

人文社科领域科学数据使用特征分析*

——基于《中国社会科学》样本论文的实证研究

□ 沈婷婷**

摘要 对《中国社会科学》期刊的论文进行内容分析,探讨我国部分人文社会科学学科在科学数据使用上的特征。重点分析研究者获取实证数据的主要来源,所要分析的数据类型,处理数据的常用方法和工具,以及数据分析完成后的表现形式,并分析一些空泛概念的名词在人文社科论文中的使用情况。根据数据分析的结果,为图书馆提供科学数据服务给出建议。

关键词 人文社会科学 科学数据 使用特征

1 引言

在大数据时代,以科学数据为主要学术资源的数据密集型科学正改变着当前的科学研究模式^[1]。人文社会科学领域的研究者已经开始关注科学数据的作用和价值,定量研究也越来越得到重视^{[2][3]}。在一些数据科学的国际会议上,人文社科的数据管理也成为研究者讨论的重点之一^{[4][5]}。

各种翔实、可靠的数据为以各种社会对象为研究主体的人文社会科学研究提供支撑,推动了社会调查方法、计量学方法、可视化方法等研究方法的应用和发展。中国管理科学与工程学会理事长李京文院士在2013年管理科学与工程学会年会暨第十一届中国管理科学与工程论坛上指出,必须不断研究大数据的形态变化规律,利用大数据来研究、认识和预测客观世界及人类自身的发展变化趋势^[6]。同时,科学数据也影响着传统人文社会科学研究方法的转型与创新。复杂的经济统计模型、大规模的社会调查、丰富的科研数据处理和分析工具等,都为人文社会科学研究的创新提供了条件。科学数据的使用使得人文社会科学的“科学性”显著增强^[7]。

虽然科学数据推动了人文社会科学的新发展,但是也有学者认为科学数据并未与人文科学结合起来,比如一些学者在从事人文研究时,常使用“绝大多数”等空泛概念的名词^[8]。在数据密集型研究模

式的背景下,人文社科领域的研究者对科学数据的使用习惯是怎样的呢?哪些人文社科学科的研究者更擅长利用数据?他们获取数据的来源、处理数据的类型和方法以及数据表现形式如何?研究者是不是由于缺乏对科学数据的使用而增加了“绝大多数”等空泛概念名词的使用呢?

本文围绕以上问题,对《中国社会科学》期刊的全文论文进行内容分析,探讨我国部分人文社会科学学科在科学数据使用上的特征。重点分析研究者获取实证数据的主要来源,所要分析的数据类型,处理数据的常用方法,以及数据分析完成后的表现形式。另外,针对学者提出“绝大多数”等空泛概念的名词在人文社科学科中使用的问题,本文将分析研究者在这些名词使用等方面的习惯。最后,根据人文社科研究者在科学数据使用上的特点,为图书馆提供科学数据服务给出建议。

2 研究方法

本文采用文献调查法采集论文资料,用内容分析方法统计论文中所用到的数据、中文数词,分析其中的问题。人文社科数据主要指调查数据、网络公开数据、政府统计数据 and 指标等^[9],本文研究的数据对象除以上这些外,还包括实验数据、文献数据和图片数据。文中讨论的实证研究数据是指狭义实证研

* 本文为教育部人文社会科学研究青年基金项目“数据素养对科学数据管理的影响及对策研究”(项目编号:14YJC870017)的研究成果之一。

** 通讯作者:沈婷婷,ORCID:0000-0001-8328-3437,shen_tt@shu.edu.cn。

究所需要分析的数据。狭义实证研究是指仅依靠统计分析法的研究^[10]。另外,本文把“绝大多数”、“差不多”、“若干”、“大量”、“无数”、“少量”等词语定义为模糊数词。

本文的数据来源是《中国社会科学》的全文论文。《中国社会科学》是综合性社会科学期刊,所刊登的论文代表我国人文社会科学领域最新的学术研究成果,也是我国人文社会科学研究的风向标。选用该期刊的全文进行分析,不仅可以了解我国最高水平人文社科学术成果在科学数据使用上的特点,而且也可以了解人文社科不同学科的学者在科学数据处理上所采用的最新方法。笔者于2014年12月通过CNKI数据库收集该期刊2010年1月至2014年6月的全部期刊论文522篇,剔除其中“编者按”之类的7篇文章后,最终得到全文数据515篇。

