基于图神经网络异构数据融合的学科新兴主题探测研究*

段庆锋 陈 红 闫绪娴 刘东霞

(山西财经大学 管理科学与工程学院 太原 030006)

摘 要:[**研究目的**]数据异构性阻碍了大数据集成分析,而异构数据的深度融合学习能够增强学科数据分析能力,为预见学科新兴主题提供有力支撑。[**研究方法**]探测分析由两部分衔接构成,一是实现多元异构学科数据深度融合的图卷积神经网络(GCN),二是旨在学科主题预测的LSTM模型。具体地,通过GCN的深度学习能力,将包含多维特征和共现关系的异构主题数据转化为同构表示向量,不但实现异构融合,更为后续预测模型提供统一数据基础;然后,将主题表示向量时间序列输入LSTM模型,预测学科主题的新兴特征,为前瞻预见学科新兴主题提供决策支持。[**研究结论**]以图书情报学为对象的实证充分检验了GCN+LSTM的设计合理性,融合模型比非融合模型在主题趋势预测中展现出明显优势。

关键词:学科新兴主题;异构数据;多维特征;共现关系;图卷积神经网络

中图分类号:G350 **文献标识码**:A **文章编号**:1002-1965(2023)12-0127-07

引用格式:段庆锋,陈 红,闫绪娴,等. 基于图神经网络异构数据融合的学科新兴主题探测研究[J]. 情报杂志,2023,42(12):127-133.

DOI:10.3969/j.issn.1002-1965.2023.12.019

Detecting Scientific Emergency Topic Based on Heterogeneous Data Fusion Using GCN

Duan Qingfeng Chen Hong Yan Xuxian Liu Dongxia

(School of Management Science & Engineering, Shanxi University of Finance & Economics, Taiyuan 030006)

Abstract: [**Research purpose**] Data heterogeneity makes large data integration analysis difficult. Deep fusion learning for data with various structures aids in improving academic data analysis capability, and support the prediction of scientific emergency topics. [**Research method**] Two components make up detection analysis: (1) Graph Convolution Network (GCN) for deep fusion with various and heterogeneous academic data. (2) LSTM model for topic prediction in academic fields. In particular, using deep learning capability of GCN, heterogeneous topics data, including multi-characteristics and co-occurrence relations, are transformed into homogeneous representation vectors, realizing heterogeneous fusion while also providing a unified data base for the subsequent prediction model. In order to anticipate the emergency characteristics of academic topics and provide decision assistance for predicting academic emergency topics, topic representation vectors are then fed into a LSTM model to predict academic emergency characteristics, giving decision assistance for predicting academic emergency topics. [**Research conclusion**] In the academic discipline of library and information science, the empirical findings support the design of GCN+LSTM model as being reasonable. In addition, the fusion model outperformed than non-fusion models.

Key words: scientific emerging topic; heterogeneous data; multidimensional features; co-occurrence relations; GCN

1 问题的提出

反映学科趋势的新兴主题是科技竞争焦点,对于科技决策者至关重要。面对学科领域交叉融合与动态演化的复杂情形,准确认知甚至前瞻预测学科新兴趋势日益困难。当前,大数据技术为洞穿表象而直达内

收稿日期:2022-12-03 修回日期:2023-03-09

基金项目:教育部人文社会科学项目"基于学术社交媒体的学科新兴趋势识别研究"(编号:20YJA870005)研究成果。

作者简介:段庆锋,男,1977年生,教授,研究方向:科技情报;陈 红,女,1972年生,教授,研究方向:创新管理;闫绪娴,女,1978年生,教授,研究方向:信息管理;刘东霞,女,1975年生,副教授,研究方向:科技情报。

在本质提供了高效工具,富含丰富信息的海量学科数据更为揭示新兴主题提供底层基础。然而,大数据往往具有不同来源、不同媒介,甚至不同结构[1],数据范畴的最大化扩展尽管有助于提升分析能力,但由此伴随的数据结构内在冲突也给分析建模带来挑战。因此,探索多元异质学科数据的融合分析与建模已经成为学界的重要研究内容。

纵观科技情报相关文献,用于学科探测的数据主要体现为两大类。一是指标型数据,反映实体对象的个体状态,如文献计量指标、altmetrics 指标等[2];二是关系型(或网络型)数据,反映实体对象个体之间的关系状态,这些关系呈现网络结构,比如主题共现、文献共引等[3]。这样两种数据呈现不同结构定义,它们在学科新兴主题探测研究中发挥不可或缺作用,通过主题指标(指标型数据)的时序分析可以从纵向揭示学科新兴状态,通过主题间横向关系模式(关系型数据)可以揭示涌现状态。可见,此两种数据从不同视角及层面反映了学科主题状态,它们相互补充且不可替代。因此,全面揭示学科规律需要以指标型数据与关系型数据的深度融合为基础,因为学科发展既表现为主题知识的个体状态动态演化,同时又存在主题间知识关联与相互影响,不论何种类型数据的缺失都可能会导致分析视角不全面,更影响学科理解的深入性。总之,把握学科新兴主题需要且离不开异构数据的集成融合。

