DOI:10.16384/j.cnki.lwas.2023.07.014

高校图书馆学科服务数据处理与分析框架构建*

林 叶 王丽艳 王月苗

摘 要 学科服务数据处理与分析是学科服务的重要工作内容,优化学科服务工作的关键在于优化学科服务数据的处理与分析流程。文章构建学科服务数据处理与分析框架,基于Python语言提出实现数据自动处理和分析的技术方案,并以浙江工商大学为例评估框架应用和技术方案实施的有效性。实证结果表明,基于该框架设计的计算机程序自动处理、分析学科服务数据的时间显著缩短,处理结果的准确性、规范性得到提升,应用效果较好。

关键词 学科服务;学科分析;学科建设;数据分析;高校图书馆

分类号 G258.6;G250.7

本文引用格式

林叶,王丽艳,王月苗.高校图书馆学科服务数据处理与分析框架构建[J].图书馆工作与研究,2023(7):69-76.

1 引言

在“十四五”规划及“双一流”建设背景下,学科建设在高校工作中的重要性日益凸显。学科服务是推动学科建设的重要手段。学科服务数据处理与分析能够为学科服务开展提供信息统计、数据分析等支撑,是学科服务的重要工作内容,所以,优化学科服务工作的关键在于优化学科服务数据的处理与分析流程。学科服务数据处理与分析主要包含方法、数据、工具3方面内容:①采用合理的分析方法,使分析结果更有效。业界相关分析方法较多,如北京大学图书馆采用的学科竞争力分析流程与方法[1]。②汇聚多维度、充足的数据,使分析结果更具全面性和客观性。数据主要来源于学科服务机构购买的数字资源。③利用合适的分析工具提高分析效率。分析工具可以选择商业软件,也可以自主开发。但商业软件一般价格昂贵,且未必支持个性化的学科分析需求。鉴于此,本研究构建了学科服务数据处理与分析框架,并使用Python语言设计计算机程序,以实现自动处理和分析学科服务数据,提高学科服务质量。

2 学科服务数据处理与分析相关研究与实践现状

2.1 研究现状

学科服务数据处理与分析相关研究主要包括以下五方面:①学科服务数据处理与分析的框架、操作流程研究。有学者探讨学科竞争力分析的流程与方法,形成可推广、可复用的学科竞争力分析架构和报告模式[1-2]。②数据源研究。P.Mongeon等[3]分析Web of Science和Scopus两个数据库的学术期刊覆盖情况,发现二者均偏向收录自然科学、工学、生物医学等学科文献,不能为社会科学、艺术人文学科评估提供数据源。匡广生等[4]提出基于图的多源数据融合框架,为多源数据融合提供

*本文系浙江工商大学学科建设管理重点项目“基于Python的我校学科数据分析系统的构建研究”(项目编号:XKJS2021005)研究成果之一。

了理论基础。李慧等[5]将科技文献、专利信息和网页数据进行多源数据融合分析。③分析指标创新优化研究。俞立平、L. Bornmann等[6-7]探讨学科分析指标的创新优化，为基于多指标的学科分析提供了理论基础。④跨学科分析方法应用研究。姚海燕等[8]采用层次分析法对学科建设各层级指标体系权重进行优化。Y. Zhang、T. Timakum等[9-10]采用深度学习方法和数据挖掘技术对全文期刊文献进行主题提取，为更广泛学科领域的主题分析、热点分析提供了参考。俞立平等[11]使用统计学方法修正了学科分析指标。⑤分析工具研究。D. Yu、杜蕾等[12-13]使用VOSviewer、CiteSpace等工具进行文献计量分析，在优化学科分析工具方面进行了探索。

2.2 实践现状

学科服务数据处理与分析的相关实践主要聚焦于以下三方面：①基于学科服务数据创建各类大学排名。如英国泰晤士高等教育THE世界大学排名、英国QS世界大学排名、美国US News世界大学排名、中国软科世界大学学术排名等[14]。②国内外数据库商基于学科服务数据提出新的评价指标。如传统的影响力指标只考虑学术影响力，而Priem定义的Altmetrics指标将社会传播热度也纳入考量[15]；Altmetric公司推出的“Altmetric it”浏览器插件可帮助研究者快速获知有多少媒体关注了某篇文献。PlumX是另一种替代计量学指标，补充了Altmetrics指标在评价网页及数据库使用量、用户收集量、引用量等方面的不足[16]。当前，国际合作成为全球科研活动的主要特征，Clarivate公司在学科规范化引文影响力(CNCI)基础上，定义了CNCI合作性创新指标(Collab-CNCI)[17]，用于评估合作科研成果中的合作贡献度。③商业公司利用计算机技术开发学科服务数据处理与分析系统。如北京恒通博联公司推出的“学科检测分析与情报服务系统”，能够自动抓取、分析学科服务数据，并根据使用者需求生成学科分析报告。