3 数据分析

3.1 总体概况

笔者对2010年1月到2014年6月的《中国社会科学》进行分析,共有全文数据515篇,把它们按研究内容分成马克思主义、哲学、社会学、管理学、人口学、政治学、法学、经济学、传播学、语言学、文学、历史学这12个学科。但由于人口学和传播学的样本数都小于5,统计的数据会在一定程度上有所失真,因而本文不统计这两个学科的数据。

首先对论文的引用数据和实证数据进行分析。除去人口学和传播学的论文,在剩余的论文中,仅引用数据而未进一步统计分析的论文168篇,占总数的33%,进行实证研究统计分析的论文113篇,占总数的22%。统计发现(见图1):除了哲学学科的论文没有引用数据外,其余学科的论文都或多或少地引用了数据。而实证数据主要集中在经济学、社会学、管理学等几个学科上,其中社会学和经济学实证研究的比例分别达到67%和73%。这里政治学的数据值得注意,虽然该学科的论文没有实证研究的数据,但引用数据的比例却较高,达到53%,显示了该学科数据使用的特点。

3.2 实证数据情况分析

本节主要对社会学、管理学、法学、经济学、语言学和文学这六个有实证数据的学科进行分析,分别调查其使用数据的类型、来源、处理方法及其表现形式。这里,语言学和文学的实证数据由于是小

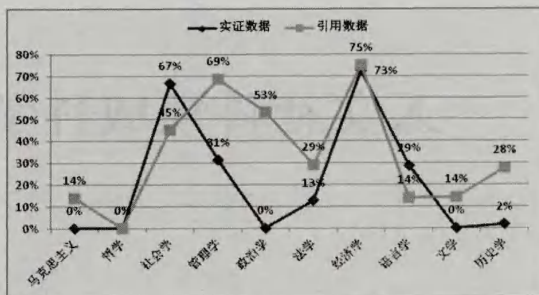


图1 各学科使用数据的论文比例

样本(样本数小于5),因而需辩证看待这两个学科的数据。

(1) 数据类型

从使用数据的类型上看,这六个学科使用的数据基本上是数值型数据,经济学还使用了部分图片数据(表1)。

表1 各学科使用数据的类型情况

学科	社会学	管理学	法学	经济学	语言学	历史学
数据类型	数值型数据	数值型数据	数值型数据	数值型数据 图片	数值型数据	数值型数据

(2) 数据来源

本文根据数据来源把实证数据分为一手数据和二手数据,其中一手数据是指研究者通过访谈、直接观察、间接观察等方式首次亲自收集并经过加工处理的数据,二手数据是指来源于他人调查和科学实验的数据^[11]。本文中,把一手数据分为调查数据、实验数据和文献资料数据;二手数据分为政府信息公开数据(如全国普查数据、各类统计年鉴)和数据管理机构(如中国社会科学调查中心 ISSS 等)的数据。

从数据的来源分析,社会学中44%的论文采用的是一手数据,而在经济学中这一比例只有15%,经济学更多的是使用政府信息公开数据等二手数据(图2)。

在对一、二手数据的进一步分析后,可以看出社会学和法学的一手数据主要来自调查数据,经济学的一手数据主要来自实验数据;对于二手数据的来源,管理学、法学和经济学主要以政府信息公开数据为主,而社会学稍微偏向于管理机构的数据(表2)。