然而,目前科技情报学界对于异构数据融合研究还不够充分,仍不能完全满足学科探测对于全景大数据的统一集成分析需求。面对指标型数据与关系型数据,已有研究大多将源于不同数据分析的结果通过集结方式实现综合分析[4],这样虽然获得全面性,但是缺乏数据的统一建模利用,分析效率与深度都受到限制。究其原因,结构上的差异阻碍了信息的深度融合。例如,在探测新兴主题场景中常用的预测模型通常只能将指标型时序数据作为输入对象,而作为非欧式数据的关系型数据难以被导入模型,这样意味着学科主题预测结果只能利用主题个体在时间纵向上的动态规律,而忽略了主题横向间存在的相互依赖与作用关系,显然重要信息维度的缺失会直接降低预测能力,严重损害主题知识发展规律的深度揭示。因此,指标型数据与关系型数据的异构融合已经成为制约学科新兴主题发现的不可忽视环节,是促进学科探测研究的重要内容。

针对上述研究缺口,本文将图神经网络应用于学科新兴主题探测领域,通过异构数据的深度融合构建全景数据驱动的学科新兴主题预测模型。卷积神经网络模型(GCN)是专门针对图结构的学习模式,能够克服传统模型对于非欧式数据(关系数据)学习能力不足问题,尤其能够将节点特征(指标数据)与节点关系(关系数据)融合学习,非常适合于异构学科数据的融合场景。具体地,构建基于 GCN 的多维特征与共现关系的融合表示学习模型,获得主题向量;然后,基于主题向量时序样本,构建基于与 LSTM 的主题趋势预测模型;最后,以图书情报学领域为例,开展实证研究,以检验方法有效性。

2 相关研究

新兴主题战略价值关键在于面向未来的趋势与影响力,因此揭示主题状态趋势的预测方法成为最常用分析工具。在学科主题预测方面,基于指标型数据和关系型数据的研究方法具有鲜明不同之处,各自形成分析范式。

基于指标型数据的预测方法。反映主题趋势的方法有增长型指数法[5]、S 型增长曲线拟合法[6]、主题聚类变化[7]等。值得注意的是,近年以深度学习为代表的机器学习理论及算法不断成熟,成为新兴主题预见的热门方法。霍朝光等[8]构建基于 LSTM 模型的学科主题热度预测模型。朱光等[9][10]融合深度神经网络模型和文献计量指标用于预测新兴主题。陈伟等[11]采用包含双重随机过程的隐马尔可夫模型预测未来技术趋势。许学国和桂美增[12]构建采用 LSTM 模型和经验模态分解 EMD 的技术主题预测,并通过 Clarivate Analytics 机构发布的年度年报告对比说明方法有效性。Xu 等[13]构建了融合多种机器学习模型的新兴主题预测识别方法。虽然这些预测模型采用了最新预测技术,但是主要基于时序特征数据分析主题趋势,对关系模式的抽取与分析不足。

基于关系数据(网络结构)的预测方法。链路预测能够预测节点连接几率,为主题分析提供结构视角方案[14]。然而,链路预测建立在拓扑特征指标之上,缺乏网络关系学习能力[15]。值得注意的是,图神经网络 GNN 克服了网络结构学习的难题,开始得到学界广泛重视。作为 GNN 典型代表的图卷积神经网络 GCN 已在命名实体识别[16]、异质链路预测[17]、舆情分析[18]、文本分类[19]、多维学科知识网络融合[20]等多个场景取得研究进展。例如,刘非凡等[21]使用深度图神经网络探测学科领域主题知识结构,该方法能够有效融合文献的文本内容特征信息以及其引用关系特征信息,提升结果精准性。Kong 等[22]使用图卷积神经网络构建技术主题收敛识别模型。张思凡等[23]提出了基于 GCN 的文献被引量预测模型。GCN 模型对网络数据的学习能力开始已经得到广泛应用,尤其同时利用节点特征与网络关系的表示学习能力更体现了分析

优势。

基于主题指标数据与关系数据的研究都已相当成熟,但融合两种类型数据的集成探测相对不足。普通神经网络虽然能够实现不同来源及类型数据的特征自学习,但是无法实现网络结构上的学习,难以将关系数据与特征数据融合学习。近年发展迅速的图神经网络实现了神经网络模型在图结构上学习,能够达成节点特征与网络关系的信息融合,生成表示向量[24]。图表示学习模型事实上实现了异构数据的信息融合与统一表达,能够为面向主题分析的异构数据集成利用与统一建模提供基础。

