一种融合用户显隐式阅读偏好的论文推荐模型

唐 浩 刘柏嵩* 黄伟明
(宁波大学信息科学与工程学院 浙江 宁波 315211)

摘 要 在海量学术文献的个性化推荐中,现存基于内容的方法以 CNN 作为特征提取工具,关注用户的显式阅读偏好,却忽略了全局语义特征,而基于图的方法通常忽略用户和论文之间的高阶关联结构信息。针对以上问题,提出一种混合推荐模型 GNPR(Graph Neural Paper Recommendation),能够学习更完整的用户显式阅读偏好及用户和论文之间的高阶关联信息。该方法使用 Word2vec 和 DCNN(Dual Convolutional Neural Network)处理文本,以双层自注意力的特征抽取模式学习文本全局特征,补充用户显式阅读偏好。针对概念、用户、论文和论文元数据等数据构建知识图谱,使用改进的图卷积网络学习用户和论文之间的高阶关联信息,从而挖掘用户隐式的阅读偏好。在 CiteULike-a 等数据集上验证了 GNPR 模型的有效性。

关键词 论文推荐 知识图谱 高阶结构信息 用户偏好 图神经网络

中图分类号 TP391 **文献标志码** A DOI: 10.3969/j.issn.1000-386x.2022.05.039

A PAPER RECOMMENDATION MODEL WITH USER EXPLICIT AND IMPLICIT READING PREFERENCES

Tang Hao Liu Baisong* Huang Weiming
(*College of Information Science and Engineering, Ningbo University, Ningbo 315211, Zhejiang, China*)

Abstract In the personalized recommendation of massive academic literature, the existing content-based methods use CNN as a feature extraction tool, focusing on the user's explicit reading preferences, but they ignore the global semantic features. The graph-based methods usually ignore the high-order association structure information between users and papers. Aiming at the above problem, this paper proposes a hybrid recommendation model GNPR, which can more effectively learn the user's explicit reading preferences and the high-level association information between the user and the paper. This method used Word2vec and DCNN to process the text, and learned the global features of the text with a two-layer self-attention feature extraction mode to supplement the user's explicit reading preference. The knowledge map was constructed for the data such as concept, user, paper and paper metadata. The improved graph convolution network was used to learn the high-order correlation information between users and papers, so as to mine users' implicit reading preferences. This paper validates the effectiveness of the GNPR model on datasets such as CiteULike-a.

Keywords Paper recommendation Knowledge graph High-order structure information User preference Graph neural network

0 引 言

对科研工作者而言,获取高相关性和高质量论文需要耗费大量时间和精力,如果存在一种可以智能协助研究人员(下文简称“用户”)高效地寻找论文的过滤工具必将受到欢迎。目前用户查找论文一般通过特定关键词直接搜索,然而每次查找后必须再经若干次的过滤才能得到较为满意的论文列表;另一种有效的方法是从参考书目或者引文数据中筛选,虽然在某种程度上提高了查询结果的相关度,但是固有问题是优质的新论文由于引文的缺失而很难被搜索到。论文推

收稿日期:2020-02-27。学术性大数据知识组织与服务标准研究项目(15FTQ002)。唐浩,硕士生,主研领域:推荐系统。刘柏嵩,研究员。黄伟明,博士生。

荐系统简化了用户查找论文的流程,促成从人找论文到论文找人的转变[1],极大地提高了论文获取的效率。目前的论文推荐方法[2]一般分为基于协同过滤(CF)、基于内容过滤(CBF)和基于图(GB)的方法。

CF 方法的主要观点是行为相似的用户对项目有相同的偏好,一般通过计算用户向量和论文向量的匹配获得分数,由于论文推荐固有的数据稀疏等问题通常表现不佳。近年来由于深度学习强大的特征学习能力,一定程度改善了 CF 方法的推荐性能,例如,Ebesu 等[3]提出一种协同记忆网络 CMN,以非线性的方式统一了全局因子模型和基于局部邻域结构的两类 CF 模型,取得了较好的论文推荐效果。然而,CF 仅基于用户和论文的交互数据,丢失大部分显式和隐式的关联信息。

CBF 技术已经较为成熟,其原理是推荐与用户兴趣相似的文章(论文或论文),关键步骤是匹配用户和文章的相关性[4]。例如,ER[1]融合内容特征和非内容的偏置,在基于内容的框架中推荐多类型的学术资源。微软学术推荐系统[5]是一种基于内容和基于图的混合系统,首先用基于内容的方法从海量文献中召回大部分相似论文,接着融合学术图谱的引用关系等推送论文,提升整个系统的推荐覆盖率和用户满意度。在论文推荐领域的 CBF 虽然简单易行,但是仅用文本的语义相似度去衡量用户兴趣使得质量难以保证[2],此外,用户的阅读行为具有很强的目的性,CBF 方法导致用户的阅读视野局限于个人掌握的背景知识范围内[6]。

