

基于多特征和深度学习的论文推荐算法研究

摘 要

学术论文推荐系统能够从海量的数字化出版物中过滤不符合用户需求的出版物，并提供其可能感兴趣的信息。随着深度学习技术的普及，利用神经网络的强大表达能力，能够更好地挖掘出论文的潜在特征，并将其转化为有效的推荐结果。

学术论文蕴含着多种属性特征。传统的论文推荐方法使用基于文本特征的推荐算法，利用文本匹配的方式进行特征学习。这种忽略了论文的文本结构信息和其他特征信息，推荐结果的相关性并不好。而根据引用关系构建学术论文网络，使用链路预测完成推荐的方法，缺乏对论文内容中文本信息的挖掘。针对以上问题，本文从多特征融合的角度，结合混合神经网络开展研究，综合考虑学术论文本身的文本特征和结构特征，以及引文网络中不同类型节点的信息和不同关系的信息。本文的主要贡献如下：

(1) 针对传统推荐方法不能充分挖掘论文内容信息问题，基于特征融合的思想，提出了一种基于混合神经网络和特征融合的论文推荐方法。构建了一种混合神经网络推荐框架，从论文的标题、摘要、关键字以及上下文结构等文本信息出发，提取论文不同特征后融合作为每篇论文的表达。使用卷积神经网络和注意力机制提取标题和关键词特征，使用双向长短期记忆神经网络提取摘要的上下文语义特征，通过融合三种不同视图的文本特征，丰富论文的表达。通过相似度的计算，最终完成论文推荐任务。在论文数据集 KIWF 和 DBLP 上，通过对比实验和消融实验对该模型进行了验证。相比基线模型，本文所提出模型的 *Recall* 和 *F1* 值分别提升了 52.4%和 47.7%，提高了论文推荐的准确性。

(2) 针对学术论文网络缺少论文属性特征以及无法充分捕捉论文节点的结构信息的问题，本文提出了一种基于多特征和图卷积网络的上下文感知论文推荐方法。使用混合神经网络框架作为论文内容的特征编码器，利用图卷积神经网络提取论文的图

特征，通过融合文本特征和图特征得到完整的论文表示，提升了论文推荐服务质量。在 AAN 数据集和 FullTextPeerRead 数据集上，对模型进行了对比实验和消融实验。从实验结果来看，本文所提出的模型相比于基线模型，*MAP* 和 *MRR* 指标都有明显提升，*Recall@10* 指标最高提升了 77.6%，表明该模型对于学术论文推荐有着较好的效果。

(3) 根据所提出的论文推荐方法，设计并实现了一个论文推荐系统原型。通过对系统需求和系统功能的分析，完成了系统的整体架构和数据库表的设计，最后实现了系统的功能界面。

关键词：推荐系统；混合神经网络；注意力机制；图神经网络；变分图自编码器

RESEARCH ON PAPER RECOMMENDATION ALGORITHM BASED ON MULTI-FEATURE AND DEEP LEARNING

ABSTRACT

Academic thesis recommendation systems are able to filter out publications that do not meet the needs from the vast number of digital publications and provide information that may be of interest. With the popularity of deep learning techniques, the powerful expressive power of neural networks can be used to better mine the potential features of papers and turn them into effective recommendation results.

Academic papers contain a variety of attribute features. Traditional paper recommendation methods use text feature-based recommendation algorithms that utilize text matching for feature learning. This ignores the text structure information and other feature information of the paper, and the relevance of the recommendation results is not good. And the method of constructing academic paper network based on citation relationship and using link prediction to complete recommendation lacks the mining of textual information in the paper content. To address the above problems, this paper carries out research from the perspective of multi-feature fusion, combined with hybrid neural networks, taking into account the textual and structural features of academic papers themselves, as well as the information of different types of nodes and different relationships in the citation network. The main contributions of this thesis are as follows:

(1) Aiming at the problem that traditional recommendation methods cannot fully exploit the information of paper contents, a paper recommendation method based on hybrid neural network and feature fusion is proposed based on the idea of feature fusion. A hybrid neural network recommendation framework is constructed to extract different features of papers from textual information such as titles, abstracts, keywords and contextual structures of papers and then fuse them as a representation of each paper. The title and keyword features are

extracted using convolutional neural networks and attention mechanisms, and the contextual semantic features of the abstract are extracted using a bidirectional long- and short-term memory neural network to enrich the representation of papers by fusing text features from three different views. The final task of paper recommendation is accomplished through the calculation of similarity. The model is validated by comparison experiments and ablation experiments on the dissertation datasets KJWF and DBLP. Compare with the baseline model, the Recall and F1 values of the proposed model in this thesis are improved by 52.4% and 47.7%, respectively, which improves the accuracy of paper recommendation.

