文章编号：2095-2163(2022)06-0060-06 中图分类号：TP391 文献标志码：A

融合知识图谱和深度学习的学术论文推荐算法

吴舒展

（湖北工程学院 数学与统计学院，湖北 孝感 432000）

摘 要：由于传统学术论文推荐方法存在推荐效果差的问题，导致推荐结果不能满足用户的检索需求，为此融合知识图谱和深度学习算法，实现对学术论文推荐方法的优化设计。收集学术论文数据，并构建相应的知识图谱。分析用户的行为偏好和检索需求。提取学术论文资源特征，利用深度学习算法实现对学术论文的分类处理。通过构建查询向量、度量相似性、生成资源推荐列表三个步骤实现学术论文推荐。通过实验对比得出结论：设计的推荐方法的召回率提高了 4.53%，且推荐结果的命中率得到明显提升。

关键词：知识图谱；深度学习；学术论文；资源推荐

Academic papers recommendation integrating knowledge graph and deep learning

WU Shuzhan

（School of Mathematics and Statistics, Hubei Engineering University, Xiaogan Hubei 432000, China）

【Abstract】 Due to the poor recommendation effect of traditional academic papers recommendation methods, the recommendation results can not meet the retrieval needs of users. In the research, knowledge graph and deep learning algorithm are integrated to realize the optimization design of academic papers recommendation methods. The data of academic papers is collected and the corresponding knowledge map is constructed. Users′ behavior preferences and retrieval requirements are also analyzed. Furtherly, the feature of academic papers resources is extracted, and the classification of academic papers is realized by deep learning algorithm. Therefore, academic papers recommendation is realized by constructing query vector, measuring similarity and generating resource recommendation list. Through experimental comparison, it is concluded that the recall rate of the designed recommendation method is increased by 4.53%, and the hit rate of the recommendation result is obviously improved.

【Key words】 knowledge graph; deep learning; academic papers; resource recommendation

0 引 言

随着科学技术的迅速发展，各个行业领域和学科产生的研究成果也在大幅增长，海量学术成果的涌现在为学者提供丰富学术论文的同时，也对论文的检索工作带来了困难和挑战。在科研人员进行相关科学研究的过程中，需要查询和引用相应的学术论文，并在前人的研究基础上开展进一步研究和优化，从而有效保证研究成果的价值和可行性。然而在实际的检索过程中，用户很难在短时间内精准地获得最具参考价值的学术论文[1]。为了解决学术论文的检索和查询问题，提出了学术论文推荐方法。

学术论文推荐方法的提出与应用，不仅提高了用户检索目标论文的速度，同时也解决了网络环境中信息过载以及信息迷航的问题。现阶段，国内外学术论文推荐方法大体可分为协同过滤方法和内容过滤方法[2]。然而上述传统的论文推荐方法主要针对的是静态的、存储在固态数据库中的学术论文，由于在线学术论文处于动态变化的状态，因此使用传统的论文推荐方法会出现推荐效果差、推荐速度慢等问题。

以解决传统学术论文推荐方法存在的问题为目的，融合知识图谱和深度学习算法，对在线学术论文推荐方法进行优化设计。将知识图谱引入到推荐方法中，可以实现论文实体之间的连接，并以此来表示不同语义论文的扩展潜在因子模型。在以往的研究工作中，基于深度学习的论文推荐方法虽提升了推荐性能，但只考虑了用户对论文的评分数据，削弱了推荐效果。融合知识图谱和深度学习算法并将其应用到学术论文推荐方法的设计工作中，以期在保证推荐性能的同时，提升推荐效果，满足用户的论文查询需求。首先通过构建论文中实体间的三元组关系表达式，构建学术论文知识图谱，再通过知识图谱嵌入式分析知识图谱中的论文的特征，并转化为低维的连续向量，结合用户的兴趣，利用深度学习的循环神经网络进行训练，根据论文的相似度实现学术论文的精准推荐。

