国内自然语言处理领域数据集引用行为分析*

徐琳宏　王凯达　张立杰
（大连外国语大学软件学院，大连 116000）

摘要：随着科学研究对数据的依赖性不断增强，分析国内自然语言处理领域内数据集的引用行为，有利于规范化数据集的构建和使用，推动国内自然语言处理领域的快速发展。选取《中文信息学报》2013—2022年的1 628篇论文为样本，通过全文本分析法，人工标注1 970条数据集引用信息，以研究文献对数据集的引用行为。研究发现：在国内自然语言处理领域研究中，引用他人数据集的论文数量逐渐增加，使用自建数据集的论文逐渐减少，并且引用数据集论文的篇均被引频次高于自建数据集论文；引用多个数据集的倾向较为明显，引用单个数据集的论文逐渐减少，并且引用2~3个数据集论文的篇均被引频次高于引用单个数据集的论文；数据集重用性较低，高被引数据集主要来源于评测。

关键词：数据集引用；数据引用；自然语言处理；高被引数据集；数据集重用
中图分类号：G353.1　　DOI：10.3772/j.issn.1673-2286.2023.11.004
引文格式：徐琳宏，王凯达，张立杰. 国内自然语言处理领域数据集引用行为分析[J]. 数字图书馆论坛，2023（11）：29-37.

当前，随着互联网技术的迅速发展与普及，人类逐步迈入“大数据时代”，科学研究对数据的依赖程度也越来越高。图灵奖获得者Jim[1]指出，科学研究正朝着“第四范式”即数据密集型科学的方向发展。对于科学研究范式的演化过程，邓仲华等[2]在他人研究的基础上详细介绍了科学研究的四种范式，即“经验科学”“理论科学”“计算科学”和“数据密集型科学”，可见科学数据的价值逐渐凸显。关于科学数据的具体含义，司莉等[3]指出，科学数据是指通过科技活动或其他方式所获取的反映客观世界的本质、特征、变化规律等的原始基本数据，以及根据不同科技活动需要系统加工整理的各类数据集。数据计量是研究科学数据的重要方式，数据计量学也应运而生。顾立平[4]明确指出，数据级别计量是指在数据共享的背景下与数据发布和数据引用相关的计量，并且数据级别计量涉及的核心概念之一便是数据引用（Data Citation）。

数据引用[5]是指作者在文献中以参考文献、脚注或文中注等方式对其所引用数据提供来源出处的做法。对于数据引用的必要性，美国地质调查局（United States Geological Survey，USGS）[6]指出引用数据可帮助提高科学可信度和重现性，确保科学透明度和对数据作者和管理员的合理问责，并帮助未来的用户识别其他人如何使用数据。数据集引用属于数据引用的一部分。

自然语言处理被称为人工智能皇冠上的明珠，目前已经成为人工智能的核心领域，在机器翻译、人机对话、信息检索和舆情监控等多个场景中发挥重要作用。随着时间的推移，自然语言处理经历了基于规则、机器学习和深度学习等多个阶段，每个研究阶段对数据的质量、数量和形态的需求也在不断变化，数据集在自然语言处理领域乃至人工智能领域已经成为众多研究开展的基础。在更长的时间跨度中，研究国内自然语言处理领域数据集的变化规律以及引用特征，不仅能够为该领域的高质量数据集的构建提供支撑，也为其他领

收稿日期：2023-08-28
*本研究得到国家自然科学基金项目“面向社交媒体的多语种文本情感分析方法研究”（编号：61806038）资助。

域的数据建设提供参考。

1 数据引用相关研究

数据集引用是数据引用的一种类型，特指具有一定规模的数据作为一个整体的引用。因此，从特定领域的数据引用、数据引用规范和数据集引用3个方面介绍数据引用相关研究。

（1）对于特定科学领域数据引用的研究。国外的研究主要集中在自然科学领域：在农业科学领域，Williams[7]研究了该领域常见的数据类型，通过科学引文索引扩展数据库中的作者搜索得到出版物，主要描述了农业科学领域中的数据重用和共享情况；在神经科学和分子生物学领域，Leitner等[8]对相关出版物研究发现，NCBI数据库中标有数据集及其相关术语的出版物的影响因子要比普通出版物的更高；Huang等[9]则对生物学蛋白质数据库（Protein Data Bank，PDB）数据进行了调查，发现虽然已经有大量文章被发表，但该数据库中的原始出版物仍在被大量引用且引用率均高于后续出版物。国内研究主要集中在社会科学领域：在图书情报领域，邱均平等[10]就国内外的研究现状做了详细介绍，对数据引用特征进行了多维度分析，发现我国目前对数据引用的重视程度不够、国内外引用数据功能差异也十分明显；史雅莉等[11]针对生命科学、地球物理学和社会科学等多个领域，就总体引用情况、引用数据类型、引用的元数据元素等做了详尽的研究。

