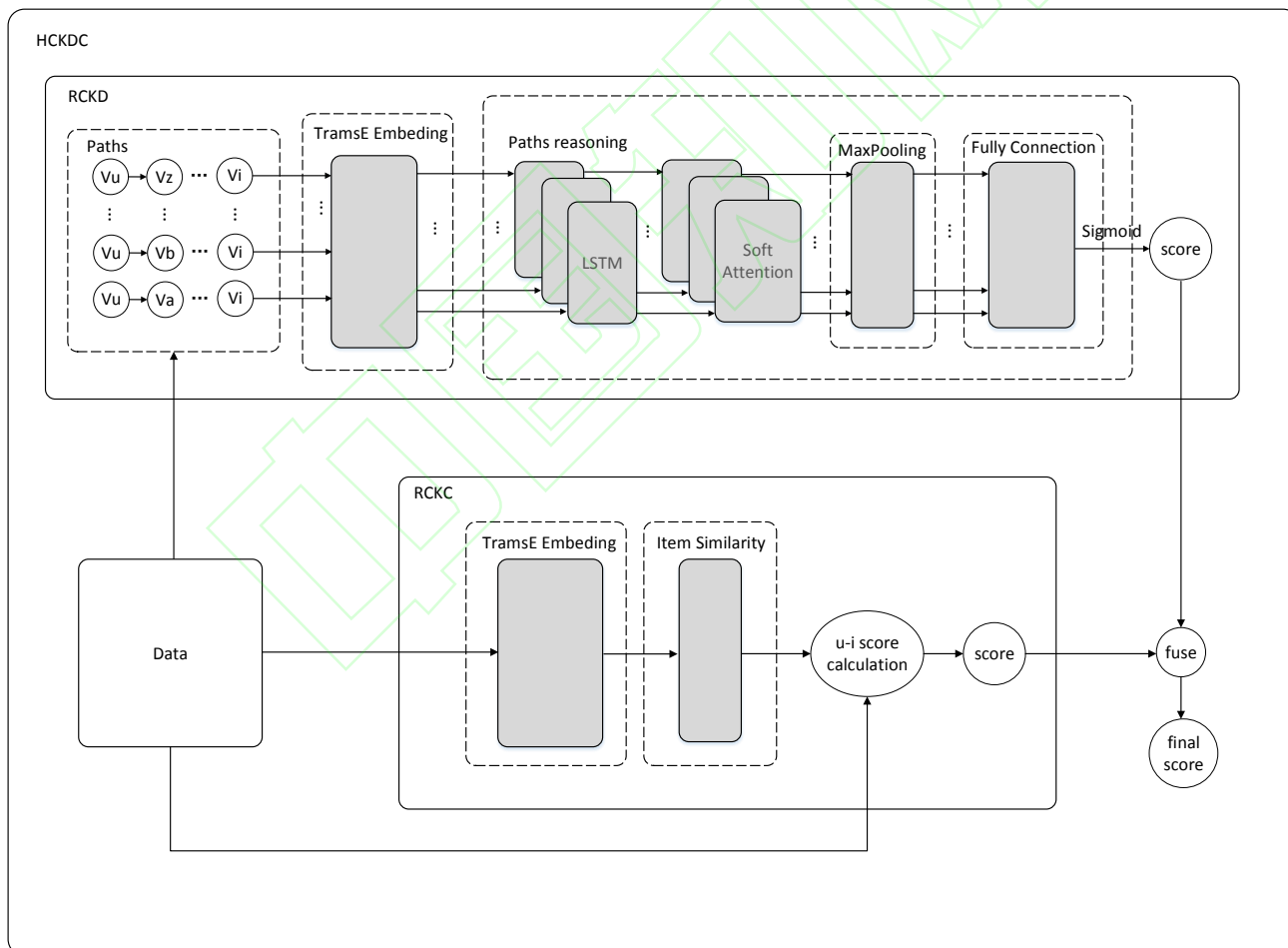


图 1 HCKDC 模型框架图

2.1 RCKD 模型

RCKD 结合知识图谱与深度学习方法, 首先获取知

识图谱推理路径,然后基于知识图谱表示学习 TransE 方法将推理路径嵌入成为向量,最后运用深度学习捕获路径推理语义进而获得预测评分。

2.1.1 推理路径的生成及推理描述

我们利用已知的所有三元组(h, r, t)构成图,然后使用构建图的方法搜索知识图谱中的推理路径。具体地,将实体映射到图中的点,关系映射为图中的边,通过指定两个实体遍历整个图就能得到两个实体的所有路径。这些路径即是两个实体之间的推理路径。图 2 是一个局部的图,u186 到 i322 的生成的推理路径总共有 6 条,分别是“u186→i1042→g3→i322”、“u186→i79→g3→i322”、“u186→i291→g3→i322”、“u186→i55→g3→i322”、“u186→i540→g3→i322”、“u186→i540→a152→i322”。

推理路径“u186→i540→g3→i322”可以表示用户 u186 喜欢电影 i540,电影 i540 属于 g3 流派,电影 i322 也属于 g3 流派,所以用户 u186 也可能喜欢电影 i322;推理路径“u186→i540→a152→i322”可以表示用户 u186 喜欢电影 i540,电影 i540 的演员有 a152,电影 i322 也被 a152 演过,所以用户 u186 也可能喜欢电影 i322。这个例子强调了连接同一个实体对的不同路径通常带有不同语义关系。通常,它们在描述用户对物品的品味方面具有不同的重要性。区别这些路径的重要性并综合这些路径的推理结果得到预测评分,这种推理过程具有较好可解释性。