作者简介：吴舒展（1988-），男，硕士，讲师，主要研究方向：网络信息组织与检索。

收稿日期：2021-12-24

1　学术论文推荐方法设计

学术论文推荐方法的设计目标是预测用户需求与学术论文之间的匹配程度，根据匹配结果生成用户的推荐列表。在实际的设计与运行过程中，以深度学习算法为基础迭代算法，知识图谱以嵌入式的方式与深度学习算法融合[3]。知识图谱模块构建的三元组表达式为：

$$k = (E, R, S) \tag{1}$$

其中，E、R 和 S 分别表示实体、关系和属性三元组集合。

1.1　构建学术论文知识图谱

按照公式(1)表示的结构，构建学术论文的知识图谱，具体的构建过程如图 1 所示。

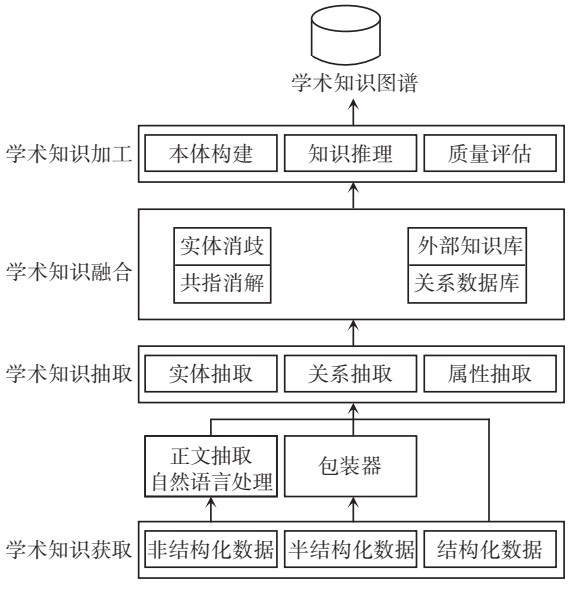

图 1　学术论文知识图谱构建流程图

Fig. 1　Academic papers knowledge graph construction flow chart

从图 1 中可以看出，采用自底向上的方式进行知识图谱的搭建，分别抽取学术论文中的实体知识和关系知识，根据实体之间的关系对其进行连接，并通过知识融合和加工，得出最终的图谱构建结果[4]。论文实体的抽取就是从文本数据集合中识别论文的命名实体，建立知识图谱中的节点。根据特定需求可以将实体分为时间类、数字类和实体类三种类型，选择合适的实体抽取目标并按照词性进行标签编辑，通过分析各个标签之间的搭配关系，实现对实体的抽取，进而创建实体模型。实体模型中，令 $\boldsymbol{M}_r$、C_t 分别为学术论文模型和论文类型，则学术论文模型结构可以表示为：

$$\boldsymbol{M}_r = \begin{matrix} \\ 作者 \\ 论文 \\ 主体 \\ 关键词 \end{matrix} \begin{bmatrix} \boldsymbol{AA} & \boldsymbol{AP} & \boldsymbol{AT} & \boldsymbol{AW} \\ \boldsymbol{PA} & \boldsymbol{PP} & \boldsymbol{PT} & \boldsymbol{PW} \\ \boldsymbol{TA} & \boldsymbol{TP} & \boldsymbol{TT} & \boldsymbol{TW} \\ \boldsymbol{WA} & \boldsymbol{WP} & \boldsymbol{WT} & \boldsymbol{WW} \end{bmatrix} \tag{2}$$

公式(2)中的任意一个子矩阵代表知识图谱中的任意 2 种论文之间的关系，例如 $\boldsymbol{AA}$ 表示学术论文作者与论文之间的从属关系，$\boldsymbol{PP}$ 为论文引用关系和相似关系。关联提取是在一句话中识别出实体对的语义关系和实体对应的属性，两者之间是相互联系的语义纽带[5]。将关系抽取结果代入到公式(2)中，实现对实体的连接。另外，在学术论文知识图谱中，定义关键词的权重为 ω_{ri}，其计算公式为：

$$\omega_{ri} = \frac{tf(r,i) \times \lg\left(\dfrac{Z}{l} + 0.02\right)}{\sqrt{\sum_{i=1}^{r}\left(tf(r,i) \times \lg\left(\dfrac{Z}{l} + 0.02\right)\right)^2}} \tag{3}$$

