面向ResearchGate的古气候文献数据采集系统的研制与应用测评

张学珍[1,2],尹　君[1],白孟鑫[1,2],李艳波[1],郑景云[1,2*]

(1. 中国科学院地理科学与资源研究所,中国科学院陆地表层格局与模拟重点实验室,北京 100101;
2. 中国科学院大学,北京 100049)

摘要:论文基于Linux平台,利用Python(V3.6)和MySQL(V5.7),开发了一套"面向ResearchGate的古气候文献数据采集系统";并且通过人工判读从全球古气候资料共享网(https://www.ncdc.noaa.gov)数据库中遴选出1450篇古气候重建论文,对关键词进行了分类汇总,初步构建了用于古气候文献检索的关键词表。依据这一关键词表,利用古气候文献数据采集系统,从ResearchGate数据库中进行了文献检索。针对来自ResearchGate的32493篇文献和来自NCDC的1450篇文献,通过时间尺度、代用资料类型、气象要素、研究地区(国家)4个维度关键词词频的对比分析,发现2套文献数据关键词词频的相对差异基本一致。这表明依据初步构建的关键词表,自ResearchGate检索获取的古气候重建文献是有效的,能反映古气候重建研究现状。如此庞大数量的研究论文为下一步收集未被NCDC收录的古气候重建结果提供了丰富的数据源。"面向ResearchGate的古气候文献数据采集系统"达到了预期设计目标。

关键词:古气候;ResearchGate;文献数据;采集系统;应用测评

过去千年是当代与未来气候环境变化的背景,其气候环境变化高分辨率数据是揭示气候系统变化规律、诊断全球增暖机理、预测预估未来情景及影响的科学基础。20世纪90年代以来,在过去全球变化(PAGES)等一系列国际研究计划组织下,建立了一批高分辨率的过去千年气候变化序列[1-3]。与仪器测量数据(简称器测数据,包括地面观测和卫星遥测两方面的数据)和气候(地球)系统模式模拟数据不同,古气候数据主要是依据自然环境产物或历史文献记载等气候代用资料,通过不同校准方法重建获得[4-5]。但是,目前代用指标的物理意义尚欠清晰,而且不同资料时空代表性、分辨率、定年精度各异,数据可靠性亟需系统评估。因而,如何全面获取这些重建数据成为当前面临的首要问题[6-7]。

目前,国际学界已建立全球古气候资料共享网(https://www.ncdc.noaa.gov),收录了2000多个地点(区)的气候环境变化代用资料。包括:冷暖、降水、旱涝等历史记录,动物、昆虫历史与考古记录,树轮宽度、密度年表,冰芯、石笋、湖泊沉积物、珊瑚等的同位素、理化指标、生物(孢粉、微生物组成)指标等,以及利用这些记录或指标重建的温度、降水/干湿/旱涝变化序列,若干时段的冰川、北极西北部及格陵兰海冰、植被(主要是全新世大暖期)分布特征。该网站成为国际古气候研究的重要数据源。但该

收稿日期:2019-05-23;**修订日期:**2019-10-09。

基金项目:国家重点研发计划项目(2017YFA0603301);国家自然科学基金项目(41430528);中国科学院重点部署项目(ZDRW-ZS-2017-4);中国科学院前沿科学重点研究项目 (QYZDB-SSW-DQC005);中国科学院青年创新促进会项目(2015038)。[**Foundation:** National Key Research and Development Program of China, No. 2017YFA0603301; National Natural Science Foundation of China, No. 41430528; Key Project of the Chinese Academy of Sciences, No. ZDRW-ZS-2017-4; Key Research Program of Frontier Sciences from CAS, No. QYZDB-SSW-DQC005; Youth Innovation Promotion Association, CAS, No. 2015038.]

第一作者简介:张学珍(1981—),男,山东济宁人,研究员,主要从事气候变化研究。E-mail: xzzhang@igsnrr.ac.cn

***通信作者简介:**郑景云(1966—),男,福建莆田人,研究员,主要从事气候变化研究。E-mail: zhengjy@igsnrr.ac.cn

引用格式:张学珍, 尹君, 白孟鑫, 等. 面向ResearchGate的古气候文献数据采集系统的研制与应用测评 [J]. 地理科学进展, 2020, 39(7): 1140-1148. [Zhang Xuezhen, Yin Jun, Bai Mengxin, et al. Development and application test of a collection system for paleoclimate research documents from ResearchGate. Progress in Geography, 2020, 39(7): 1140-1148.] DOI: 10.18306/dlkxjz.2020.07.007

网站收集的重建序列仅涉及不到2000篇研究论文，远低于本领域期刊的发文量。并且各类型代用资料数量差异较大，空间分布也极为不均。从代用资料类型看，树轮资料占75%以上，从资料空间分布看，北半球资料约占85%，特别是欧美地区的资料密度远大于其他地区。由此可见，该网站收集的数据资源仍很有限，尚有大量的代用资料和重建结果尚未被收录。因而，亟需研发古气候文献数据的采集技术，以获取公开发表但未被收录的古气候数据。