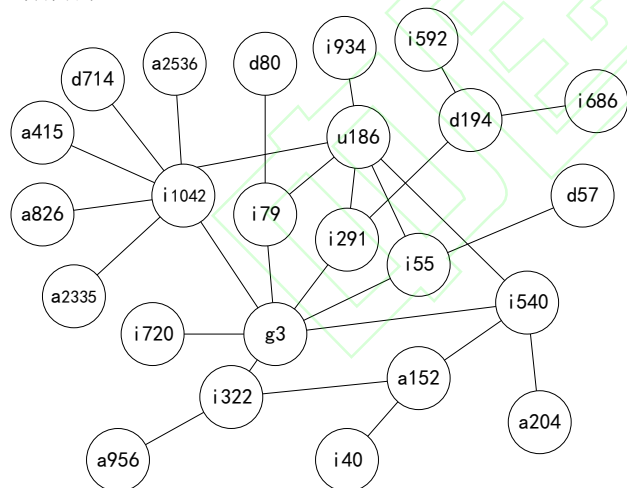


图 2 知识图谱局部图

RCKD 模型就是使用深度学习的方式模拟这些路径的推理过程,再对这些路径的重要性加以区分,最后综合这些路径的推理语义得到预测评分。

2.1.2 推理路径嵌入向量的生成

深度学习在自然语言处理、图像处理等领域的巨大

突破很大程度得益于其领域的强大表示学习能力。因此,我们基于知识表示学习 TransE 算法得到知识图谱推理路径嵌入向量。TransE 的基本方法是给定三元组(h, r, t),TransE 用关系 r 的向量 \mathbf{l}_r 作为头实体向量 \mathbf{l}_h 和尾实体向量 \mathbf{l}_t 之间的平移。TransE 的损失函数定义为:

$$\text{loss}_r(h, t) = \|\mathbf{l}_h + \mathbf{l}_r - \mathbf{l}_t\|_{L1/L2} \quad (1)$$

即 $\mathbf{l}_h + \mathbf{l}_r$ 到 \mathbf{l}_t 的 L1 或 L2 距离。

在实际训练过程中,为了加强知识表示的区分能力,TransE 采用了最大间隔方法,定义了优化目标函数:

$$\text{loss} = \sum_{(h,r,t) \in S} \sum_{(h',r',t') \in S^-} \max(0, f_r(h, t) + \gamma - f_{r'}(h', t')) \quad (2)$$

其中, S 是合法三元组的集合, S^- 为错误三元组的集合, γ 为合法三元组得分与错误三元组得分之间的间隔距离。

错误三元组也不是随机产生的。为了选取有代表性的错误三元组,TransE 将 S 中每个三元组的头实体、关系和尾实体其中之一随机替换成其他实体或关系来得到 S^- 。