(2) To address the problems that academic paper networks lack paper attribute features and cannot adequately capture the structural information of paper nodes, this thesis proposes Paper Recommendation Method Based on Multi-feature and Graph Convolutional Networks. Using a hybrid neural network framework as the encoder of paper content, graph features of papers are extracted using graph convolutional neural networks, and a complete paper representation is obtained by fusing text features and graph features to improve the quality of paper recommendation services. Comparative experiments and ablation experiments of the model are conducted on the AAN dataset and FullTextPeerRead dataset. From the experimental results, the proposed model in this thesis has significantly improved MAP and MRR compared with the baseline model, and the highest improvement of Recall is 77.6%, indicating that the model has better effect for academic paper recommendation.

(3) Based on the proposed paper recommendation method, a paper recommendation system prototype is designed and implemented. Through the analysis of system requirements and system functions, the overall architecture of the system and the design of database tables were completed, and finally the functional interface of the system was realized.

KEY WORDS: Paper Recommendation; Hybrid Neural Network; Attention Mechanism; Graph Neural Network; Variational Graph Auto-Encoders

目 录

摘 要	I
ABSTRACT	III
目 录	V
第 1 章 绪论	1
1.1 研究背景及意义	1
1.2 国内外研究现状	2
1.2.1 传统推荐算法	2
1.2.2 基于深度学习的推荐算法	5
1.2.3 基于图神经网络的推荐算法	6
1.3 本文研究内容	7
1.4 本文组织结构	8
第 2 章 相关理论和技术	10
2.1 推荐系统	10
2.1.1 基于内容的推荐算法	10
2.1.2 基于协同过滤的推荐算法	11
2.1.3 混合推荐算法	14
2.1.4 相似度计算	15
2.2 深度学习相关理论技术	16
2.2.1 卷积神经网络	16
2.2.3 循环神经网络	17
2.2.4 注意力机制	20
2.3 图神经网络相关理论	21
2.3.1 图卷积神经网络	21
2.3.2 图自编码器	22
2.4 本章小结	24
第 3 章 基于混合神经网络和特征融合的论文推荐模型	25
3.1 词向量模型	25
3.2 混合神经网络和特征融合的推荐模型	27
3.2.1 问题定义	28

3.2.2 数据输入	28
3.3.3 混合神经网络模型	29
3.3.4 输出层	32
3.4 实验与分析	33
3.4.1 数据集	33
3.4.2 数据预处理	33
3.4.3 基线模型	34
3.4.4 实验环境和评价指标	35
3.4.5 实验结果及分析	36
3.5 本章小结	39
第 4 章 基于多特征和图卷积网络的论文推荐模型	40
4.1 文献关系网络	40
4.2 多特征和图卷积网络的推荐模型	41
4.2.1 图卷积网络(GCN)层	42
4.2.2 变分图自动编码器(VGAE)	43
4.2.3 特征融合及推荐模块	43
4.3 实验及分析	44
4.3.1 数据收集和预处理	44
4.3.2 实验设计	45
4.3.3 实验环境和评价指标	46
4.3.4 实验结果和分析	46
4.4 本章小结	50
第 5 章 推荐系统原型的设计与实现	52
5.1 系统需求分析	52
5.2 系统架构设计	53
5.3 数据库设计	54
5.4 系统功能与展示	56
5.4 本章小结	59
第 6 章 总结与展望	60
6.1 总结	60
6.2 展望	61
参考文献	62
致 谢	69
攻读学位期间取得的研究成果	70

