第 60 卷 第 3 期
2024 年 5 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 60, No. 3
May, 2024

DOI:10.13232/j.cnki.jnju.2024.03.005

多视角网页分类数据集构建及性能评估

孙辰星[1],刘 伟[1],卢 彬[1],梁诗宇[1],诸云强[2],甘小莺[1*]

(1.上海交通大学电子信息与电气工程学院,上海,200240;2.中国科学院地理科学与资源研究所,北京,100101)

摘 要:网页分类是互联网数据挖掘中的一项重要任务,在信息搜索、推荐系统和知识发现等领域发挥着关键作用.然而,现有的公开网页数据集缺乏多视角信息,难以适用于蕴含复杂特征的网页分类任务.针对上述问题,基于"收集-处理-标注"构建流程,提出一个涵盖文本语义、网页结构等多视角特征的网页数据集Web-Minds,该数据集包含600余个门户网站下的21828条网页.首先,在开放互联网中通过关键词检索采集得到相关网页数据;其次,使用网页解析工具对收集的数据中的文本、DOM结构树、关键词等多视角信息进行提取与清洗;最后,采用大语言模型与"人在回路"的联合标注策略,形成网页类型与网页主题两种标签.在此基础上,针对Web-Minds数据集,测试评估了机器学习、文本分类和网页分类多种算法,结果表明,综合利用多视角特征能有效提升算法的准确率,和仅应用单视角特征相比,在网页类型和主题分类任务上,准确率分别提升了5.49%和5.61%.

关键词:网页数据集,网页分类,文本分类,数据挖掘,深度学习

中图分类号:TP301 **文献标志码:**A

Multi-view webpage classification dataset construction and evaluation

Sun Chenxing[1], Liu Wei[1], Lu Bin[1], Liang Shiyu[1], Zhu Yunqiang[2], Gan Xiaoying[1*]

(1. School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China;
2. Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, China)

Abstract: Webpage classification is an important task in Internet data mining, playing a crucial role in information retrieval, recommendation systems, and knowledge discovery, etc. However, existing public webpage datasets suffer from limitations such as scarcity, single sources and insuffcient information, which hinder the development of webpage classification techniques. To address these issues, we propose a publicly available dataset for webpage classification called Web-Minds, incorporating multi-view features by designing a three-step process of "collection-processing-annotation". Specifically, the relevant webpage data are collected and integrated from the open Internet. Then, a webpage parsing tool is employed to extract and clean multi-view information from the collected data, including text, structure, keywords, etc. We design a large language model and a "human-in-the-loop" annotation strategy to assign two types of labels, namely webpage type and webpage topic. Furthermore, we establish an algorithmic evaluation benchmark based on the Web-Minds dataset, containing such methods as machine learning, text classification, and webpage classification. The results demonstrate that compared to using single-view features alone, the comprehensive utilization of multi-view features significantly improves algorithm accuracy, with an increase of 5.49% and 5.61% in webpage type and topic classification tasks, respectively.

Key words: webpage dataset, webpage classification, text classification, data mining, deep learning

基金项目:国家重点研发计划(2022YFB3904204),国家自然科学基金(62272301,42050105,62020106005,62061146002,61960206002)
收稿日期:2023-11-11

*通讯联系人,E-mail:ganxiaoying@sjtu.edu.cn

互联网技术的飞速发展使各式各样的网页成为获取信息的主要来源,如今,网页种类复杂多样,网页数量呈爆炸性增长,极大地刺激了网页数据挖掘的蓬勃发展.作为网页数据挖掘的一项基本任务,网页分类旨在依据内容、形式对网页进行归类.在互联网搜索[1-3]领域,应用网页分类可以大幅提高搜索结果的质量,在诸如网页推荐[4-5]、开放数据发现[6-7]等实际应用领域,网页分类同样发挥着重要的作用.

随着大规模网页文本语料库的出现以及深度学习技术在自然语言处理等领域的发展,网页分类任务在过去几年取得了很大的进步.本文梳理了近十年网页分类相关的系列工作,对其应用数据集和模型算法进行分析.数据集方面,当前网页分类算法的数据集使用比例如图1所示,公开数据项目ODP提供的DMOZ-50[8]、卡内基梅隆大学提供的WebKB[9]以及Kushmerick[10]提供的AD数据集得到了广泛应用,但72.89%的研究者偏向使用个人收集的数据进行实验,而这些数据往往采集方式模糊且不开源,难以形成统一的测评基准.另外,随着网页分类研究的不断深入,算法性能不断提高,如图2所示,Deng et al[11]和Kipf and Welling[12]提出的算法在DMOZ-50,WebKB及AD数据集上达到了95%以上的准确率.