3　主题共现矩阵与多维指标构建

3.1　数据来源及处理

数据来源于两部分:一是 WoS 数据库;二是 altmetrics 网站平台。其一,WoS 是主流文献题录数据库,收录文献能够反映学科前沿,是主题探测的常用数据源。图书情报学为实证领域,是典型复合型学科,前沿信息技术不断引入并融入图情场景,新兴主题不断活跃涌现,复杂动态的学科前沿非常适合用于检验学科探测方法。检索策略为查询代表性期刊文献,包括 Scientometrics、Journal of the Association for Information Science and Technology、Journal of Informetrics、Information Processing & Management、Information & Management。具体地,采用上述检索策略可以得到文献类型为 article 的文献记录共计 6326 条,其时间跨度为 8 年(2013-2020)。其二,网站 altmetric. com 是目前主流的 altmetrics 服务提供商,具有开源免费、指标丰富、覆盖率高等优点,完全满足数据采集需要。具体地,采用 python 爬虫工具,以文献 DOI 号为线索爬取 altmetrics 数据,经过多轮数据清洗,最终获得 3448 条匹配记录。

3.2　主题共现矩阵

主题词是分析基础,从文献关键词、标题及摘要当中通过分析程序抽取获得,经过多轮数据清洗、筛选、检验从中挑选出 250 个高频主题词。对于备选主题词,定义 250×250 的共现矩阵 A,其任意元素 $A_{ij} \in \{0, 1\}$,当主题 i 和 j 存在共现关系,则 A_{ij} 取值为 1,否则为 0。共现关系是揭示主题语义模式的重要工具,如果两个主题词共同出现在同一篇学术文献之中,则认为两者存在共现关系。按照年份计算,由此得到共 8 年(2013-2020)的主题共现矩阵时序数据。

3.3　主题多维特征

为了全面揭示主题状态,分别构建文献热度、引用热度、社交热度三个主题指标。这些指标是学科知识形成传播演变的不同视角揭示,涵盖了学科主题在文献媒介、引用媒介、社交媒介的多维特征体现。首先,主题的文献热度 D_i 定义为包含主题 i 的学术文献数量。学术文献是学科主题形成与传播的关键主要载体,主题出现的次数越多,反映其关注程度越高,文献媒介热度亦越高。

其次,从引用视角,定义主题引用热度 C_i 为

$$C_i = \sum_j \text{cited}_j * I_{ij} \tag{1}$$

其中变量 cited_j 表示学术文献 j 的被引数,I_{ij} 表示指标变量,反映了主题 i 是否出现在文献 j,其定义为:

$$I_{ij} = \begin{cases} 1, \text{if topic } i \text{ occuredindocument } j \\ 0, \text{otherwise} \end{cases} \tag{2}$$

被引数是反映文档学术价值的最常见和经典的文献计量指标,通常认为被引越高,则学术影响力越高。主题是文档内容的总结凝练,引用实质上也对相关主题的指向。因此,主题引用热度某种程度上亦反映了主题的学术影响力,体现了学者关注程度,取值越大,说明主题越具有学术价值。

最后,从社交媒介视角,构建主题社交热度指标 A_i。借鉴段庆锋等[25]提出的社交媒介关注指标,其定义为:

$$A_i = \sum_j \text{altmetrics}_j * I_{ij} \tag{3}$$

其中变量 altmetrics_j 表示学术文档 j 的替代计量指标,I_{ij} 表示指标变量,定义见公式(2)。主题社交热度指标建立在替代计量指标基础之上,是主题获得的社交媒体关注数量累积,反映主题在网络环境的关注热度。替代计量指标是文献计量指标的补充和扩展,刻画了更加广泛的学术影响力。主题社交热度指标越高,说明该主题在社交媒体上获得越高的关注和传播,尤其反映了社交网络层面的热度。

由此,指标集(D,C,A)构成三维特征,250 个主题形成 250×3 的特征矩阵 X。特征矩阵反映了主题的多维特征,这些特征源于不同媒介,构成多源数据。多维特征矩阵 X 与共现矩阵 A 具有不同结构形式,共同组成了多源异构数据,并作为异构数据融合模型的输入

4　模型构建

4.1　整体框架

GCN 模型具有强力的网络结构学习能力,并可以实现关系数据与特征数据的拟合与自学习。基于 GCN 模型获得的主题表示向量,采用 LSTM 模型预测新兴主题。整体上由两部分模型组合而成,一是用于图表示学习的 GCN 模型,二是用于新兴主题预测的 LSTM 模型,如图 1 所示。

