面向开放互联网的科学数据挖掘与理解

卢 彬[1] 甘小莺[1] 甘 雨[1] 唐 顾[1] 马婷晏[1] 吴律文[2]
赵 泽[1] 傅洛伊[2] 金 梦[2] 王新兵[1] 周成虎[3]

[1](上海交通大学信息与电子工程学院|集成电路学院 上海 200240)
[2](上海交通大学计算机学院 上海 200240)
[3](中国科学院地理科学与资源研究所 北京 100101)

摘 要 随着数据观测、采集手段的发展,科学大数据正快速增长,并推动着科研范式变革。然而,科学数据分散在互联网中各类数据仓储与个人数据库中形成了“数据孤岛”,难以有效整合与关联科学数据。为此,本文提出了一种面向开放互联网的科学数据挖掘与理解方法,通过机器阅读各类互联网数据资源,自动识别科学数据并结构化抽取关键字段,实现对科学数据的高效发现与管理。具体来说,本文融合网页多视角信息设计了网页筛选器 WebFilter,通过融合网页 DOM 树的结构共现与语义相关实现对网页级特征理解与分类;此外,本文设计了基于节点异构关联的网页阅读器 WebReader,通过异构图网络的消息传递对网页关键信息进行结构化抽取,形成科学数据画像。本文采用了多个公开数据集进行实验性能评估:在网页分类方面,本文提出的 WebFilter 相较于基线模型准确率提升了 1.39%到 3.71%、$F1$ 分数提升了 1.42%到 4.10%;在网页信息抽取方面,本文提出的 WebReader 平均提升 1.40%,在少训练样本情况下性能提升显著。更进一步,基于本文技术研究成果研制了面向地球科学领域的开放科学数据系统 DataExpo,汇聚百万科学数据并提供了数据多维查询、地图查询等数据服务,已应用于“深时数字地球”国际大科学计划,推动了地球科学领域数据驱动范式研究。

关键词 科学数据;网页数据挖掘;AI for Science;文本图神经网络;信息检索;自然语言处理
中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2026.00015

Towards Scientific Data Mining and Understanding on the Internet

LU Bin[1] GAN Xiao-Ying[1] GAN Yu[1] TANG Gu[1] MA Ting-Yan[1] WU Lv-Wen[2]
ZHAO Ze[1] FU Luo-Yi[2] JIN Meng[2] WANG Xin-Bing[1] ZHOU Cheng-Hu[3]

[1](School of Information and Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240)
[2](School of Computer Science, Shanghai Jiao Tong University, Shanghai 200240)
[3](Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101)

Abstract With the development of data observation and collection methods, scientific big data is growing rapidly and driving paradigm shifts in research. However, scientific data, scattered across various data warehouses and personal databases on the internet, forms “data silos”, making it difficult to effectively integrate and correlate scientific data. To address this problem, this

收稿日期:2025-02-18;在线发布日期:2025-10-29。本课题得到国家自然科学基金青年学生基础研究项目(博士研究生)(No. 623B2071)、国家自然科学基金面上项目(No. 62272301)、国家自然科学基金创新研究群体项目(No. T2421002)、国家资助博士后研究人员计划(No. GZB20250806)资助。**卢 彬**,博士,博士后,中国计算机学会(CCF)会员,主要研究领域为图机器学习、AI for Science。E-mail:robinlu1209@sjtu.edu.cn。**甘小莺(通信作者)**,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为物联网数据挖掘、时空感知计算。E-mail:ganxiaoying@sjtu.edu.cn。**甘 雨**,硕士研究生,主要研究领域为网页数据挖掘。**唐 顾**,博士研究生,主要研究领域为推荐系统、文本图神经网络。**马婷晏**,硕士研究生,主要研究领域为科学数据、图神经网络。**吴律文**,博士研究生,主要研究领域为推荐系统、文本图神经网络。**赵 泽**,博士研究生,主要研究领域为知识图谱、图机器学习。**傅洛伊**,博士,副教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为网络信息结构化表达与计算。**金 梦**,博士,副教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为物联网、AI for Science。**王新兵**,博士,特聘教授,博士生导师,中国计算机学会(CCF)杰出会员,主要研究领域为物联网、大数据、AI for Science。**周成虎**,博士,研究员,博士生导师,中国科学院院士,主要研究领域为地理信息系统、地球科学智能。

paper proposes a scientific data mining and understanding method on the open internet, which automatically identifies scientific data and extracts key fields in a structured manner, enabling efficient discovery and management of scientific data through machine reading techniques. Specifically, this paper integrates multi-view web information to design the WebFilter, which understands and classifies webpage-level features by combining the structural co-occurrence and semantic relevance of the webpage's DOM tree. Additionally, this paper designs the WebReader based on DOM node heterogeneous associations, which performs structured extraction of key webpage information through message passing in a heterogeneous graph network, forming a scientific data profile. Several publicly available datasets were used for experimental performance evaluation: in the case of webpage classification, the proposed WebFilter improved accuracy by 1.39% to 3.71% and the $F1$ score by 1.42% to 4.10% compared to the baseline model; in the case of webpage information extraction, the proposed WebReader improved performance by an average of 1.40%, with significant improvement under scenarios with few training samples. Furthermore, based on the technological research achievements of this paper, an open scientific data system, DataExpo, was developed for the field of Earth sciences. It aggregates millions of scientific data and provides data services such as multi-dimensional queries and map queries. DataExpo has been applied in the "Deep-Time Digital Earth" international big science program, advancing research in data-driven paradigms in the field of Earth sciences.

Keywords scientific data; web data mining; AI for Science; text-attributed graph neural network; information retrieval; natural language processing

1 引 言

人工智能技术的快速发展在基础科学研究领域正掀起变革性影响[1-2]，AI for Science 正逐渐推动形成科学发现的"第五范式"[3]。科学数据作为人工智能的核心要素，是推动科学研究的基石。特别是开放科学(Open Science)倡议下，快速增长的科学数据共享[4]有助于更广泛的数据融合、更公开的数据校验、更全面的学术传播。近年来，我国《科学数据管理办法》出台也确立了"开放为常态、不开放为例外"的原则，鼓励科学数据开放共享。然而，目前开放科学数据分散在互联网上多源异构的不同数据仓储及大量科学家个人构建的数据库中，形成了"数据孤岛"。如何进行统一化的数据发现与管理，并提供高效的开放科学数据服务是促进下一代科学数据基础设施的重要问题。

科学数据发现与管理是一个学科交叉问题，受到了包括计算机科学、情报学、信息管理与信息系统等多学科领域的关注。(1)科学数据发现：互联网上的各类数字资源数量庞大，科学数据资源的占比极小，科学数据发现旨在基于特定主题(关键词)对互联网上相关科学数据资源进行查询与汇聚。受限于数字资源的复杂性，早期成果主要通过 API 接口集成对少量的垂直特定领域数据资源进行汇聚，形成了如 DataCite[5]、DataONE[6]、DataMed[7]等集成数据检索平台。然而，更多长尾科学数据缺乏 API 进行直接工具化集成，使得科学数据库的规模受限。因此，一项代表性研究成果是 Google Dataset Search[8-9]，其基于谷歌大规模网页数据资源，根据部分网页开发时编写的 Schema @type[10]标准字段对网页进行识别，筛选出数据集页面进行汇聚。基于互联网的开放资源，显著扩大了数据发现的范围。然而，统计结果发现大约 70%的网站并不提供 Schema，甚至有研究表明 61%包含 Schema.org/Dataset 的网站并非数据网站[11]，这使得基于 Schema 标准的数据发现方法并不可靠。(2)科学数据管理：2016 年 Wilkinson[12]等人首次提出了开放科学数据的 FAIR 原则，即 Findable(可发现)、Accessible(可访问)、Interoperable(可互操作)、Reusable(可重用)，旨在建立可供人类和机器共同读取和使用的元数据标准，对科学数据进行规范化管理。为了响应 FAIR 原则倡议，Dryad[13]、PANGAEA[14]、Zenodo[15]等数据平台联合学术期刊共同推动开放