GB 方法不考虑用户的行为和论文的内容,而是将用户和论文抽象化为图上的节点,在论文推荐领域常用的拓扑图类别有引文网络、社交网络和其他异构信息网络等。例如,Cai 等[7]将推荐的若干相关要素表示在同一个图上,使用图表示学习的方法计算推荐列表,例如将查询人员、查询文本、论文、作者、实体之间的关系构建成异构网络,或使用作者、论文和发表场地组成的书目网络[8]。然而,GB 方法显然浪费了用户个性化的特征和内容特征,而基于引文网络的推荐因为新论文的被引数较少面临冷启动问题。

在表示学习技术帮助下,辅助信息(side information)可以有效缓解上述的论文推荐问题,知识图谱正因为其包含的丰富实体和关系,被认为是一种十分优良的辅助信息。例如 Zhao 等[9]构建概念图谱跨越用户与项目的知识鸿沟,从知识图谱上抽取符合用户认知模式的概念路径帮助研究人员获取目标知识; Frederick 等[10]通过映射专业术语到外部知识图谱 DPpedia,用于扩展查询手稿(摘要)的特征生成排序列表。然而现存的方法却未能考虑知识图谱上用户实体和论文实体的高阶关联关系。

综合以上问题,本文重点关注以下两个方面:(1)在用户历史交互稀疏的前提下,如何推荐给用户相关的论文;(2)在已知有限的领域知识,研究人员如何全面获取自身领域相关的论文。在分析现有研究成果的基础上,本文提出一种混合的推荐模型(GNPR)。首先,为了取得用户更完整的显式阅读偏好,DKN[6]利用多通道 CNN 获取文本特征的启发,模型首先使用 Word2vec 和多通道 CNN 处理文本。由于 CNN 的表示方法重点关注文本的局部特征,句子的全局特征被忽略无法得到全面的用户显式阅读偏好,提出一种双层自注意力特征抽取模式补充用户显式的阅读偏好。其次,在外部知识库的帮助下,从论文文本内容中抽取概念,与用户和论文、论文元数据等构建成知识图谱。最后,为了有效挖掘用户的隐式阅读偏好,鉴于图神经网络可以有效获取高阶关系[11-14],文本使用改进的图卷积网络学习用户和论文之间的关联。本文工作的贡献如下。

(1) 提出一种新的论文推荐模型,混合了基于图的推荐和基于内容的推荐。其中,用户显式阅读偏好由文本局部特征和文本的全局特征组成,提出双层自注意力机制来建模全局性特征。

(2) 用论文的非结构化数据、半结构化元数据和 LOD 数据构建知识图谱。为解决在建模高阶关系的图卷积网络不考虑关系类型的问题,提出以关系类型为权值的邻域聚合方式,以获得用户隐式阅读偏好的部分。

(3) 经过在真实数据集 CiteULike-a 和学术推荐应用日志数据的验证,与传统推荐模型和融合知识图谱的模型相比,本文模型在准确率和点击概率方面有不错的提升。

1 问题描述和任务定义

1.1 问题描述

假设论文推荐系统中包括 N 位用户 $U=\{u_1,u_2,\cdots,u_N\}$ 和 M 篇论文 $P=\{p_1,p_2,\cdots,p_M\}$,根据用户的历史交互,对于用户 $u\in U$ 与论文 $p\in P$ 的交互情况可表示为:

$$y_{up}=\begin{cases}1 & \text{用户 } u \text{ 与论文 } p \text{ 有交互}\\ 0 & \text{其他}\end{cases} \tag{1}$$

用户的交互行为可以是隐式反馈或者显式评分,本文选择更贴近实际场景的隐式反馈。根据式

(1) 的表示可以得到用户-论文的交互矩阵 $\boldsymbol{Y} \in \mathbf{R}^{m \times n}$，$y_{ij} \in \boldsymbol{Y}$ 表示第 $i(i=1,2,\cdots,m)$ 位用户与第 $j(i=1,2,\cdots,n)$ 篇论文的交互结果。关于本文涉及的图数据，首先将 $\boldsymbol{Y}$ 转换为用户-论文二部图 G_1，图上的边代表用户与项目的交互情况，其次在 G_1 的基础上加入更多节点(例如概念、关键词和实例等)以及它们对应的关系，形成知识图谱 $G=<V,E>$，其中: V 表示实体; E 表示关系集合。本文的文本数据 $C=\{T,A,K\}$，其中: T 为论文标题; A 是摘要; K 是关键词。