多源异构数据。多维指标与共现关系分别反映主

题的不同数据，蕴含互补信息，都能够为主题趋势预测提供信息支撑。将两种异构数据同时用于预测模型以提升预见能力是研究的逻辑出发点。

数据融合模型。采用 GCN 模型实现异构主题数据的融合，得到用于趋势预测的主题表示向量。GCN 模型设定中，节点属性通过多维特征加以刻画，节点关系通过共现矩阵加以表现。

预测模型。LSTM 预测模型以融合多源异构数据的主题时序向量为输入，输出为反映是否为新兴主题的二元标签。

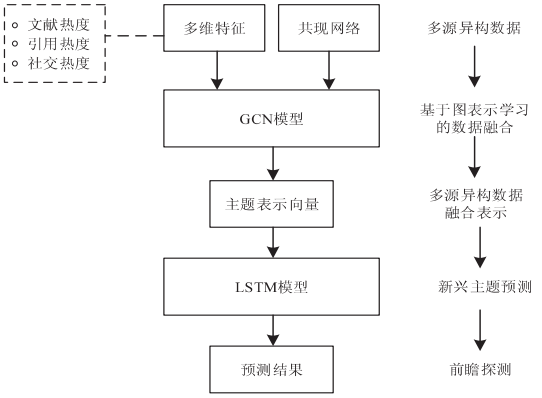

图 1 融合多维特征和共现关系的学科新兴主题探测框架

4.2 基于 GCN 的主题异构数据融合表示学习

图卷积神经网络最早于 2017 年由 Kipf 和 Welling 提出[26]，以图数据为样本，对节点关系进行表示学习，基于拉普拉斯矩阵的图上卷积运算是算法核心，理论扎实且性能优异。虽然，以 DeepWalk、Node2Vec 为代表的图神经网络模型能够在网络结构上进行学习，但采用的随机采样策略存在信息丢失问题[27]。GCN 模型克服了上述问题，通过图上卷积运算将节点表示为低维实向量，该向量是针对网络关系与节点特征的学习输出结果。

a. 模型架构。自编码结构是神经网络分布式表示学习常用框架，整体上包括编码器（encoder）和解码器（decoder）两部分，如图 2 所示。编码器部分由两层 GCN 模块串联而成，负责将共现矩阵 A 和多维特征 X 构成的异构数据转化为表示向量 Z，解码器部分负责基于表示向量 Z 实现网络还原。中间向量 Z 就是需要的主题表示结果，期望该向量能够尽可能地学习得到网络关键特征。

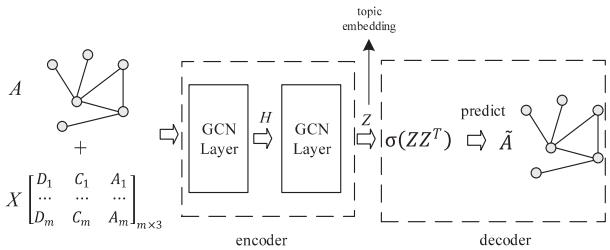

图 2 基于 GCN 的主题表示向量学习模型

b. 编码器部分。两层 GCN 模块连接组成。研究指出 GCN 模型不需要堆叠多层就可以取得不错效果[28]。具体地，公式（4）和（5）给出了 GCN 模块的正向传播过程。

$$H = ReLU\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}X W^{(1)}\right) \tag{4}$$

$$Z = \hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H W^{(2)} \tag{5}$$

其中，$\hat{A} = A + I$，A 为主题共现矩阵，I 为单位矩阵；$\hat{D}$ 是度矩阵，定义为 $\hat{D}_{ii} = \sum_j \hat{A}_{ij}$；X 为主题多维特征；$W^{(1)}$ 和 $W^{(2)}$ 为神经网络权重。GCN 模型中对称归一化拉普拉斯矩阵 $\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}$ 是图卷积操作的关键，形成焦点节点的嵌入向量 Z。该向量是从节点多维特征与共现关系之中提取信息得到的结果，用于表征主题词节点。

c. 解码器部分。解码器部分旨在基于主题向量 Z 进行网络结构还原。链路预测理论认为节点相似度越高，建立连接几率越大。基于此，定义节点连接概率矩阵 $\tilde{A}$ 为解码器输出，如公式（5）所示。

$$\tilde{A} = sigmod(Z\,ZT) \tag{5}$$

d. 模型训练。模型期望得到富含表征信息的表示向量 Z，即要求矩阵 A 与 $\tilde{A}$ 的差异最小化。将模型任务视为二分类问题，损失函数定义为输入 A 与输出 $\tilde{A}$ 形成的交叉熵。基于拟合模型，可以将主题多维特征和共现矩阵组成的异构数据学习转化为稠密实向量 Z。