科学数据共享,并设计了规范的元数据字段信息要求作者上传数据时手动填写。汇聚多源互联网数据资源的 Google Dataset Search 则是基于 Schema 标准的多个字段自动化提取文件格式、下载地址、DOI 标识符、发布日期等并进行规范化。然而,由于不同数据网页开发与维护情况不同,不同科学数据的元数据存在着字段缺失、质量参差不齐、格式不统一的问题[16]。基于上述分析发现,由于互联网上的资源规模庞大、信息组织异构,现有研究依旧面临着科学数据资源发现不够全、科学数据管理不够自动化的难题。

为此,本文提出了面向互联网的开放科学数据挖掘与理解方法。如图 1 所示,针对搜索引擎获取的大量网页资源,本文通过设计深度学习模型自动筛选得到科学数据页面,并进行网页理解得到标准化、丰富的元数据信息用于刻画科学数据,支撑灵活多样的检索与查询。具体来说,首先本文提出了融合网页多视图信息的网页表征模型 WebFilter 对互联网上的各类资源进行分类,通过挖掘利用网页的文本信息、HTML 结构及网页的渲染信息进行网页级表征,从纷繁复杂的网页中筛选出科学数据网页;其次,提出了网页结构感知的网页理解算法 WebReader,进一步通过机器阅读对网页中的元素进行向量化表示,实现数据网站中的核心元数据字段进行通用化提取,形成科学数据画像;最后,在公开数据集上对于算法进行性能验证,同时以地球科学领域为例,将所提方法进行了实际落地应用,开发了数据巡航系统 DataExpo(https://dataexpo.deep-time.org),发现并汇聚了超过百万地球科学元数据,涵盖近 3 万个机构,数据源 IP 覆盖 120 余个国家和地区,支持对数据网页超过 10 个多模态字段对信息抽取,形成元数据画像。

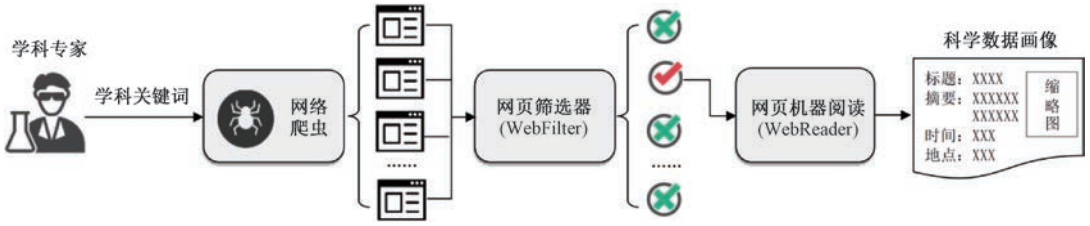

图 1　面向开放互联网的科学数据挖掘与理解整体流程图

本项研究工作支撑了以数据共享和知识发现为目标的深时数字地球(Deep-time Digital Earth,DDE)国际大科学计划[17],预期通过数据和知识驱动推动地球科学研究范式的变革。

具体而言,本文的主要贡献如下:

(1)针对开放互联网上科学数据资源分散混杂的问题,提出了科学数据挖掘与理解方法,突破传统 API 集成与 Schema 规则理解范式,通过机器自动阅读互联网资源,自动识别科学数据并结构化抽取关键字段,实现对科学数据的高效发现与管理。

(2)本文提出的网页筛选器 WebFilter 与网页阅读器 WebReader 通过融合网页的 DOM 树结构信息与节点多视图特征语义,实现了网页级与节点级的向量化表征与分类,在 WebKB 等三个公开数据集网页分类相较于最优基线算法 F1 分数提升 1.42%到 4.10%,网页信息抽取在少样本标注需求下性能提升 1.92%到 2.33%。

(3)构建了面向地球科学领域的开放科学数据系统 DataExpo,面向地层学、沉积学、古地理等 18 个地学学科方向发现超百万科学数据,实现了面向异质网页数据的标题、摘要、发布时间、地点、机构等字段的统一结构化理解,并提供数据多维查询、数据地图查询等在线服务。

本文共分为 6 节,各节的组织结构如下:第 1 节为引言,对本文研究背景与意义进行概述性介绍;第 2 节为相关工作,详细阐述了与本文研究有关的国内外研究进展;第 3 节介绍本文提出的技术方法,并在第 4 节使用多个公开数据集进行实验验证;第 5 节介绍基于本研究的技术成果,在地球科学领域的应用与 DataExpo 系统研制情况;第 6 节为本文的总结与展望。

2　相关工作

在本节中,对本文涉及的相关代表性工作进行概述,分析现有研究脉络及面临的技术挑战。

2.1　开放科学数据发现与管理

开放科学环境下,科学数据的开放共享有助于促进全球科学的共同发展。现有开放科学数据平台建设的一条技术路线便是利用不同数据平台的 API 集成,来关联不同数据仓储的数据。例如,Open Data Portal Watch [18]聚焦政府数据汇聚了超过

260 个数据仓储的元数据，DataMed [7] 则是聚焦生命科学领域，汇聚了约 75 个数据仓储的元数据，Socrata [19] 聚焦城市数据，汇聚了超过 200 个数据仓储。这一方法的局限性便是极依赖知识先验与科学家经验，集成搜索的范围是有限的，难以拓宽科学研究更广的数据查询范围。为了解决这一问题，谷歌公司开发了 Google Dataset Search（GDS）[8-9] 对互联网上的各类数据资源进行索引。然而，GDS 仅支持包含 Schema 标准的网站，但对于更多广泛的科研实验室、小型科研项目、个人维护的科学数据网站等长尾数据均会出现大量的遗漏。为此，DSDD[20] 提出了采用网络爬虫与网页数据分析的方法对一个特定领域的数据进行自动化发现，DSDD 通过一批用户给定的种子网站，通过获取其前向与反向链接来进行持续探索，并结合词频－逆向文件频率 TF-IDF [21] 对网页文本进行编码及使用支持向量机（SVM）进行分类。然而，DSDD 采用的文本分类模型性能是不足的，引入了大量的误判与遗漏，同时如何对数据进行高效的管理是其缺乏研究的内容。

2.2 网页数据表示学习

早期对于网页数据表示学习聚焦文本表征，即通过网页解析获取网页的文本信息，后采用面向自然语言的统计机器学习（如词袋模型、词频－逆向文件频率 TF-IDF[21]、词向量化模型 Word2Vec[22]）、卷积神经网络模型（如文本卷积神经网络 TextCNN[23]、DCNN[24] 等）、基于 Transformer 的语言模型（BERT[25]、RoBERTa[26]、XLNet[27]，T5[28]）等进行理解。

然而，网页除了文本信息还有丰富的多视图信息，包括超链接 URL、HTML 的 DOM（Document Object Model）树结构等。超链接 URL 的文本信息虽短，却关联了共享域名信息的多个网页。如图 2 所示是 HTML 的 DOM 树结构，它进一步给文本信息赋予段落结构，模型可以识别出标题、数据集摘要等重要的文本信息来对网页进行表示，主动忽略如广告、外链、推广等无关噪音信息。为了对上述多视图信息进行刻画，RiSER[29] 通过词向量嵌入和 XPath 嵌入分别得到文本内容与 HTML 结构的表征，利用两层 LSTM 实现文本与结构联合编码实现融合表示。DOM-LM[30] 将 DOM 树分割成多个子树，每个子树保留重要的上下文信息，通过在词向量的基础上增加网页 HTML 节点的属性与位置结构信息，扩展了 BERT 等语言模型的能力，使其能够

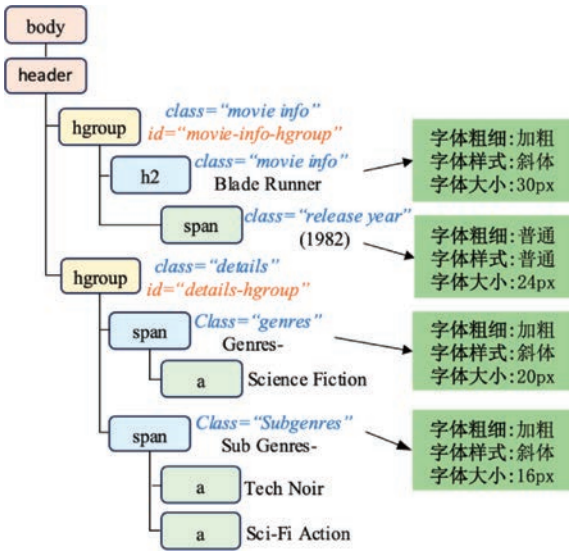

图 2 网页的 DOM 树结构示例

处理 HTML 文档的结构化特性。WebFormer[31] 优化 HTML 标签的向量表征，为每个 DOM 节点引入 HTML 标记，这些标记不仅包含文本内容，还包含了节点的结构信息，如父节点、子节点和兄弟节点的关系等，并设计了多个注意力模式，包括 HTML 标记之间的注意力、HTML 标记与文本标记之间的注意力等，以此帮助模型更好地理解网页的布局结构。