Web of Science是全球最大的科研论文共享平台，已有大量研究利用该平台提供的论文进行了整合分析[8-10]，发现了新的科学规律。但是，该平台属于付费平台，并且不支持用户自行开发自动检索工具。ResearchGate是一个科研社交网络服务网站，旨在推动全球范围内的学术交流和科学合作。目前用户总量超过1000万，几乎遍布全球各个国家。用户可以联系同行，分享科研成果，了解研究动态，以及交流想法。据不完全统计，通过ResearchGate分享的研究论文总计超过1亿篇。大量国际同行分享的研究成果为我们全面收集古气候研究文献(数据)提供了可能。因而，本文拟面向ResearchGate研制古气候文献数据采集系统，并进行初步的应用测评。

1 系统设计思路与实现

1.1 设计思路

从已发表论文中采集古气候重建数据包括2个独立环节：① 从网络数据库(如：ResearchGate)中采集古气候研究论文，核心功能是基于一定的搜索算法，由计算机自动发现、定位并下载相关研究论文，在本地建立具有专业领域特色的文献数据库；② 从论文中(图和表)采集古气候重建数据，核心功能是基于计算机图形学领域的相关算法，由计算机从图和表中自动“读取”古气候重建数据。

本文拟研制的文献数据采集系统主要针对第一个环节。该系统需具备数据定位、数据检索、数据解析、格式转换和入库等基本功能。具体需求包括：实时、稳定、高效地收集特定源的数据；把收集到的数据进行解析，并进行数据格式转换；同时，系统应具有良好的可移植性、安全性、可伸缩性、负载平衡能力和易维护性。

据此设计的系统逻辑结构如图1所示。主要包括4个模块，分别是：数据源定位模块、数据采集模块、数据解析模块、数据格式转换和入库模块。其中，数据源定位模块的主要功能是根据采集任务的描述，查找并定位数据源。数据采集模块的主要功能是从指定的位置将数据高效、无损地传回本地。数据解析模块的主要功能是从数据流中“摘录”预定义的各字段内容，以备写入数据表。数据格式转换和入库模块的主要功能是将各字段的内容，转换成数据库指定的数据格式，进而写入数据库保存。根据ResearchGate提供的文献数据内容，本系统数据库设计包括7个字段，分别是：题目(title)、期刊名(journal)、发表日期(published date)、作者(authors)、摘要(abstract)、doi和URL。

考虑到拟采集的数据类型复杂多样与用户技术水平不一，同时需为未来的升级发展预留空间，系统设计遵循如下原则：① 普适性原则。针对每种来源的数据类型，设计了不同解析器，确保能够正确解析各种来源的数据。② 灵活可移植。采用具有较高灵活性的系统开发语言和数据库平台，确保

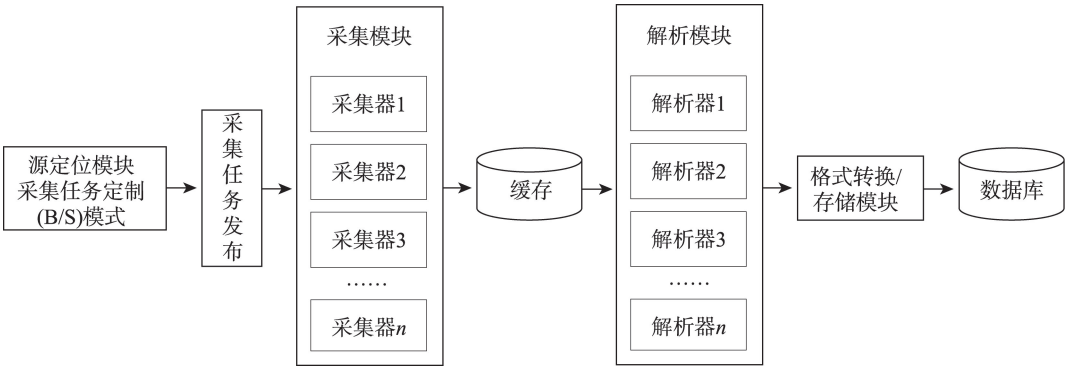


图1　系统逻辑结构图

Fig.1　Logical structure of the data collection system

数据采集系统能够在多种Linux和UNIX环境运行。③易维护性原则。代码简洁,模块鲜明,注释清晰,方便日后维护。④可扩展性原则。模块化设计,接口注释清晰,方便日后的升级。

1.2 技术实现

根据上述设计思路和原则,本系统以Linux/UNIX为开发平台,以Python(V3.6)为开发语言,确保系统具有较高的灵活性和可移植性,可以在任何装有Python解析器的计算机上运行,而不依赖于操作系统。系统依赖的函数库包括:requests、bs4、pymysql。本系统采用MySQL(V5.7)数据库,易于各应用程序数据共享,提高整体系统的性能,方便扩展。系统建议采用的硬件环境配置如下:Intel Xeon X5650 2.66 G处理器,8 G及以上内存,1000 G硬盘,或者以上配置。