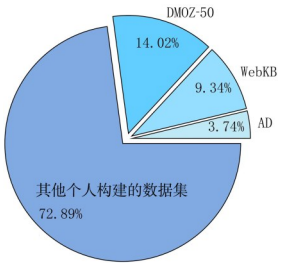

图1　网页分类数据集的使用率分布

Fig. 1　Distribution of usage rates of webpage datasets

在信息科学技术与自然科学互促发展的趋势下,由我国科学家主导的国际大科学计划"深时数字地球"项目[13]正在借助网页挖掘技术构建一个全球共享的地学数据网站检索平台,其中,数据网站的分类是平台构建的核心技术.针对此问题,本文联合领域专家选取一批数据集网站的正负样例进行实验,如图2所示,各类算法性能均有20%左右的明显下降.这可能是因为WebKB[9]等公开数据集的结构关系简单,所以基于文本信息进行表征即可达到较高的网页分类性能.然而,开放域中存在大量结构复杂、布局多样的网页.通过对比网页结构标签节点个数,本文对WebKB与测试数据进行了网页结构复杂度的对比分析.如图3所示,测试数据网页的平均结构标签节点个数为573,约为WebKB的5倍,即前者具有更高的结构复杂度.因此,融合更多视角的网页信息(例如DOM结构树)有利于更好地刻画开放域网页特征,提升分类任务性能,然而,现有公开数据集均未提供DOM结构树等网页特征,一个涵盖多视角信息的网页分类数据集亟待提出.

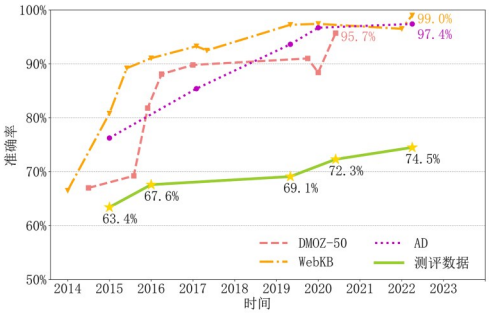

图2　网页分类算法在不同数据集上的表现

Fig. 2　Performance of webpage classification algorithms on different webpage datasets

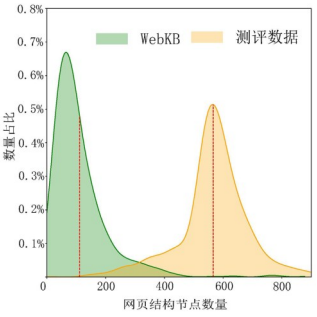

图3　WebKB与测评数据网页结构标签节点的分布

Fig. 3　HTML tag distribution of WebKB and evaluation data

基于以上分析,本文提出一个用于网页分类的多视角网页数据集Web-Minds(Webpage with Multi-View Information Dataset),收集了来自600余个门户网站的21828条相关网页.构建流程包括数据收集、数据处理及数据标注三个步骤,提供如图4所示的纯文本、DOM结构树、关键词等多视角网页表征信息.作为网页分类数据集,Web-

Minds数据标签内容包括网页类型信息,即数据网页与非数据网页以及网页主题信息,有地质学、地球物理学、地理学和地质资源四个主题. 和现有数据集相比,Web-Minds更注重网页多样性以及网页和文本不同的结构多样性. 此外,作为一个公开网页数据集,Web-Minds的每个样本都经过先验知识标注及专家验证,确保数据真实可靠. 为了评估各网页分类算法在Web-Minds上的表现,本文针对Web-Minds提供的两类标签设计了网页类型分类与网页主题分类两种任务. 同时,为了证明使用多视角信息能提升网页分类算法的性能,分别采用单视角信息与多视角信息进行对比实验,后者准确率比前者提升5.61%. 最后,针对开放域中域名分布偏移与类别不均衡问题进行了广泛研究与深入分析,为研究人员后续的模型设计与性能调优提供数据参考.

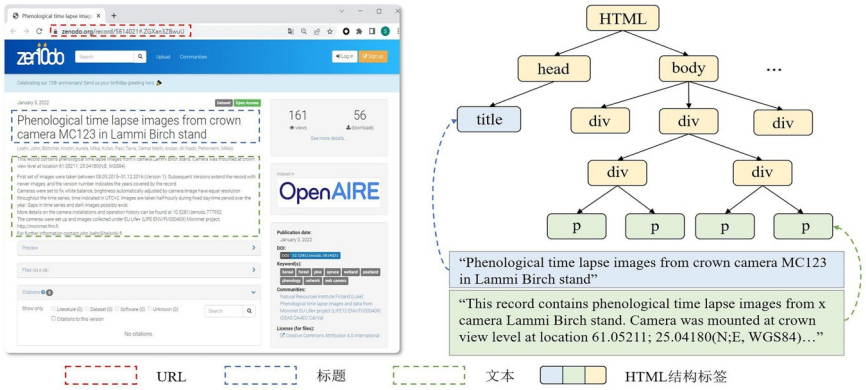

图4 Web-Minds数据集提供的网页URL链接、纯文本、DOM结构树、网页标题、网页关键词信息

Fig. 4 Web page URL links, plain text, DOM structure tree, web page titles, web page keyword information from Web-Minds dataset

本文的主要贡献如下.

(1)提出一个多视角网页分类数据集Web-Minds,其网页数据来源多样,包含来自600多个门户网站的20000多条网页数据,标注信息真实可靠,提供专家标注的网页类型和网页主题信息.