（2）数据引用规范性的相关研究。由于当前缺乏公认的科学数据引用标准，且科学数据引用发展还不成熟，该方向也是当前的一个热点。从讨论机器的可访问性出发，Starr等[12]提供了学术数据引用和数据存储的操作指南，提出了实现机器可访问性的5项建议。在数据出版过程中，Costello等[13]认为应该有一个公认的出版流程，于是提出了一个涉及编辑和技术质量控制的分阶段出版流程。黄如花等[14]对国外科学数据引用规范性进行了研究，调查了DataCite网站上的相应机构以及Web of Knowledge中的众多领域，从科学数据引用的指南性文件出发，为我国科学数据引用标准和规范的制定提供了参考。王丹丹[15]对科学数据规范引用的关键问题进行了探析，从如何引用、引用什么和何时引用等方面调查了ICPSR Bibliography of Data-Related Literature数据库的引用情况。

（3）对数据集引用的相关研究。杨波等[16]研究了生物信息学领域中国学者在科学软件和数据集引用等方面的行为特点，提出一种对科学软件和数据集引用特征与科学文献的关联分析方法，并指出科学软件和数据集与科学文献的质量之间存在密切相关关系。Zhao等[17]调查了*PLoS ONE*，对其提及和引用的数据集进行了分析，发现不同学科数据集引用与提及区别很大，正式的数据集引用较少，作者更倾向于URL的引用方式。杨宁等[18]在生物信息学领域构建了生物信息学引文分类数据集，用深度学习和机器学习模型进行了数据集引用识别方法的研究。

综上所述，国内外数据引用研究取得了众多进展，但仍存在诸多的问题。一方面，国内数据引用的众多研究集中在社会科学领域，特别是图书情报领域，对自然科学领域的数据集引用行为研究较少；另一方面，国内外的大部分研究集中于宏观层面，比如引用规范政策、引用主体、总体引用行为等研究，微观层面的数据集引用实证研究还有待加强。因此，本文对自然语言处理领域数据集引用的行为特征进行研究，并从微观层面通过被引频次对高被引数据集的特点进行分析总结。

2 数据获取

2.1 数据来源

研究对象是国内自然语言处理领域对数据集的引用行为，具有较强的领域相关性，因此选取专注于自然语言处理领域的期刊《中文信息学报》。它是中国中文信息学会会刊，重点刊登中文信息处理基础理论与应用技术研究的学术论文。从中国知网（CNKI）中下载该期刊2013—2022年共10年的所有论文，得到1 876篇，剔除其中的撤稿、书籍介绍、综述、通知等非研究型论文248篇，最终得到论文1 628篇。

2.2 数据标注

通过全文本分析法对1 628篇论文进行数据集引用行为的识别与标注。首先，从下载的论文中人工提取数据引用的上下文。然后，设计标注的内容，制定标注规范，采用双人标注的方式，由两名硕士研究生进行标注：一名人员的研究方向是科学计量学与文本情感分析，另一名人员的研究方向是知识图谱，两人的研究内容都与自然语言处理密切相关，对相关数据集具体特

征和应用领域都有一定的了解。为保证数据标注的质量，每个人在进入标注组前都要先学习标注规范，完成试标注，在标注质量合格后开始正式标注。最终，标注结果的Kappa值为0.88，说明结果具有高度一致性。对于存在问题的数据进行组内讨论，经过审核校对后完成全部标注过程。

（1）标注内容。由于数据集引用尚未有可靠的机器标识方法，且目前数据集引用未形成规范标准，参考了前人的工作，总结了一套本领域的数据集引用标注内容（见表1），通过人工识别数据集引用。

表1　数据集标注内容及描述

标注内容	取值范围	描　述
数据集数量	[0，∞）	0代表当前论文未使用任何数据集
数据集类型	自建、引用和混合	引用：引用公开数据集或者引用前人构建的数据集； 自建：作者自主获取原始数据自行构建的数据集； 混合：同时使用了引用和自建两种类型的数据集
引用上下文	文本长度500字以内	施引文献中提及数据集的引用内容
数据集信息		数据集的名称和来源等信息
引用形式	基本引用形式和复合引用形式	基本引用形式包括参考文献、脚注URL、正文说明；复合引用形式包括参考文献+脚注URL、正文说明+正文URL、正文说明+脚注说明、参考文献+正文URL