研制过程重点解决了如下3个技术难点:

(1) 各数据源格式不一致的问题

由于不同来源的数据格式不一致,导致数据解析器的通用性较差。因此需要根据数据源的结构特征,有针对性地研发解析器,使得解析器与数据源结构类型一一对应,确保解析过程的准确无误,并正确转换为古气候文献数据库标准格式。

(2) 网络不稳定问题

网络不稳定时有发生,导致连接URL地址过程或数据接收过程异常中断,从而降低数据采集效率,增加采集成本。针对这一问题,需要在软件设计过程中,嵌入等待重试机制,并且重试要求有一定的时间间隔,以减少对方服务器的负载。

(3) 数据的重复采集问题

从不同维度对同一数据源进行数据采集时,普遍存在重复采集的问题,致使数据库中记录冗余,不仅增加不必要的存储空间,而且降低了数据库的运行效率。针对这一问题,本系统利用数据来源URL地址的唯一性特征,利用URL作为数据库记录的唯一标识,建立数据表的主索引,并且将数据获取分为2个步骤:第一个步骤仅获取URL,然后将此URL与数据库中已有的URL进行匹配,如果该URL已经存在,那么则跳过该URL,获取下一个URL,如果该URL不存在,则进行第二步,启动解析器解析这个URL指向的网页内容,并且经过解析后存入数据库。通过这一过程,既避免数据重复采集问题,又避免了数据遗漏,以最高效的方式获取最全面的数据。

2 应用测评

2.1 关键词表

关键词及其组合是获取目标文献的主要依据。为编制一个指示性强并且相对精简的关键词表,我们对全球古气候资料共享网公开数据对应研究论文的关键词进行了系统的整理分析。通过人工判读,从全球古气候资料共享网获取了共计1450篇文献数据,进而对这1450篇论文涉及的关键词进行分类汇总。首先,将关键词分为5个一级类,依次是:时空类、领域类、主题类、资料类、方法类。进而,对每个大类进一步细分,共计包括20个二级分类,详见表1。

我们旨在检索气候变化重建序列或者代用资料序列的研究论文,因而在检索过程中重点采用了主题类、时空类和资料类关键词。为尽可能提高检索效率,简化检索式的结构,在应用测评环节,对表1所示的关键词表进行了精简。精简后的关键词表包括2类关键词:一是主题类,二是时空与资料类,并且只采用了出现频次较高的关键词,大幅压缩了关键词数量,减少了数据采集的耗时。其中,主题类关键词包括了出现频次较高的12个关键词,分别是temperature、precipitation、rainfall、drought、flood、dryness/wetness index、Palmer Drought Severity Index、monsoon、ice、snow、glacier、glacial。时空与资料类关键词包括了出现频次较高的290个关键词。在检索过程中采用主题类与时空资料类关键词逻辑"与"关系的检索式,利用上文建立的古气候文献数据采集系统从ResearchGate上进行检索。具体地,分别抽取一个主题类关键词和一个时空资料类关键词,建立逻辑"与"检索式进行检索,从ResearchGate网站数据库中共计检索出743840篇研究论文,并下载了这些文献数据,建立本地格式化的古气候文献数据库。为突出"千年""重建"方面的研究,再利用reconstruct、millennium、Medieval Warm Period、Little Ice Age构建逻辑"或"检索式,进行了二次筛选,共计筛选出32493篇文献。

2.2 关键词词频的对比分析

为评价面向ResearchGate的古气候文献数据采集系统的有效性,本文针对其采集到的32493篇文献(以下简称:ResearchGate数据)及通过人工判读从全球古气候资料共享网获取的1450篇文献(以下简称:NCDC数据),分别分析了主要关键词的出现