(2)Web-Minds提供网页语义文本信息与结构信息,包括网页纯文本信息、网页标题、DOM结构树、网页关键词等一系列多视角网页属性信息,全方位刻画网页特征.

(3)Web-Minds支持多种网页数据挖掘任务,为研究人员提供数据支撑. 本文在网页类型与主题分类上通过先进的分类算法进行性能评估,证明多视角特征对于网页分类性能有显著增益.

1 相关工作

1.1 网页数据集 网页数据集是网页分类的基础,广泛应用于网页数据挖掘. DMOZ是全球学术志愿者建立并维护的公共开放目录项目,DMOZ-50[8]是来自DMOZ网站的50个子数据集,包含3~10个类别,如艺术、运动、科学、购物等. DMOZ-50的内容以纯文本为主,其网页数据主要为门户网站首页内容信息. WebKB[9]来源于卡内基梅隆大学语言学习实验室主导的世界知识库项目,其网页数据来自四所高校计算机科学系,根据内容分为学生、教师、员工、系、课程、项目和其他,其中常用版本为课程网页与非课程网页. MGC数据集[14]收集开放互联网上的1539个英文网页,这些网页被标记为博客(Blog)、个人主页(Personal)、诗歌(Poetry)等. AD数据集[10]包含3279个网页,分458个广告网页与2821个非广告网页,目前的公开版本为预处理后得到特征向量表示,包含网页URL链接、超链接跳转信息和图片链接三种特征. 以上公开数据集均以网页文本或URL链接为主,常用于文本分类、文本理解等相关任务.

1.2 网页分类算法 近年来,网页分类问题已被国内外学者广泛研究. Kocayusufoglu et al[15]提出垃圾邮件分类模型RiSER,通过对邮件文本内容与布局结构联合编码训练分类器来实现对垃圾邮件的识别与过滤. Alrashed et al[16]提出DC-F算法,利用网页标题、描述等元数据信息训练多层感

知机来鉴别谷歌数据集标签的真假. 基于卷积神经网络的WebCNN[6]主要依赖网页的URL链接与文本特征,用于开放数据发现中的数据网页分类任务.

由于网页的多视角特性,一些多视角学习[17−18]方法被专门设计用于网页分类任务. Jing et al[19]和Wu et al[20]提出一种半监督的多视角学习方法,通过学习不同视角间与视角内的特征关联来强化网页表征. Jia et al[21]利用多视角间的一致性和互补性,设计了半监督多视角深度对比表征学习框架,通过对抗相似性约束与损失来实现对网页多视角的综合利用,并解决视角间冗余问题. 最近,Kipf and Welling[12]设计了半监督的多视角图卷积网络SMGCN,为每个视角获得最佳的图结构,并通过图卷积神经网络来学习多视角表征,提升网页分类性能.

2　Web-Minds数据集

2.1　数据集构建　Web-Minds的构建流程如图5所示,分三个步骤:数据收集,即利用领域专家提供的关键词在开放域进行搜集,并去除失效网页、垃圾网页等,得到初步的相关网页;数据处理,即将原始网页经过处理获取所需各类网页属性信息;数据标注,即专家进行数据标签标注.

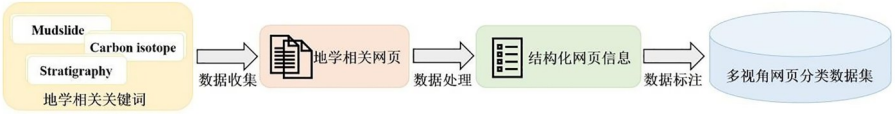

图5　Web-Minds的构建流程

Fig. 5　The construction pipeline of Web-Minds

2.1.1　数据收集　由于开放域网页信息具有多样性和隐蔽性,业外人士想要大量获取精准的、可靠的特定专业网页较困难. 因此,Web-Minds在构建初期咨询了地球科学学科的国内外专家,经过去重、名词去复数等词级别操作,整理获得了一个精确可靠的学科关键词库,包含五千多个相关关键词,如泥石流(Mudslide)、碳同位素(Carbon isotope)、地层学(Stratigraphy)等,同时涵盖了地球科学下的地质学、地球物理学、地理学等子学科. Web-Minds以该关键词库为依据,在开放域海量的网页中进行搜索整合,收集了大量相关网页,记为$\mathcal{S}$. 由于网页质量参差不齐,存在无法访问、源代码不完整、实际内容与检索标题不匹配等问题,通过检查错误代码、分析源代码完整性、人工访问等方式,最终获取有效网页数据$\mathcal{W}_s$,包含来自616个门户网站的21828个网页.

2.1.2　数据处理　在数据收集阶段获取的有效网页数据$\mathcal{W}_s$仅含有网页基本信息,如网页链接(URL)与网页源代码(HTML),而Web-Minds希望为使用者提供更多可以表征网页的属性,包括网页标题、网页纯文本、网页结构(DOM树)、网页关键词. 然而,在获取以上各类信息时,由于$\mathcal{W}_s$中网页结构复杂多样,无法通过规则匹配等方式获取网页属性信息,因此,本文设计了两步法进行数据处理,分别为源数据解析与元数据清洗.