通过 TransE 对所有三元组的训练,每一个实体 k 可以表示成为一个向量 \mathbf{e}_k 。一对实体 u-i 第 m 条路径的嵌入表示如公式 (3) 所示。

$$\mathbf{p}_{uim} = [\mathbf{e}_{uim1}, \mathbf{e}_{uim2}, \dots, \mathbf{e}_{uimm}, \dots] \quad (3)$$

其中 \mathbf{e}_{uimm} 表示用户 u 到物品 i 的第 m 条路径中的第 n 个实体的嵌入向量。

2.1.3 预测评分的生成

在这一部分,模型首先通过 LSTM(Long Short-Term Memory)^[23]和 soft attention^[24]机制捕获路径的推理语义,再通过 maxpooling 操作区分不同路径推理语义的重要性,最后用全连接汇集池化向量经过 sigmoid 函数生成预测评分。

有推理路径“u186→i540→g3→i322”,这个路径表示用户 u186 喜欢电影 i540,电影 i540 属于 g3 流派,电影 i322 也属于 g3 流派,所以用户 u186 也可能喜欢电影 i322。这个推理过程是一步一步得到的,可以采用循环神经网络 RNN (Recurrent Neural Network) 捕获推理的语义。我们选取的是有效解决了 RNN 梯度消失和梯度爆炸的改进网络模型长短期记忆网络 LSTM。

LSTM 单元如图 3 所示,它有三种类型的门控,分别为:输入门、遗忘门和输出门,分别使用 i 、 f 和 o 来表示。此外 LSTM 中还有记忆单元 c 和输出值 h 。

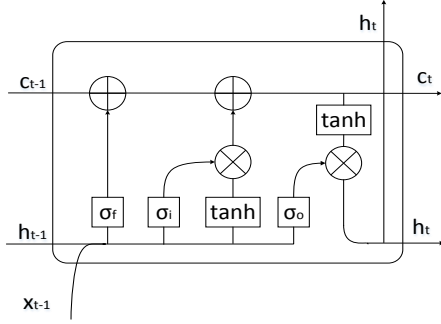


图 3 LSTM 单元

LSTM 迭代的第 t 时刻各值的计算公式由 (4) ~ (8) 所得:

$$i_t = \sigma(A_{xi}x_t + A_{hi}h_{t-1} + b_i) \quad (4)$$

$$f_t = \sigma(A_{xf}x_t + A_{hf}h_{t-1} + b_f) \quad (5)$$

$$o_t = \sigma(A_{xo}x_t + A_{ho}h_{t-1} + b_o) \quad (6)$$

$$c_t = f_t \odot c_{t-1} + \quad (7)$$

$$i_t \odot \tanh(A_{xc}x_t + A_{hc}h_{t-1} + b_c)$$

$$h_t = o_t \odot \tanh(c_t) \quad (8)$$

其中 A 和 b 分别为 LSTM 的权重矩阵和偏置向量。

\odot 表示 Hadamard 乘积。

为了避免网络过拟合,我们对每条路径都采用了同一个 LSTM 捕获语义。已嵌入的路径向量 p_{uim} 输入进 LSTM 中,能够得到 LSTM 推理中每一次迭代的推理语义,实体对 $u-i$ 的第 m 条路径的输出语义就可以表示为公式 (9)。

$$h_{uim} = [h_{uim1}, h_{uim2}, \dots, h_{uimn}, \dots] \quad (9)$$

为了使最后推理的语义与推理过程中每次迭代得到的推理语义关联性更强,我们使用了 attention 机制。我们选取的是 soft attention 机制。这一层能够求得 LSTM 每一次迭代推理得到的语义对最后一次迭代推理语义的影响度,并总结出这个路径的最终语义。假设对于实体对 $u-i$ 的第 m 条路径有 l 个实体,对于 LSTM 第 n 次迭代推理语义的 attention 权重 W_{uimn} 可以用公式 (10) ~ (11) 得到。