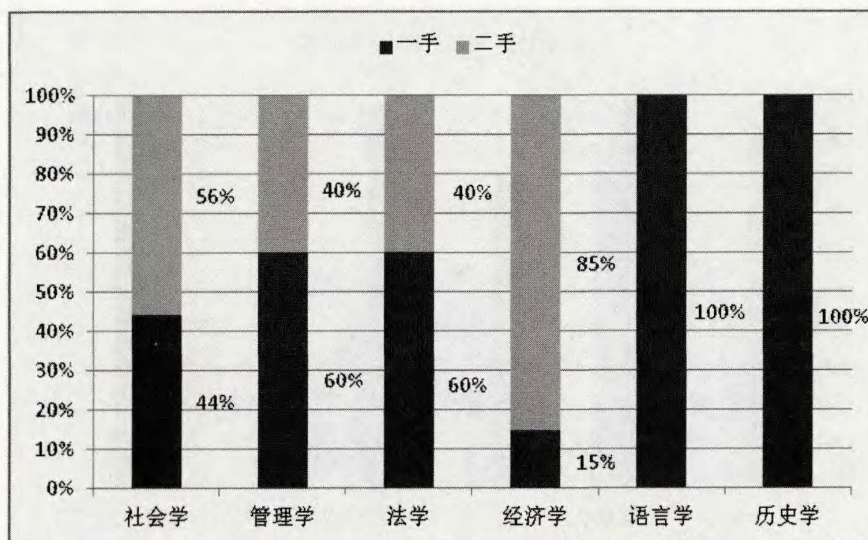


图2 六学科一、二手数据使用比例

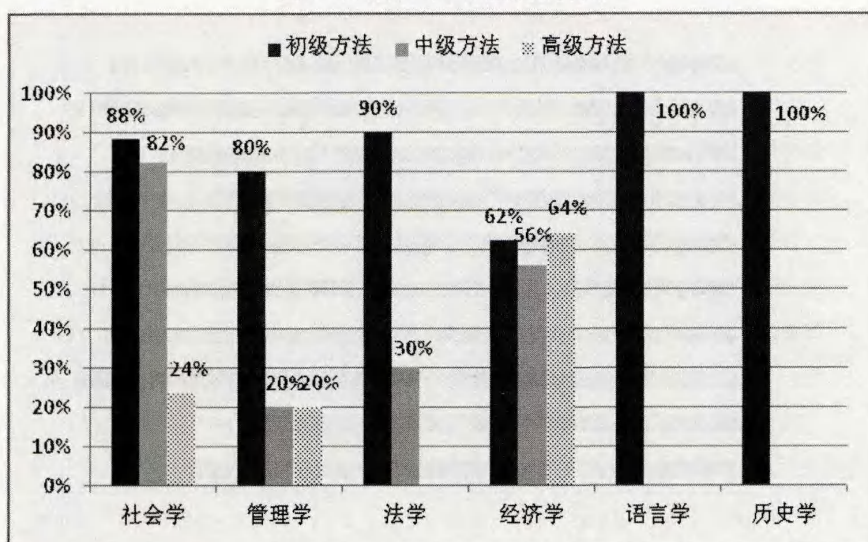


图3 六学科数据处理方法比例

表2 六学科一、二手数据的来源

学科	一手数据			二手数据	
	调查数据	实验数据	文献数据	政府信息公开数据	管理机构数据
社会学	87%	0%	13%	42%	58%
管理学	33%	33%	33%	100%	0%
法学	50%	17%	33%	75%	25%
经济学	33%	67%	0%	81%	19%
语言学	0%	50%	50%	—	—
历史学	0%	0%	100%	—	—

(3) 处理方法

本文把数据处理方法分为初级方法、中级方法和高级方法。初级方法是指平均数、频数、方差、标准差等描述性统计方法；中级方法是指回归分析、参数估计、假设检验、相关分析等统计方法；高级方法则是指模型计算等高等数学方法。

从数据处理方法上看，这六个学科的大部分论文都使用了描述性统计方法之类的初级方法，社会学比较注重对回归分析、相关分析、假设检验等中级方法的运用，而经济学则更擅长运用高等数学方法（图3）。