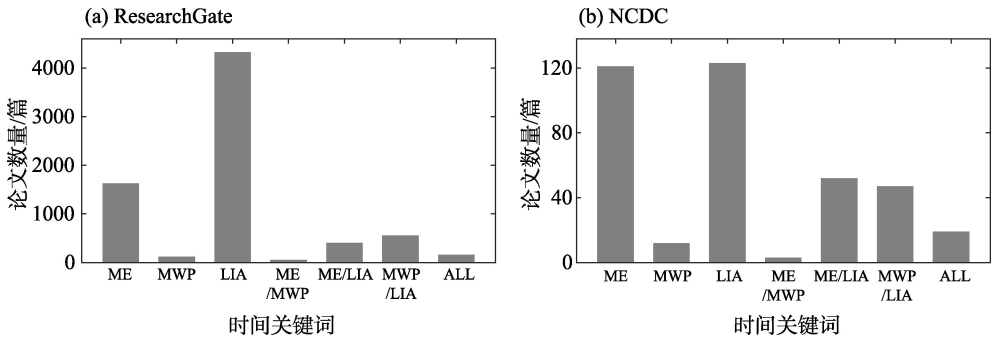
表1　古气候文献关键词分类

Tab.1　Classifications of keywords from paleo climate research articles

一级分类	二级分类	数量/个	示例
时空类	时间尺度	52	last millennium; Little Ice Age; Medieval Warm Period等
	时间分辨率	18	interdecadal variability; centennial scale; high-resolution等
	空间尺度	378	Northern Hemisphere; western North America; China等
主题类	气候	154	climate reconstruction; climatic trend; Holocene climate等
	温度	60	paleotemperature; reconstructed temperature; air temperature proxy等
	降水	23	paleo precipitation; rainfall reconstruction等
	旱涝	33	drought event; flood intensity; megadrought等
	季风	12	Asian monsoon; Indian monsoon; Australian summer monsoon等
	海气	78	atmospheric circulation; El Niño-Southern Oscillation; Pacific Decadal Oscillation等
	太阳	11	solar activity; solar forcing等
	火山	8	volcanic forcing; global volcanicity等
	碳循环	27	carbon cycle; carbon isotope等
资料类	历史文献	10	historical climatology; historical documents; historical records等
	冰雪	33	ice sheets; ice core; glacial maximum等
	树轮	88	chronology; tree rings; dendroclimatology等
	其他(花粉、石笋、钻孔、湖泊)	209	pollen; lake sediment; stalagmite等
方法类	模型模拟	20	climate model; proxy-model comparison; CMIP5 models等
	统计诊断	42	regression; synoptic climatology analysis; singular spectrum analysis等
	序列重建	32	regional proxy-based reconstructions; time-series analysis等
领域类		71	climate dynamics; biogeography; human ecology等

频次，并对2套文献数据中关键词的频次进行了对比。本文分别从时间尺度、代用资料类型、气象要素、研究地区(国家)4个维度分别选取主要关键词，进而开展2套文献数据中关键词的频次对比。

在时间尺度维度上，本文选取了千年(millennium/millennia, ME)、中世纪暖期(Medieval Warm Period, MWP)、小冰期(Little Ice Age, LIA)3个关键词，分别统计了3个关键词独立出现及其组合出现的频次。如图2所示，2套文献数据中时间维度主要关键词词频的相对多少基本一致。LIA的频次最高，其次是ME，再次是ME与LIA的组合及MWP与LIA的组合，ME与MWP组合的频次最低。这一特征与实际研究状况相符。因为LIA距现在最近，代用资料相对较为丰富，并且LIA是现代全球变暖的“参照背景”，广受研究者的关注，因而涉及LIA的研究论文远多于其他时段的研究论文。而MWP则距离现代较远，代用资料相对匮乏，因而涉及MWP的研究论文相对较少。



注：ME为千年(millennium/millennia)，MWP为中世纪暖期(Medieval Warm Period)，LIA为小冰期(Little Ice Age)。

图2　ResearchGate和NCDC文献数据中主要时间关键词及其组合的频次

Fig.2　Frequency of mainly temporal keywords in the ResearchGate and NCDC datasets

2套数据中ME的出现频次尚有一定差异。在ResearchGate数据中，ME的频次大致约为LIA的1/3，而NCDC文献数据中，ME与LIA的频次几乎相当。这可能是因为NCDC收录的重建结果一般是比较重要的研究进展。时间序列足够长，通常是覆盖过去千年，是研究进展是否重要的潜在判定标准之一。而ResearchGate则是囊括了更大量的重建结果，受限于代用资料数量，真正能够覆盖过去千年的重建结果比较少。因而，NCDC数据中ME的频次相对较高，而ResearchGate数据中ME的频次则相对较低。

在代用资料类型维度上，本文选取古气候重建较为常用的5类代用资料关键词，分别是：树轮(tree ring)、文献(documents)、石笋(stalag)、冰芯(ice core)、湖泊/沉积/花粉(lake/sediment/pollen)。如图3所示，2套文献数据中主要代用资料关键词的词频相对差异基本一致。树轮和湖泊/沉积/花粉的词频相对较高，并且远高于其余各资料类型的频次；其次是冰芯；再次是石笋和文献。词频差异的这一特征与研究实况相符。树轮和湖泊/沉积/花粉类代用资料分布较为广泛，被大量用于古气候重建研究，例如：Cook等[11-12]重建过去2000 a北美和欧洲夏季帕尔默干旱指数(PDSI)使用的代用资料全部是树轮；Pauling等[13]重建过去500 a欧洲高分辨率格网化降水使用的大部分代用资料也是树轮；Trouet等[14]重建过去1500 a北美温带地区30 a分辨率的温度变化使用的代用资料主要是花粉(孢粉)。历史文献、石笋和冰芯的分布范围较为有限，保持至今的历史文献主要出现在西欧和东亚，以及南美的部分地区，石笋和冰芯的形成条件较为苛刻，仅形成于特定环境[15-19]。因而据此进行的古气候重建研究也相对较少。