$$s_{uimn} = h_{uiml} \cdot h_{uimn} \quad (10)$$

$$W_{uimn} = \frac{\exp(s_{uimn})}{\sum_{k=1}^l \exp(s_{uimk})} \quad (11)$$

这条路径最终经过 soft attention 处理后得到的最终语义可由公式 (12) 得到。

$$q_{uim} = \sum_{n=1}^l h_{uimn} W_{uimn} \quad (12)$$

为了区分路径推理语义的不同重要程度,我们选择

了池化操作。池化操作能够关注向量最重要的特征或综合的特征,符合我们需要关注路径语义不同重要性的要求。经实验验证 maxpooling 效果优于 avgpooling,我们选取了 maxpooling 池化操作。最后通过全连接层汇聚所有路径的池化向量,经过 sigmoid 函数产生 $u-i$ 的预测评分 D_{ui} 。结果可由公式 (13) 所得,其中的 mpooling_con 的 mpooling 和 con 分别代表 maxpooling 操作和全连接操作。

$$D_{ui} = \frac{1}{1 + e^{-\text{mpooling_con}(q_{uim}, \dots, q_{uin}, \dots)}} \quad (13)$$

2.2 RCKC 模型

协同过滤方法虽然不能有效地解释推荐结果,但有较高的推荐准确率。为了使协同过滤具有更好的解释性,我们提出了 RCKC 模型。RCKC 结合知识图谱与协同过滤方法,采用 TransE 方法得到知识图谱中实体语义表示向量,再利用协同过滤思想根据向量的相似性推荐与用户喜欢物品语义相近的物品。

根据 2.1.2 每个实体都可以用 TransE 方法得到对应的语义表示向量。RCKC 首先根据这些语义表示向量计算每个物品和其他物品的语义相似性。对于物品 i 和物品 j 的语义相似性分数用公式 (14) 得到:

$$I_{ij} = I_{ji} = \frac{1}{1 + \|e_i - e_j\|_1} \quad (14)$$

其中 I 是整个语义相似矩阵, I_{ij} 表示物品 i 和物品 j 的语义相似性分数。 $\|e_i - e_j\|_1$ 表示 $e_i - e_j$ 的 L1 范数结果,即:将 $e_i - e_j$ 向量的每个分量的绝对值相加。

假设用户 u 历史评分集 T_u , 需要预测的物品集为 E_u , 对于 $\text{item}_k \in T_u$, $\text{item}_i \in E_u$, 用户 u 对物品 item_k 的评分为 R_{uk} , 物品 item_k 和 item_i 语义相似性分数表示为 I_{ki} , 预测用户 u 对物品 i 的评分为 C_{ui} 。用户 u 对物品 i 的预测评分由公式 (15) 所得。

$$C_{ui} = \sum_{\text{item}_k \in T_u} I_{ki} \cdot R_{uk} \quad (15)$$

RCKC 模型中预测评分列表生成方法如算法 1 所示。

算法 1 RCKC 评分列表生成算法

输入 历史评分集 T , 需要预测的用户物品集 E , 语义相似矩阵 I 。

输出 RCKC 预测评分列表 C 。

1. for $E_u \in \{E_1, \dots, E_n\}$ and $T_u \in \{T_1, \dots, T_n\}$ do:
2. for $\text{item}_i \in E_u$ do:
3. for $R_{uk} \in T_u$ do:

4. $C_{ui} \leftarrow C_{ui} + R_{uk} \times I_{ki}$
5. end for
6. end for
7. end for

其中 $E_u = \{item_1, item_2, \dots\}$, $T_u = \{R_{u1}, R_{u2}, \dots\}$ 。

RCKC 的推荐的预测评分是利用知识表示学习的语义相似性推荐与用户历史喜欢最接近物品, 具有较好的可解释性。

2.3 HCKDC 模型

HCKDC 模型即是融合了 RCKD 模型和 RCKC 模型的总模型。我们的方法借鉴了 Wei Zhao 等人在论文[16]中的第一种融合策略。具体融合策略为: 将矩阵分解和 RNN 推荐对应预测评分相加, 经过 sigmoid 函数产生最终的预测评分。用户 u 和物品 i 的 RCKC 预测评分是 C_{ui} , RCKC 模型的预测评分是 D_{ui} , 最后融合生成的预测评分由公式 (16) 得到。