1.2　任务定义

本文推荐模型的目标是基于图特征和内容特征预测用户 $u \in U$ 对论文候选论文 $p \in P$ 的兴趣，即给定知识图谱 G 和文本 C，预测用户 u 对论文 p 的潜在兴趣。目标函数为 $\tilde{y}_{u,p}=\mathcal{F}(u,p \mid \Theta,G,C)$，其中: $\tilde{y}_{u,p}$ 表示用户 u 对论文 p 的点击概率大小; Θ 是训练时的模型参数。本文中使用的变量符号和函数定义如表 1 所示。

表 1　本文使用的符号

符号	描述	符号	描述
u,U	用户，用户集	$\boldsymbol{u}_1,\boldsymbol{u}_2$	文本局部和全局特征向量
p,P	论文，论文集	$\boldsymbol{u}_3$	G 上用户节点的向量
v,r	实体，关系	$\boldsymbol{u}_{et}$	显式阅读偏好向量
G_1,G	二部图，知识图谱	$\boldsymbol{e}_v,\boldsymbol{e}_{N_v},\boldsymbol{e}_r$	实体、邻域、关系向量
$\tilde{y}_{u,p}$	预测的分数	$\boldsymbol{u}_{imt}$	隐式阅读偏好向量
$s_{p_i^u,p_j}$	p_i^u 与 p_j 的匹配权重	$\boldsymbol{u},\boldsymbol{p}$	用户和论文表示向量
a_{ti}^{self}	第 i 个分词权重	$\boldsymbol{W}_x$	参数($x=1,2,\cdots,L_2+2$)
a_{ti}^{self2}	第 t 个标题的权重	$\boldsymbol{b}_x$	偏置($x=1,2,\cdots,L_2$)
L_1	GCN 层数	$a_x(*)$	激活函数($x=1,2,\cdots,L_2$)
L_2	MLP 层数		

2　推荐方法

2.1　GNPR 框架

GNPR 模型框架结构如图 1 和图 2 所示。用户的阅读偏好向量表示包括两个部分的计算: 显式阅读偏好和隐式阅读偏好。用户显式阅读偏好又由文本的局部特征和文本的全局特征组成，如图 1 所示。隐式阅读偏好包含在 GCN 对知识图谱处理后的用户节点向量中，因此将用户和论文及论文的相关元数据抽象为概念知识图谱上的节点，如图 2 所示。最后，推荐计算是使用多层感知机(Multi-Layer Perceptron，MLP)学习匹配函数并输出相关度得分。

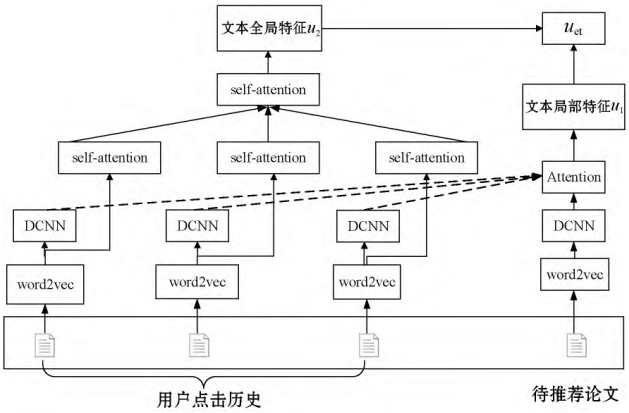

图 1　GNPR 框架之用户显式阅读偏好

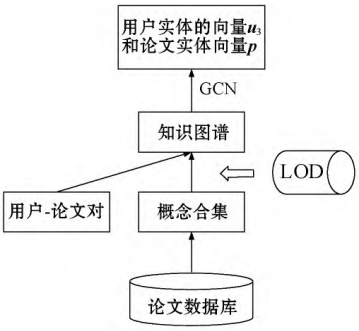

图 2　GNPR 框架之用户隐式阅读偏好

2.2　用户显式阅读偏好

从文本局部特征和文本全局特征两个方面综合用户的显式阅读偏好。基于论文标题(或关键词)和摘要的语义，文本局部特征旨在找出句子中最重要的分词特征，文本全局特征则关注于整个句子的语义特征。