（2）标注规范。针对施引文献中实际使用的数据集进行标注，不标注在文中仅提及而未使用的数据集。数据引用识别过程中除了明确指明数据集名称的样例外，也有很多文献在引用的上下文中并未明确指明数据集的具体信息，还需对以下特殊情况认真判别：①没有明确指明“数据集”，而是用“数据”“语料”“语料集”等替代。比如句子“本文实验数据来自于……发布的数据”“本文使用的语料来自……有关的评价文本”“本文使用了中英文两个语料集”中的“数据”“语料”“语料集”都可能属于数据集的范畴；②特殊的数据集，如“语料库”“树库”“知识库”“现实网络”“情感词典”“知识图谱”等，许多研究者都将其视为数据集或从中抽取数据，将这些都归为数据集。

2.3　数据概况

通过对1 628篇期刊论文的标注，总共获得2 642条有关数据集的上下文信息，共计1 970条数据集引用信息、1 180个数据集。其中，引用数据集的论文有744篇，自建数据集的有585篇，混合数据集（引用+自建数据集）的论文有213篇，具体信息见表2。

表2　数据集引用行为描述性统计分析

数据集情况		引用形式	
类　型	论文数量/篇	类　型	论文数量/篇
引用数据集	744	参考文献	678
自建数据集	585	脚注URL	201
混合数据集（引用+自建）	213	正文说明	1 049
引用0个数据集	671	参考文献+脚注URL	19
引用1个数据集	529	正文说明+正文URL	11
引用2个数据集	217	正文说明+脚注说明	11
引用3个及以上数据集	211	参考文献+正文URL	1

3　研究结果分析

本节通过数据集类型和数量的动态分析、数据集引用形式的分析、高被引数据集的特征分析、数据集与论文篇均被引频次关系分析4个方面来对国内自然语言处理领域引用行为进行分析。

3.1　数据集类型和数量的动态分析

3.1.1　数据集类型的动态分析

论文使用的数据集一般分为两种类型：一部分是引用的公开数据集或者引用的前人构建的数据集，这

部分称为引用数据集；另一部分则是作者自己爬取数据标注，自行构建的数据集，称为自建数据集。具体到实验中，按使用的数据集类型可以将论文划分为仅引用公开数据集、仅使用自建数据集和混合使用数据集。混合数据集论文是指既引用已有的数据集，又自建了部分数据集的论文。1 628篇论文中使用了数据集的论文共有1 542篇，各类型论文占比情况如图1所示。

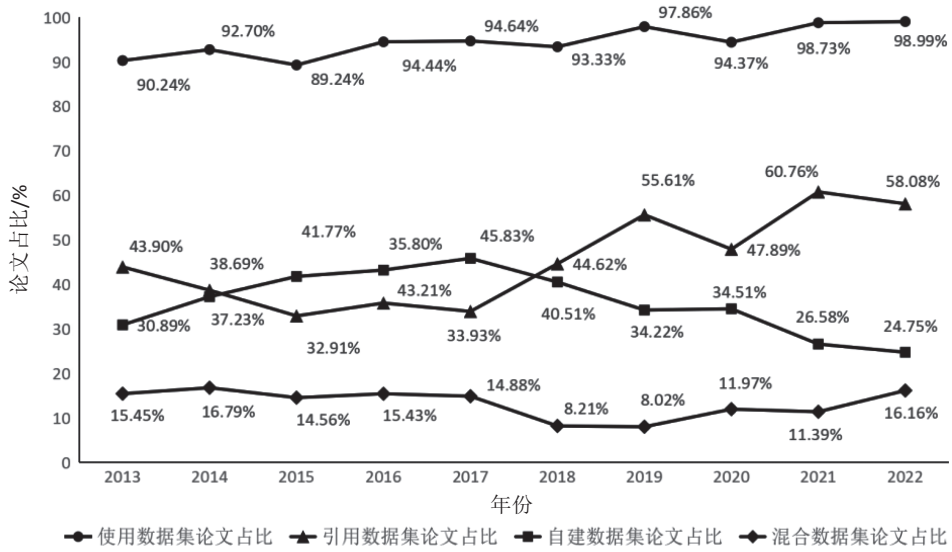

图1 使用、引用、自建和混合数据集的论文占比

使用数据集的论文占比较高，10年中有9年都在90%以上，2021年和2022年均达到了98%以上，可见我国自然语言处理领域对数据集的依赖程度较高。从图1可观察到引用数据集与自建数据集论文占比随时间的变化趋势。在2013—2018年，引用数据集与自建数据集论文占比差距并不大，而从2019年开始，两者的占比差距明显扩大，到了2022年，引用数据集论文占比已经由2017年的33.93%提升到58.08%，而自建数据集论文占比由之前的45.83%下降到24.75%。混合数据集论文的比例也从2017年开始先下降后上升。这说明近年来领域内的学者更多地倾向于引用他人数据集，而不是自建数据集。这可能有两方面原因：一是随着模型复杂性的增加，需要的数据量也不断增大，自建数据集人力成本较高；另一方面，自然语言处理领域学者搭建模型完成相关任务时，经常需要在相同数据集上比较实验结果来验证论文的创新性，这也进一步提升了引用数据集论文的比例。