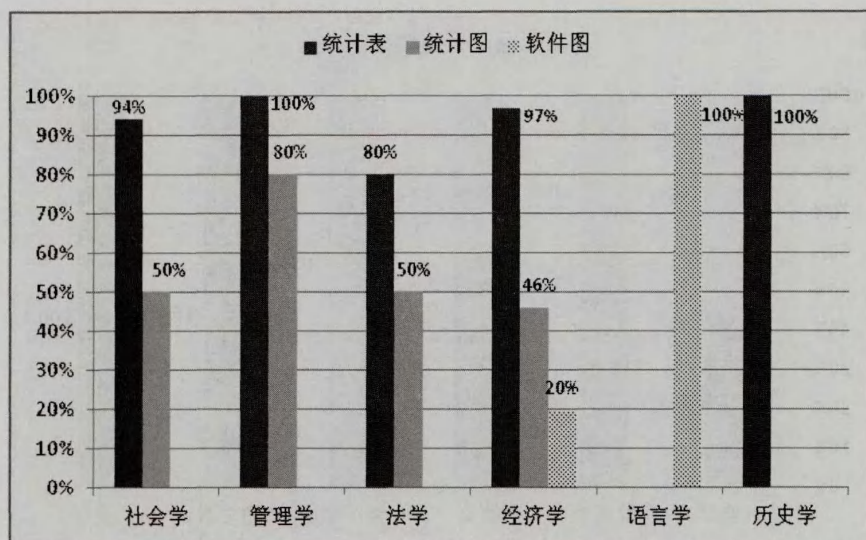


图4 六学科数据表现形式比例

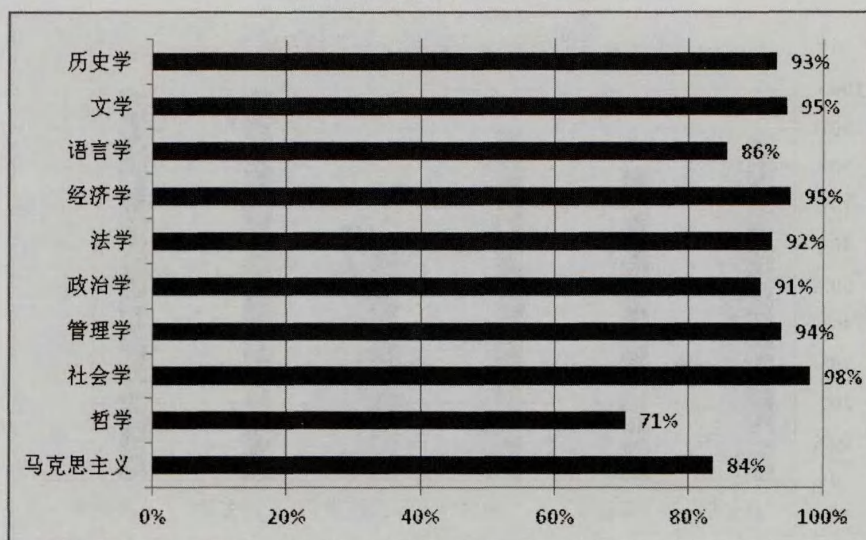


图5 模糊数词在部分人文社科论文中的使用比例

(4) 表现形式

本文把数据的表现形式分为统计表、统计图和特定软件绘制的图形这三类。

从数据的表现形式来看,比较常用的是统计表现形式,折线图、柱状图和散点图之类的统计图在管理学、社会学和法学中的应用也比较广泛。除此之外,经济学和语言学还有一些利用特定软件绘制的图形(图4)。

3.3 模糊数词使用情况分析

本节分析部分人文社科的论文使用“绝大多数”、“差不多”、“若干”、“大量”、“无数”、“少量”、“很

多”和“很少”这八个模糊数词的习惯,以及这些模糊数词在部分人文社科学科论文中的使用情况。

首先统计以上这八个模糊数词在一些人文社科学科论文中使用的比例(图5)。笔者发现社会学、经济学和文学这三个学科使用到以上八个模糊数词的论文比例最高。相对来说,哲学的论文中出现这八个模糊数词的比例稍微小些。

在统计2010年到2014年模糊数词的篇均使用次数后发现,人文社科的学者在模糊数词的使用上趋于稳定,各年篇均使用次数基本在2.0到2.5的区间内,除2012年和2013年的数值波动相对较大