(1) 局部特征抽取器 DCNN。借鉴 DKN[13] 和 GNewsRec[15] 模型利用 CNN[16] 构建文本特征抽取器的观点，本文提出局部文本特征提取器 DCNN，其结构如图 3 所示。针对标题和摘要呈现论文特征的差异性，并分别以标题和摘要输入并行且具有独立参数的 CNN，输出的结果作为论文的局部特征。假设标题和摘要的矩阵表示为 $\boldsymbol{T}=[w_{t1},w_{t2},\cdots,w_{tn}] \in \mathbf{R}^{d \times n}$ 和 $\boldsymbol{A}=[w_{a1},w_{a2},\cdots,w_{am}] \in \mathbf{R}^{d \times m}$，矩阵的列是由 Word2vec 训练的词向量，$d$ 是向量的维数。DCNN 中的 CNN 有一个卷积层和池化层，对任意论文 p，若经过 CNN 处理的标题和摘要输出向量为 $\tilde{\boldsymbol{T}}$ 和 $\tilde{\boldsymbol{A}}$，DCNN 的输出为 $e(p)=[\tilde{\boldsymbol{T}};\tilde{\boldsymbol{A}}] \in \mathbf{R}^D$，$D$ 是输出向量的维数。

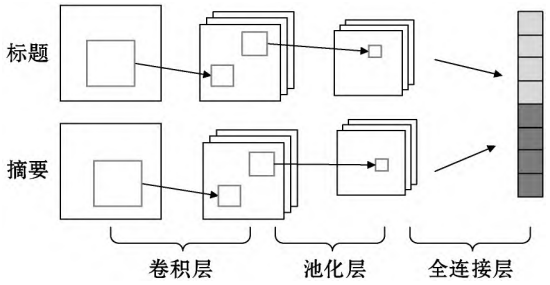

图 3　DCNN 模型的结构

为了判断待推荐论文在多大程度上符合用户历史兴趣，模型基于注意力机制区分历史论文对用户兴趣的贡献。假设用户 u 的点击历史集合为 $\{p_1^u, p_2^u, \cdots, p_k^u\}$，经 DCNN 处理后的点击历史为 $\{e(p_1^u), e(p_2^u), \cdots, e(p_k^u)\}$，待推荐论文为 $e(p_j)$。最后，将两者的拼接向量输入到神经网络中做非线性变换，由 Softmax 函数计算归一化的权重得分：

$$s_{p_i^u, p_j} = \mathrm{softmax}(DNN(e(p_i^u), e(p_j))) = \frac{\exp(DNN(e(p_i^u), e(p_j)))}{\sum_{i=1}^{k} \exp(DNN(e(p_i^u), e(p_j)))} \tag{2}$$

则从用户历史的文本局部特征中得到的显式阅读偏好表示如下：

$$\boldsymbol{u}_1 = \sum_{i=1}^{k} s_{p_i^u, p_j} \times p_i^u \tag{3}$$

(2) 全局特征抽取器。针对文本全局特征，GNPR 模型采用自注意力机制(self-attention)处理由标题、关键词和摘要组成的短文本。自注意力机制有效获取句子的长距离依赖，在考虑全局信息情况下判定词语的重要程度，文本使用度量函数为句子中的每个分词 $f_{\mathrm{self}}(\boldsymbol{w}_i)$ 计算得分，用于表示分词在句子中的重要性，此时不需要任何额外的指引信息(guider)。以计算第 t 个标题的表示为例，第 i 个分词的权重为：

$$a_{ti}^{\mathrm{self}} = \frac{\exp(f_{\mathrm{self}}(\boldsymbol{w}_{ti}))}{\sum_{k=1}^{n} \exp(f_{\mathrm{self}}(\boldsymbol{w}_{tk}))} \tag{4}$$

则考虑全局依赖的句子表示为：

$$S_t^{\mathrm{self}} = \sum_{i=1}^{n} a_{ti}^{\mathrm{self}} \boldsymbol{w}_{ti} \tag{5}$$

为了在用户点击历史集合中找到用户对论文 t 的点击意图受其他论文的影响，模型将全局特征抽取器设计为双层的自注意力层形式，若用户 u 的点击历史数为 h，则从用户历史的文本全局特征中得到的显式阅读偏好表示如下：

$$\boldsymbol{u}_2 = \sum_{t=1}^{h} a_t^{\mathrm{self2}} \times S_t^{\mathrm{self}} \tag{6}$$

式中：a_t^{self2} 是句子的权重。将 2.1 节的局部特征和全局特征拼接并输入到全连接网络中，获得用户的显式阅读偏好 $\boldsymbol{u}_{\mathrm{et}} = \boldsymbol{W}_1 [\boldsymbol{u}_1; \boldsymbol{u}_2]$，其中 $\boldsymbol{W}_1 \in \mathbf{R}^{D \times 2D}$。

2.3 用户隐式阅读偏好

为了建模用户的隐式阅读偏好，本节构造由用户、论文、论文元数据和相关概念组成的知识图谱。针对知识图谱的表示，本文使用改进的图卷积网络学习融合高阶信息的用户节点表示隐含着用户潜在的兴趣。