4.3 基于 LSTM 的学科新兴主题预测

采用 LSTM 模型，从主题历史状态数据中学习动态演化规律，并基于拟合模型预测主题未来新兴状态。GCN 模型输出的表示向量 Z 融合了源于主题多维特征与共现关系信息，能够充分表征主题状态，此主题向量时间序列被用于预测模型。长短期记忆模型 LSTM 是一种典型的循环神经网络，优点在于通过包括遗忘门、输入门、输出门的模型机制克服了训练过程中的梯度爆炸及消失问题，能够更有效地从序列数据中捕捉特征。

预测模型主要包括 4 大部分：输入层、LSTM 层、全连层和输出层，如图 3 所示。模型通过前 T 年向量数据预测第 T+1 年主题是否呈现新兴状态。为了获得训练集样本标签，采用突发性检测算法（Burst Detection）[29]，对主题的新兴状态序列进行二元标注，若某主题第 t 年处于新兴状态则将其标注为 1，否则标注为 0。以损失最小为优化目标，采用随机梯度下降法，

通过多轮次迭代,可以得到拟合效果满意的估计模型。

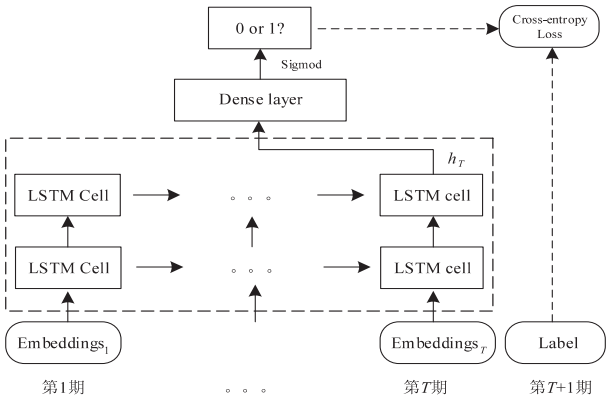

图 3　基于 LSTM 的新兴主题预测模型

5　结果分析

5.1　模型比较

为了说明设计合理性,需通过与其他基准模型的比较,检验融合模型的预测性能优势。在设计策略方面,有 3 个关键环节需要考虑,一是是否采用数据融合策略(融合数据 v. s. 非融合数据);二是采用何种融合模型(GCN v. s. 其他 GNN);三是采用何种预测模型(LSTM v. s. GRU)。由此,通过不同模型设计策略方式组合,形成除本文模型(模型 7)之外的其他 6 个参考模型用于性能比较。对于这些监督学习模型,选取 5 个常用的模型评估指标进行分析,包括 Accuracy、Precision、Recall、F1、AUC,通过多个指标的综合研判有助于全面揭示模型优劣势。采用 10 折交叉验证法划分训练集与测试集,基于此开展模型预测性能比较,见表 1。

表 1　不同模型性能比较

编号	模型	Accuracy	Precision	Recall	F1	AUC
1	x+LSTM	0.780	0.471	0.302	0.368	0.605
2	GCN(no x)+LSTM	0.800	0.565	0.345	0.428	0.597
3	GCN+GRU	0.820	0.577	0.566	0.617	0.727
4	Node2Vec+LSTM	0.828	0.625	0.471	0.538	0.698
5	GAT+LSTM	0.856	0.667	0.642	0.654	0.728
6	GraphSAGE+LSTM	0.860	0.821	0.434	0.568	0.704
7	GCN+LSTM	0.860	0.750	0.656	0.699	0.732

a. 融合数据 v. s. 非融合数据。本文模型 7 采用了多维特征与共现关系的融合数据,模型 1 只采用多维特征数据,模型 2 只采用共现关系数据。模型设计方面,模型 1 直接采用多维特征 X 数据进行预测模型训练与测试;模型 2 将 GCN 模型中输入特征 X 设定为全 1 向量,相当于只采用关系数据进行预测。通过比较可以看出,模型 7 在各个指标上都全面优于其他 2 个模型。结果支持了最初的判断,即多源异构数据有助于提升分析预测能力,将包含互补信息的多维特征与共现关系数据融合表示,能够为新兴主题趋势认知提供更有效的数据支撑,这也是本文新兴主题探测优势的关键所在。