外,其余各年篇均使用次数的数值趋于2.3(图6)。

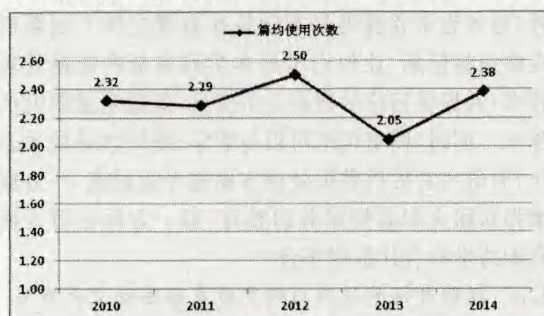


图6 模糊数词各年篇均使用次数

随后统计这八个模糊数词的总使用次数及其篇均使用次数(图7)。笔者发现,“大量”这个词在论文中的总使用次数最多,而且篇均使用次数也最高(达3.14),这说明人文社科的研究者普遍喜欢使用这个词,使用范围也比较广。其次是“很多”这个词,研究者也比较喜欢使用。而“若干”这个词,虽然总使用次数不算很大,但其篇均使用次数却很高,这说明该词的使用范围比较集中。同样的现象也发生在“绝大多数”这个词上。最后讨论一下“差不多”这个词。虽然,胡适先生提出中国人是“差不多先生”,凡事马马虎虎,不求精确,但是这个词在人文社科领域的学术论文中却很少使用,其总使用次数和篇均使用次数都是这八个词中最低的。

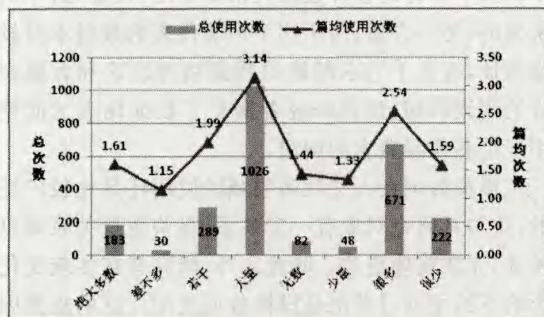


图7 模糊数词的总使用次数及其篇均使用次数

这八个模糊数词在部分社科论文中使用分布情况见表3。表3中“—”表示该词未在该学科中使用,“√”表示该词在该学科中有使用,“★”表示该词不但在该学科中使用,而且其使用的频率更高,即每一列的三个“★”分别代表使用该词的论文比例最高的前三个学科。从表3可以看出,管理学的论文对模糊数词的使用频率较高,有“差不多”、“若干”、“大量”等六个数词在其学科论文中高频使用,其次是

社会学。相比之下,哲学和语言学的论文对模糊数词的使用频率相对低一点。

表3 模糊数词在部分人文社科论文中使用分布

	绝大多数	差不多	若干	大量	无数	少量	很多	很少
马克思主义	★	√	√	√	★	√	√	√
哲学	√	√	√	√	√	√	√	√
社会学	★	★	√	√	√	★	★	★
管理学	√	★	★	★	√	★	★	★
政治学	√	—	√	√	★	—	√	√
法学	★	√	★	√	√	√	√	√
经济学	√	√	√	★	√	√	√	√
语言学	√	—	√	√	—	—	√	√
文学	√	★	√	★	★	√	★	√
历史学	√	√	★	√	√	★	√	★

4 结果讨论

本节依据以上数据分析我国部分人文社会科学学科的学者在科学数据使用上的特点,为图书馆今后开展科学数据服务给出针对性的建议。

4.1 人文社科学科对数据的使用主要由研究对象决定

我们知道,社会学和经济学是以社会现象或经济现象为研究对象的学科,是偏向量化的科学。从上一节图1的数据也可以看出,大多数论文(67%的社会学论文和73%的经济学论文)都使用了基于数据的实证研究方法。在这些学科里,研究者尊重数据,把数据当作研究的主要素材,通过调查统计和模型计算使数据和学科发展有效结合在一起,推动了新知识和新规律的发现。而且,这些学科研究者的数据意识也比较强,对数据需求也更为迫切,获取数据和处理数据的能力更强。比如经济学,这个学科的论文有较高的实证数据使用比例,也非常注重政府公开数据等二手数据的获取,二手数据的使用率比其他学科的使用率更高(图2)。这就是由于研究者数据意识强烈,数据需求迫切,从而使他们获取数据的渠道要比其他学科的研究者更广。