2.3 网页信息结构化抽取

网页的 HTML 结构可以通过层次关系转换得到 DOM 树结构，基于树上的搜索可以定位不同的网页标签与文本信息。同时，网页信息结构化抽取问题被转换为 DOM 树上的节点分类问题，即网页的元数据信息对应于 DOM 树上的某个节点。为此，FreeDOM [32] 率先通过建模 DOM 树的结构关联并融合节点文本特征，实现对网页信息的抽取。SimpDOM [33] 则进一步给出了“伙伴节点”、“朋友节点”的概念，对节点本身信息与局部邻接节点的信息进行联合表征提升表达能力。随着以 Transformer 结构为主体的大语言模型技术的发展，MarkupLM [34] 将网页 DOM 树结构信息也转换为词元，随着文本词元一同送入 Transformer 模型中，实现统一的结构与文本信息理解。进一步，WIERT[35] 考虑网页的视觉信息，将网页渲染信息的 CSS 树与 DOM 树进行融合，提升模型对于字体大小、颜色等的考虑。Xu 等人[36] 提出将网页截图作为视觉信息加入网页表征，并通过多项预训练任务进行模态对齐，但网页截图数据获取难度大，标注成本高，使得场景泛化能力受限。综上所述，现有的网页信息结构化抽取研究通过增加不同模态信息来

增强节点表达,但大都将网页视为多段长文本序列或独立节点进行处理,缺乏对于网页 HTML 天然的树状结构信息及样式相似性特点的建模,全局信息理解能力弱。

3　科学数据机器挖掘与理解方法

在本节中,针对现有相关工作的分析,本文提出了面向开放互联网的科学数据机器挖掘与理解方法,整体技术路线如图 3 所示。系统的输入是学科专家给定的关键词,利用搜索引擎以"学科关键词+dataset/data/数据库"作为输入进行查询,通过网络爬虫的方式得到一批候选网页。由于互联网上的数据资源较为复杂,既包括了科学数据网站,也包括了大量无关网页信息,例如科技新闻报道、机构介绍、学者主页等。因此,本文提出了融合网页多视图信息的网页筛选器 WebFilter,通过融合网页的 DOM 树结构与特征属性关联,实现对网页级表征与分类。进一步,针对筛选得到的数据网页,提出了网页结构感知的网页阅读器 WebReader,通过获得网页节点级表征对不同元数据字段进行提取,实现对科学数据的结构化理解。

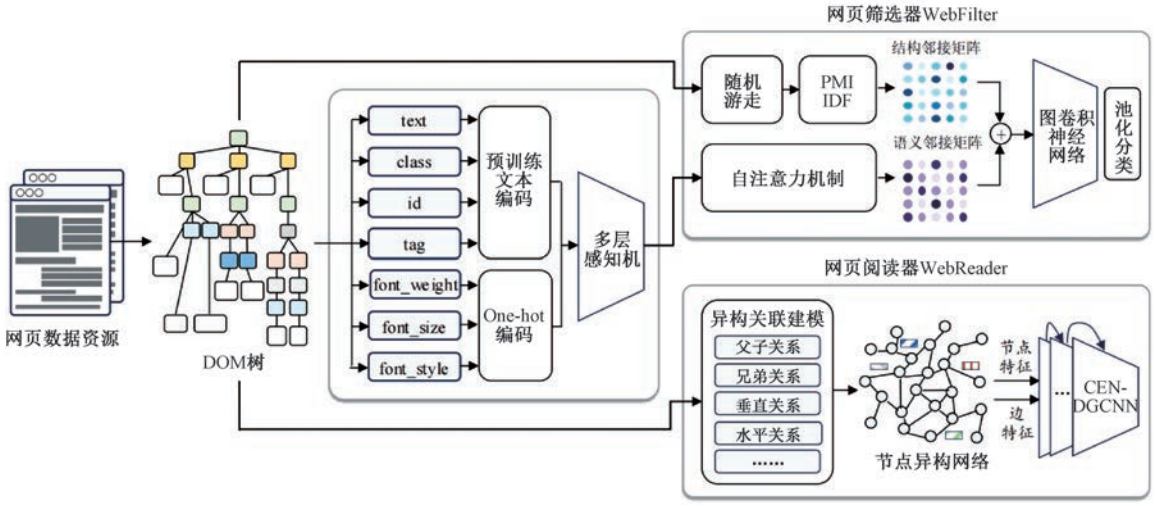

图 3　面向开放互联网的科学数据挖掘与理解技术路线图

3.1　任务与符号定义

网页筛选器 WebFilter 的输入是一批爬虫获取的网页集合 $W=\{w_1,w_2,\cdots,w_N\}$,其中 w_i 代表了每个网页样本,N 代表了网页样本的总数。给定网页分类类别 $\mathcal{C}^W=\{C_1^W,\cdots,C_{K_W}^W\}$,WebFilter 模型 $\mathcal{F}(\cdot)$ 的目标是判别网页样本的类别归属。

对于每个网页 w_i 根据其 DOM 树结构可以转换为一张如图 2 的结构图 $\mathcal{G}_i^s=(V_i,E_i^s)$,节点集合 $V_i=\{v_1^i,v_2^i,\cdots,v_{m_i}^i\}$ 代表了 DOM 树的节点,节点总数为 m_i。边集合 E_i^s 代表了网页 DOM 树的结构组织关系。网页阅读器 WebReader 模型 $\mathcal{R}(\cdot)$ 的目标是根据核心元数据字段类别 $\mathcal{C}^V=\{C_1^V,\cdots,C_{K_V}^V\}$,如标题、作者、摘要、机构、缩略图等,对节点类别进行分类。

3.2　网页筛选器 WebFilter

3.2.1　基于多视图信息的网页节点表征

网页 w_i 的 HTML 信息经过转换得到一张结构图后,图中的任意节点 v_j^i 蕴含了如表 1 所示的多个特征,这些特征蕴含了节点的文本语义以及在结构中的样式信息,均对于判断网页内容类型具有重要作用。进一步,根据其文本类型或枚举类型,分别采用预训练 BERT 语言模型进行文本向量编码或 one-hot 编码,然后进行特征拼接得到节点的初始特征 $X_j^i\in\mathbb{R}^d$,其中 d 是特征拼接后的特征长度。网页 w_i 所有节点的特征组成了 $X^i\in\mathbb{R}^{m_i\times d}$。

3.2.2　网页级多视图信息融合表征

为了得到网页级的特征表示,进一步根据节点的 DOM 树结构与文本语义相关性进行图节点特征的消息传递,从而得到网页级表征。

首先考虑网页 DOM 树的结构邻近特性,通过在结构图 $\mathcal{G}_i^s$ 上进行长度为 K_{rw} 步随机游走,得到一组元路径,通过点互信息算法(Pointwise Mutual Information,PMI)得到元路径上两两节点的关系。以节点 v_p^i 和 v_q^i 为例,两节点的共现相关性为

$$PMI(v_p^i,v_q^i)=\log\frac{p(v_p^i,v_q^i)}{p(v_p^i)p(v_q^i)}。$$

其中,$p(v_p^i)$ 与 $p(v_q^i)$ 分别表示节点 v_p^i 和 v_q^i 的元

路径数量，$p(v_p^i,v_q^i)$ 表示两节点同时出现的元路径数量。通过 PMI 可以有效平衡高频节点和低频节点的贡献权重。在网页数据表征问题中，考虑到网页 HTML 中/html、/head、/body 等模板式节点出现频率极高但缺乏有效网页差异化特性，故采用 TF-IDF 算法中的逆向文件频率 IDF 限制此类节点的权重。从而得到 PMI-IDF 指数 ρ，对于任意两节点 v_p^i 和 v_q^i 计算为

$$\rho(v_p^i,v_q^i)=PMI(v_p^i,v_q^i)*\log\frac{|D|}{|m:(v_p^i,v_q^i)\in d_m|+1}。$$