不过，ResearchGate数据中树轮的频次相对低于NCDC数据。在ResearchGate数据中，树轮的频次大致相当于湖泊/沉积/花粉类关键词频次的1/3，而NCDC中树轮的频次则略高于湖泊/沉积/花粉类关键词的频次。这一差异可能是因为NCDC收录的重建结果比较注重定量重建，树轮(宽度、密度)指标与气候要素的关系相对比较清晰，在气候变化的定量重建中应用较为广泛。湖泊/沉积/花粉类代用资料分布范围较广，数量较多，但与气候要素的定量关系较为复杂，很多是不明晰的，较少用于气候变化的定量重建。由此造成了NCDC数据中树轮的频次略高于湖泊/沉积/花粉类关键词的频次，而ResearchGate数据中树轮的频次则明显低于后者。

在气象要素维度上，本文分析了7类主要气象要素关键词，分别是温度(temperature)、降水(precipitation/rainfall)、旱涝(drought/flood)、干湿(dryness-wetness index/Palmer Drought Severity Index)、季风(monsoon)、冰雪(ice/snow)、冰川(glacier/glacial)。如图4所示，在这2套文献数据中，各要素关键词词频的相对高低基本一致。温度和冰雪的频次均较高，降水和冰川的频次位居其次，随后依次是旱涝、季风和干湿。词频差异的这一特征与研究实况一致。在古气候研究中，温度、降水重建是研究重点，旨在理解现代全球变暖的背景、历史地位(相似型)、成因机制等。冰雪和冰川一方面是重要的研究要素[20-21]，同时也是重建温度、降水等气象要素的重要代用资料[22-25]。因而，温度、降水和冰雪、冰川类关键词的频次较高。旱涝、干湿和季风的研究相对较少，因而这一类关键词的频次也较低。

不过，2套数据中温度与降水关键词频次的相对高低尚存一定差异。在ResearchGate数据中，温度的频次略高于降水；在NCDC数据中，温度的频

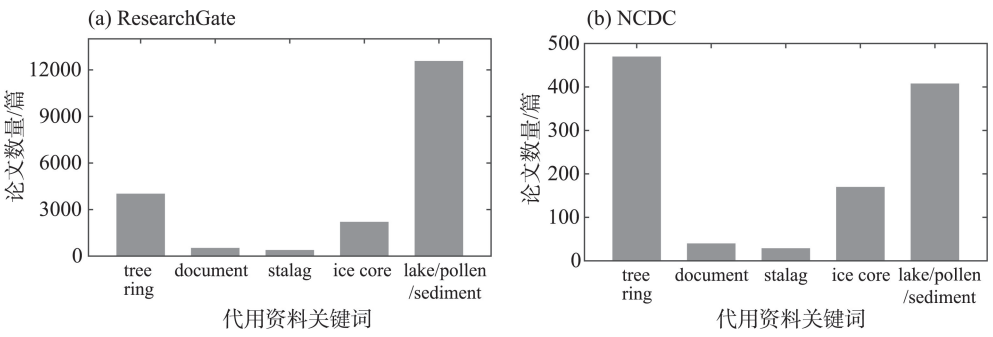


图3 ResearchGate和NCDC文献数据中主要代用资料类型关键词的频次

Fig.3 Frequency of mainly proxy data type keywords in the ResearchGate and NCDC datasets