相比之下,马克思主义、哲学和文学这些学科是依靠思维逻辑的研究学科,主要是对传统文献资料的探讨,很少讨论实证或经验现象,因而对实证数据的使用就不是很多,偶尔会引用一些数据,数据来源也只是依靠文献。这些学科的研究方法仍较为传

统,主要以文献为主,对数据使用的需求不是很迫切。

科学数据服务作为图书馆的一项创新服务,首先应该明确服务的对象。诸如以上这些人文社科学科,如果图书馆对其全面铺开科学数据服务,那可能效果就并不是很理想,但如果图书馆首先对社会学和经济学的研究者提供科学数据服务,为他们提供数据获取、处理、共享方面的服务,则可能会取得事半功倍的效果。因而,笔者建议,针对人文社科的科学数据管理服务,首先可以把社会学和经济学的研究者作为主要服务对象,把政治学的研究者作为潜在服务对象,在服务得到一定认可后,逐步向其他学科展开。对于主要服务对象,图书馆可以提供常规的检索、收集、存储、分析等数据情报服务和数据技术服务;而对于潜在服务对象,则可先向研究者提供科学数据在该学科中创新应用的情报信息,待研究者对科学数据服务需求提高后,再提供常规数据服务。

4.2 人文社科学科对实证数据的处理以初级方法为主

在人文社科类论文的实证研究中,为了让读者了解数据和研究对象的特征,进而更好地了解统计分析结果,研究者根据不同的研究目的和研究对象会选取不同的数据处理方法。人文社科的研究者对数据处理采用的方法以描述性统计等初级方法为主,初级方法在法学、社会学、管理学等学科的实证数据研究中占了较大比例(图3)。

相对而言,一些实证研究比较多的学科在数据处理方法上则更为多样。比如社会学和经济学,这两个学科的研究者除了掌握描述性统计等初级方法外,对回归分析、参数估计、假设检验、相关分析等一些中级方法也运用自如,而经济学的研究者使用高等数学方法进行模型计算的能力更为突出,数据的表现形式也多种多样,除了常用的统计图表外,还有由各种特定软件绘制的图形,这些都显示出他们优异的数据素养。

人文社科研究者在数据处理方法上的使用,一方面由论文的研究目的和对象决定,另一方面也由研究者的数据素养决定。对于前者,可能超出了图书馆的服务范围,但要改善研究者的数据素养,图书馆还是可以有所作为的。笔者建议图书馆为研究者提供有针对性的数据处理方面的开放课程等信息,

帮助研究者掌握更多关于数据处理的知识。除此之外,也可为研究者提供本学科在数据应用上创新研究的情报信息,让他们了解本学科最新的数据处理技术,并提供相应的数据分析工具,促进定量研究的深入。同时,图书馆也可以与数学、统计学等院系合作,为研究者提供数据处理方面的专业讲座,一方面增强对研究者数据素养的教育,另一方面也可为研究者跨学科合作提供平台。

4.3 模糊数词的使用与研究对象和传统文化有关

模糊数词的使用在人文社科领域比较常见,并且各学科在使用上并没有体现出明显的差异。诸如在社会学和经济学这些数据使用比较广泛的学科里,研究者并没有因为使用了数据而减少对模糊数词的使用,反而模糊数词在这两个学科中的使用要比其他学科更多(图5),而且各年模糊数词的篇均使用次数也较稳定,没有很大波动(图6)。因而,模糊数词的使用并不能说明科学数据没有与人文科学的发展结合起来,这主要还是与研究者的研究对象和受传统文化影响下的用语习惯有关。

比如“大量”这个词,该词使用总量和篇均使用量在这八个模糊数词中都是最高的,说明人文社科的研究者普遍习惯使用该词。然而进一步分析显示,该词在经济学中的使用频率最高。一方面,经济学论文中存在的各种数量关系需要用“大量”这个词来表述,另一方面,由于这个词所代表的数量本身较难考证,再加上力求精确的数据精神缺乏和大概而言的用词习惯,使得研究者摒弃了数据化表达而使用了这些模糊概念的词语。