2.1.2.1　源数据解析　早期网页分类工作通常仅使用网页链接w_{URL}作为分类依据,忽视网页内部包含的大量信息,如网页内部文本信息、网页排版格式信息等,局限性较大,难以对网页进行精准分类. 随着大规模网页文本语料库的出现以及深度学习技术在自然语言处理等领域的发展,众多文本分类算法都取得了很好的效果,基于网页纯文本的网页分类应运而生. 然而,上述方式仍然舍弃了网页的排版格式信息,难以获取文本之间的结构联系. 为了同时获取多种网页属性信息,Web-Minds对源代码w_{HTML}进行逐条解析,分别获取网页纯文本信息w_{TEXT}、网页结构信息w_{DOM}以及含于纯文本信息中的网页标题w_{TITLE}和网页关键词w_{KEY}. 解析后的网页数据集形式为$\mathcal{W}_p=\left\{w\middle|w=(w_{\mathrm{TEXT}},w_{\mathrm{DOM}},w_{\mathrm{URL}},w_{\mathrm{TITLE}},w_{\mathrm{KEY}})\right\}$.

2.1.2.2　元数据清洗　经过源数据解析后的数据集有良好的结构规范性,然而,将网页纯文本信息、网页标题信息与网页关键词信息作为文本类信息会存在冗余、符号错乱、排版等问题,对后续使用带来负面影响. 因此,本文根据地球科学专家与信息科学专家联合提出的数据格式要求,对

数据集 $\mathcal{W}_p$ 中的元数据进行了清洗整理:(1)正则去冗余:编写正则表达式去除前缀符、结尾符、换行符等冗余信息,并利用集合运算进行数据去重;(2)非法字符转换:针对存在的非法字符,如%20 ampersand \ / : * ?等,采取字符强制转换策略,转换为合法且易处理的格式;(3)标准格式归一化:为了实现数据的一致性与可比性,对网页元数据信息进行标准格式的归一化处理. 最终获取结构化网页信息数据集 $\mathcal{W}$.

2.1.3 数据标注 经过数据收集与数据处理后,网页信息数据集 $\mathcal{W}$ 已经可以提供良好的网页信息资源用于网页数据挖掘算法应用. 为了进一步服务网页分类任务,Web-Minds对每个网页 w 提供两种类别标注,分别是网页类型与网页主题,可实现不同网页分类任务. 值得注意的是,Web-Minds中的每个标注均得到了学科专家严格验证,具有真实性和可靠性.

2.1.3.1 网页类型标注 研究人员采集数据往往需要花费较多的人力物力资源,例如,花岗岩数据需要专业人员前往花岗岩分布地区采集数日,冰川数据则可能需要采集数年甚至更久,因此,建立一个全球可共享的科学数据集网站有重要意义. Web-Minds对每个网页进行类型标注,标注内容为(数据网页,非数据网页),采用网站级别为主、网页级别为辅的标注方式,标注结果的示例如图6所示. 由于网页分布服从帕累托准则,出现频次最高的20个门户网站涵盖了75%的网页,剩余25%网页分散在近600个门户网站中. 针对此现象,对较高频门户网站下的网页进行随机采样,采样结果基本可以代表全网站下网页类型,提高了标注效率;对较低频的门户网站则进行逐网页专家标注,确保标注结果准确无误.