$$L_{ui} = \frac{1}{1 + e^{-(\alpha C_{ui} + \beta D_{ui} + b)}} \quad (16)$$

其中 α , β 分别为 RCKC 预测评分 C_{ui} 和 RCKD 的预测评分 D_{ui} 的比例系数。 b 为偏置项。

融合策略的训练过程自动根据 RCKC 和 RCKD 预测得分的准确性调整其对最终模型的影响程度, 并优化 RCKC 和 RCKD 的内部参数, 具有很好可解释性。

最后, 用户的推荐列表通过用户对物品最后预测评分从大到小排序得到, 用户的推荐列表生成过程可用算法 2 描述。

算法 2 HCKDC 推荐生成算法

输入 需要预测的物品集 E , RCKC 推荐预测评分列表 C , RCKD 预测评分列表 D , RCKC 比例系数 α , RCKD 比例系数 β , 偏置项 b 。

输出 HCKDC 推荐列表 L 。

1. for $E_u \in \{E_1, E_2, \dots, E_n\}$ do:
2. for $item_i \in E_u$ do:
3. $L_{ui} \leftarrow \alpha \times C_{ui} + \beta \times D_{ui} + b$
4. end for
5. sort(L_u)
6. end for

其中 $E_u = \{item_1, item_2, \dots\}$ 。

3 实验与分析结果

3.1 数据集

本文使用的数据集是 MovieLens-100K 数据集。在

此基础上, 我们在 IMDB 上爬取到了电影流派、导演、演员信息作为辅助信息。这些辅助信息扩展了知识库, 使知识图谱能获得更好的性能。

MovieLens-100K 数据集包含 943 个用户, 1682 个电影, 100000 个评分。去除缺失数据后的数据如表 1 所示。这些数据映射到知识图谱后, 总共包含 7746 个实体、8 种关系和 202183 个三元组。其中的三元组不包含测试集中正样本集的评分与被评分关系三元组。

表 1 去除缺失数据后的数据集

数据名称	数量
用户	943
电影	1675
评分数据	99975
导演	1154
演员	3948
电影流派	26

3.1.1 数据划分

经测试, 在协同过滤中测试了考虑评分值和不考虑评分值, 推荐效果差距很小, 甚至考虑评分后还有推荐效果下降的趋势。这会在 3.5 中通过实验证明。因此我们的实验中没有考虑评分值, 即: 用户对某个电影有评分, 就假设为该用户喜欢该电影, 评分置为 1。该用户不喜欢的电影评分设为 0。

为了使实验具有可比性, 我们对所有的实验都采用相同的训练集和测试集。产生的训练集的正样本和测试集的正样本是对加入辅助信息后的 99975 个评分数据以 4:1 的比例进行拆分得到的。测试集的正样本的作用是检验模型产生的推荐列表中的物品是否准确。

对于训练集中负样本的选取, 我们采用的是随机抽取。为使模型能够学到更多的负样本信息, 我们对每个用户抽取的负样本数据是训练集正样本的 120%。

由于随机选取了一些负样本作为训练集, 如果不把这些负样本纳入测试集。这会引起相较于协同过滤等方法, 得到的测试集中可能包含的负样本更少, 预测结果有不真实性。为了避免这样的情况, 我们将除了训练集中正样本的其他所有样本作为测试集。

3.2 路径抽取

论文[20]已证明, 越短路径对推荐结果越重要且路径长度超过 5 会引入噪声。为加速路径抽取过程和训练过程, 本文对每一对 user-item 最多只挖掘 5 条最短的路径, 即最多挖掘 5 条长度为 3 的路径。

路径抽取过程是首先抽取训练时所用到的正负样本的路径, 再抽取用户到其他物品的路径。

3.3 评价指标

我们采用的评价指标包括 precision@N 和 MMR@N , precision@N 表示每个用户推荐列表前 N 项在测试集正样本中出现的概率的均值。 precision@N 的 N 取值包括 1、5、10。