浙江师范大学学位论文诚信承诺书	71
浙江师范大学学位论文独创性声明	73
学位论文使用授权声明	73

第 1 章 绪论

1.1 研究背景及意义

中国互联网络信息中心(CNNIC)于 3 月 2 日在京发布第 51 次《中国互联网络发展状况统计报告》显示,截至 2022 年 12 月,我国网民规模达 10.67 亿,互联网普及率达 75.6%^[1]。随着互联网、大数据等信息技术的迅速发展,学术论文作为科研人员之间最重要的交流媒介,其以前所未有的速度被发表、传播在网络上,根据截至 2012 年出版年的数据,全球所有学科的科学出版物产量正以每年约 3%的速度增长^[2],2018 年 STM(International Association of Scientific, Technical and Medical Publishers)报告最近显示,论文数量加速增长至每年 4%,期刊增长超过 5%^[3]。在计算机科学开放数据库系统 DBLP 中,2022 年收录的各类出版物总量已经超过 600 万篇,数量是 10 年前的 3 倍多,相比较 2021 年增加了 50 万篇,如图 1.1 所示。与此同时,研究人员数量和研究方向也在不断增加。过多的信息带来了严重的“信息过载”问题。学术成果数量的不断膨胀,使得论文的检索工作变得更加困难和繁琐。研究人员需要花费大量时间和精力在论文的检索上,来确保他们能够获得准确的文献信息。在此过程中繁杂的筛选工作,导致了科研效率的低下,不利于科研产出,因此如何有效地收集并阅读相关的文献越来越受到学界的重视。

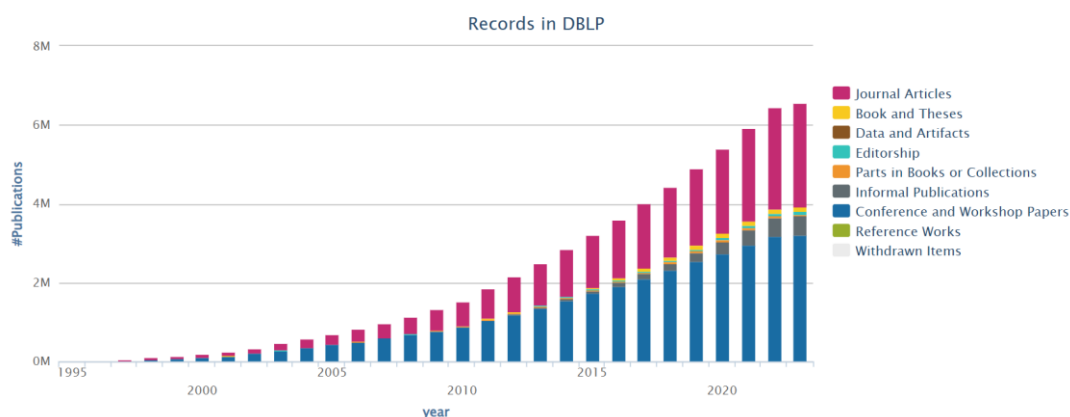


图 1.1 DBLP 不同出版物类型的出版物总数变化趋势¹

¹ <https://dblp.org/statistics/recordsindbpl.html>

推荐系统(Recommendation System)通过个性化的方式向用户提供感兴趣的内容,能够有效地过滤掉无用的信息,有效解决了“信息过载”等问题。论文推荐系统旨在帮助研究人员的减轻学术论文的“信息过载”与“信息迷航”等问题,准确挖掘研究者的阅读兴趣,进而为研究人员提供精准的论文推荐,为了更好地推荐研究文献,越来越多的研究者投入到论文推荐系统的开发中,以更有效的方式搜索和检索资料,提高科研的效率。

传统的推荐系统依赖于用户的历史行为数据,通过对用户的历史行为数据进行建模,挖掘用户的行为偏好,并将其与相关物品进行匹配,从而向用户推荐感兴趣的物品。学术论文推荐一定程度上借鉴了经典商品推荐的思想,包括基于协同过滤的推荐、基于内容的推荐等等。学术论文本身具有大量的文本信息,其中蕴含着论文的各种特征,因此基于文本内容的推荐方法非常适合论文推荐。随着深度学习技术的发展,得益于神经网络模型强大的表示能力,借助深度学习对论文的文本特征进行深层语义挖掘也愈发流行。学术论文包含多种属性特征,除了基于内容的推荐方法之外,根据论文之间的引用关系构建学术论文网络,利用图神经网络技术挖掘论文特征也是论文推荐领域比较常用的推荐方法。

目前,在学术论文推荐的相关研究中,还存在着对论文的文本特征挖掘不充分,以及缺乏对论文其他属性特征的利用问题。本文以此为基础,对学术论文的多属性特征和用户偏好进行深入研究,采用混合神经网络、注意力机制对论文的文本信息进行深入挖掘,学习上下文深层语义,同时结合学术论文网络,丰富论文的特征表示,缓解论文数据的属性单一问题,为科研人员提供准确的论文推荐,提高科研效率。

1.2 国内外研究现状

1.2.1 传统推荐算法

推荐系统是一个横跨数据挖掘、推理算法、机器学习和自然语言处理等多种研究方向的新研究领域。推荐系统基于对人类行为的模拟而设计,其核心在于通过推荐算法对特定数据信息进行深入分析和处理,最终将处理结果向有相关需求的用户推荐^[4]。近年来,不同的学者从不同的角度为推荐算法进行分类,以便初学者更好的入门推荐系统,根据推荐算法所用数据的不同,传统的推荐算法分为三类^[5]:基于内容的推荐(Content-Based Recommendation, CB)^[6]、基于协同过滤的推荐(Collaborative Filtering-