(1) 概念知识图谱构建。开放互联数据(Linked Open Data，LOD)储存丰富的实体和关系构成的关联知识。本文从论文中获取的概念和其他实体构建成面向推荐的知识图谱 G，目的是关联已知概念和未知的相关概念，并通过高阶关联关系融入到用户和论文表示中，以更好地建模用户隐式的阅读偏好，从而缓解用户-项交互数据稀疏问题。

针对标题、摘要、关键词等数据，本文首先提取 TF-IDF 权重较高的词，然后链接 LOD 中准确的概念以及若干跳邻域，例如与链接实体处于上下位关系的概念、概念的实例等等。关于用户与论文之间的关系，若存在交互，在用户和论文实体之间设置“交互”关系边。以上的概念部分子图融合用户-论文二部图 G_1 组成最终的知识图谱 G，其中包含实体类型：用户、论文、概念和论文的其他元数据等；关系类型：用户与论文的交互关系、论文的引用关系、概念与论文的从属关系、实例与概念的 $\mathrm{is_a}$ 关系、主题与论文从属关系、概念之间的上下位关系。因此，通过知识图谱的组织形式，论文的标题、摘要、关键词中蕴含的语义可通过该概念之间的上下位关联显示出来，而论文其他元数据(如参考文献)则通过论文之间的引用关系保持关联。

(2) 基于改进 GCN 的知识图谱表示。针对知识图谱的表示学习，本文使用输出结果包含了实体间高阶关系的图卷积网络(GCN)。首先，GCN 通过传播嵌入的方法获得用户和论文的分布式表示，其中包含用户潜在的阅读偏好，即上文提及的用户隐式兴趣。然而，GCN 一般处理的方式是将知识图谱当成无向图，忽略对关系类型的区分，因此本文预先考虑用户对所有关系的隐含偏好分布[11,13]。以下是计算单个 GCN 层的某节点 v 嵌入的一般形式：

$$\boldsymbol{h}_{N_v} = f_{\mathrm{agg}_N}(\{\boldsymbol{e}_v, \boldsymbol{e}_{N_v}\}) \tag{7}$$

$$\boldsymbol{h}_v = \sigma(\boldsymbol{W}_2 \cdot \boldsymbol{h}_{N_v} + \boldsymbol{b}_1) \tag{8}$$

式中：$f_{\mathrm{agg}_N}: \mathbf{R}^d \times \mathbf{R}^d \rightarrow \mathbf{R}^d$ 表示邻域聚合函数用于聚合来自邻域的信息，本文使用文献[11]提到的函数 *Concat aggregator*；$\boldsymbol{e}_v$ 和 $\boldsymbol{e}_{N_v}$ 分别是实体 v 和 v 的邻域的向量表示。

本文模型在邻域的计算方式中融入了关系类型，即每一个邻域实体对邻域表示的贡献度取决于用户和关系的匹配值，例如，用户更喜欢通过引用关系查找论文，则图上的嵌入传播方向则受到相应的影响。假设邻域为 $N_v = \{(r, ent) \mid (v, r, ent) \in G, (ent, r, v) \in G\}$，本节定义一个映射 $f_{\mathrm{ur}}: \mathbf{R}^D \times \mathbf{R}^D \rightarrow \mathbf{R}$(例如内积)计算用户 u 和关系 r 的匹配值。因此邻域的向量表示为：

$$\boldsymbol{e}_{N_v} = \sum_{(r, ent) \in N_v} f_{ur}(\boldsymbol{u}, \boldsymbol{e}_r) \times ent \tag{9}$$

经过 L_1 层邻域的聚合后，论文节点的向量表示为

e_p,用户节点的向量表示,即用户的隐式阅读偏好表示为 u_{imt}。论文最终的表示向量为 $p = e_p$,用户的最终的向量表示为 $u = W_3 [u_{\text{et}}; u_{\text{imt}}]$,$W_3 \in \mathbf{R}^{D \times 2D}$。

2.4　用户-论文交互建模

现存的深度学习对推荐模型的侧重点分为两方面: 侧重用户与项目的表示学习,侧重用户与项目的交互建模。与先前的研究不同,通过上文的介绍可知模型 GNPR 已经对用户和论文进行了学习表示,接下来利用训练好的用户向量和论文向量进行推荐预测。在交互建模阶段,本文基于用户-论文交互对,拼接训练好的用户向量与用户交互历史的论文向量,以作为交互建模层的输入 x_1,在 L_2 次非线性变换后得到预测分数。由以上的计算可知,用户最终表示为 u,候选论文向量表示 p,将最终用户与论文向量输入到交互建模层 MLP 中进行计算:

$$\begin{aligned} x_1 &= [u; p], \\ x_2 &= a_1(W_4 \cdot x_1 + b_2) \\ &\cdots \\ x_{L_2} &= a_{L_2-1}(W_{L_2+2} \cdot x_{L_2-1} + b_{L_2}) \\ \tilde{y}_{u,p} &= \sigma(a_{L_2}(h_{L_2}^T x_{L_2})) \end{aligned} \tag{10}$$