MMR@N 定义如公式 (17) 所示。

$$\text{MMR@N} = \frac{1}{N} \sum_{i=1}^N \sum_{v_j \in \text{test}(u_i)} \frac{1}{\text{rank}(u_i, v_j)} \quad (17)$$

其中我们设的 N 为 10, v_j 是在 top- N 推荐列表中正确出现在测试集正样本的项。 $\text{rank}(u_i, v_j)$ 表示 v_j 在用户 u_i 推荐列表的位置数。

3.4 模型设置

考虑到所有 user-item 路径数量的不确定性,在池化操作中如果用相同池化窗口和池化步长会引起数据维度不一致。所以需要根据路径数量动态地调整池化窗口和池化步长,大小都为(paths_size, 1)。

为保证 RCKD 有效性,我们设置了预训练次数,即在不融合的情况下单独训练 RCKD 的次数。实验中 RCKD 训练 6 次的时候效果最好,所以我们将预训练次数设置为 6 次。 α , β , b 初始化都为 0。LSTM 隐藏单元数为 64,学习率为 0.1,优化方法采用随机梯度下降。

针对 RCKC 和 RCKD 不同推荐列表的预测评分差异性,再考虑到 RCKD 模型的预测评分属于[0,1]范围。我们首先得到 RCKC 模型的所有推荐结果,再将 RCKC 得到的推荐列表的从前到后以 1 到 0 递减的顺序重新设置评分。

3.5 实验结果及分析

本文在公开数据集 MovieLens-100K 进行测试,比较方法包括近期先进的可解释推荐方法:Zhu Sun 等人的 RKGE 模型、Wang 等人提出的 RippleNet。此外,我们还和经典的协同过滤算法进行了对比。实验结果如表 2 和表 3 所示。

表 2 实验结果 1

推荐方法	RippleNet	RKGE	RCKD	RCKC	HCKDC
precision@1	0.1548	0.1580	0.2142	0.2163	0.2545
precision@5	0.1379	0.1094	0.1777	0.1864	0.2023
precision@10	0.1300	0.0956	0.1577	0.1687	0.1836
MRR@10	0.4108	0.3399	0.5259	0.5493	0.6086

表 3 实验结果 2

推荐方法	User-based -CF	User-based -CF-CR	Item-based -CF	Item-based -CF-CR
precision@1	0.2174	0.2163	0.2238	0.2280
precision@5	0.1968	0.1968	0.1892	0.1908
precision@10	0.1668	0.1667	0.1679	0.1674
MRR@10	0.5627	0.5620	0.5538	0.5568

在数据 3.1.1 小节中有提到,协同过滤考虑评分值和不考虑评分值推荐效果差距很小。甚至考虑评分后还有推荐效果下降的趋势。表 3 中的 User-based-CF 表示基于用户的协同过滤; User-based-CF-CR 表示考虑评分值的基于用户的协同过滤; Item-based-CF 表示基于物品的协同过滤; Item-based-CF-CR 表示考虑评分值的基于物品的协同过滤。从表 3 来看基于用户的协同过滤考虑评分值时有轻微的下降,基于物品的协同过滤考虑评分值时也会在 precision@10 有所下降。

从表 2 和表 3 可以看出,本文提出的 HCKDC 的推荐准确性在所有比较方法中最高,协同过滤、RCKC、RCKD 其次, RippleNet 和 RKGE 最差。可以发现可解释推荐算法 RippleNet 和 RKGE 虽然对其推荐可解释性较好,但推荐效果低于协同过滤。并且,在同时具有较好可解释性情形下, RippleNet 和 RKGE 推荐效率也低于我们的 RCKD 和 RCKC 单个模型。此外,很容易观察到 RCKC 和 RCKD 在融合之后推荐准确率得到进一步提高,这证明了融合策略的有效性。本模型的可解释性已在第 2 章有了较好的描述,从比较方法上来看,我们的模型取得了最高的推荐准确率,这证明了本模型能够在具有较好可解释性的情况下拥有更高的推荐准确率,具有先进性。