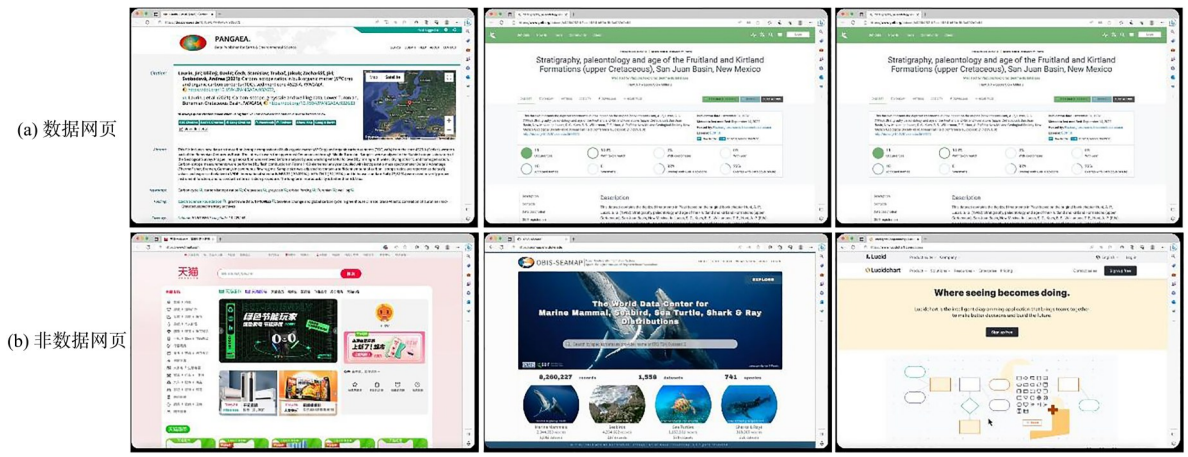

图6 数据网页与非数据网页标注的样例

Fig.6 Annotations for dataset and non-dataset webpages

2.1.3.2 网页主题标注 地球科学是个庞大的学科,存在众多分支学科. Web-Minds在数据收集阶段依据学科关键词库,因此每个网页都存在对应的子学科属性,可实现基于子学科的多分类任务. Web-Minds对每个网页进行主题标注,标注内容为地质学、地球物理学、地理学和地质资源学,但由于每个门户网站均涵盖不同的主题页面,因此,和网页类型标注相比,主题标注无法利用门户网站关系来简化标注流程. 为了获取准确可靠的标注信息,选择大语言模型与学科专家共同标注的策略,借助学科专家标注的少量数据对GPT-3.5-turbo进行上下文学习,对大批量数据进行标注,再由专家纠正错误标注,实现"大模型赋能+专家在回路"式数据标注. 另外,由于子学科之间存在交叉的必然性,专家进一步对所有模糊网页进行最终评判,保证数据标签的唯一性、可靠性.

在数据标注阶段,网页类型标签 l_c 与网页主题标签 l_s 组成网页对应的标注信息 l,构成Web-Minds的标注集 $\mathcal{L}$. 其中,网页类型标签包括地学数据集网站与非地学数据集网站,网页主题标签包括地质学、地球物理学、地理学和地质资源学.

2.2 数据集统计信息 Web-Minds作为开放域

网页数据分布下的多视角网页数据集,由涵盖多类网页属性的网页信息集合 $\mathcal{W}$ 与包含两类标签的标签集 $\mathcal{L}$ 对应组成,即:

$$\text{Web-Minds}=\{x\mid x=(w,l);w\in\mathcal{W},l\in\mathcal{L}\}$$

Web-Minds提供丰富的多视角网页属性及可靠的网页标注. 首先,Web-Minds已收集来自616个门户网站下共计21828条网页,其整体分布如图7所示. 由图可见,超过80%的域名仅涵盖约25%的网页,显示出明显的长尾分布现象,不同于其他网页数据集近似均匀的域名分布,说明Web-Minds更贴近开放域网页分布. 对网页作进一步分析,35.16%的网页包含Schema.org[22]格式信息,可以提供更多的网页相关标准信息,例如数据发布机构、数据采样时间等,供研究人员使用,同时表明该网页具有更高的质量与可信度.

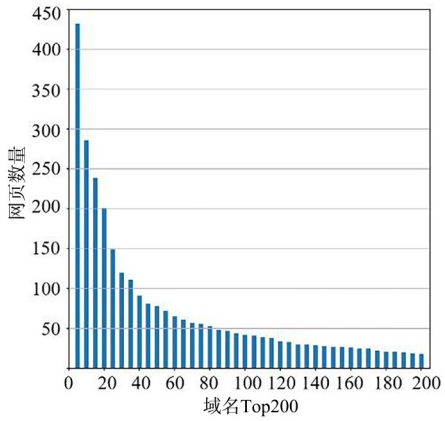

图7　Web-Minds门户网站频次分布图

Fig. 7　Frequency distribution of Web-Minds websites

表1展示了Web-Minds提供的网页类型与主题两种标注信息. 网页类型指数据网页与非数据网页,二者占比分别为54.27%和45.73%. 网页主题涵盖地质学、地球物理学、地理学和地质资源学,占比依次为58.7%,24.1%,9.6%和7.6%. 可以发现,不同主题的网页分布呈现不均衡的趋势,服从开放域网页的分布规律.

Web-Minds针对每个网页提供了丰富的属性信息,如表2所示,包含网页URL链接、网页标题、网页关键词、网页纯文本信息以及网页结构信息. 和其他网页数据集相比,Web-Minds更注重多视角信息,能帮助研究者还原网页本身. 同时,多视角信息也可以支撑更多网页.

表1　Web-Minds数据集的标签分布

Table 1　Distribution of Web-Minds dataset labels

标签	类别	占比
网页类型	数据网页	54.3%
	非数据网页	45.7%
网页主题	地质学	58.7%
	地球物理学	24.1%
	地理学	9.6%
	地质资源学	7.6%

表2　Web-Minds与现有数据集的对比

Table 2　Comparison between Web-Minds and other datasets

数据集	网页种类	网页数量	网站数量	网页链接	标题	关键词	纯文本	DOM结构树
DMOZ-50[8]	10	6458	432	√	√		√	
WebKB[9]	2	1051	16	√			√	
MGC[14]	20	1539	—	√				
AD[10]	2	3279	—	√				
Web-Minds	2/4	21828	616	√	√	√	√	√

数据挖掘下游任务助力研究人员设计算法与性能评估,Web-Minds数据集的详细信息与数据文件可通过https://github.com/sjtu-scx/web-minds获取下载.