3.1.2 数据集数量的动态分析

有数据集引用的论文有957篇（引用和混合数据集论文），统计调查显示，单篇论文引用数据集的数量分布为1~16个，其中引用1个数据集的论文数量最多，占比达55.3%，引用2个数据集的论文占比为22.7%，引用3个数据集的论文占比为9.1%，这三者总共占近90%，即单篇论文主要引用1~3个数据集。

把引用1个数据集的称为单数据集论文，引用2个及以上数据集的称为多数据集论文。根据2013—2022年多数据集论文与单数据集论文占比情况（见图2）可以发现：多数据集论文占比由2013年的35.62%增加到2022年的52.38%，呈明显的上升趋势；而单数据集论文占比则由2013年的64.38%减少到2022年的47.62%，呈明显的下降趋势。特别是在2016年与2019年这两个时间节点，多数据集论文占比迅速上升。结合自然语言处理领域的发展情况，推测这可能是因为随着深度学习技术的发展，特别是2015年长短时记忆网络的引入以及2018年BERT、GPT、ELMO等大语言模型的出现，模型的参数量和复杂度不断增加，学者需要更多的数据集来训练和测试模型。

3.2 数据集引用形式的分析

3.2.1 数据集引用形式的总体特征

957篇有数据集引用的论文中，总计有1 970条具体的数据集引用信息。根据不同引用形式的数据集占比（见图3）可知，在自然语言处理领域中大部分论文以正

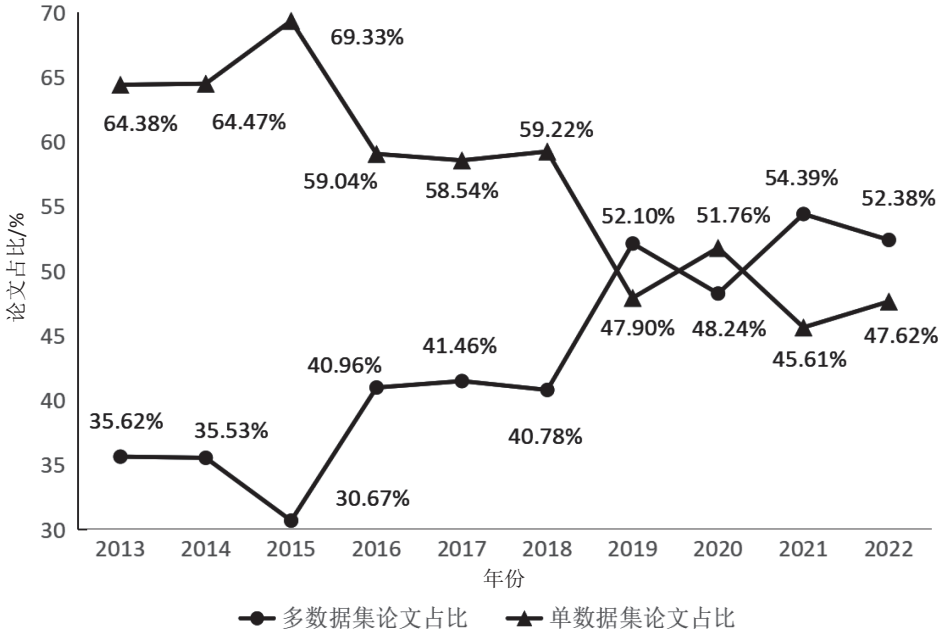

图2　多数据集论文与单数据集论文占比情况

文说明的方式引用数据集，占比达53.25%，其次是通过参考文献形式引用，占比为34.42%，通过脚注URL的形式引用的数据集占比为10.20%，这三者为该领域主要数据集引用形式，总占比高达98%左右。剩余的参考文献+脚注URL、正文说明+正文URL、正文说明+脚注说明、参考文献+正文URL等形式均为复合引用形式，仅占2%左右。