式中: W_i、b_i 和 σ 分别表示第 i 层感知器的权重矩阵、偏置向量和激活函数。

为了有效地训练 GNPR 模型,从隐式反馈中为特定用户采样未交互的论文作为负样本,数量和正样本相同。例如,一个训练样本可以表示为,其中 x 是预测是否单击的候选论文。对于每个正样本,$y = 1$,否则 $y = 0$。文本使用交叉熵损失(cross-entropy)作为损失函数:

$$\mathcal{L} = -\{\sum_{x \in \triangle^+} y_{u,p} \log \tilde{y}_{u,p} + \sum_{x \in \triangle^-} (1 - y_{u,p}) \log(1 - \tilde{y}_{u,p})\} + \lambda \| W \|_2 \tag{11}$$

式中: $\triangle^+$ 是正样本集合; $\triangle^-$ 是负样本集; $\lambda \| W \|_2$ 是 L2 正则项。

3　实验设计与分析

本节给出实验设计细节和相应的结果,为了证明本文模型的有效性,本次实验用它与基准模型进行比较。实验将从下面两个研究问题(Research Question, RQ)来分析实验。

RQ1: 在用户-论文的交互记录十分稀疏的前提下,如何有效推荐论文? 即与基准模型比较,本文模型在稀疏数据集的实验效果是否超过 state-of-the-art 的性能?

RQ2: 在已知有限的领域知识,研究人员如何获取更多样的论文? GNPR 模型的组成部分对模型的影响是什么,特别是知识图谱的融入对实验结果是否有提升?

3.1　实验设计

(1) 数据集和预处理。论文推荐数据集使用 CiteULike-a 和学术推荐应用的日志(文中称为 APPData 数据集)。CiteULike 是一个在线论文存储与分享平台,允许用户创建自己感兴趣的论文集合,选择该平台数据的理由在于用户主观创建的论文集很大程度体现用户真实的阅读偏好,而且提供了论文的标题和摘要等元数据。CiteULike-a 是文献[17]从该平台收集并预处理后的隐式反馈数据集; 而 APPData 是部署在学术机构的推荐应用,实验中的数据集是用户与论文交互后产生的点击日志。

数据集预处理: 针对知识图谱构建,首先依次对文本内容进行清洗和概念抽取,最终挑选权重较高的名词性术语; 依次将术语链接到外部知识库 Xlore 得到半结构数据,接着对其清洗和预处理得到三元组,统计如表 2 所示。最后,按照 7∶2∶1 比例将数据集划分成训练集、验证集和测试集,其中验证集用于优化超参数。

表 2　数据集的各项统计

数据集	用户	项目	历史交互	数据稀疏率/%
CiteULike-a	5 551	16 980	204 986	99.78
APPData	5 000	20 000	189 141	99.81

(2) 基准方法。BPRMF: 基于贝叶斯后验优化的个性化排序的矩阵分解,本文使用用户-论文交互矩阵 Y。

NeuMF[18]: 一种 NCF 框架的实例,在用户和项目的嵌入层组合了广义矩阵分解(GMF)和 MLP,本文使用与 BPRMF 相同输入。

CML[19]: 一种度量学习算法,同时编码了用户的偏好以及用户-用户、项-项的相似性,本文使用与 BPRMF 相同输入。

KGAT[13]: 在知识图谱上显式地建模用户和项目的高阶关系,使用注意力的聚合方法。

DKN[6]: 基于内容的深度学习推荐框架,它融合多通道 CNN 对论文的语义层和知识层的表示。在本文中,将内容 C 的特征作为语义层特征,知识图谱 G 的特征作为知识层特征。

(1) 评估指标。准确率(precision)表示推荐列表

预测为真的论文占推荐列表的比例; 召回率(recall) 是覆盖率的评价指标,表示推荐列表中预测为真的论文占所有与论文相关论文数的比例。F1-score 是准确率和召回率的加权平均,其数值越大越准确,计算方式如下:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{12}$$

AUC 为 ROC 曲线下方的面积。ROC 曲线的横坐标是预测结果的假阳性率,纵坐标是真阳性率。

3.2 实验结果(RQ1)