2.3　数据集特征对比　将Web-Minds与其他网页分类数据集进行对比. 如表2所示,和DMOZ-50[8],WebKB[9],MGC[14],AD[10]相比,Web-Minds包含更多的网页,域名分布更广泛,且具有显著的长尾现象,更加符合开放域网页的分布规律. 另外,MGC与AD仅提供网页链接信息,WebKB提供网页链接与纯文本信息,DMOZ-50提供网页链接、纯文本信息以及网页标题信息,均未提供网页结构相关信息. 与现有数据集相比,Web-Minds提供了更多视角的网页属性信息,拓展了网页分类算法的设计空间. 同时,Web-Minds可支持更多网页数据挖掘任务,如标题生成、关键词生成、网页信息提取、网页问答等.

3　实验结果与分析

为了评估诸多网页分类算法在Web-Minds上的表现,针对Web-Minds提供的两类标签,设计网页类型分类与网页主题分类两种任务,并对

类别不均衡与域名分布偏移问题展开讨论. 实验旨在验证数据集的可用性与多视角特征的增益,同时为后续网页分类研究提供基准指标参考.

3.1 实验设置

3.1.1 数据集划分 在网页类型与主题分类实验中,Web-Minds数据集中的样本被随机切分为训练集、验证集与测试集,比例为6∶1∶3. 在域名分布偏移研究中,根据网页域名信息分为训练集、验证集与测试集,比例同前. 原则上训练集与后两者中的样本来源于不同域名,以模拟开放域应用场景分布.

3.1.2 实验环境 网页分类算法均基于Pytorch深度学习框架实现,采用Adam优化器对网络进行参数更新,实验设备为NVIDIA GeForce GTX 3090 GPU.

3.1.3 评估指标 采用准确率(Accuracy,*Acc*)、精确率(Precision,*Pre*)、召回率(Recall,*R*)和*F*1分数(*F*1-score,*F*1)对算法进行评估. 对于主题分类,考虑到其类别不均衡性,采用Micro-Recall(Micro-*R*)与Micro-*F*1进行评估.

3.2 基准算法

选用机器学习方法、文本分类算法与网页分类算法对Web-Minds进行多维评估,具体算法如下.

(1)机器学习方法

LR(Logistic Regression):逻辑回归模型.

SVM(Support Vector Machine):支持向量机模型.

(2)文本分类算法

BERT[23]:是基于Transformer架构的预训练语言模型,利用掩码语言模型生成深层双向语言表征,在自然语言处理多个任务中取得了最优性能.

RoBERTa[24]:是BERT的调优版本,有更大的模型参数量、更大的批容量和更多的训练数据.

XLNet[25]:是一种自回归语言模型,利用双流自注意力机制对上下文信息进行建模.

(3)网页分类算法

RiSER[15]:使用Word2Vec与LSTM对垃圾邮件中的文本与对应的XPath进行编码,对二者隐向量拼接后用于垃圾邮件分类.

DC-F[16]:是谷歌学者提出的利用网页URL链接与短文本描述信息进行数据集网页分类的算法.

SMGCN[12]:针对网页多视角特征构建多个关系图,使用图卷积网络提取多视角信息,并通过注意力机制加权多图贡献.

Fusion:使用BERT与LSTM[26]对网页中的文本信息与DOM结构信息分别进行编码和特征融合后训练分类器.

3.3 实验结果分析

3.3.1 网页类型分类 针对网页类型分类任务进行多种基准算法的评估实验,实验结果如表3所示. 由表可得:(1)融合网页多视角信息的深度学习方法准确率与召回率最优,例如SMGCN[12]和Fusion;(2)BERT[23]等预训练的自然语言模型虽然对文本具有强大的嵌入表征能力,取得了较优的召回率,但由于缺少网页结构特征,其性能略低于多视角的深度学习方法;(3)RiSER[15]与DC-F[16]虽然利用了网页中的URL、文本等特征,但受限于其编码器的性能,表现不佳.

表3 多种基准算法在本文Web-Minds上的网页类型分类性能

Table 3 Performance of webpage classification by different benchmark algorithms on our Web-Minds

方法	*Acc*	*Pre*	*R*	*F*1-score
LR	65.26%	70.44%	70.52%	0.7048
SVM	68.31%	72.42%	75.04%	0.7371
BERT	76.65%	82.18%	86.53%	0.8430
RoBERTa	76.56%	81.89%	86.12%	0.8395
XLNet	75.08%	82.03%	85.89%	0.8392
RiSER	70.33%	76.54%	79.21%	0.7785
DC-F	70.29%	79.34%	83.27%	0.8126
SMGCN	78.04%	83.11%	87.76%	0.8537
Fusion	82.14%	84.89%	90.75%	0.8772

网页类型分类中,综合利用多视角特征的方法明显优于只使用单视角特征的方法,证明了网页DOM结构特征的重要性. 网页文本仅能表达网页内容的部分语义信息,无法精确刻画网页的布局特征,这在一定程度上限制了分类准确性.