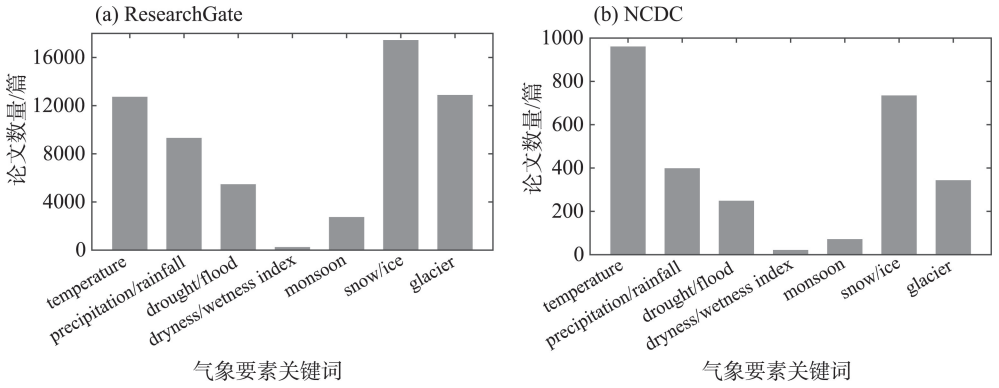


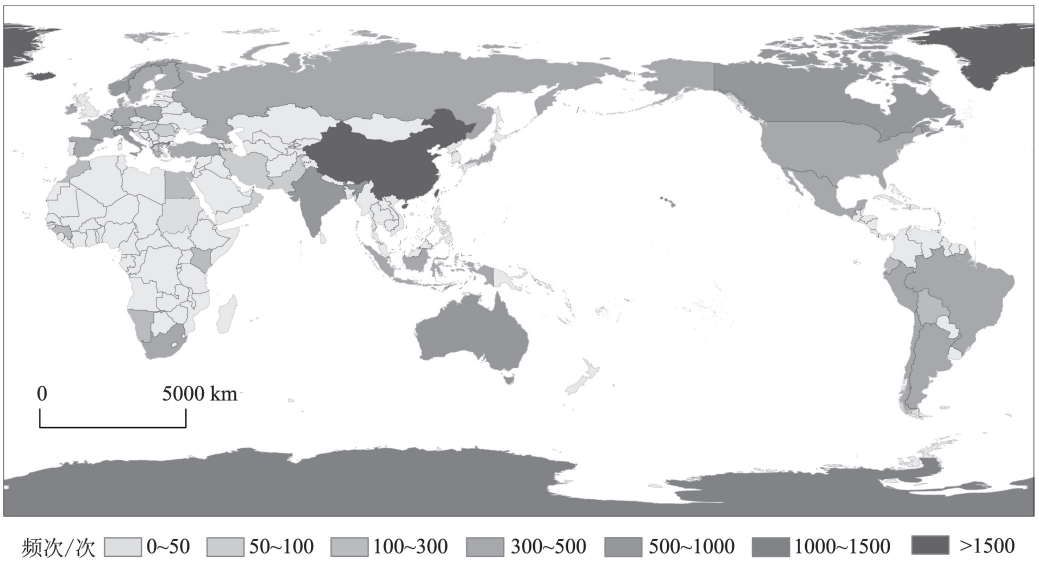
图4　ResearchGate和NCDC文献数据中主要气象要素关键词的频次

Fig.4　Frequency of mainly meteorological element keywords in the ResearchGate and NCDC datasets

次超过降水频次的1倍多。主要原因可能是温度的定量重建相对较多，而降水的定量重建相对较少，造成NCDC收录的温度重建结果较多，降水重建结果相对较少，因而温度的词频大幅超过降水的词频，同时高于冰雪类关键词的频次。在气候变化中，降水与温度是同等重要的气象要素；同时，由于ResearchGate涉及的样本量巨大，较NCDC更有说服力。因而，在古气候研究中，降水与温度的研究引起了同样的重视。

在研究地区维度上，本文分析了每个国家及主要地区名称出现的频次。国家和地区名称词频的高低基本反映了其古气候研究数量的多寡。如图5所示，在ResearchGate数据中，关于中国和格陵兰的研究数量位居前列，其中尤以中国的研究更多，超过位于第二位的南极研究的1/3多，是位居第三梯队美国和加拿大研究的3~4倍。在NCDC数据中，各国家名称的词频排序与ResearchGate基本一致，关于格陵兰的研究数量位居首位，超过南极研究数量的1/3多，是美国和加拿大研究数量的2~3倍(图6)。

但是，关于中国研究数量的排序，2套文献数据有所不同(图6)。在ResearchGate中，中国的研究数量位居首位，比格陵兰研究数量还多。然而，在NCDC中，中国的研究数量位居第三位，比南极少，略高于加拿大。这可能与NCDC收录数据的程序有关。NCDC收录数据需要数据作者提出申请，并



注：本图中世界全图基于自然资源部标准地图服务网站下载的审图号GS(2016)1667号的标准地图制作，底图无修改。

图5　ResearchGate文献数据中国家和地区名称的频次

Fig.5　Frequency of country and region names in the ResearchGate dataset

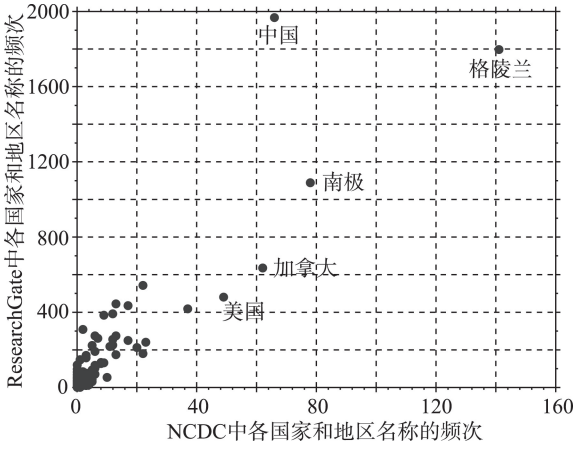


图6 ResearchGate和NCDC文献数据中国家和地区名称频次的相关关系

Fig.6 Correlation of frequencies of country and region names in the ResearchGate and NCDC datasets

提供格式化的数据说明(元数据)文档和数据文档，然后再经科学委员会审核。因为涉及数据分享，并且需要按照对方格式准备相关文档，中国科学家主动提出收录申请的积极性可能较低，由此导致NCDC数据中关于中国的研究成果相对较少。而ResearchGate数据中，与文章发表量保持一致，不需要作者花费额外的精力，导致中国的研究成果相对较多。这也从一个侧面反映了NCDC的被动收录方式获取的数据远少于实际的研究(重建)成果。因而，为更全面地收集重建数据，弥补NCDC的不足，亟需主动采集已经经过同行评审且公开发表，但未被NCDC收录的古气候重建数据。由此可见，本文研制的主动数据采集系统是非常有必要的。