b. GCN v. s. 其他 GNN 模型。GCN 是图神经网络 GNN 的一种,与其他图神经网络类型进行比较,能够说明选用 GCN 的合理性。这里选用 3 种 GNN 模型作为对比模型,包括 Node2Vec、GAT、GraphSAGE。Node2Vec 模型原理源于词嵌入方法,在 DeepWalk 基础上进一步优化了采样效果,采用了不同的随机游走方式;GAT 模型则在 GCN 基础上加入了注意力机制,具有更强自适应能力;GraphSAGE 则是一种 inductive 式学习,相对于 GCN 模型 transductive 式学习的有限扩展,能够将已训练模型灵活地应用于未知新数据。分别采用上述 3 种模型作为主题向量学习模块,而主题预测部分保持固定,由此得到采用 Node2Vec 的模型 4、采用 GAT 的模型 5,采用 GraphSAGE 的模型 6。将本文模型 7 与上述 3 个模型进行比较,可以发现它们的准确率和 AUC 差距不大,而召回率差距最为明显。其中,模型 4 表现最差,可能受到 Node2Vec 算法采样随机性的制约影响。基于 GAT 的模型 5 和基于 GraphSAGE 的模型 6 在各项指标上各有优劣,但仍与基于 GCN 的模型 7 存在差距。模型 6 有很高的分类精度,但召回率很低。综上所述,基于 GCN 的模型 7 表现最佳,GCN 在新兴主题预测中表现优于其他 3 种对比模型,说明图特征提取能力优异的 GCN 模型是合适的图学习模型选择。

c. LSTM v. s. GRU。模型 3 改用 GRU 作为预测模块,图表示学习部分保持不变。GRU 模型结构相对简单,采用重置门和更新门实现对历史数据的遗忘和更新,与 LSTM 模型同属循环神经网络,都是当前应用广泛的时间序列模型。通过比较可以看出,模型 3 明显劣于本文模型 7,尤其在预测分类精度和召回率方面水平较低。由此,LSTM 模型比 GRU 模型更加适用于新兴主题的前瞻探测任务,LSTM 模型的选用具有合理性。

综上所述,比较结果说明 GCN 与 LSTM 的模型组合方案具有相对优势,能够在出准确率之外的指标上保持领先,展现出良好预测能力。良好性能表现也反映了基于融合数据模型设计的合理性,一方面,GCN 模型对异构数据展现出良好的深度信息融合能力,尤其将主题多维指标特征与共现关系实现深度信息融合,形成表征能力强的表示向量,该向量同时包含了主题个体特征与横向关联特征,这种信息丰富度的最大化为学科分析提供了有力数据基础;另一方面,LSTM 模型负责接收融合数据并从中学习获得趋势预测能力,其与 GCN 模型形成的组合方案中表现出良好适配性,耦合良好的模型设计有助于充分挖掘融合数据的

内在丰富内涵,形成面向学科预见的一体化模型。

5.2 识别分析

进一步结合学科背景检验新兴主题预测有效性。具体地,将 2017 — 2020 年数据导入拟合模型,预测各个主题 2021 年的新兴涌现程度。通过主题预测,从中筛选出潜在新兴主题,揭示这些主题的学术价值与学科启示,可以检验异构融合数据预测模型在学科探测场景的应用性。

表 2 给出模型预测出的涌现潜力前 20 名主题,是重点考察分析对象。虽然通过预测年份文献资料回溯分析可以比较检验上述主题的涌现特征,但这些主题是否真正反映学科复杂动态及未来趋势,甚至具有远期的科技战略价值,需要综合分析并多方求证,尤其需要借助专家的丰富经验形成学科主题洞见。这里期望筛选出的主题不仅具有预期的高增长潜力,更希望这些主题刚刚兴起——因为及早发现涌现不久的主题才更具决策价值。因此,为了判断主题的新鲜程度,表 2 同时标出了主题最早出现的年份(限于样本跨度 2013—2020 年)。

基于以上思路,依据主题新鲜度(最早出现年份),表 2 中主题可划为两大类。一类是以知识表达(knowledge representation)、科研评估(research assessment)等为代表的热门主题,通常出现时间较久,有的甚至远早于样本年份,基本成为近 10 年图情领域的主流热点之一,它们虽然相对新鲜度不够高,但依然保持较高的学科活跃度和生命力,尤其随着学科发展其内涵及外延不断变化,甚至呈现阶段性涨落,持续推动学科创新前行。另一类则是新鲜涌现的新兴主题(其涌现年份以粗体标示),这些主题不但体现学科前沿最新变化,更暗示学科领域的发展新方向。从内容上,这些新兴主题进一步细分为两种。一是新兴信息技术推动下的情报学应用,包括区块链(blockchain)、物联网(Internet of things)、词嵌入(word embedding)、深度学习(deep learning)、链路预测(link prediction),这些新兴主题是学科交叉融合的典型体现,近年信息技术的重大突破都快速进入并影响图情领域面貌,从方法工具层面极大地推动并改变着该学科方法范式,而且这种影响是长期深远的,值得学者高度关注;二是反映新研究议题的新冠疫情(covid-19),作为爆发于 2019 年的全球性重大公共卫生事件对全球社会各个层面带来了巨大深远影响,同样图情学科亦做出了及时重点关注。检索样本期刊发现该主题论文发表分布为 7 篇(2020 年)、67 篇(2021 年)、50 篇(2022 年),可见其突发性涌现特征明显,也成为模型对其 2021 年的高突发性预测的有力印证,而且这种活跃性在随后年份也得到了延续,凸显了作为高价值新兴主题对学科未来的巨大影响力。这些新兴主题的学术价值日益获得领域学者的关注认同,其高涨学术影响趋势也能够通过新近学术文献得到证据支撑,说明了模型在学科预见场景的有效应用。