3.3.2 网页主题分类 针对网页主题分类任务来测试多种基准算法的性能,实验结果如表4所示. 与网页类型分类结果相似,Fusion由于综合

表4　多种基准算法在本文Web-Minds上的网页主题分类性能

Table 4　Performance of webpage topic classification by different benchmark algorithms on our Web-Minds

方法	*Acc*	Micro-*R*	Micro-*F*1
LR	48.37%	58.25%	0.5633
SVM	51.35%	60.42%	0.6267
BERT	68.75%	76.35%	0.8430
RoBERTa	69.59%	77.21%	0.7690
XLNet	68.03%	75.88%	0.7478
RiSER	62.87%	69.47%	0.7064
DC-F	65.74%	73.27%	0.7265
SMGCN	70.12%	77.76%	0.7709
Fusion	74.36%	81.79%	0.8021

利用了网页文本与DOM结构特征,取得了74.36%的准确率,优于其他基准模型.由于缺少网页结构信息,BERT,RoBERTa与XLNet等预训练模型和Fusion相比,性能下降了5%～6%.

Web-Minds中四个主题类别样本的数量不平衡,所以其分类准确率参差不齐.针对这一问题,采用三种常用的类别不均衡策略来优化类别分布与模型参数更新过程.样本生成与下采样分别对应增加少样本类别中网页数量与减少多样本类别中网页数量,损失重加权采用Lin et al[27]的Focal loss,通过修改损失函数对不同类别样本赋予不同的权重来优化模型参数.实验结果如图8所示,分析发现:(1)Focal loss损失重加权的结果最优,尤其是在Fusion模型上,原因是在训练期间改变了四个类别的权重,并强化了对难区分样本的学习;(2)通过生成相似的样本和调整样本比例来提高性能,但由于样本信息有限,改进不够显著;(3)尽管下采样平衡了不同类别的样本数量,但其随机丢失了部分关键信息,降低了分类性能.

网页主题分类任务中,Web-Minds在提供网页文本、DOM结构和语义等信息的同时,其样本类别分布不均的特性真实反映了模型处理类别不平衡数据的能力.表5展示了进行网页主题分类时多视角分类Fusion模型在四种分布不均衡类别上的性能.分析发现:(1)由于地质学与地球物理学样本占比较高,分别为58.7%和24.1%,在不同样本不均衡策略下,分类准确率较高,地理学

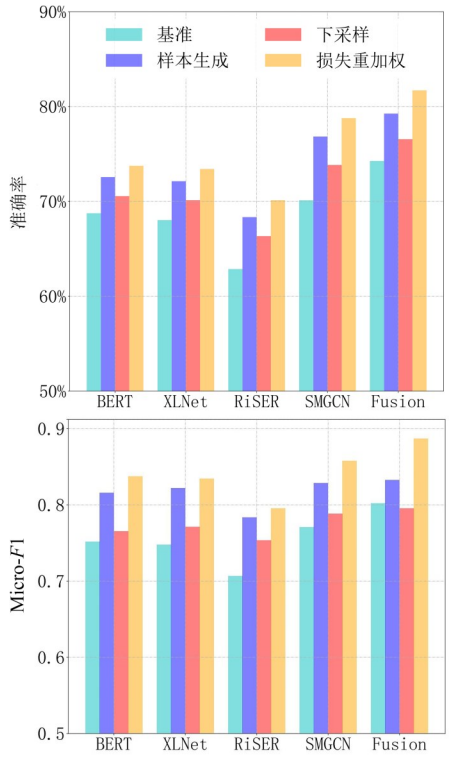

图8　网页主题分类中类别不均衡的实验结果

Fig. 8　Performance of different topics with category imbalance

表5　Fusion模型对分布不均衡的网页主题进行分类的准确率比较

Table 5　Classification accuracy of different topics with category imbalance by Fusion model

类别	下采样	样本生成	损失重加权
地质学	78.36%	82.24%	84.17%
地球物理学	77.35%	81.92%	83.67%
地理学	68.71%	74.35%	78.17%
地质资源学	68.95%	74.21%	77.70%
平均	75.56%	79.26%	82.90%

与地质资源学两类样本较少,模型在这两类上性能有所下降;(2)和下采样与样本生成策略相比,损失重加权将样本分布与预测概率纳入优化过程,对于样本偏少的类别准确率的提升更显著.这种评估可以更全面准确地体现模型对少数类别的分类效果,而不是仅仅关注整体准确率.

3.3.3　域名分布偏移分析　域名分布偏移是指将网页分类技术应用于开放互联网过程时,待分类网页与模型观测数据来源不一致的现象,是归纳学习研究的基础问题.Web-Minds根据域名信

息来划分训练数据与测试数据,原则上保证两者数据来源不同,即具备分布偏移性.

针对网页类型与网页主题分类任务,选取BERT,XLNet,RiSER,SMGCN和Fusion算法模型,实验结果如图9所示.分析发现,测试样本与训练样本来源于不同网站,其内容形式有较大差异,特征分布存在明显偏移,所以各算法在Web-Minds分布偏移数据集上的性能均有不同程度的下降.进一步对比,Fusion与SMGCN的下降幅度较小,因为这两种模型均利用文本与DOM结构信息进行网页表征,充分学习相似网页间的语义特征与布局特征关联,提升了模型在分布偏移场景下的分类性能.