3 结论与讨论

通过上述分析，我们发现2套文献数据中不同维度关键词词频的相对差异基本一致，由此表明依据由NCDC文献分析建立的关键词表，利用“面向ResearchGate的古气候文献数据采集系统”检索出的文献是有效的。该系统检索出的32493篇文献是NCDC提供的1450篇文献的20倍多，并且关键词词频的对比分析表明前者可能更全面的反映了历史气候重建研究的现状。如此庞大数量的研究论文为下一步获取未被NCDC收录的古气候重建结果提供了丰富的数据源。“面向ResearchGate的古气候文献数据采集系统”达到了预期设计目标。

不过，由于采集系统使用的是关键词精确匹配算法，因而务必建立完整全面的关键词表才能确保检索的全面性。这对关键词表的全面性和科学性提出了较高的要求。然而，由于研究者撰写论文过程中的措辞风格不一，建立一个完整关键词表几乎是不可能的事情；另外，一个特别冗长的关键词表会导致搜索过程耗时较长，降低搜索效率。因而，在改进版的采集系统中，有必要采用模糊匹配算法，降低对关键词表全面性的要求，同时提高搜索效率。另外，本文对NCDC提供的1450篇“样板”文献关键词的分析尚是初步的，为配合采集系统高效的搜索，下一步需要对“样板”文献关键词及其匹配关系进行深入分析，以建立精简且高效的核心关键词表。同时，由于解析器具有较强的针对性，通用性较差，因而本系统仅能用于从ResearchGate数据库中发现并下载相关研究论文。面向中文文献数据库(比如中国知网)的采集系统需要另行研制。本文对各维度关键词词频的分析是相互独立的，为深入揭示古气候研究的现状，有必要从多维度联合的角度，揭示各维度关键词组合的频次分布特征。最后，本文研制的文献数据采集系统的功能定位是：从网络数据库中，发现、定位并下载相关研究论文；从有效论文(图、表)中，如何由计算机自动“读取”古气候重建数据(结果)服务于古气候的科学研究将是下一步的研制重点。

参考文献(References)

[1] PAGES2k Consortium. Continental-scale temperature variability during the past two millennia [J]. Nature Geoscience, 2013, 6(5): 339-346.

[2] Solomina O N, Bradley R S, Jomelli V, et al. Glacier fluctuations during the past 2000 years [J]. Quaternary Science Reviews, 2016, 14: 60-90.

[3] PAGES Hydro2k Consortium. Comparing proxy and model estimates of hydroclimate variability and change over the Common Era [J]. Climate of the Past, 2017, 13(12): 1851-1900.

[4] 葛全胜, 等. 中国历朝气候变化 [M]. 北京: 科学出版社, 2011. [Ge Quansheng, et al. Climate change in Chinese dynasties. Beijing, China: Science Press, 2011.]

[5] 刘时银, 等. 冰川观测与研究方法 [M]. 北京: 科学出版社, 2012. [Liu Shiyin, et al. Glacier observation and research methods. Beijing, China: Science Press, 2012.]

[6] Franke J, Bronnimann S, Bhend J, et al. A monthly global paleo-reanalysis of the atmosphere from 1600 to 2005 for

studying past climatic variations [J]. Scientific Data, 2017, 4: 170076. doi: 10.1038/sdata.2017.76.

[7] PAGES2k Consortium. A global multiproxy database for temperature reconstructions of the common era [J]. Scientific Data, 2017, 4: 170088. doi: 10.1038/sdata.2017.88.

[8] Malcevschi S, Marchini A, Savini D, et al. Opportunities for web-based indicators in environmental sciences [J]. PLos One, 2012, 7(8): e42128. doi: 10.1371/journal.pone.0042128.

[9] Hu Y, Han Y, Zhang Y, et al. Information extraction and spatial distribution of research hot regions on rocky desertification in China [J]. Applied Sciences-Basel, 2018, 8 (11): 2075. doi: 10.3390/app8112075.

[10] 张学珍, 赵彩杉, 董金玮, 等. 1992—2017年基于荟萃分析的中国耕地撂荒时空特征 [J]. 地理学报, 2019, 74 (3): 411-420. [Zhang Xuezhen, Zhao Caishan, Dong Jinwei, et al. Spatio-temporal pattern of cropland abandonment in China from 1992 to 2017: A Meta-analysis. Acta Geographica Sinica, 2019, 74(3): 411-420.]

[11] Cook E R, Woodhouse C A, Eakin C M, et al. Long-term aridity changes in the western United States [J]. Science, 2004, 306: 1015-1018.

[12] Cook E R, Seager R, Kushnir Y, et al. Old world megadroughts and pluvials during the common era [J]. Science Advances, 2015, 1(10): e1500561. doi: 10.1126/sciadv.1500561.

[13] Pauling A, Luterbacher J, Casty C, et al. Five hundred years of gridded high-resolution precipitation reconstructions over Europe and the connection to large-scale circulation [J]. Climate Dynamics, 2006, 26(4): 387-405.