其中，$tf(r,i)$ 表示第 i 个关键词在论文 r 中出现的频度；Z 表示学术论文总数；l 表示包含关键词 i 的论文数量[6]。

知识合并主要是针对结构化数据的整合，在进行了知识抽取和知识融合后，得到了一系列的事实表达，需要进行知识加工，才能最终形成结构化、网络化的知识系统。知识点中心度参数计算方法如下：

$$I_i = \alpha\delta_i + \gamma\sum_{j\in\pi_\pi^1}\delta_j + \gamma^2\sum_{j\in\pi_\pi^2}\delta_j + \cdots + \gamma^m\sum_{j\in\pi_\pi^m}\delta_j \tag{4}$$

其中，δ 表示知识点的贡献量；α 和 γ 分别表示知识点自身及其各阶邻居知识点的贡献程度；$\sum_{j\in\pi_\pi^m}\delta_j$ 表示知识点 i 的 m 阶邻居知识点及对知识点 i 中心度的总贡献量。通过构建学术论文知识图谱，深度挖掘论文的特征，并突出了论文之间的关联性。

1.2　分析用户需求和兴趣

用户需求的分析可以通过用户输入的检索或查询词条直接读出，根据用户的基本信息和输入的检索词在学术论文中进行匹配[7]。而用户兴趣是在用户使用学术论文平台一段时间后，通过对用户的历史行为数据进行分析，得到用户兴趣。用户兴趣由主题偏好、学科偏好和关键词偏好三个部分组成，其中用户 u_i 对某个主题 t_k 的兴趣值可以表示为：

$$P_{u_it_k} = \sum_{j=1}^{J} A_{ij}T_{jk} \tag{5}$$

其中，A_{ij} 表示的是在知识图谱环境下，用户对论文产生操作行为对应边的权值，而 T_{jk} 为论文属于主题 t_k 设定阈值的权值[8]。同理可以得出用户对

关键词和学科兴趣的量化分析结果。

1.3 提取学术论文特征

为了提升用户检索词条与学术论文匹配任务的处理速度,提取学术论文的特征,并以特征向量的形式输出。这里,词频特征也就是某一个给定的词语在学术论文中出现的次数,其表达式为:

$$TF_c = \frac{T_c}{T} \tag{6}$$

其中,T 和 T_c 分别为学术论文中的总词数和单词 c 在学术论文中出现的次数。由于学术论文数据量较多,因此在词频特征提取过程中可能会出现提取偏差,为此引入了逆文档词频的概念,在逆文档词频特征的提取过程中,认为一个单词在一篇学术论文中出现的频率越高,则该词在所有论文中出现的频率越低,表明该单词在指定学术论文中的主题突出性[9]。融合词频和逆文本频率指数,可以反映出整个资源库中单词特征的大众化程度,从而过滤出论文中的关键词特征。除了关键词外,学术论文的权威度、引用量、时新度、论文质量等也能够在一定程度上反映论文特征,其特征向量表达式为:

$$\begin{cases} Authority = \frac{1}{2}Level + \frac{1}{2}Cite \\ Cite = \frac{Cites}{\max Cite} \\ Recentness = \frac{12 \times (Y - \min Y) + (M - \min M)}{12 \times (\max Y - \min Y) + (\max M - \min M)} \\ Quality = \frac{1}{3}Authority + \frac{1}{3}Popularity + \frac{1}{3}Recentness \end{cases} \tag{7}$$

其中,*Level* 和 *Cite* 分别为学术论文的发刊级别和被引量的量化结果;*Cites* 和 max *Cite* 对应的是论文被引量和论文来源数据库中最大的被引量;*Recentness* 为论文发表时间距离最早发表时间和最晚发表时间的月份数的比值;Y 和 M 表示年份和月份[10]。另外,变量 *Popularity* 表示的是学术论文的热度。使用相同的方式对特征向量进行提取与融合,最终得出学术论文的综合特征提取结果。