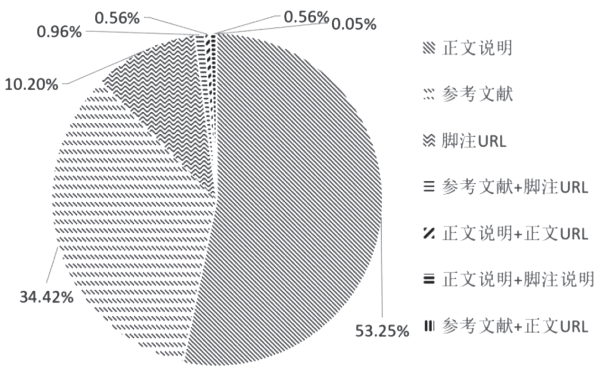

图3　不同引用形式的数据集占比

3.2.2　数据集引用形式与引用规范性的关系

2017年我国发布《信息技术 科学数据引用》，规定了科学数据通用引用格式为：作者.名称（版本）.创建机构[创建机构]，创建时间，传播机构[传播机构]，传播时间，唯一标识符；解析地址。但由于上述标准所列信息比较复杂，我国自然语言处理领域数据集引用没有完全遵守此格式。于是参考丁楠等[5]的定义，即数据引用是指作者在文献中以参考文献、脚注或文中注的方式，对其所引用数据提供来源出处的做法，把规范的数据集引用行为总结为两点：第一，有实际使用的数据集，不仅仅提及数据集；第二，有明显的引用标志（参考文献上标或脚注上标）。那么按此定义可以把参考文献、脚注URL、参考文献+脚注URL、参考文献+正文URL这4类可追溯的引用形式定义为规范引用，其他则为不规范引用。规范引用的论文总占比为46%左右，可见仍有超过一半论文的数据集引用行为不规范，加强数据集引用的规范性仍然十分有必要。从时间维度来看，2013—2022年的规范引用论文占比从25%上升到49%左右，整体规范性呈上升趋势。

3.3　高被引数据集的特征分析

3.3.1　高被引数据集的总体特征分析

邱均平等[19]对高被引论文的定义是数据库中被引频次排名前1%的论文，借鉴这一定义，把高被引数据集定义为被引频次排名前1%的数据集。研究涉及被引数据集共计1 180个，故选取被引频次排名前12的数据集作为高被引数据集，具体信息如表3所示。（特别说明：HowNet的来源较模糊，发布人是董振东、董强，董振东曾就职于中国科学院计算机语言信息工程研究中心，但

表3 高被引数据集特征

序 号	数据集名称	被引频次/次	来 源	发布者	发布年份	归 属	任务类型
1	HowNet	30	HowNet网站	董振东、董强	1999—2000	中国	多种
2	ACE 2005	23	语言数据联盟	宾夕法尼亚大学	2006	美国	抽取
3	NIST MT2005	21	NIST MT2005评测	美国国家标准与技术研究院	2005	美国	生成
4	NIST MT2002	19	NIST MT2002评测	美国国家标准与技术研究院	2002	美国	生成
5	NIST MT2006	18	NIST MT2006评测	美国国家标准与技术研究院	2006	美国	生成
6	北大1998年人民日报语料	17	北京大学计算语言学研究所	北京大学	2002	中国	序列标注
7	NIST MT2004	16	NIST MT2004评测	美国国家标准与技术研究院	2004	美国	生成
8	SemEval2014 restaurant	16	SemEval2014评测	国际计算语言学学会	2014	全球性组织	分类
9	SemEval2014 laptop	16	SemEval2014评测	国际计算语言学学会	2014	全球性组织	分类
10	NIST MT2003	15	NIST MT2003评测	美国国家标准与技术研究院	2003	美国	生成
11	NIST MT2008	15	NIST MT2004评测	美国国家标准与技术研究院	2008	美国	生成
12	汉语框架语义知识库	15	山西大学	山西大学	未知	中国	多种

该中心是否为发布者不可知。调查文献中有作者提到引用的数据集为HowNet 2000版，但有些文献未提到版本。考虑到HowNet第1版公布于1999年，故将其发布时间定为1999—2000年。汉语框架语义知识库于2004年开始建立，公布时间未知。）

从来源、发布者、发布时间、归属以及任务类型总结国内自然语言处理领域高被引数据集的特征如下。

（1）高被引数据集主要来自于相关评测。调查中有8个数据集来自2个不同类型的评测：一个评测为2002—2006年NIST MT系列评测，它属于一个年度性的机器翻译评估活动；另外一个评测为SemEval2014评测，它是一个国际性语义测评。评测数据集具有较高的被引频次可能是因为评测过程得到了许多相应领域学者的关注。此外，该类数据集一般具有稳定的共享方式和详细数据说明文档，这也是评测数据集影响较大的原因。