其中，$|D|$ 是采样的元路径总数，$m:(v_p^i,v_q^i)\in d_m$ 表示包含节点 v_p^i 和 v_q^i 的元路径。通过计算两两节点的 PMI-IDF 指数可以良好地表示 DOM 树中具有特殊结构含义节点的结构共现特性，从而得到节点的结构邻近性邻接矩阵 $A^s\in\mathbb{R}^{m_i\times m_i}$，对任意节点 v_p^i 和 v_q^i 的 PMI-IDF 系数大于超参数阈值 ξ，两节点间存在一条边，否则不存在。考虑不同网页的结构差异性，不同阈值设置结果具有差异性，在本文中我们设置 $\xi=0$，其表示了 $p(v_p^i,v_q^i)>p(v_p^i)p(v_q^i)$ 的通用条件。

其次考虑节点特征的语义相关性，针对节点的多维特征采用自注意力机制对节点间的特征相关性进行度量。对于输入的网页节点特征 X^i 分别进行线性变换得到查询变量 $Q=X^iW_Q$ 和键变量 $K=X^iW_K$，随后对查询变量与键变量进行点积与缩放，并应用 softmax 函数后得到注意力分数，从而得到因子。进而加权得到网页多视图信息融合的邻接矩阵 $\boldsymbol{A}=\lambda\boldsymbol{A}^s+(1-\lambda)\boldsymbol{A}^f$，其中 λ 是比例调和因子，用于平衡结构与语义对于消息传递的影响。

表 1 DOM 树中节点特征及编码方式

特征类型	编码方式
标签	节点 HTML 的 84 类标签 one-hot 编码
属性类	节点属性类，文本向量模型编码
元素标识符	元素唯一标识符，文本向量模型编码
字体粗细	普通、加粗、变细三种类型 one-hot 编码
字体样式	普通、斜体、倾斜三类 one-hot 编码
字体大小	10 种字体大小 one-hot 编码
文本	文本向量模型编码

最终，通过图卷积神经网络实现节点的消息传递，本文采用两层图神经网络对图上信息进行聚合，得到融合表征

$$Z=\widetilde{A}ReLU(\widetilde{\boldsymbol{A}}XW_0)W_1。$$

其中，$\widetilde{\boldsymbol{A}}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$ 为归一化后的邻接矩阵，$\boldsymbol{D}$ 为度矩阵。为了得到网页级的表征，本文首先设计了一个独立的[CLS]虚拟节点作为全图表征聚合节点，图上的其他节点则按照广度优先搜索 DOM 树展开行程节点序列，进一步通过单层 Transformer 模型的自注意力机制来得到[CLS]节点的表征向量，作为全图的表征向量。进一步，通过多层感知机与 Softmax 非线性激活函数作为输出层进行分类。

3.2.3 网页筛选器 WebFilter 训练过程

针对给定的网页分类类别 $\mathcal{C}^W$，每个训练样本都有一个给定的标签(包含一类标签为其他，即不代表任何数据网页的字段信息)，并采用 one-hot 向量进行表示，其中只有正确类别的位置为 1，其余位置为 0。模型输出的是对于每个类别的预测概率，并通过交叉熵损失进行训练。与此同时，为了保证语义相似性邻居的稀疏性，通过 $L1$ 范数对 A^f 进行约束，从而得到 WebFilter 的训练损失

$$\mathcal{L}_{\mathcal{F}}=\alpha\,\mathcal{L}_c+(1-\alpha)\|\boldsymbol{A}^f\|。$$

其中，$\mathcal{L}_c$ 为交叉熵损失，α 为调整因子。

3.3 网页阅读器 WebReader

3.3.1 网页内节点异构关联建模

网页信息的结构化抽取的本质是对网页 DOM 树中的各类节点进行分类，即网页中的不同部分信息归属于标题、摘要、作者、发布时间、空间描述等各类数据字段。针对已经分类得到的数据网站，其网页节点的初始表征通过 3.2.1 节中所提方法得到。

为了进一步刻画网页内节点的异构关联，我们对 3.2.2 节的图建模方法进行了改进。相较于直接通过随机游走得到的结构共现性，本文提出了网页节点异构网络 $g_i^{\mathcal{H}}=(V_i,E_i^{\mathcal{H}})$，其中 V_i 依旧表示了网页 w_i 的 DOM 树节点，$E_i^{\mathcal{H}}$ 则表示了异构连边。具体本文定义了基于节点在 DOM 树中的连接关系，定义了父子关系、兄弟关系、祖先关系，并从前端渲染角度定义了垂直关系、水平关系、相同字体、相同背景。特别说明，父子关系、兄弟关系是指网页渲染前 DOM 树上节点间的结构关系，分别表示从属、并列结构关系；垂直关系、水平关系是指网页渲染后呈现的视觉排布效果，分别表示视觉上下、左右排列关系。不同的异构关联通过 one-hot 编码得到了每个边的属性 $\boldsymbol{\epsilon}_i^{\mathcal{H}}\in\mathbb{R}^{|E_i^{\mathcal{H}}|\times d_E}$。

3.3.2 基于异构图关联的消息传递

针对网页节点异构网络 $\mathcal{G}_i^{\mathcal{H}}$ 中的异构关联，无法直接使用同质图神经网络对不同连边关系进行刻画，需要将边属性进行协同考虑。同时，由于 DOM 树的树状结构，网络的平均路径长度较大，需要更深

的图神经网络进行消息传递,实现兼顾全局与局部的信息共享。因此,我们首先定义了对网络中的任意节点 v_p 的第 $l+1$ 层图神经网络的表征的消息传递机制满足:

$$x_p^{l+1}=\sigma\Big(\sum_{q\in N(p)}\mathrm{Agg}(x_p^0,x_p^{l-1},x_p^l,x_q^l,e_{pq}^l)\Big)。$$

其中,$\sigma(\cdot)$ 表示非线性激活函数,Agg 是聚合函数,N_p 是节点 v_p 的邻居集合,节点 v_p 的特征为 $x_p^l\in\mathbb{R}^d$,节点 v_p 与节点 v_q 的连边 $e_{pq}^l\in\mathbb{R}^{d_E}$。为了避免由于图神经网络堆积带来的过平滑问题,采用残差连接与密集连接的思想同时聚合初始层 x_p^0 与上一层的节点特征 x_p^{l-1}。本文采用了 CEN-DGC-NN 模型[37]的聚合函数进行特征融合,该模型同时对网络的节点特征、边特征进行更新,同时采用残差和密集连接的思想来聚合初始层和上一层的特征,通过跨层连接输出,有效缓解过度平滑并增加了网络深度。对于每一层图神经网络,第 l 层图神经网络的输出为

$$X^l=\sigma\cdot\Big[\prod_{p=1}^{P}M^l(\alpha_p^l(X^{l-1},E_p^{l-1})W^lX^{l-1},W^lX^{l-2},W^lX^0)\Big]。$$

其中,$M^l(X_1,X_2,X_3)=\xi X_1+\eta X_2+\theta X_3$ 表示聚合残差及初值的聚合函数,第一项包括了对第 $l-1$ 层特征的变化,特别是系数 $\alpha_p^l(X^{l-1},E_p^{l-1})$ 融合了节点与异构连边的注意力特征,对于相连的节点 i 与节点 j 其注意力权重的运算方式为

$$\alpha_{ijp}^l(X^{l-1},E_{ijp}^{l-1})=\frac{\exp(\mathrm{LeakyReLU}(WX_i^{l-1}\,\|\,WX_j^{l-1}))}{\sum_{k\in N_i}\exp(\mathrm{LeakyReLU}(WX_i^{l-1}\,\|\,WX_k^{l-1}))}E_{ijp}^{l-1}。$$