1.4 利用深度学习算法划分学术论文类型

利用深度学习算法中的循环神经网络,实现学术论文的分类处理。循环神经网络的学习迭代原理如图 2 所示。

在实际的论文分类处理过程中,将提取的特征向量作为输入项在 t 时刻输入到循环神经网络中,经过隐藏层处理后输出为 $\boldsymbol{s}_t$,在输出层输出 $\boldsymbol{o}_t$[11]。那么隐藏层和输出层的学习处理函数如下:

$$\begin{cases} \boldsymbol{s}_t = f(\boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{W}\boldsymbol{s}_{t-1} + b) \\ \boldsymbol{o}_t = g(\boldsymbol{V}\boldsymbol{s}_t + c) \end{cases} \tag{8}$$

其中,$\boldsymbol{x}_t$ 为循环神经网络的输入项;f 和 g 为激活函数;b 和 c 是隐藏层和输出层的偏置量,取值为常数;$\boldsymbol{U}$、$\boldsymbol{W}$ 和 $\boldsymbol{V}$ 为神经网络不同层级之间的权重矩阵。

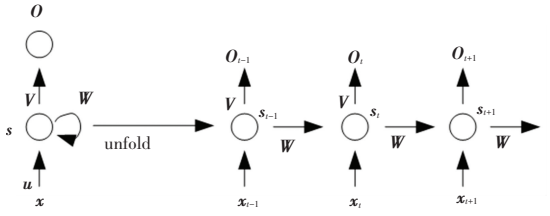

图 2 循环神经网络学习原理图

Fig. 2 Schematic diagram of recurrent neural network learning

1.5 实现学术论文推荐

1.5.1 构建查询向量

由于用户的查询检索需求不同,因此通过知识图谱构建并结合深度学习训练而生成的查询向量也存在差别,用户输入的查询词条类型包括:学术论文作者、名称、主题和关键词。在知识图谱嵌入层,构建的查询元素由上述 4 个部分信息共同组成,并转化为向量表达,其表达式为:

$$\boldsymbol{q} = [q_a, q_p, q_t, q_w] \tag{9}$$

查询矢量 $\boldsymbol{q}$ 是由不具有完整语义信息的不同单词组成的,在实际的查询过程中只要求一个向量值不为空即可,将构建的查询向量作为学术论文推荐的输入词条,输入到推荐运行程序中[12]。

1.5.2 度量论文的相似性

提取的知识图谱中论文特征向量用 $\boldsymbol{\mu}$ 来表示,在论文类型划分环境下,从 2 个方面进行论文相似性度量,一个是知识图谱中查询向量与学术论文的相似性,另一个则是知识图谱中用户兴趣与学术论文的相似性。则相似度的度量结果为:

$$Sim(\boldsymbol{\mu}, \boldsymbol{\theta}) = \frac{\sum_{i=1}^{n} \mu_i \cdot \theta_i}{\sqrt{\sum_{i=1}^{n} \mu_i^2} \cdot \sqrt{\sum_{i=1}^{n} \theta_i^2}} \tag{10}$$

其中,$\boldsymbol{\theta}$ 为输入的知识图谱中查询向量或用户兴趣分析向量。

1.5.3 生成学术论文推荐列表

生成的学术论文推荐列表中,约束前 20 个推荐论文必须与输入的知识图谱中的查询向量有关,且相似度不得低于 70%。按照相似性度量结果由大

到小的顺序进行论文排列，得出学术论文的最终推荐结果。

2　推荐效果测试实验分析

2.1　搭建实验环境

实验采用 FloyHub 作为训练和推荐效果测试平台，测试环境中包含一台服务器和多台计算机设备，实验环境配置见表 1。

研究指出，由于设计的学术论文推荐方法应用了知识图谱和深度学习算法，因此需要在实验环境的基础上嵌入相应的运行程序插件，保证 2 种技术的协同运行。

2.2　准备学术论文数据样本

实验所采用的论文数据样本可由多所高等院校图书馆提供，而且还可以利用网络爬虫，在多个学术与教学网络中获取学术论文、学术会议等类型的论文样本数据。本文实验所用的学术论文数据样本是由本地 2 所高校图书馆提供，准备的论文数据样本包含中文、英文等多种语言，通过解析与统一化操作后，得出实验数据样本见表 2。