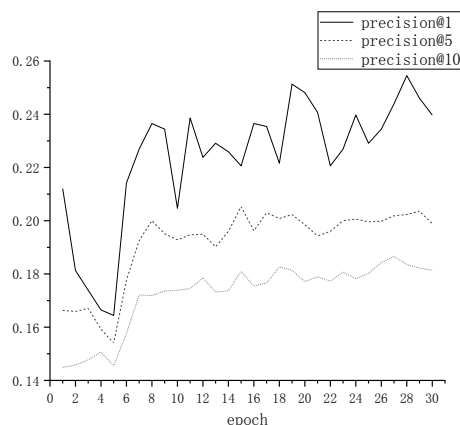


图 4 HCKDC 训练过程中 precision 的变化

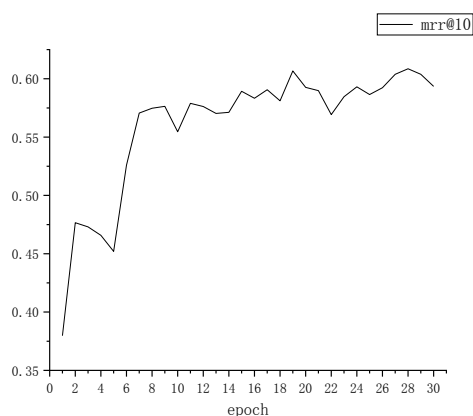


图 5 HCKDC 训练过程中 MRR@10 的变化

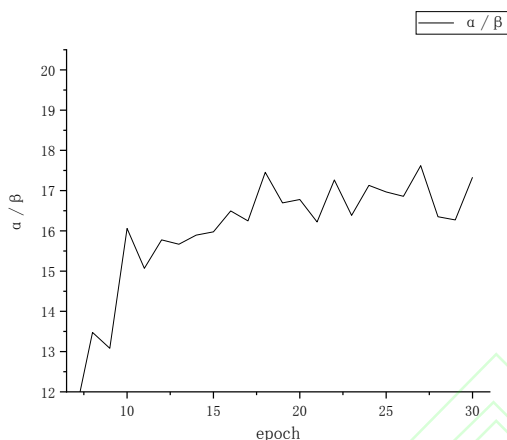


图 6 HCKDC 训练过程中 α/β 的比值的变化

HCKDC 的训练过程如图 4、图 5、图 6 所示。图 4 和图 5 中可以发现 MRR@10 和 precision 在 epoch 为 28 左右时最好。从图 6 中观察到,当训练次数达到 15 次左右时,训练出模型中的 RCKC 和 RCKD 的比例基本维持在 16.5 左右,可见 RCKD 对整个模型的影响程度很小。但 RCKD 和 RCKC 融合后整体推荐效率提高,说明 RCKD 有效修正了融入知识图谱的协同过滤的推荐结果。

4 结束语

通过该研究,我们对知识图谱推荐、协同过滤推荐和混合推荐有了更深的认识。本文提出了一种结合了知识图谱、深度学习和协同过滤的混合推荐模型 HCKDC。该模型由 RCKD 和 RCKC 组成,RCKD 通过将知识图谱中的推理用深度学习模拟得到预测评分,RCKC 将知识图谱表示学习结合协同过滤思想得到预测评分。HCKDC 将对应的两种推荐得到的同一用户对同一物品不同预测评分分别乘以系数,相加再加上一个偏置项,得到的结果作为用户对该物品的最终喜好程度。最后通过对用户

按物品预测评分从高到低排序得到最终的物品推荐列表。实验结果表明相较于近期先进的可解释模型 RKGE 和 RippleNet,该模型在具有可解释性情形下能达到更高的推荐准确率,具有先进性。未来,我们将尝试把电影海报等信息加入进该推荐系统中。