其中,E_{ijp}^{l-1} 表示第 $l-1$ 层节点 v_i 与节点 v_j 在特征通道 p 上的边特征。通过这一运算,实现了融合异构图关联的消息传递。

最终通过多层感知机与 Softmax 非线性激活函数作为输出层进行分类。

3.3.3　网页阅读器 WebReader 训练过程

由于网页节点标签的不平衡性,目标节点只占网页文本节点中的极少部分,因此网页阅读器 WebReader 选择 Focal Loss[38]作为网页节点分类任务的损失函数,其通过调节不同类别的权重因子让模型更关注难分类的样本,从而提升模型在样本不平衡时带来的训练困难的问题。具体损失函数的表达式为

$$\mathcal{L}_{\mathbb{R}}=\sum_{y=1}^{K_V}-w_y(1-p_y)^{\gamma}\log(p_y)。$$

其中,p_y 表示模型对于种类 y 的预测概率,w_y 为种类 y 的类别权重,由样本在训练数据中的反比例决定,使得模型将会更加关注这些少数类别的样本。γ 作为调节因子,用于让模型更关注难分类的样本,从而进一步解决样本不平衡可能导致的训练问题。

在本章中,我们采用公开数据集对所提出的科学数据机器挖掘与理解方法 WebFilter 和 WebReader 进行性能验证。

4　实验结果与分析

4.1　数据集说明

为了分别验证 WebFilter 的网页分类能力,以及 WebReader 的网页信息提取能力,我们均采用了广泛使用的公开数据集进行训练与测试。针对网页分类任务,采用了 WebKB[39](四所高校计算机科学系相关 6 类网页,约 8000 个网站)、SWDE[40]8 个不同领域的热门网站,总计 80 个网站和 124291 个网页)和 WebCLS[41](Kaggle 数据平台抓取的 16 个类别的 1408 个网页记录)数据集进行实验。针对网页信息提取任务,采用了 SWDE[40] 和 Klarna 商品网页数据集[42](来自 8175 个真实电子商务网站的 51701 个手动标记的产品页面)进行验证。SWDE 网页在 8 种不同类型网页下设置了 3－5 个属性作为结构化数据提取目标,而 Klarna 商品网页数据集希望可以提取 5 个网页元素,包括两个动作元素(购买按钮与购物车按钮)及 3 个信息元素(产品价格、名称与图片)。

4.2　网页分类性能评价

网页分类性能评价指标采用准确率($Acc.$)和 $F1$ 分数来进行评估,其中 $F1$ 分数是精确率和召回率的调和平均数,以综合评估两者的性能。数据集的训练、验证、测试集均按照 6∶2∶2 随机划分。

为了全面对比评估本文所提出的基于网页结构和语义的多视角表征学习方法,采用了一系列基线方法进行性能对比,包括仅利用文本属性的 BiLSTM[43]、TextCNN[23]、BERT[25]、RoBERTa[44],以及融合网页多视角特征的 RiSER[29]、DC-F[45]、SMGCN[46]、GROWN+UP[47]。对于所有数据集和基线算法,本文使用了固定的 10 个随机种子,并取每个评估指标的平均值与标准差作为最终结果。为保证公平和一致性,所有算法都采用相同的数据集分割和随机种子,参数配置也相同。

表 2 记录了网页分类任务在不同基线模型与本

文所提的 WebFilter 模型上的性能对比，并得到如下结论：首先，本文提出的 WebFilter 模型能够有效提升在网页分类任务上的性能。相较于文本处理算法和多视角网页处理算法，在准确率上提升 1.39% 到 3.71%，$F1$ 分数提升 1.42% 到 4.10%。其次，WebKB 和 SWDE 数据集结构信息较为单一，网页以文本作为主要是信息表达载体，仅依赖于纯文本方法（如 RoBERTa）已经可以取得较好的性能，多视角方法在准确率和 $F1$ 分数上提升幅度较小。对于结构特征相对丰富的 WebCLS 数据集，仅利用文本单一视角的方法由于忽略了网页结构以及标签信息，其表征能力有限，网页分类性能不如综合利用多视角信息的 GROWN＋UP 和 本文提出的 WebFilter 模型。

表 2 网页分类任务在 WebKB，SWDE，WebCLS 数据集上的性能评价 (%)

数据集	WebKB		SWDE		WebCLS	
模型/性能指标	*Acc.* ↑	*F1* ↑	*Acc.* ↑	*F1* ↑	*Acc.* ↑	*F1* ↑
BiLSTM[43]	85.26±0.44	86.44±0.82	94.12±0.12	94.74±0.20	84.37±0.45	86.37±0.72
TextCNN[23]	85.52±0.64	87.17±1.25	93.21±0.24	93.57±0.30	84.29±0.59	86.68±1.01
BERT[25]	91.23±0.34	91.15±0.35	97.86±0.08	98.01±0.21	92.88±0.29	94.08±0.52
RoBERTa[44]	92.75±0.35	91.47±0.47	97.67±0.12	98.31±0.27	92.68±0.30	94.12±0.52
RiSER[29]	85.96±1.13	88.35±2.21	95.15±0.34	95.87±0.78	92.06±0.59	93.61±0.92
DC-F[45]	87.21±0.65	87.51±1.47	95.25±0.23	95.96±0.54	90.23±0.84	91.24±1.12
SMGCN[46]	89.67±0.79	92.32±0.34	96.24±0.74	97.01±0.58	90.02±0.34	91.61±1.56
GROWN＋UP[47]	91.78±0.52	94.01±0.43	98.12±0.47	98.22±0.72	93.94±0.76	94.69±0.46
WebFilter	**94.83±0.21**	**96.24±0.38**	**99.50±0.12**	**99.73±0.25**	**97.56±0.23**	**98.74±0.64**

为了进一步验证 WebFilter 模型各个模块的有效性，从节点特征、多视图信息融合（结构邻近性、语义相关性）、网页级池化方法等设置了如图 4 所示的消融实验进行验证，其中平均池化是指使用平均池化代替本文提出的基于注意力的池化模块。实验结果表明每个模块的设置对于模型的有效性具有明显提升，特别是两类视图联合的信息融合才能充分表达网页级的表征。具体来说，由于标签属性提供了纯文本之外的其他信息，对整体实验结果有一定提升。其次，结构图和语义图作为特征增强模块，从两个视角构建 DOM Graph 节点间的信息传递，增强模型的表征能力。进一步地，在 WebKB 数据集上，由于多数网页结构特征较为简单，语义图的作用相较于结构图更显著。最后，特征融合模块默认为全图平均值池化，本实验中将 Transformer 层作用于特征融合，采用序列化方式融合多节点特征，更进一步提升网页分类准确性。

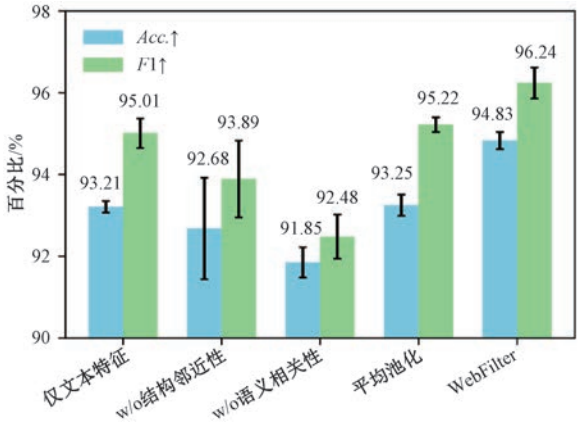

图 4 WebFilter 在 WebKB 数据集上的消融实验

最后，我们以 WebKB 数据集为例进行参数敏感性分析实验，包括结构图和语义图的权重系数 λ、分类损失和语义图稀疏性损失的权重系数 α 和结构图随机游走路径长度 K_{rw} 三个超参数，如图 5 所示。(1)结构图和语义图的权重系数：随着 λ 从 0 到 1 逐渐变大，模型准确率和 $F1$ 分数均为先升高后降低。由此可见，结构图和语义图在失去任何一个时均会降低模型性能，且 WebKB 数据集上当 $\lambda=0.25$ 时效果最佳，即语义图起到主导作用。(2)分类损失