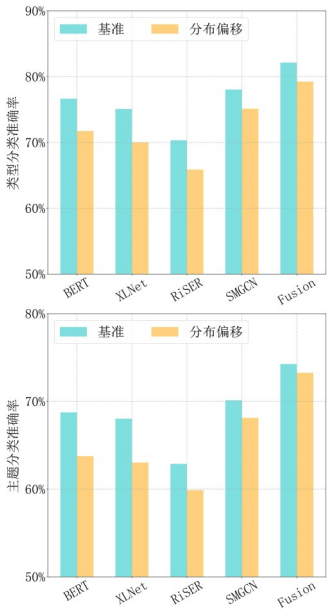

图9 不同算法对存在域名分布偏移的网页进行分类的准确率比较

Fig.9 Accuracy of webpage classification with domain distribution shift by different algorithms

本文提供的具备域名分布偏移性的Web-Minds数据集,旨在为归纳学习研究与网页分类算法的实际应用提供重要数据来源.

4 结论

本文提出了一个面向多视角网页分类的公开数据集Web-Minds,旨在为网页分类研究提供高质量数据来源.Web-Minds包含600余个门户网站的21828条网页,同时提供多视角的网页语义文本与DOM树结构信息,全方位刻画网页特征.在网页类型分类与主题分类上分别使用多种基准分类算法进行评估,证明多视角特征可以显著提升分类任务的性能,为模型设计和性能调优提供数据参考.后续将持续更新Web-Minds,提供更大规模、更多网页内容属性信息的数据集.

参考文献

[1] 程学旗,靳小龙,王元卓,等.大数据系统和分析技术综述.软件学报,2014,25(9):1889－1908.(Cheng X Q, Jin X L, Wang Y Z, et al. Survey on big data system and analytic technology. Journal of Software, 2014,25(9):1889－1908.)

[2] 寇菲菲,杜军平,石岩松,等.面向搜索的微博短文本语义建模方法.计算机学报,2020,43(5):781－795.(Kou F F, Du J P, Shi Y S, et al. Microblog short text semantic modeling method for search. Chinese Journal of Computers,2020,43(5):781－795.)

[3] Chapman A, Simperl E, Koesten L, et al. Dataset search: A survey. The VLDB Journal, 2020, 29(1): 251－272.

[4] Wang X, Huang T L, Wang D X, et al. Learning intents behind interactions with knowledge Graph for recommendation//Proceedings of the Web Conference 2021. Ljubljana, Slovenia: ACM, 2021: 878－887.

[5] Xie X, Sun F, Liu Z Y, et al. Contrastive learning for sequential recommendation//Proceedings of the IEEE 38th International Conference on Data Engineering. Kuala Lumpur, Malaysia: IEEE, 2022: 1259－1273.

[6] Lu B, Wu L W, Yang L N, et al. DataExpo: A one-stop dataset service for open science research// Companion Proceedings of the ACM Web Conference 2023. Austin, TX, USA: ACM, 2023: 32－36.

[7] Castelo S, Rampin R, Santos A, et al. Auctus: A dataset search engine for data discovery and augmentation. Proceedings of the VLDB Endowment, 2021, 14(12): 2791－2794.

[8] Onan A. Classifier and feature set ensembles for web page classification. Journal of Information Science, 2016, 42(2): 150－165.

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

[10] Kushmerick N. Learning to remove internet advertisements//Proceedings of the 3rd Annual Conference on Autonomous Agents. Seattle, WA, USA: ACM, 1999: 175－181.

[11] Deng L, Du X, Shen J Z. Web page classification based on heterogeneous features and a combination of multiple classifiers. Frontiers of Information Technology & Electronic Engineering, 2020, 21(7): 995－1004.

[12] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net, 2017.

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

[14] Vidulin V, Luštrek M, Gams M. Using genres to improve search engines//Proceedings of the International Workshop: Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing. Borovets, Bulgaria: INCOMA Ltd., 2007: 45－51.

[15] 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, CA, USA: ACM, 2019: 886－895.

[16] 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. Springer Berlin Heidelberg, 2021: 338－356.

[17] Wu C H, Wu F Z, An M X, et al. Neural news recommendation with attentive multi-view learning//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: AAAI Press, 2019: 3863－3869.

[18] Li S, Li W T, Wang W. Co-GCN for multi-view semi-supervised learning//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, NY, USA: AAAI Press, 2020: 4691－4698.

[19] Jing X Y, Wu F, Dong X W, et al. Semi-supervised multi-view correlation feature learning with application to webpage classification//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, CA, USA: AAAI Press, 2017: 1374－1381.

[20] Wu F, Jing X Y, Zhou J, et al. Semi-supervised multi-view individual and sharable feature learning for webpage classification//Proceedings of the World Wide Web Conference. San Francisco, CA, USA: ACM, 2019: 3349－3355.

[21] Jia X D, Jing X Y, Zhu X K, et al. Semi-supervised multi-view deep discriminant representation learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(7): 2496－2509.

[22] 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.

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

[24] Liu Y H, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach. 2019, arXiv:1907.11692.

[25] Yang Z L, Dai Z H, Yang Y M, et al. XLNet: Generalized autoregressive pretraining for language understanding//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2019: 517.

[26] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735－1780.

[27] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 2980－2988.

（责任编辑　杨可盛）