表 1　实验环境参数配置表

Tab. 1　Experimental environment parameters configuration table

设备名称	硬件配置	软件程序	操作系统
计算机	CPU：Intel 酷睿 m3-7Y30，1.00 GHz 内存：4 GB	Python 3.5.2 Anaconda 5.0.0	Windows XP 64 位
服务器	CPU：Intel 酷睿 i7-6700HQ，260 GHz 内存：16 GB GPU：NVIDIA GeForce GTX 960 M 显存：2 GB	Python 3.5.2 Anaconda 5.0.0 数据库：MySQL Community　Server 5.7.25	Ubuntu　16.04 LTS

表 2　学术论文数据样本

Tab. 2　Academic papers data samples

资源类型	中英文样本数量	实验选择样本数量	涉及作者数量	被引论文数量	引用关系	时间跨度	更新模式
图书	7 004 528	325 244	574 562	11 439	94 973	2015～2020	每周更新 1 次
专利	621 792	364 372	468 754	15 626	15 957	2017～2020	每周更新 2 次
会议论文	43 660	12 566	25 497	9 765	8 254	2014～2020	每天更新 2 次
学术论文	7 846 985	354 056	435 762	213 765	162 381	2009～2021	实时更新
学术新闻	110 520	27 480	357 568	152 478	785 624	2015～2021	实时更新

另外，根据高校图书馆的学术论文的历史评论记录和查询行为等条目，在实验环境中导入 100 458 条评论记录和行为记录。将准备的所有论文数据样本上传到实验环境中，上传界面如图 3 所示。

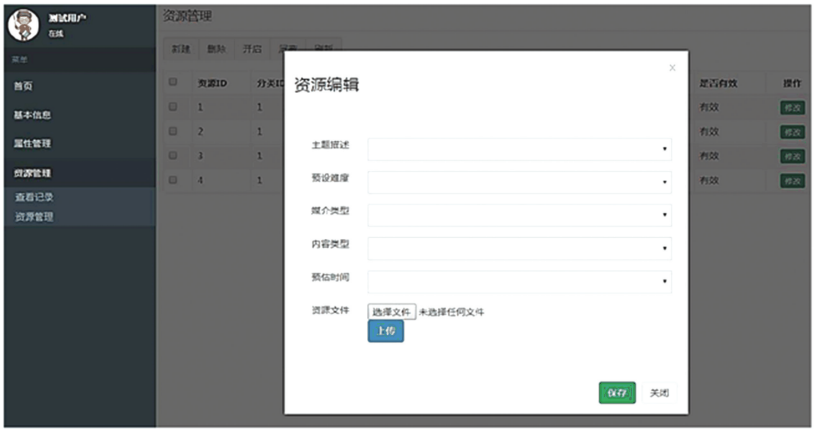

图 3　论文数据样本上传界面

Fig. 3　Thesis data samples upload interface

2.3 设置推荐效果评价指标

实验设置命中率和召回率作为实验的评价指标,命中率越高的推荐列表,证明推荐方法的推荐效果更好。召回率为被引用的论文在前 N 个推荐论文中占比。计算方式分别为:

$$\begin{cases} HR@k = \dfrac{NumberofHits@k}{|GT|} \\ Recall@k = \dfrac{R_p \cap R_q}{R_q} \end{cases} \tag{11}$$

其中, $NumberofHits$ 和 $|GT|$ 分别为用户在推荐列表中的点击次数和测试集合;参数 R_p 表示实际被引用的论文集合; R_q 表示推荐学术论文集合。

2.4 描述推荐效果测试过程

为了形成实验对比,分别设置传统的推荐方法和文献[9]推荐方法作为实验的 2 个对比方法,并将所有的推荐方法以程序代码的形式导入到实验环境中。按照用户的需求输入目标检索词,为了保证实验结果的可信度,输入的多个检索词形成实验的多个组别,并通过计算评价指标的平均值得出最终推荐效果的仿真测试结果。研究中,论文设计推荐方法的输出推荐结果如图 4 所示。