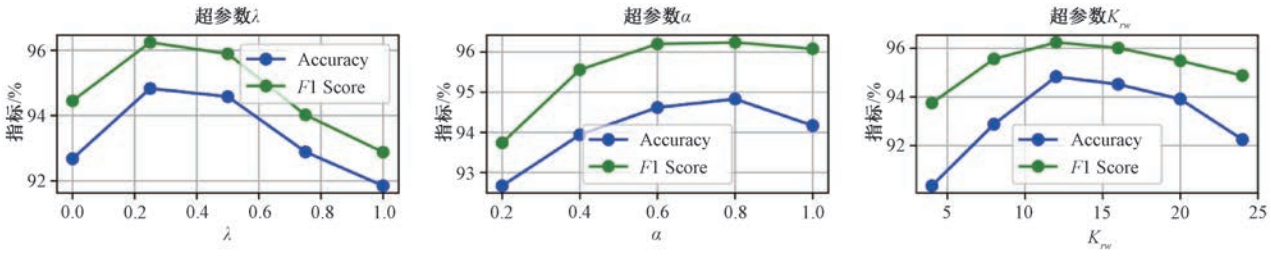

图 5 WebFilter 在 WebKB 数据集上的超参数实验

和语义图稀疏性损失的权重系数:当 α 较小时,模型更关注语义图的学习,相反则模型分类损失占据主导作用,综合来看 $\alpha=0.8$ 时性能最佳。(3)结构图随机游走路径长度:当 K_{rw} 比较小时,结构图只会关注到很小的局部结构,反之则会由于视野变大引入额外的噪声,在 WebKB 数据集上设置为 $K_{rw}=12$ 时结果最佳。

4.3　网页信息结构化抽取性能评价

针对网页信息抽取的指标采用了 F1 分数进行评价。为全面评估所提算法的有效性,实验中选择了倾斜堆叠模型 SSM[48],离散特征的节点分类方法 FreeDOM[32] 和 SimpDOM[33],基于 Transfor-mer 架构的 DOM-LM[30]、WebFormer[31]、Markup-LM[34] 和 WIERT[35] 方法进行对比。

在网页级节点信息抽取任务中,数据集划分方式与网页分类任务不同,由于目标为衡量算法在不同结构网页间的准确性与泛化性能,从而可以减少网页模版的标注数量。因此在数据集划分时选择不同的网站种类作为训练/验证集、测试集,并选择不同的比例进行多次实验从而验证模型在不同网页间的泛化性能。例如,对于 SWDE 数据集中,选择 k 个网站作为训练集/验证集、余下网站作为测试集,其中训练集和测试集的比例为 8∶2。对于 Klarna 商品网页数据集,由于其网站数目众多,因此通过比例 p 代替数量进行数据集划分。

表 3 展示了网页信息抽取的任务性能,并得出如下结论:WebReader 在两类数据集的所有设置中均取得了最优效果,相较于次优模型性能平均提升 1.40%,展现了所提方法的有效性与跨网页的泛化性能。通过对比可以发现,各类方法随着种子网站数量的上升,性能表现均有所提升,这说明当模型在更多网页模版中训练之后,在面对新网页时会显示出更强的泛化性。在 $k=1$ 与 $p=0.1$ 的少标注样本设置下,WebReader 在 SWDE 和 Klarna 数据集上的性能提升分别为 1.92%、2.33%,优势明显。由于 Klarna 商品网页数据集的网站个数更多,这一趋势也更为显著。对比基于离散节点分类的方式 FreeDOM、SimpDOM 与基于预训练方法 DOM-LM、MarkupLM 和 WIERT 的表现,可以发现在 SWDE 数据集上两类方法各有优劣,前者通过节点的有效表征实现良好的分类性能,后者即使没有针对节点级别的处理,但由于其预训练数据庞大,网络结构丰富,在处理该问题时可以取得较好的效果。在 Klarna 商品数据网页中,由于网页模版数量更多,使得基于 Trans-former 的方法展示出了更好的鲁棒性。本文所提出的 WebReader 算法结合了两者之间的优势,在文本表征方面使用预训练 BERT 模型来保证向量表征能力以及不同网站文本之间的泛化性,且节点表征与边表征中均加入了与节点相关的离散信息,使得模型更好地理解网页结构。

表 3　网页信息抽取任务在 SWDE,Klarna 商品网页数据集上的性能评价

数据集	SWDE					Klarna 商品网页数据集				
模型/种子网站	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$p=0.1$	$p=0.2$	$p=0.3$	$p=0.4$	$p=0.5$
SSM[48]	62.53	64.50	69.29	72.70	74.87	52.88	56.63	60.39	63.89	65.33
FreeDOM[32]	82.32	86.36	90.49	91.29	92.56	75.33	78.26	81.47	85.50	88.93
SimpDOM[33]	83.06	88.96	91.63	92.84	93.75	77.02	80.53	84.88	88.97	92.48
DOM-LM[30]	80.63	85.77	89.20	91.57	93.28	79.84	83.90	88.66	91.25	94.10
WebFormerl[31]	81.08	85.32	90.37	91.80	92.99	79.30	84.02	88.97	91.08	93.77
MarkupLM[34]	83.22	87.53	90.28	92.87	94.03	81.08	86.88	90.25	93.01	94.80
WIERTI[35]	83.66	87.91	91.54	93.02	94.14	80.23	85.80	89.54	92.77	94.53
WebReader	**85.27**	**89.88**	**92.53**	**94.50**	**96.12**	**82.97**	**87.24**	**91.85**	**94.10**	**95.44**

在当今大语言模型逐渐普及,并在多类自然语言处理及其他领域中取得最优表现的背景下,为验证网页节点级信息抽取任务的研究价值以及本文所提出方法的必要性,我们使用 SWDE 数据集在开源/闭源大模型上进行测试验证其效果。由于大语言模型输入文本长度的限制以及成本问题,因此首先对网页 HTML 文本进行处理,仅保留其文本节点的 Xpath 信息和文本信息,形成一个由字典组成的列表,字典的键为每个文本节点的 Xpath,值为其对应内容,并编写以下提示语用于大语言模型的推理:

Prompt: You're an expert at web data mining as well as information retrieval, and you dabble in a wide range of information in any field. Here is all the text from a web page that I provided to you, along with its corresponding Xpath information: 'xpath1':'text1','xpath2':'text2',…

I would like you to use your understanding of

web pages and your refactoring skills to extract the ＜attribute＞ that corresponds to the ＜vertical＞ in this web page. Please note that you can only return information about one text node with a json format like "＜attribute＞":"＜your answer＞", which is one of the values in the dictionary above, and be sure not to return too much information.

其中＜vertical＞表示所选网页的领域，例如书籍，＜attribute＞表示提取目标，例如书籍的名称或价格，实验使用闭源大语言模型 GPT-4 以及开源模型 Llama3-70B-Instruct 与我们提出的 WebReader（$k=3$）进行对比，其在 SWDE 中各个领域的 F1 分数如表 4 所示，其结果表明，即使是目前较为优秀的大语言模型，在网页级节点信息抽取任务上也很难达到令人满意的水平，本文对其原因进行案例分析，发现大语言模型的回复中经常加入总结类语言、与任务无关的解释，或产生错误与幻觉。由此可见，即使在大语言模型时代，针对网页级节点信息抽取的研究依然存在重要价值。

表 4 大语言模型在 SWDE 数据集上的信息抽取能力表现

模型	auto	university	camera	movie	job	book	restaurant	nbaplayer	平均性能
Llama3-70B	45.22	42.97	52.03	56.77	51.20	38.47	47.29	35.81	**46.23**
GPT-4	42.01	48.77	55.64	52.30	60.77	37.61	49.84	34.87	**47.73**
WebReader($k=3$)	90.83	93.70	95.01	95.55	94.35	89.37	91.79	89.65	**92.53**

5 DataExpo 系统及在地球科学领域应用

基于上述技术研究成果，本文提出的方法可以有效地对网页数据进行分类，筛选出相关的科学数据，并通过信息抽取对不同网站科学数据建立元数据画像，实现统一的科学数据管理。

在“深时数字地球”国际大科学计划的号召下，基于本文提出的技术研制了面向地球科学领域的学术科学元数据库及相应的 DataExpo 系统（https://dataexpo.deeptime.org/）。图 6 展示了系统的主要界面包括主页、数据多维查询页、数据地图查询页。