[14] Trouet V, Diza H F, Wahl E R, et al. A 1500-year reconstruction of annual mean temperature for temperate North America on decadal-to-multidecadal time scales [J]. Environmental Research Letters, 2013, 8(2): 024008. doi: 10.1088/1748-9326/8/2/024008.

[15] 葛全胜, 方修琦, 郑景云. 中国历史时期气候变化影响及其应对的启示 [J]. 地球科学进展, 2014, 29(1): 23-29. [Ge Quansheng, Fang Xiuqi, Zheng Jingyun. Learning from the historical impacts of climatic change in China. Advances in Earth Science, 2014, 29(1): 23-29.]

[16] 葛全胜, 郑景云, 郝志新. 过去2000年亚洲气候变化(PAGES-Asia2k)集成研究进展及展望 [J]. 地理学报, 2015, 70(3): 355-363. [Ge Quansheng, Zheng Jingyun, Hao Zhixin. PAGES synthesis study on climate changes in Asia over the last 2000 years: Progresses and perspectives. Acta Geographica Sinica, 2015, 70(3): 355-363.]

[17] 田立德, 姚檀栋. 青藏高原冰芯高分辨率气候环境记录研究进展 [J]. 科学通报, 2016, 61(9): 926-937. [Tian Lide, Yao Tandong. High-resolution climatic and environmental records from the Tibetan Plateau ice cores. Chinese Science Bulletin, 2016, 61(9): 926-937.]

[18] Atsawawaranunt K, Comas-Bru L, Mozhdehi S A, et al. The SISAL database: A global resource to document oxygen and carbon isotope records from speleothems [J]. Earth System Science Data, 2018, 10(3): 1687-1713.

[19] Duan W H, Cheng H, Tan M, et al. Timing and structure of Termination II in North China constrained by a precisely dated stalagmite record [J]. Earth and Planetary Science Letters, 2019, 512: 1-7.

[20] 许艾文, 杨太保, 王聪强, 等. 1978—2015年喀喇昆仑山克勒青河流域冰川变化的遥感监测 [J]. 地理科学进展, 2016, 35(7): 878-888. [Xu Aiwen, Yang Taibao, Wang Congqiang, et al. Variation of glaciers in the Shaksgam River Basin, Karakoram Mountains during 1978–2015. Progress in Geography, 2016, 35(7): 878-888.]

[21] 张震, 刘时银, 魏俊锋, 等. 东帕米尔高原昆盖山跃动冰川遥感监测研究 [J]. 地理科学进展, 2018, 37(11): 1545-1554. [Zhang Zhen, Liu Shiyin, Wei Junfeng, et al. Monitoring a glacier surge in the Kungey Mountain, eastern Pamir Plateau using remote sensing. Progress in Geography, 2018, 37(11): 1545-1554.]

[22] Shi H, Wang B, Cook E R, et al. Asian summer precipitation over the past 544 years reconstructed by merging tree rings and historical documentary records [J]. Journal of Climate, 2018, 31(19): 7845-7861.

[23] Mills S C, Grab S W, Rea B R, et al. Shifting westerlies and precipitation patterns during the Late Pleistocene in southern Africa determined using glacier reconstruction and mass balance modelling [J]. Quaternary Science Reviews, 2012, 55: 145-159.

[24] Heyman B M, Heyman J, Fickert T, et al. Paleo-climate of the central European uplands during the last glacial maximum based on glacier mass-balance modeling [J]. Quaternary Research, 2013, 79(1): 49-54.

[25] Wang J, Cui H, Harbor J M, et al. Mid-MIS3 climate inferred from reconstructing the Dalijia Shan Ice Cap, north-eastern Tibetan Plateau [J]. Journal of Quaternary Science, 2015, 30(6): 558-568.

Development and application test of a collection system for paleoclimate research documents from ResearchGate

ZHANG Xuezhen[1,2], YIN Jun[1], BAI Mengxin[1,2], LI Yanbo[1], ZHENG Jingyun[1,2*]

(1. Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: A collection system for paleoclimate research documents (CSPD) was developed in this study using Python (V3.6) and MySQL (V5.7) on the Linux platform. Meanwhile, 1450 research papers of paleoclimate from the National Climate Data Center (NCDC) were manually selected. The keywords from these papers were classified and, then, a keyword list for the research paper collection was prepared. Using the CSPD with the keyword list, we collected 32493 paleoclimate research papers from ResearchGate. To verify the validity of CSPD with the keyword list, we counted the frequencies of four categories of keywords from the 32493 paleoclimate research papers from ResearchGate and from the 1450 papers from NCDC, respectively. Then, the frequencies from the two document datasets were compared. The four categories of keywords refer to the dimensions of temporal scale, type of proxy data, meteorology factors, and study area. We found that the frequencies of the four categories of keywords match well for the two document datasets. This result suggests that the CSPD together with the keyword list is a valid method and the resulting document dataset represents the status of paleoclimate research. A large number of paleoclimate research documents from ResearchGate would work as a great source of paleoclimate reconstruction results, which have not been fully included by NCDC. The CSPD reached the design objective.

Keywords: paleoclimate; ResearchGate; document data; collection system; application test