（2）高被引数据集的发布者主要为国家研究院、大学、国际学术组织。有6个数据集来自直属美国商务部的美国国家标准与技术研究院，有3个来自不同的大学，有2个来自国际学术组织国际计算语言学学会。从其发布者不难看出，高被引数据集的创建门槛较高，90%都是由专门的机构发布，这些机构通常能提供稳定共享方式，且在领域内有较大影响。

（3）高被引数据集发布时间普遍较早。除汉语框架语义知识库未调查到时间，在2008年之前发布的数据集有9个，最近发布年份为2014年。这可能是因为早期建立的数据集引用时间窗口较长，也有可能是因为早期的数据集建立时，相关研究方向还处于起步阶段，数据缺乏，因而数据集得到领域内相关学者的更多关注。

（4）数据集的任务类型集中于生成与分类任务。任务类型参考了王诚文等[20]概括的4种分类任务，包括序列标注、分类、抽取和生成任务，此外单独分类了适用于各种任务的数据集。生成任务主要集中于其子任务机器翻译，有6个数据集；分类任务则集中于其子任务情感分析，有2个数据集。

3.3.2 数据集的重复引用情况分析

在国内自然语言处理领域，高被引数据集只是一小部分，还有众多的数据集被引频次较少，甚至一次也没有被其他人引用过。如表4所示，被重复引用的数据集占比达21.7%左右，没有被重复引用的数据集占78.3%，被引超过3次的数据集仅占6.5%。由此可见，大量数据集在创建后并没有引起领域内学者的广泛关注，在科学研究中的价值有限。

表4 数据集被引频次与占比

被引频次/次	1	2	3	4	5	6~10	11~30
数据集数量/个	924	134	45	17	13	27	20
占比/%	78.3	11.4	3.8	1.4	1.1	2.3	1.7

3.3.3 高被引数据集与数据集来源的相关性分析

为了分析高被引数据集与数据集来源（来自英文评测）之间的关系，采用相关性统计方法研究两者之间的相关性。两个变量分别是“高被引数据集”（X）与“数

据集来源（来自英文评测）”（*Y*）。对于*X*取值，高被引数据集为1，非高被引数据集为0；对于*Y*取值，来自英文评测的数据集为1，其他为0。

选取被引频次排名前100的数据集，包括12个高被引数据集以及88个非高被引数据集，利用SPSS软件进行Spearman相关性分析，分析结果如表5所示。可以看出相关性系数为0.508，显著性值小于0.05，即这两个变量之间有统计学意义上的中度相关性。为了验证结果的稳健性，还选取了被引频次排名前50的数据集，包括12个高被引数据集以及38个非高被引数据集，利用SPSS软件进行Spearman相关性分析，分析结果如表6所示。可以看出相关性系数为0.719，显著性值小于0.05，即这两个变量之间有统计学意义上的高度相关性。从上述结果可知，高被引数据集与数据集来源（来自英文评测）之间存在一定的相关性，来自英文评测的数据集更容易成为高被引数据集。

表5　高被引数据集与数据集来源（来自英文评测）相关性分析（100条数据）

变　量		*X*	*Y*
X	相关系数	1.000	0.508*
	显著性（双尾）		0.000
	N	100	100
Y	相关系数	0.508*	1.000
	显著性（双尾）	0.000	
	N	100	100

注：*表示在0.05级别（双尾），相关性显著。

表6　高被引数据集与数据集来源（来自英文评测）相关性分析（50条数据）

变　量		*X*	*Y*
X	相关系数	1.000	0.719*
	显著性（双尾）		0.000
	N	50	50
Y	相关系数	0.719*	1.000
	显著性（双尾）	0.000	
	N	50	50

注：*表示在0.05级别（双尾），相关性显著。

3.4　数据集与论文篇均被引频次关系分析

3.4.1　论文使用数据集的方式与论文被引频次关系

从2013—2022年总体被引频次数据（见表7）来看，引用数据集会对论文的篇均被引频次有一定影响。当论文未使用数据集时，在所有类型论文中篇均被引频次最低，为7.59次；当使用了自建的数据集时，论文篇均被引频次为9.97次，较低；当使用了混合数据集时，论文篇均被引频次为10.78次，较高；当完全使用引用的数据集时，论文篇均被引频次达到了12.46次，为最高。即引用了数据集的论文篇均被引频次高于自建数据集的论文，高于未使用数据集的论文，这可能说明数据集在学术研究中的重要性和价值。

表7　论文使用数据集的方式与论文被引频次概况

类　型	论文数量/篇	总被引频次/次	篇均被引频次/次
未使用数据集	86	653	7.59
自建数据集	585	5 833	9.97
混合数据集（引用+自建）	213	2 296	10.78
引用数据集	744	9 272	12.46