总之,本文模型不但展现良好预测性能,更能够给出颇具学科启发价值的新兴主题,体现应用价值。尤其值得注意的是,本文方法能够及时敏感地发现以 covid-19、blockchain 为代表的新鲜高价值主题,其时间短、数量相对少但具有高增长潜力,这些特征通常给探测任务带来困难。这里,面向未来的学科趋势探测建立在异构数据驱动的学习预测能力之上,结合新鲜程度有助于将高价值新兴主题从包括热门主题的干扰信号中进一步细分甄别。

表 2 高潜力学科新兴主题列表

序号	主题	突发性预测(2021 年)	最早出现年份(2013-2020 年)
1	covid-19	0.771	2020
2	blockchain	0.770	2019
3	knowledge representation	0.688	2013
4	research assessment	0.686	2013
5	artificial intelligence	0.686	2014
6	community detection	0.685	2014
7	social sciences and humanities	0.680	2013
8	intellectual structure	0.680	2013
9	social media	0.679	2013
10	social influence	0.678	2013
11	Internet of things	0.673	2015
12	machine learning	0.671	2013
13	word embedding	0.671	2018
14	deep learning	0.671	2018
15	natural language processing	0.671	2013
16	link prediction	0.652	2015
17	big data	0.608	2013
18	open access	0.608	2013
19	information security	0.572	2014
20	topic evolution	0.558	2013

6 结 论

多维特征与共现网络互为异构数据,虽然都是主题探测依赖的数据基础,但难以融合用于新兴主题分析。针对异构数据融合难题,借助 GCN 模型对网络结构的信息提取能力,实现主题多维特征与共现网络的融合表达,并获得融合异构数据的主题表示向量。基于此,以该融合向量为输入,利用 LSTM 时序模型,预测学科主题的新兴状态涌现趋势。以图书情报学为领域开展实证研究,对 GCN+LSTM 的模型框架进行性能比较和结果检验,以验证本文方法的有效性。

本文探测方法表现出良好新兴主题预测能力,具有应用价值。a. 研究说明融合模型的预测能力优于非融合模型,预测能力的提升主要来源于异构数据融合

而获得的信息丰富度。将应用广泛但难以集成分析的计量指标与共现网络数据进行深度图学习,以获得蕴含全息特征的主题嵌入向量,该向量能够更好地捕捉主题内在规律,不但能够用于发现新兴主题,而且具有更加广阔的应用场景,例如主题演化、主题语义等。b. 研究说明图卷积神经网络对于网络结构关系具有极强学习捕捉能力,有助于深刻把握学科领域的复杂性,能够将其应用扩展至其他对象,比如引用网络、学术合作等。另外,虽然链路预测及社会网络指标也被应用于主题分析,但他们缺乏图卷积神经网络拥有的图学习能力,更无法实现异构数据的多元融合。本文贡献在于借助图卷积神经网络实现以往难以集成利用的异构数据融合,通过 GCN+LSTM 的模型设计为更加全面深刻把握学科主题新兴规律提供有益探索。

虽然方法科学性与适用性得到有力支撑,但也存在局限性。本文选择了主题共现进行融合分析,但还有其他反映主题状态的重要网络关系,比如语义关联、主题引用等,这些数据的综合集成分析值得下一步深入探讨。另外,非结构化的文献全文蕴含更丰富主题信息,将其与网络关系的深度融合也是有待探索方向。

参 考 文 献

[1] 段庆锋, 闫绪娴, 陈　红, 等. 基于媒介比较的学科新兴主题动态识别——altmetrics 与引文数据的融合方法[J]. 情报学报, 2022, 41(9): 930-944.

[2] 李　静, 徐路路. 基于机器学习算法的研究热点趋势预测模型对比与分析——BP 神经网络、支持向量机与 LSTM 模型[J]. 现代情报, 2019, 39(4): 23-33.

[3] 黄　璐, 朱一鹤, 张　嶷. 基于加权网络链路预测的新兴技术主题识别研究[J]. 情报学报, 2019, 38(4): 335-341.