首先介绍与所有基线相比的总体性能,表 3 和表 4 分别是在 CiteULike-a 数据集和 APPData 上与所有模型对比的实验结果,加粗表示最好性能。

表 3 CiteULike 数据集的实验结果

模型	*Prec@ N*		*Rec@ N*		F1-score		*AUC*
	$N=5$	$N=10$	$N=5$	$N=10$	$N=5$	$N=10$	
BPRMF	0.104	0.199	0.113	0.194	0.133	0.196	0.508
NeuMF	0.130	0.204	0.120	0.189	0.125	0.134	0.515
CML	0.145	0.262	0.188	0.127	0.164	0.142	0.534
KGAT	0.230	0.204	0.220	0.289	0.225	0.219	0.556
DKN	0.346	0.362	0.383	0.327	**0.364**	0.342	0.603
GNPR	**0.374**	**0.459**	**0.438**	**0.487**	0.344	**0.432**	**0.703**

表 4 APPData 数据集的实验结果

模型	*Prec@ N*		*Rec@ N*		F1-score		*AUC*
	$N=5$	$N=10$	$N=5$	$N=10$	$N=5$	$N=10$	
BPRMF	0.296	0.190	0.129	0.153	0.180	0.170	0.532
NeuMF	0.300	0.236	0.157	0.246	0.206	0.241	0.557
CML	0.389	0.311	0.245	0.266	0.301	0.287	0.595
KGAT	0.300	0.266	0.287	0.377	0.293	0.312	0.588
DKN	0.451	0.472	0.500	0.426	0.474	0.448	0.606
GNPR	**0.488**	**0.599**	**0.571**	**0.635**	**0.526**	**0.616**	**0.686**

通过结果的对比可以得出: 首先,通过比较本文模型和其他基准模型的结果,本文模型在两个数据集的 F1-score 和 AUC 分别优于大部分基线。与 BPRMF、NeuMF 和 CML 的比较结果说明在数据稀疏(见表 1) 的情况下 GNPR 模型性能没有遭受较大影响,即可以更好地缓解数据稀疏问题,可能的原因是本文模型挖掘了更丰富的内容特征和知识图谱特征。而 DKN 虽然同样有足够的内容特征,但是 GNPR 模型结果较好可能原因在于考虑全局文本特征。KGCN 和 GNPR 都使用了图神经网络,实验差别的原因可能在于融合了内容特征。实验还发现,所有基于内容的模型都比基于 CF 的模型具有更好的性能。原因是基于 CF 的方法在数据稀疏的论文推荐场景性能受影响。本文模型是一个混合模型,结合了基于内容的方法和基于图的方法的优点,对于缺少点击历史的论文,可以通过内容和图上的关联。

3.3 模型分析(RQ2)

以 APPData 数据集为例,对 GNPR 变体的实验结果进行比较,以证明本文的模型设计在以下方面的有效性: (1) 内容特征可以达到的实验效果; (2) 双层注意力机制对用户显式阅读偏好的影响; (3) GCN 的高阶关系与普通表示学习方法对实验的影响; (4) MLP 对建模交互的影响。实验结果如表 5 所示。三个设置的详细信息如下:

(1) 删除图的特征(Remove Graph Future,RGF) : 只保留用 2.2 节描述的用户显式阅读偏好模块对实验结果的影响。

(2) 删除自注意力(Remove Self-Attention,RSA) : 除去 2.2 节描述用户显式阅读偏好模块中文本全局特征对实验结果的影响。

(3) 用 TransE 替换 GCN(With TransE,WTE) : 用 TransE 代替 2.3 节描述用户隐式阅读偏好模块中的 GCN 后对实验的影响。

(4) 用内积替换 MLP(With Inner-Product,WIP) : 内积代替 2.4 节描述的用户-项交互计算对结果的影响。

表 5 GNPR 变体的实验结果

模型	*Pre@* 10	*Rec@* 10	F1-score	*AUC*
RGF	0.299	0.294	0.296	0.575
RSA	0.404	0.439	0.421	0.521
WTE	0.462	0.427	0.424	0.553
WIP	0.504	0.589	0.543	0.688
GNPR	0.599	0.635	0.616	0.703

可以看出: (1) GNPR 表现最好,表明模型的不同成分的有效性; (2) 缺少文本的全局特征对结果影响较大; (3) 知识图谱的嵌入对结果提升较大和 GCN 算法在本文中性能比 TransE [20] 更优。

参数分析: GNPR 模型涉及多个参数的选择,接下来以 APPData 为例,考虑 GCN 的层数 L_1 和 MPL 的层数 L_2 对评估指标 F1-score 和 *AUC* 的影响,结果如表 6 所示。