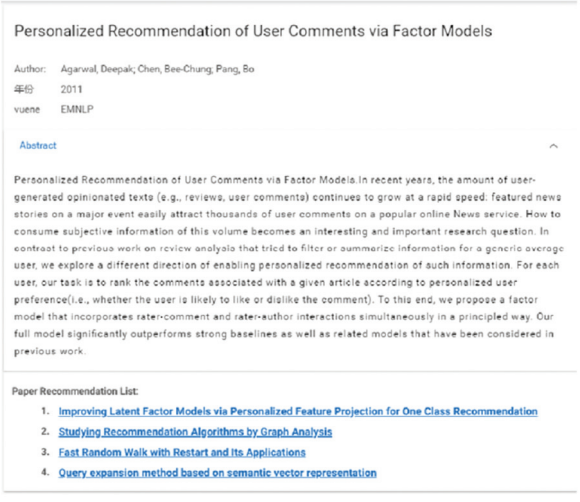

图 4 学术论文推荐页面

Fig. 4 Academic papers recommendation page

2.5 测试实验结果对比分析

利用相关数据的记录与统计,运算得出推荐召回率的量化测试结果见表 3。

通过对表 3 中数据的处理,进一步得出 3 种推荐方法的平均召回率分别为 91.57%、93.18% 和 96.10%。由此可见,设计方法的召回率更高,即实际引用结果在推荐结果中的占比较高。同时,还给出了推荐结果命中率指标测试结果,如图 5 所示。

表 3 学术论文推荐召回率测试结果

Tab. 3 Academic papers recommendation recall test results

方法	测试指标	用户输入检索词				
		机械制造	网络调度	图像处理	通信传输	设备控制
传统	R_p	1 740	1 854	2 903	2 815	2 274
推荐	$R_p \cap R_q$	1 146	1 323	2 164	2 271	1 985
方法	R_q	1 359	1 404	2 518	2 405	2 007
文献[9]	R_p	1 740	1 854	2 903	2 815	2 274
提出推	$R_p \cap R_q$	1 201	1 377	2 384	2 391	2 003
荐方法	R_q	1 382	1 459	2 577	2 515	2 064
设计	R_p	1 740	1 854	2 903	2 815	2 274
推荐	$R_p \cap R_q$	1 379	1 462	2 573	2 480	2 015
方法	R_q	1 403	1 528	2 647	2 652	2 103

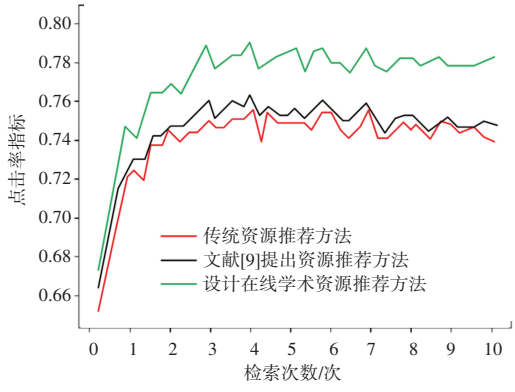

图 5 推荐结果命中率对比曲线

Fig. 5 Recommendation results hit rate comparison curve

从图 5 中可以直观地看出,应用设计方法得出推荐结果的命中率更高,即用户的满意度较高。

3 结束语

目前学术界对基于关键词的学术论文推荐的研究,多是从词义层面上进行优化,并没有考虑到不同文章中不同词义类型的差异。通过知识图谱和深度学习算法的应用,直接提升学术论文的推荐效果,并在一定程度上间接地满足用户对学术论文的需求,有助于提高科研人员的科研效率,拓宽科研视野,把握相关研究的新趋势。

参考文献

[1] 康雁, 李涛, 李浩, 等. 融合知识图谱与协同过滤的推荐模型[J]. 计算机工程, 2020, 46(12):79-85,93.

[2] 熊回香, 景紫薇, 杨梦婷. 在线学术资源中知识图谱的应用研究综述[J]. 情报资料工作, 2020, 41(03):61-68.

(下转第 71 页)