根据地球科学学科专家提供的大规模地球科学学科知识体系[49]作为系统的关键词输入，整个关键词库涵盖了如地层学、沉积学、古地理、矿物学、地质制图等 18 个学科方向，并进一步基于 WebFilter 发现了领域内超百万地球科学数据。以关键词“Geochemistry（地球化学）”为例，表 5 是 DataExpo 系统发现并返回的相关地球科学数据，其中既包括了一些科学数据平台，还包括了一些政府组织构建的公开数据库、科学文献关联的数据库等，数据库的空间范围也包括了中国、美国、英国等，还有更多长尾的网站也被搜寻发现并链接在 DataExpo 系统中，供用户一站式检索。

（a）系统主界面 （b）数据多维查询 （c）数据地图查询

图 6 数据巡航 DataExpo 系统界面

表 5 地球化学关键词下的高频数据网站域名

网站链接	数据平台
doi.pangaeade	地学数据共享发布平台 PANGAEA
data.mendeley.com	开放数据存储库 Mendeley Data
data.gov.au	澳大利亚政府开放数据门户
www.sciencebase.gov	美国地质调查局科学元数据存储库
www.gbif.org	全球生物多样性信息设施 GBIF
frontiersin.figshare.com	Frontiers 期刊的开放数据共享平台
data.tpdc.ac.cn	国家青藏高原科学数据中心
figshare.com	Figshare 开放研究数据存储库
geolsoc.figshare.com	伦敦地质学会开放数据存储库
earthreforg	地球科学参考数据和模型门户网站

进一步，表 6 对比了一些地球科学领域的热点话题关键词在 Google Dataset Search（GDS）与本文研制的 DataExpo 系统的检索结果，可以对比发现 DataExpo 技术在地球科学领域的检索结果更多、更加专业，也在实际与地球科学家开展合作的过程中获得了认可。以花岗岩为例，在 GDS 上一共检索到 181 个数据库，而在 DataExpo 上可以检索到 2,383 个数据库，这是由于大量花岗岩数据库由个人科学家构建，缺少标准化 Schema 并未纳入 GDS

表 6　不同关键词 Google GDS 与 DataExpo 检索结果数量对比

关键词	Google GDS 检索结果数量	DataExpo 系统检索结果数量
granite(花岗岩)	181	2,383
zircon(锆石)	163	3,871
wildfire(野火)	142	502
landslide(滑坡)	157	1,391
tsunami(海啸)	132	687
glacial lake(冰川湖)	129	245
trace fossils(遗迹化石)	122	1,192

系统中。我们提出的方法通过高效的机器阅读,实现了更高的数据召回。

针对获取的地球科学数据,针对地学数据管理的需要,我们采用网页解析与 WebReader 技术对地学数据网站构建元数据画像,对数据网站的标题、摘要、关键词、缩略图、时间、地点、发布机构等各类重要字段进行抽取。基于上述字段在数据多维查询时,可以对检索结果进行约束和可视化分析(图 6—中)。同时,进一步通过时空解析将不同数据投影在地图上便于地学科学家进行地图检索(图 6—右),图 7 展示了一个地图检索的元数据结果及其与原始网页的对应关系。

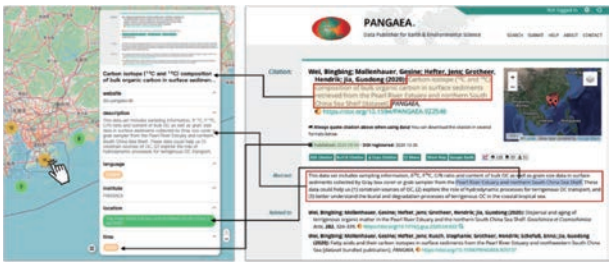

图 7　地学数据网页信息抽取案例

6　总结与展望

大数据与人工智能时代,科学数据的快速增长正推动着科学研究范式的变革。尽管科学数据的数量快速增长,但是科学家发现研究相关数据依旧面临着数据分散、检索效率低的问题。本文从计算机领域视角出发,提出了面向开放互联网的开放科学数据挖掘与理解方法:设计了网页筛选器 WebFilter 对互联网上获取的各类信息进行分类,筛选得到相关的科学数据资源;进一步设计了网页阅读器 WebReader 对网页中的重要字段信息进行提取,实现对科学数据资源的细粒度理解。公开数据集实验表明我们所提方法均取得了最优性能表现,在网页信息抽取方面优于 GPT—4 等大模型提取效果。同时,基于技术研究成果,面向地球科学领域研制了 DataExpo 系统,汇聚全球超百万地学科学数据,并支撑了数据检索查询等各类数据服务。

面向未来,随着大模型技术的普及与广泛应用,如何利用大模型增强领域理解能力、消除幻觉,实现对网页进行高效准确的理解,并进一步构建智能体进行任务编排与调度是一个值得探索的方向。同时,未来工作将聚焦面向科学数据的智能服务、多模态科学数据融合开展研究,促进科学数据的综合应用,推动以地球科学为代表的 AI for Science 领域取得突破。

致　谢　感谢为本文稿件提供宝贵修改意见的审稿人与编辑。感谢国家自然科学基金、国家资助博士后研究人员计划对本文工作的资助。本论文是“深时数字地球”(Deep-time Digital Earth, DDE)国际大科学计划的系列成果之一。

参　考　文　献

[1] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596(7873): 583-589

[2] Lam R, Sanchez-Gonzalez A, Willson M, et al. Learning skillful medium-range globalweather forecasting. Science, 2023, 382(6677): 1416-1421

[3] Li Guojie. Intelligent research (AI4R): The fifth research paradigm. Bulletin of the Chinese Academy of Sciences, 2024, 39(1): 1-9 (in Chinese)
(李国杰. 智能化科研(AI4R):第五科研范式. 中国科学院院刊, 2024, 39(1): 1-9)

[4] Wang J L, Li Y, Wang S Q, et al. Global impact analysis of FAIR principles and suggestions for their implementation strategies. Chinese Science Bulletin, 2024, 69(9): 1183-1191 (in Chinese)
(王卷乐,李扬,王淑强,等. FAIR 原则全球影响分析及其实施策略建议。科学通报,2024,69(9):1183-1191)

[5] Brase J. Datacite-a global registration agency for research data//2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology. Beijing,China,2009: 257-261

[6] Michener W, Vieglais D, Vision T, et al. Dataone: Data observation network for earth—preserving data and enabling innovation in the biological and environmental sciences. D-Lib Magazine, 2011, 17(1/2): 12

[7] Chen X, Gururaj A E, Ozyurt B, et al. Datamed—an open source discovery index for finding biomedical datasets. Journal

of the American Medical Informatics Association, 2018, 25 (3): 300-308

[8] Brickley D, Burgess M, Noy N. Google dataset search: Building a search engine for datasets in an open web ecosystem//Proceedings of the World Wide Web Conference. San Francisco, USA, 2019: 1365-1375

[9] Benjelloun O, Chen S, Noy N. Google dataset search by the numbers//Proceedings of the International Semantic Web Conference. Cham, Switzerland, 2020: 667-682

[10] Guha R V, Brickley D, Macbeth S. Schema.org: evolution of structured data on the web. Communications of the ACM, 2016, 59(2): 44-51

[11] KRUTIL J, KUDĚLKA M, SNÁŠEL V. Web page classification based on Schema.org collection//2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN). 2012: 356-360

[12] Wilkinson M D, Dumontier M, Aalbersberg I J, et al. The fair guiding principles for scientific data management and stewardship. Scientific Data, 2016, 3(1): 1-9

[13] Datadryad. Dryad digital repository. 2024. https://datadryad.org/

[14] Pangaea. Pangaea-data publisher for earth & environmental science. 2024. https://www.pangaea.de/

[15] Zenodo. Zenodo. 2024. https://zenodo.org/

[16] Luo P C, Wang J M, Nie L. Research progress on unified discovery platforms for open scientific datasets. Journal of the China Society for Scientific and Technical Information, 2022, 41 (6): 637-650 (in Chinese)
(罗鹏程,王继民,聂磊。开放科学数据集的统一发现平台研究进展。情报学报,2022,41(6):637-650)

[17] Wang C, Hazen R M, Cheng Q, et al. Thedeep-time digital earth program: data-driven discovery in geosciences. National Science Review, 2021, 8(9): nwab027

[18] Neumaier S, Umbrich J, Polleres A. Lifting data portals to the web of data//LDOW@WWW. 2017