模糊数词在人文社科领域的使用还是比较广泛的,这与其研究对象有一定联系,有些数量关系难以考证,无法精确量化。除此之外,研究者在传统文化影响下的用词习惯也是模糊数词使用广泛的重要因素之一。笔者建议图书馆在为人文社科学者提供科学数据服务时,要考虑到不同学科研究对象的影响,既要提倡数据文化,又要遵循人文精神。平时要注重收集有学术价值的科学数据,并为研究者提供方便查询的数据平台。与此同时,也可以根据研究者的需求,帮助他们收集和统计相关研究所需要的特定数据,以尽量减少模糊数词的使用。但作为科学数据服务的提供方,图书馆也应尊重原有的人文社科研究方法,使科学数据作为一种补充材料,与现有的人文社科研究结合,相得益彰。

5 结语

随着大数据时代的到来,大数据分析方法为人文社会科学研究提供了新的研究空间和研究可能^[7]。人文社会科学领域也会有越来越多的研究者使用科学数据,同时也引起更多图书馆员关注并研究人文社会科学领域的科学数据管理问题^{[12][13]}。高校图书馆应抓住这一契机,根据研究者不同的数据需求和使用特征,为他们提供更多、更有效的科学数据服务。

当然,本文的研究也有一定的局限性。由于《中国社会科学》是一份综合性期刊,一些跨学科论文给学科分类带来了一些困难,不可避免地存在一定偏差。另外,虽然综合性期刊为研究的广度提供了有利条件,但同时也给研究的深度带来困难,无法更详细、深入地探讨。因而,笔者后续将会对某一学科进行深入研究,以更好揭示其数据使用特征。

参考文献

- 1 海伊,坦斯利,托尔.第四范式:数据密集型科学发现[M].潘教峰,张晓林等译.北京:科学出版社,2012:181-187
- 2 King G. Ensuring the Data-Rich Future of the Social Sciences[J]. Science, 2011, 331: 719-721
- 3 陈云松,吴晓刚.“复制性研究”:社会科学定量分析新趋势[J].

评价与管理,2012(4):47

- 4 A Review of the U. S. Global Change Research program's Draft Strategic Plan [EB/OL]. [2014-9-22]. http://www.nap.edu/catalog.php?record_id=13330
- 5 UK e-science All Hands Meeting 2011[EB/OL]. [2014-9-22]. <http://www.allhands.org.uk/>
- 6 杨怡.大数据在人文社科领域有广泛应用前景[N].中国社会科学报,2013-11-6(2)
- 7 孙建军.大数据时代人文社会科学如何发展[N].光明日报,2014-7-7(11)
- 8 韩晗.论“大数据”与人文研究的转向[J].晋阳学刊,2014(3):22-25
- 9 The University of Michigan. Data-PASS[EB/OL]. [2014-9-14]. <http://www.data-pass.org/>
- 10 乔坤,马晓蕾.论案例研究法与实证研究法的结合[J].管理案例研究与评论,2008,1(1):62-67
- 11 Hox J J, Boeije H R. Data Collection, Primary vs. Secondary[J]. Encyclopedia of Social Measurement, 2005(1): 593-599
- 12 Mooney H. Citing data sources in the social sciences: do authors do it[J]. Learned Publishing, 2011, 24(2): 99-108
- 13 彭建波.北美人文社会科学数据管理的实践及其启示[J].大学图书馆学报,2013(6):33-37,87

作者单位:上海大学图书馆,上海,200444

收稿日期:2014年11月26日

The Characteristic of Research Data Application in Humanities and Social Science: An Empirical Study of Publications in Social Science in China

Shen Tingting

Abstract: This paper analyses the full texts in Social Sciences in China, and discusses research data application characteristic of Humanities and Social Sciences in our country. The focus has been put on the main sources of the empirical data, data types, common methods and tools for processing data, and display forms of the data analyzed. Some vague concepts in the use of Humanities and Social Science's papers have also been discussed. According to data, the paper then gives library some advices on providing research data services for researchers.

Keywords: Humanities and Social Sciences; Research Data; Application Characteristic