Based Recommendation, CF)^[7]和混合推荐^[8], 如图 1.2 所示:

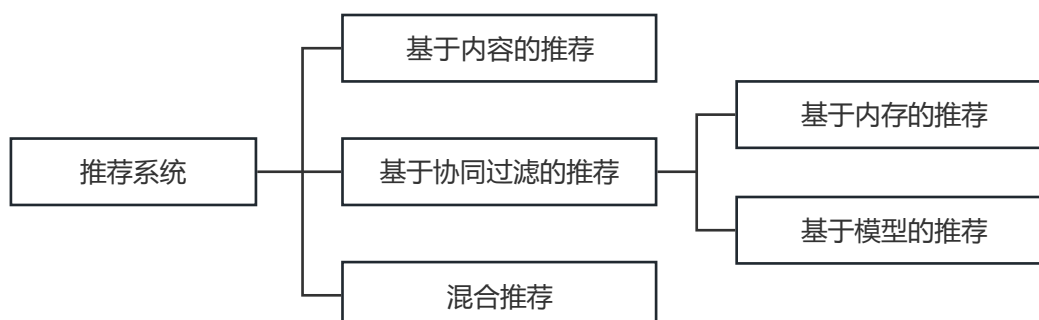


图 1.2 传统推荐算法分类

基于内容的推荐算法(CB)是根据用户感兴趣的物品的特征,推荐与该项目雷似的物品。例如在电影推荐中,若用户喜欢《血战钢锯岭》,那么 CB 方法可能会推荐给用户《敦刻尔克》。其算法名称源于其基于内容的推荐特性。基于内容的推荐算法不依赖于用户的评价意见,而是基于项目的内容信息进行推荐。使用机器学习方法对项目的内容特征进行建模,提取用户的兴趣偏好。在论文推荐系统中,项目指的是数字数据库中的论文,而用户则是研究人员。

在 CB 方法中,首先收集研究者的论文。引用研究人员的论文或其他一些信息可以用来建立他的用户画像。研究者的偏好和兴趣可以通过从研究者的研究领域中提取关键词来表示。论文推荐系统可以从论文的标题、摘要和内容中提取关键词来表示这些论文。这些候选论文可以从数字图书馆检索到。然后,论文推荐系统计算研究人员简介和候选论文之间关键词的相似度,并对其进行排名,将相似度较高的候选论文推荐给研究人员。

Amami 等人^[9]对论文的摘要内容使用 LDA(Latent Dirichlet Allocation)主题建模,计算论文摘要的核心主题与研究者的主题表示之间的相似度,从而进行论文推荐。刘君良等人^[10]总结了基于内容的过滤推荐算法中常用的几种相似度计算方法,并描述了词向量模型 TF-IDF 和 n-gram 的主要内容。樊加倍等人^[11]提出一种基于多层注意力匹配的引文推荐方法。使用不同的注意力机制关注论文的不同层次结构,挖掘学术论文的深层语义特征,取得了较好的论文推荐效果。

基于协同过滤的推荐算法是基于用户对物品的历史行为数据,利用相似用户或相似物品的评分来预测用户对未评分物品的评分,从而为用户进行推荐。协同过滤推荐算法根据数据维度可以分为基于用户的协同过滤和基于物品的协同过滤。相比于 CB,

CF 具有一些不同的优势。比如可以不考虑推荐论文的内容,因为推荐方法取决于用户做出的评级,而不考虑他们属于哪种项目。此外,推荐给用户的项目可能与用户的当前研究不相关,因为相似性是在用户之间的关系之间测量的。

Sugiyama 等人^[12]使用协同过滤的思想考虑论文之间的引用关系,将引用关系看作一种评分,采用基于物品的协同过滤方法进行论文推荐。孔功胜^[13]为了缓解推荐系统中的冷启动和可解释性问题,使用主题模型和矩阵分解方法获得论文的显式和隐式主题特征,使用随机梯度下降法得到用户对论文的预测值。Liu 等人^[14]提出了一种基于引文上下文的协同过滤方法。利用引用上下文,采用关联挖掘技术确定共出现引用论文。每一篇引用论文都由其他引用论文表示。然后,对论文的进行两两比较,计算引用论文之间的相似性。最后,利用与目标论文相似的被引论文的被引量预测目标论文的引用概率。