参考文献

- [1] Zhu Wenyue, Liu Wei, Liu Zongtian, et al. News Personalized Recommendation Based on Event Ontology[J].Computer Engineering,2019,45(06):267-272+279.
- [2] Resnick P, Iacovou N, Suchak M, et al. GroupLens: An Open Architecture for Collaborative Filtering of Netnews[C]// Acm Conference on Computer Supported Cooperative Work. 1994.
- [3] Sarwar B, Karypis G, Konstan J, et al. Item-based Collaborative Filtering Recommendation Algorithms[C]// International Conference on World Wide Web. 2001.
- [4] 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, Las Vegas, Nevada, USA, August 24-27, 2008. ACM, 2008.
- [5] Koren Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model[C]. knowledge discovery and data mining, 2008: 426-434.
- [6] Ekstrand M D. Collaborative Filtering Recommender Systems[J]. Acm Transactions on Information Systems, 2007, 22(1):5-53.
- [7] Zhang Y, Chen X. Explainable Recommendation: A Survey and New Perspectives[J]. arXiv: Information Retrieval, 2018.
- [8] Ricci F, Rokach L, Shapira B, et al. Recommender Systems Handbook[M]. Springer US, 2011.
- [9] Pazzani M J. Content-based recommendation systems[C]// Adaptive Web. Springer-Verlag, 2007.
- [10] Guan Sai-ping, Jin Xiao-long, Jia Yan-tao, et al. Knowledge Reasoning Over Knowledge Graph: A Survey[J].Journal of Software,2018,29(10):2966-2994.

官赛萍,靳小龙,贾岩涛,王元卓,程学旗.面向知识图谱的知识推理研究进展[J].软件学报,2018,29(10):2966-2994.

- [11] Salakhutdinov R , Mnih A , Hinton G . [ACM Press the 24th international conference - Corvalis, Oregon (2007.06.20-2007.06.24)] Proceedings of the 24th international conference on Machine learning - ICML \07 - Restricted Boltzmann machines for collaborative filtering[C]// International Conference on Machine Learning. ACM, 2007:791-798.
- [12] Huang P S , He X , Gao J , et al. Learning deep structured semantic models for web search using clickthrough data[C]// Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013.
- [13] He X, Liao L, Zhang H, et al. Neural Collaborative Filtering[J]. arXiv: Information Retrieval, 2017.
- [14] Zhang S, Yao L, Sun A, et al. Deep Learning Based Recommender System: A Survey and New Perspectives[J]. ACM Computing Surveys, 2019, 52(1).
- [15] Wu Xiyu, Chen Qimai, Liu Hai, et al. Collaborative Filtering Recommendation Algorithm Based on Representation Learning of Knowledge Graph[J]. Computer Engineering, 2018, 44(02):226-232+ 263.
吴玺煜,陈启买,刘海,贺超波.基于知识图谱表示学习的协同过滤推荐算法[J]. 计算机工程, 2018, 44(02):226-232+263.
- [16] Zhao W, Wang B, Ye J, et al. Leveraging Long and Short-term Information in Content-aware Movie Recommendation.[J]. arXiv: Information Retrieval, 2017.
- [17] Sun Z, Yang J, Zhang J, et al. Recurrent knowledge graph embedding for effective recommendation[C]. conference on recommender systems, 2018: 297-305.
- [18] Wang H , Zhang F , Wang J , et al. RippleNet: Propagating User Preferences on the Knowledge Graph for Recommender Systems[C]// the 27th ACM International Conference. ACM, 2018.
- [19] Shi Y , Larson M , Hanjalic A . Collaborative Filtering beyond the User-Item Matrix: A Survey of the State of the Art and Future Challenges[J]. ACM Computing Surveys (CSUR), 2014, 47(1):1-45.
- [20] Sun Y , Han J , Yan X , et al. Pathsirn: Meta path-based top-k similarity search in heterogeneous information networks[J]. Proceedings of the Vldb Endowment, 2011, 4(11):992-1003.
- [21] Liu zhiyuan, Sun Maosong, Lin Yankai, et al. Knowledge Representation Learning: A Review[J]. Journal of Computer Research and Development, 2016, 53(2):247-261.
刘知远, 孙茂松, 林衍凯, et al. 知识表示学习研究进展[J]. 计算机研究与发展, 2016, 53(2):247-261.
- [22] Bordes A, Usunier N, Garcíadurán A, et al. Translating Embeddings for Modeling Multi-relational Data.[C]// International Conference on Neural Information Processing Systems. 2013.
- [23] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8):1735-1780.
- [24] Xu K, Ba J, Kiros R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[C]. international conference on machine learning, 2015: 2048-2057.