其原因可能有以下几点：第一，数据集可以增加论文的可信度和可重复性。当研究人员使用公开可用的数据集进行研究时，其他研究者可以更容易地验证和重现实验结果。因此，相对于自建数据集，引用数据集的论文可能被视为更可靠和具有相关性的研究。第二，在国内自然语言处理领域，数据集是重要的资源，使用它进行研究是领域内专家比较认可的方式。因此，与不使用数据集的论文相比，使用数据集的论文获得了更多的引用。

3.4.2　论文数据集引用数量与论文篇均被引频次关系

对2013—2022年所有论文数据集引用数量与相应论文篇均被引频次进行统计，绘制了图4，发现论文篇均被引频次随着论文数据集引用数量的增加先增加后减少，在数据集引用数量为3个的时候，论文的篇均被引频次达到最大值。但需要说明的是，在数据集引用数量≤6个时，样本论文数量都在20篇以上，而在数据集引用数量＞6个时，样本论文数量缩减到10篇以下，样本数较少，因此结论参考性有限。发生这种现象可能是因为引用2~3个数据集时，论文的复杂性、深度和可靠性增加，吸引更多学者引用，但当引用更多的数据集时，数据集的搜集变得相对困难，论文的复现也更困难，这可能导致论文被引频次的下降。

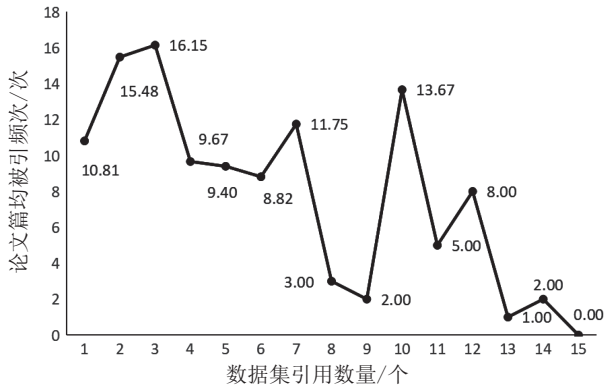

图4　不同数据集引用数量下论文篇均被引频次

4　结论

本研究通过调研我国自然语言处理领域《中文信息学报》的1 628篇论文，揭示该领域数据集引用情况，获得以下结论。

（1）国内自然语言处理领域引用他人数据集的论文逐渐增加，使用自建数据集的论文逐渐减少，并且引用数据集论文的篇均被引频次高于自建数据集论文。在论文数据集的使用方式上，引用数据集与自建数据集的论文占比变化较为明显，呈现相反的变动趋势；在论文的篇均被引频次上，引用数据集会对论文篇均被引频次产生较为明显的正向影响。这说明该领域的研究随着时间的推移，更加重视与前人研究的对比，研究内容和方法的继承性逐步增强。

（2）国内自然语言处理领域论文倾向于引用多个数据集，引用单个数据集的论文逐渐减少，并且引用2~3个数据集论文的篇均被引频次高于引用单个数据集的论文。在数据集引用数量上，多数据集论文占比与单数据集论文占比变化较为明显，在一篇论文中引用多个数据集的行为增多；在论文篇均被引频次方面，其整体随着数据集引用数量的增加先增加后减少。由此可见，为了更好地推动国内自然语言处理领域的科学研究，高水平和大规模数据集的构建和共享势在必行。

（3）国内自然语言处理领域数据集重用性较低，高被引数据集主要来源于评测。在调查的数据集中，有超过3/4的数据集处于未被重复引用的状态。为此，相关部门应为数据集的获取提供便利，提高数据集引用的规范性，提供数据存储平台，并鼓励将数据发布于平台上，建立相应的数据共享政策。数据集是否高被引与是否来源于评测高度相关，且高被引数据集主要由国家机构及学术组织发布。因此为推动我国高质量数据集的发展建设，应组建相应的机构，推动学术组织发展，并鼓励开展评测。

本文对国内自然处理领域的数据集引用行为进行调查，指出了国内该领域数据集引用的特征以及问题，调查了高被引数据集的一些特征，有助于科研工作者明晰本领域的数据集引用情况，为进一步完善数据集引用规范、促进数据集的共享和重用以及为构建高质量的数据集提供依据。本文存在以下几点不足：一是主要基于自然语言处理领域展开研究，结论在其他领域是否成立有待验证；二是人工标注的成本限制调研文献的数量和标注范围。今后将继续拓宽研究的领域，探索不同领域间数据集引用行为的差异。

参考文献

[1] JIM G. On eScience：a transformed scientific method[M]// TONY H，STEWART T，KIRSTIN T. The Fourth Paradigm：Data Intensive Scientific Discovery. Redmond：Microsoft Research，2009：19-33.