[4] 卢　超, 侯海燕, Ying D, 等. 国外新兴研究话题发现研究综述[J]. 情报学报, 2019, 38(1): 97-110.

[5] 宋欣娜, 郭　颖, 席笑文. 基于专利文献的多指标新兴技术识别研究[J]. 情报杂志, 2020, 39(6): 76-81.

[6] 曹艺文, 许海云, 武华维, 等. 基于引文曲线拟合的新兴技术主题的突破性预测——以干细胞领域为例[J]. 图书情报工作, 2020, 64(5): 100-113.

[7] 刘敏娟, 张学福, 颜　蕴. 基于核心词、突变词与新生词的学科主题演化方法研究[J]. 情报杂志, 2016, 35(12): 175-180.

[8] 霍朝光, 霍帆帆, 董　克. 基于 LSTM 神经网络的学科主题热度预测模型[J]. 图书情报知识, 2021(2): 25-34.

[9] 朱　光, 刘　蕾, 李凤景. 基于 LDA 和 LSTM 模型的研究主题关联与预测研究——以隐私研究为例[J]. 现代情报, 2020, 40(8):38-50.

[10] Liang Z T, Mao J, Lu K, et al. Combining deep neural network and bibliometric indicator for emerging research topic prediction [J]. Information Processing & Management, 2021, 58(5): 1-18.

[11] 陈　伟, 林超然, 李金秋, 等. 基于 LDA-HMM 的专利技术主题演化趋势分析——以船用柴油机技术为例[J]. 情报学报, 2018, 37(7): 732-741.

[12] 许学国, 桂美增. 基于深度学习的技术预测方法——以机器人技术为例[J]. 情报杂志, 2020,39(8): 53-62.

[13] Xu S, Hao L Y, An X, et al. Emerging research topics detection with multiple machine learning models[J]. Journal of Informetrics, 2019, 13(4): 1-19.

[14] Huang L, Chen X, Ni X X, et al. Tracking the dynamics of co-word networks for emerging topic identification[J]. Technological Forecasting and Social Change, 2021, 170: 1-14.

[15] 段庆锋, 陈　红, 刘东霞, 等. 基于 LSTM 模型与加权链路预测的学科新兴主题成长性识别研究[J]. 现代情报, 2022, 42(9): 37-48,142.

[16] 景慎旗, 赵又霖. 面向中文电子病历文书的医学命名实体识别研究——一种基于半监督深度学习的方法[J]. 信息资源管理学报, 2021,11(6):105-115.

[17] Wang X, Chai Y, Li H, et al. Link prediction in heterogeneous information networks: An improved deep graph convolution approach[J]. Decision Support Systems, 2020, 141(113448): 1-12.

[18] Zhao M, Yang J, Zhang J, et al. Aggregated graph convolutional networks for aspect-based sentiment classification[J]. Information Sciences, 2022, 600: 73-93.

[19] 张晓丹. 改进的图神经网络文本分类模型应用研究——以 NSTL 科技期刊文献分类为例[J]. 情报杂志, 2021, 40(1): 184-188.

[20] 李　慧, 胡吉霞. 一种基于图卷积自编码模型的多维度学科知识网络融合方法[J]. 图书情报工作, 2020, 64(18): 114-125.

[21] 刘非凡, 张　爽, 罗双玲, 等. 基于深度图神经网络方法的领域知识结构探测[J]. 情报学报, 2021, 40(11): 1209-1220.

[22] Kong D, Yang J, Li L. Early identification of technological convergence in numerical control machine tool: A deep learning approach[J]. Scientometrics, 2020, 125: 1983 - 2009.

[23] 张思凡, 牛振东, 陆　浩, 等. 基于图卷积嵌入与特征交叉的文献被引量预测方法:以交通运输领域为例[J]. 数据分析与知识发现, 2020, 4(9): 56-67.

[24] 范　涛, 王　昊, 吴　鹏. 基于图卷积神经网络和依存句法分析的网民负面情感分析研究[J]. 数据分析与知识发现, 2021, 5(9): 97-106.

[25] 段庆锋, 潘小换. 利用社交媒体识别学科新兴主题研究[J]. 情报学报, 2017, 36(12): 1216-1223.

[26] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[C]. ICLR 2017, 2017: 1-14.

[27] Zeb A, Saif S, Chen J, et al. Complex graph convolutional network for link prediction in knowledge graphs[J]. Expert Systems with Applications, 2022, 200: 1-16.

[28] 吴　博, 梁　循, 张树森, 等. 图神经网络前沿进展与应用[J]. 计算机学报, 2022, 45(1): 35-68.

[29] Kleinberg J. Bursty and hierarchical structure in streams[J]. Data Mining and Knowledge Discovery, 2003,7(4):373-397.

(责编:王平军;校对:贺小利)