表 6　APPData 数据集的 GCN 和 MLP 层数变化

模型	层数	*Pre*@10	*Rec*@10	F1-score	*AUC*
GCN	$L_1=1$	0.472	0.549	0.536	0546
	$L_1=2$	**0.599**	**0.635**	**0.616**	**0.686**
	$L_1=3$	0.492	0.619	0.594	0.585
MLP	$L_2=1$	0.504	0.589	0.543	0.588
	$L_2=2$	0.536	0.611	0.578	0.593
	$L_2=3$	**0.599**	**0.635**	**0.616**	**0.686**

4　结　语

本文提出一种混合的端到端的推荐模型 GNPR。首先，自注意力机制考虑了文本的全局特征，融合 CNN 后的双层特征抽取模式可以获取用户更完整的显式阅读偏好。从论文中抽取概念并链接外部知识库，通过概念之间的关联寻找研究人员、论文和概念之间的潜在相关性，此时知识的融入有效缓解了数据稀疏性，图神经网络通过在图上传播嵌入高阶结构信息，可以有效地挖掘出用户的隐式阅读偏好。在真实的论文推荐数据集 CiteULike-a 和学术推荐应用的实验结果表明，本文提出的论文推荐模型在 F1-score 和 AUC 指标上明显优于基线方法。

参　考　文　献

[1] 尹丽玲，刘柏嵩，王洋洋. 跨类型的学术资源优质推荐算法研究[J]. 情报学报，2017，36(7)：715－722.

[2] Bai X，Wang M，Lee I，et al. Scientific paper recommendation：a survey[J]. IEEE Access，2019，7：9324－9339.

[3] Ebesu T，Shen B，Fang Y，et al. Collaborative memory network for recommendation systems[C]//41st International ACM SIGIR Conference on Research & Development in Information Retrieval，2018：515－524.

[4] Li X，Chen Y F，Pettit B，et al. Personalised reranking of paper recommendations using paper content and user behavior[J]. ACM Transactions on Information Systems，2019，37(3)：1－23.

[5] Kanakia A，Shen Z Z，Eide D，et al. A scalable hybrid research paper recommender system for Microsoft Academic[C]//World Wide Web Conference，2019：2893－2899.

[6] Wang H W，Zhang F Z，Xie X，et al. DKN：Deep knowledge-aware network for news recommendation[C]//World Wide Web Conference，2018：1835－1844.

[7] Cai X Y，Han J W，Pan S R，et al. Heterogeneous information network embedding based personalized query-focused astronomy reference paper recommendation[J]. International Journal of Computational Intelligence Systems，2018，11(1)：591－599.

[8] Cai X Y，Zheng Y，Yang L B，et al. Bibliographic network representation based personalized citation recommendation[J]. IEEE Access，2019，7：457－467.

[9] Zhao W D，Wu R，Liu H T，et al. Paper recommendation based on the knowledge gap between a researcher´s background knowledge and research target[J]. Information Processing & Management，2016，52(5)：976－988.

[10] Frederick A G，Bálint D，András B，et al. Global citation recommendation using knowledge graphs[J]. Journal of Intelligent & Fuzzy Systems，2018，34(5)：3089－3100.

[11] Wang H W，Zhao M，Xie X，et al. Knowledge graph convolutional networks for recommender systems[C]//World Wide Web Conference，2019：3307－3113.

[12] Wang H W，Zhang F Z，Wang J L，et al. Exploring high-order user preference on the knowledge graph for recommender systems[J]. ACM Transactions on Information Systems，2019，37(3)：1－26.

[13] Wang X，He X N，Cao Y X，et al. KGAT：Knowledge graph attention network for recommendation[C]//25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining，2019：950－958.

[14] Sha X，Sun Z，Zhang J，et al. Attentive knowledge graph embedding for personalized recommendation[EB]. arXiv：1910.08288，2019.

[15] Hu L M，Li C，Shi C，et al. Graph neural news recommendation with long-term and short-term interest modeling[J]. Information Processing & Management，2020，57(2)：102－142.

[16] Kim Y. Convolutional neural networks for sentence classification[C]//2014 Conference on Empirical Methods in Natural Language Processing，2014：1746－1751.

[17] Wang C，David M B. Collaborative topic modeling for recommending scientific articles[C]//17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，2011：448－456.

[18] He X N，Liao L Z，Zhang H W，et al. Neural collaborative filtering[C]//26th International Conference on World Wide Web，2017：173－182.

[19] Hsieh C K，Yang L，Cui Y，et al. Collaborative metric learning[C]//26th International Conference on World Wide Web，2017：193－201.

[20] Bordes A，Usunier N，Garcia D A，et al. Translating embeddings for modeling multi-relational data[C]//27th Annual Conference on Neural Information Processing Systems，2013：2787－2795.