[2] 邓仲华，李志芳. 科学研究范式的演化：大数据时代的科学研究第四范式[J]. 情报资料工作，2013（4）：19-23.

[3] 司莉，邢文明. 国外科学数据管理与共享政策调查及对我国的启示[J]. 情报资料工作，2013（1）：61-66.

[4] 顾立平. 数据级别计量：概念辨析与实践进展[J]. 中国图书馆学报，2015，41（2）：56-71.

[5] 丁楠，丁莹，杨柳，等. 我国图书情报领域数据引用行为分析[J]. 中国图书馆学报，2014，40（6）：105-114.

[6] USGS. Data citation[EB/OL]. [2023-06-01]. https://www.usgs.gov/data-management/data-citation.

[7] WILLIAMS S C. Data practices in the crop sciences：a review of selected faculty publications[J]. Journal of Agricultural & Food Information，2012，13（4）：308-325.

[8] LEITNER F，BIELZA C，HILL S L，et al. Data publications correlate with citation impact[J]. Frontiers in Neuroscience，2016，10：419.

[9] HUANG Y H，ROSE P W，HSU C N. Citing a data repository：a case study of the protein data bank[J]. PLoS ONE，2015，10（8）：e0136631.

[10] 邱均平，肖博轩，徐中阳，等. 国内外图书情报领域数据引用特征的多维度分析[J]. 情报理论与实践，2022，45（9）：44-50，36.

[11] 史雅莉，司莉. 我国科研人员数据引用行为特征分析[J]. 情报理

论与实践，2019，42（6）：36-41.

[12] STARR J，CASTRO E，CROSAS M，et al. Achieving human and machine accessibility of cited data in scholarly publications[J]. PeerJ. Computer Science，2015，1：e1.

[13] COSTELLO M J，MICHENER W K，GAHEGAN M，et al. Biodiversity data should be published，cited，and peer reviewed[J]. Trends in Ecology & Evolution，2013，28（8）：454-461.

[14] 黄如花，李楠. 国外科学数据引用规范调查分析与启示[J]. 图书馆学研究，2016（10）：2-9.

[15] 王丹丹. 科学数据规范引用关键问题探析[J]. 图书情报工作，2015，59（8）：42-47，53.

[16] 杨波，王雪，苏娜. 不同文献集中中国学者引用软件和数据集的特征比较研究[J]. 图书情报工作，2017，61（14）：109-115.

[17] ZHAO M N，YAN E J，LI K. Data set mentions and citations：a content analysis of full-text publications[J]. Journal of the Association for Information Science and Technology，2018，69（1）：32-46.

[18] 杨宁，张志强. 融合全文信息的科学数据正式引用识别方法研究[J]. 情报理论与实践，2022，45（2）：191-197.

[19] 邱均平，马瑞敏. 引文索引的功能与科学评价：以美国《基本科学指标》引文数据库为例[J]. 评价与管理，2006，4（2）：1-8.

[20] 王诚文，董青秀，穗志方，等. 自然语言处理评测数据集质量评估研究[J]. 中文信息学报，2023，37（2）：26-40.

作者简介

徐琳宏，女，博士，副教授，硕士生导师，研究方向：情感分析和引文分析，E-mail：qingniao1203@163.com。
王凯达，男，硕士研究生，研究方向：引文分析。
张立杰，女，硕士，讲师，研究方向：数据库、数据流、数据挖掘。

Analysis of Dataset Citing Behaviors in the Field of Natural Language Processing in China

XU LinHong　WANG KaiDa　ZHANG LiJie
(School of Software, Dalian University of Foreign Languages, Dalian 116000, P. R. China)

Abstract: With the increasing dependence of scientific research on data, investigating the reference behavior of datasets in the field of natural language processing (NLP) in China is conducive to promoting the standardized construction and citation of datasets and the fast development of this field. This paper selects 1 628 papers from the *Journal of Chinese Information Processing* from 2013 to 2022 as samples and the citation information of 1 970 datasets is manually marked through full-text analysis to study the citation behavior of datasets in the literature. In the field of NLP research in China, the number of papers citing others' datasets is gradually increasing, while the number of papers using self-built datasets is decreasing. Furthermore, the average citation frequency of papers citing datasets is higher than that of papers using self-built datasets. There is a tendency to cite multiple datasets, and the number of papers citing a single dataset is decreasing. Moreover, the average citation frequency of papers citing 2 to 3 datasets is higher than that of papers citing a single dataset. Dataset reusability is relatively low, and highly cited datasets primarily come from evaluations.

Keywords: Dataset Citation; Data Citation; Natural Language Processing; Highly Cited Dataset; Dataset Reuse

（责任编辑：王玮）