[19] Socrata. The socrata open data api. 2024. https://dev.socrata.com

[20] Zhang H, Santos A, Freire J. Dsdd: Domain-specific dataset discovery on the web//Proceedings of the30th ACM International Conference on Information & Knowledge Management. Queensland, Australia, 2021: 2527-2536

[21] Luhn H P. The automatic creation of literature abstracts. IBM Journal of Research and Development, 1958, 2(2): 159-165

[22] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space//Proceedings of the 1st International Conference on Learning Representations, ICLR 2013. Scottsdale, USA,2013

[23] Kim Y. Convolutional neural networks for sentence classification//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014. Baltimore, USA, 2014: 1746-1751

[24] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014. Baltimore, USA,2014,1: 655-665

[25] Devlin J, Chang M, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019. Minneapolis, USA,2019,1: 4171-4186

[26] Liu Z, Lin W, Shi Y, et al. A robustly optimized bert pretraining approach with post-training//Proceedings of China National Conference on Chinese Computational Linguistics. Huhhot, China, 2021: 471-484

[27] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding//Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019. Vancouver, BC, Canada. 2019: 5754-5764

[28] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 2020, 21(140): 1-67

[29] Kocayusufoglu F, Sheng Y, Vo N, et al. Riser:Learning better representations for richly structured emails//Proceedings of the World Wide Web Conference. San Francisco, USA, 2019: 886-895

[30] Deng X, Shiralkar P, Lockard C, et al. Dom-lm: Learning generalizable representations for html documents. arXiv preprint arXiv:2022

[31] Wang Q, Fang Y, Ravula A, et al. Webformer: The webpage transformer for structure information extraction//Proceedings of the ACM Web Conference. Lyon, France, 2022: 3124-3133

[32] Lin B Y, Sheng Y, Vo N, et al. Freedom: A transferable neural architecture for structured information extraction on web documents//Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Virtual, USA, 2020: 1092-1102

[33] Zhou Y, Sheng Y, Vo N, et al. Simplified dom trees for transferable attribute extraction from the web. arXiv preprint arXiv:2101.02415 (2021)

[34] Li J, Xu Y, Cui L, et al. Markuplm: Pre-training of text and markup language for visually rich document understanding//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022. Dublin, Ireland, 2022: 6078-6087

[35] Li Z, Shao B, Shou L, et al. WIERT: web information extraction via render tree//Thirty-Seventh AAAI Conference

on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023. Washington, USA, 2023: 13166-13173

[36] Xu H, Chen L, Zhao Z, et al. Hierarchical multimodal pre-training for visually rich webpage understanding//Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM 2024. Merida, Mexico, 2024: 864-872

[37] Zhou Y, Huo H, Hou Z, et al. Co-embedding of edges and nodes with deep graph convolutional neural networks. Scientific Reports, 2023, 13(1): 16966

[38] Lin T, Goyal P, Girshick R B, et al. Focal loss for dense object detection//Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017. Venice, Italy, 2017: 2999-3007

[39] Chen X, Chen S, Xue H, et al. A unified dimensionality reduction framework for semi-paired and semi-supervised multi-view data. Pattern Recognit., 2012, 45(5): 2005-2018

[40] Hao Q, Cai R, Pang Y, et al. From one tree to a forest: a unified solution for structured web data extraction//Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011. Beijing, China, 2011: 775-784

[41] Prettenhofer P, Stein B. Cross-Language Text Classification using Structural Correspondence Learning// Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics (ACL 2010). Uppsala, Sweden, 2010: 1118-1127

[42] Hotti A, Risuleo R S, Magureanu S, et al. The klarna product page dataset: Web element nomination with graph neural networks and large language models. Transactions on Machine Learning Research, 2024, 2024: 1-17

[43] Schuster M, Paliwal K K. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681

[44] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized BERT pretraining approach. CoRR, 2019, abs/1907.11692

[45] Alrashed T, Paparas D, Benjelloun O, et al. Dataset or not? A study on the veracity of semantic markup for dataset pages//Proceedings of the 20th International Semantic Web Conference (ISWC 2021). Virtual, 2021: 338-356

[46] Wu F, Jing X, Wei P, et al. Semi-supervised multi-view graph convolutional networks with application to webpage classification. Information Sciences, 2022, 591: 142-154

[47] Yeoh B, Wang H. GROWN+UP: A "graph representation of a webpage" network utilizing pre-training//Proceedings of the 31st ACM International Conference on Information & Knowledge Management. Atlanta, USA, 2022: 2372-2382

[48] Carlson A, Schafer C. Bootstrapping information extraction from semi-structured web pages//Lecture Notes in Computer Science: Vol. 5211 Machine Learning and Knowledge Discovery in Databases, European Conference, ECML/PKDD 2008, Antwerp, Belgium, 2008: 195-210

[49] Shi S Z, Lü H R, Dong S C, et al. Editing platform for geoscience knowledge system. Geological Journal of China Universities, 2020, 26(4): 384-394. DOI:10.16108/j.issn1006-7493.2020019 (in Chinese)
(石顺中,闾海荣,董少春,等。地球科学知识体系编辑平台. 高校地质学报, 2020, 26(4): 384-394. DOI: 10.16108/j.issn1006-7493.2020019)

LU Bin, Ph. D., postdoctoral researcher. His research interests include graph neural network, AI for Science.

GAN Xiao-Ying, professor, Ph. D. supervisor. Her research interests include Internet of Things data mining, spatio-temporal computing.

GAN Yu, master candidate. His research interests is webpage data mining.

TANG Gu, Ph. D. candidate. His research interests include recommender system, text-attributed graph neural network.

MA Ting-Yan, master candidate. Her research interests include scientific data, graph neural network.

WU Lv-Wen, Ph. D. candidate. Her research interests include recommender system, text-attributed graph neural network.

ZHAO Ze, Ph. D. candidate. His research interests include knowledge graph, graph neural network.

FU Luo-Yi, Ph. D., associate professor, Ph. D. supervisor. Her research interests include network analysis, network representation.

JIN Meng, Ph. D., associate professor, Ph. D. supervisor. Her research interests include Internet of Things, AI for Science.

WANG Xin-Bing, Ph. D., distinguished professor, Ph. D. supervisor. His research interests include Internet of Things, Big Data, AI for Science.

ZHOU Cheng-Hu, Ph. D., professor, Ph. D. supervisor. His research interests include geographic information system, AI for geoscience.

Background

The rapid development of artificial intelligence technology is bringing transformative impacts to the field of basic scientific research. Scientific data, as a core element of artificial intelligence, is the cornerstone driving scientific research. However, currently, open scientific data is scattered across multiple data repositories on the internet and numerous personal databases created by individual scientists, forming "data silos." The challenge of unified data discovery and management, along with providing efficient open science data services, is a crucial issue for advancing the next generation of scientific data infrastructure.

Due to the complexity of online resources, early efforts mainly relied on API integrations to aggregate data resources from a few vertical and specific fields, resulting in integrated data search platforms such as DataCite, DataONE, and DataMed. However, many long-tail scientific datasets lack APIs for direct integration, limiting the scale of scientific databases. Another representative effort is Google Dataset Search, which uses Google's large-scale web data resources and identifies web pages based on Schema @ type standard fields written during the development of some web pages, aggregating dataset pages. However, statistics reveal that approximately 70% of websites do not provide Schema, and some studies have shown that 61% of websites containing Schema.org/Dataset are not data websites, making schema-based data discovery methods unreliable.

This paper proposes an open scientific data mining and understanding method for the internet. For the large number of web resources obtained by search engines, we designed a deep learning model to automatically select scientific data pages, conduct web page understanding to extract standardized, enriched metadata to characterize scientific data, and support flexible and diverse search and queries. Using the Earth sciences as an example, we applied the proposed method and developed the DataExpo system (https://dataexpo.deep-time.org/), which discovered and aggregated over one million Earth science metadata, covering nearly 30,000 institutions, with data source IPs from over 120 countries and regions. It supports information extraction from more than 10 multimodal fields on data web pages and forms metadata profiles.

This research supports the Deep-time Digital Earth (DDE) international big science program, which aims to promote the transformation of Earth science research paradigms through data and knowledge-driven approaches.

This work was supported by National Natural Science Foundation of China (No. 623B2071, 62272301, T2421002), Postdoctoral Fellowship Program of CPSF under Grant Number No. GZB20250806.