早期的论文推荐工作主要使用基于内容和基于协同过滤的方法。但是,随着研究的发展,研究人员发现基于协同过滤的模型受限于数据稀疏和冷启动问题,而基于内容的模型主要关注于论文之间的内容相关性,不能有效利用论文的语义知识,可能面临内容过少等问题^[15]。

混合推荐算法(Hybrid Recommendation)。为了提高推荐结果的准确性,获得更好的性能,一些科学论文推荐系统结合了两种或多种推荐技术,将更符合需求的论文推荐给研究人员^[16]。混合推荐的明显优势是:混合推荐系统可以组合使用不同的推荐技术和来自不同来源数据的信息。

吴磊等人^[17]将科研人员的标签信息作为科研人员的偏好,增加数据集中的负样本的比例,使用矩阵分解技术将“标签-论文”矩阵特征融入“科研人员-论文”,以缓解数据的稀疏性问题。Wang 等人^[18]提出了一个使用科学社交网络的社交信息的混合文章推荐系统。将改进的基于内容过滤的推荐模型和改进的协同过滤推荐模型分别结合社会标签和社交信息进行预测,然后将这两个模型的结果通过一定的权重结合成参数可调的混合模型,保留了 CB 和 CF 优越性的同时最大限度地减少了这两种方法各自的局限性,进一步提高科学论文推荐的性能。

传统的论文推荐方法主要还是集中在科研人员与论文之间交互信息的挖掘上,忽略了论文本身具备的丰富属性特征。因此本文从论文内容出发,主要使用深度学习技术充分提取论文的属性特征,完成论文推荐任务。下一节介绍深度学习的推荐方法。

1.2.2 基于深度学习的推荐算法

深度学习是机器学习领域的发展和延伸。深度学习算法具备自主学习的能力,能够处理复杂的问题,从多个维度分析大规模高维数据中的复杂结构,类似于人类的学习方式。近年来,在机器翻译、自然语言处理、图像识别等领域,基于深度学习的表征学习方式已经得到广泛使用^[19]。在推荐系统领域,研究人员渐渐应用深度学习技术对数据进行建模。因此,如何有效结合深度学习技术和推荐技术,并深入研究,成为了一个新的研究方向。深度学习技术通过将用户与物品之间的交互信息转化成低维度的向量,从而发现用户与物品之间的潜在特征。除此以外,深度学习技术能够解决传统推荐方法中的一些问题,能够有效地提升推荐系统的准确率^[20]。

近年来,我们见证了神经推荐模型发展的重大进展,由于神经网络强大的表示能力,这些模型概括并超越了传统推荐模型^[21]。如 Gundogan 等人^[22]提出了一种基于深度学习的混合论文方法,使用文档相似性、层次聚类 and 关键词提取三种不同方法的组合,将不同领域的论文根据其主题进行分组,并根据输入的查询将具有高语义相似性的论文呈现给用户。Kong 等人^[23]为了解决计算论文相似度需要手动设计特征的缺陷,基于文本和网络表示设计了新的论文的特征提取方式,对引文网络中论文的向量表示学习。使用 Doc2vec 方法表示文本信息, Struct2vec 方法表示引文网络结构,利用引文网络统一文本信息和结构信息,将通过网络结构学习到的图嵌入作为论文的代表。Tang 等人^[24]将基于主题的引用推荐形式化,其中使用两层受限玻尔兹曼机(Restricted Boltzmann Machines)对论文内容的主题分布进行建模,以预测可能的引用。

卷积神经网络(Convolutional Neural Network, CNN)^[25]具备的参数共享和局部连接的特点使其具有强大学习计算能力,它的卷积和池化操作,对文本数据特征的提取和表示学习也具有很好的效果。如 Zheng 等人^[26]提出了一种深度合作神经网络(DeepCONN),利用评论文本对用户的行为物品的特性联合建模,最后利用因子分解机(Factorization Machine, FM)来捕捉用户行为和物品属性的特征及交互联系进行推荐。基于 CNN 的推荐相较于传统的推荐模型能够更加关注文本中的关键词信息。吴俊超等人^[27]为了缓解论文数据稀疏性问题,使用卷积神经网络将学术论文的文本特征融入到异质信息网络中,丰富了论文的多样化表示,缓解数据稀疏问题。

循环神经网络(Recurrent Neural Network, RNN)^[28]得益于对顺序信息处理的优势,在推荐领域也得到了广泛的研究。然而在实际应用中,研究者们发现基础的 RNN 在

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/368027006061006111>