综上所述，现有理论研究较少涉及学科服务数据的深度处理和分析内容，如对于作者贡献、院系贡献、机构贡献的分析，以及二级机构地址规范化处理、作者姓名规范化处理等均需进行深度的数据处理才可实现。实践领域，有些商业软件虽具有部分学科服务数据处理和分析功能，但在数据源的精准度、全面性和分析功能方面仍有诸多不足，不能完全满足学科服务人员的需求，加之商业分析软件往往价格昂贵，很多机构也没有足够资金购买。鉴于此，本研究设计了通用、系统的高校学科服务数据处理与分析框架，旨在厘清学科分析中数据深度处理和分析的流程；基于Python语言提出实现数据自动处理和分析的技术方案，以快速处理和分析学科服务数据，提升学科服务工作效率。

3 学科服务数据处理与分析框架

学科服务数据处理与分析框架如图1所示。本研究采用的数据来源于Web of Science数据库和Incites数据库，这两个数据库覆盖学科范围广、数据量充足，是众多高校图书馆进行学科服务数据处理与分析的必选数据源[18]，因此，本研究采用上述两个数据库数据，以便分析框架可适用于大多数高校。本研究以采集数据作为框架的输入点，数据经过深度处理和分析后输出结果，经过测试和评估确认其可用后，学科服务人员可直接从结果中提取信息制作分析报告。

3.1 数据采集与预处理

学科服务人员从Incites数据库和Web of Science数据库下载所需数据。因为这两个数据库的数据格式不一致，需要对其进行归并处理，即处理为同一字段、同一格式的数据并合并成一个数据表。在技术实现上，本研究采用Python程序语言进行归并处理。Python语法简单，数据处理能力强，特别是其中的Pandas

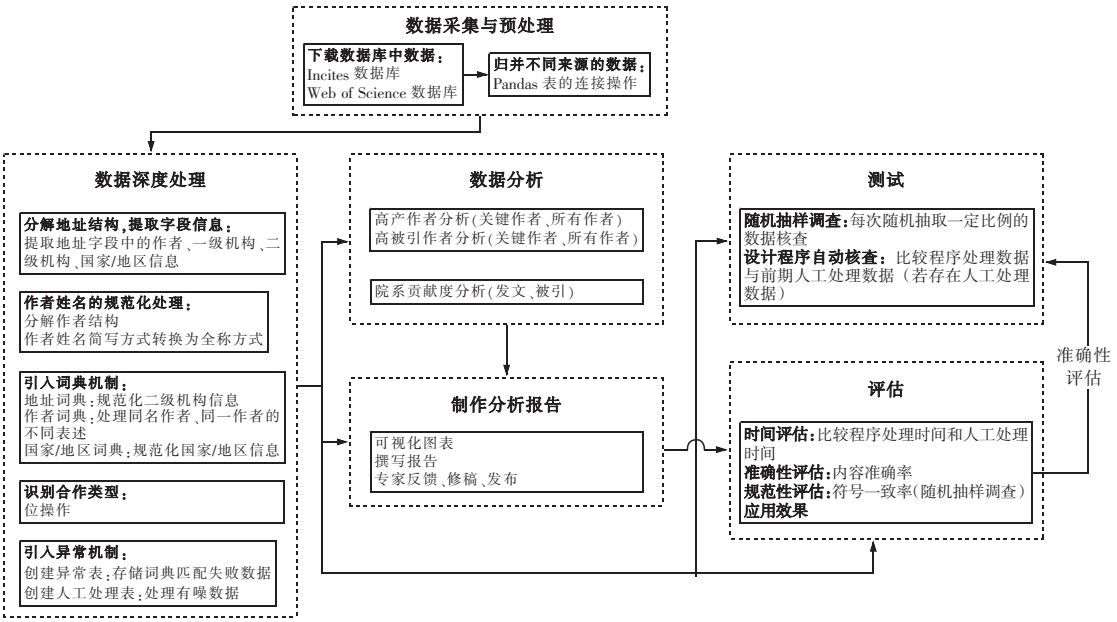

图 1　学科服务数据处理和分析框架

数据包提供了高效操作大型数据集所需的工具[19]，数据处理归并等操作非常简单和高效，可以将“WOS 入藏号”字段作为唯一标识符，对多个数据表进行连接操作。

3.2　数据深度处理

(1)分解地址结构，提取字段信息

Web of Science 数据库中的“C1”字段代表作者及其详细地址，“RP”字段代表通讯作者及其详细地址。在学科分析中，经常需要对这两个字段进行分解，从中提取作者(包括通讯作者)、一级机构(院校)、二级机构(院系)、国家/地区等信息，如图 2 所示。由图 2 可知，“C1”“RP”两个字段由若干个地址块组成，不同地址块之间用“;”连接。根据这两个字段的结构，将作者信息及其地址信息提取到一一对应的关系数据表中。

(2)作者姓名的规范化处理

“RP”字段中的通讯作者姓名为简写方式(如“张小三”显示为“Zhang，XS”)，“C1”字段中的作者姓名为全称方式(如“张小三”显示为“Zhang，XiaoSan”)，Web of Science 数据库中“AF”字段用作者全称表示。为便于处理数据，需要对作者名称的表述方式进行规范化处理，

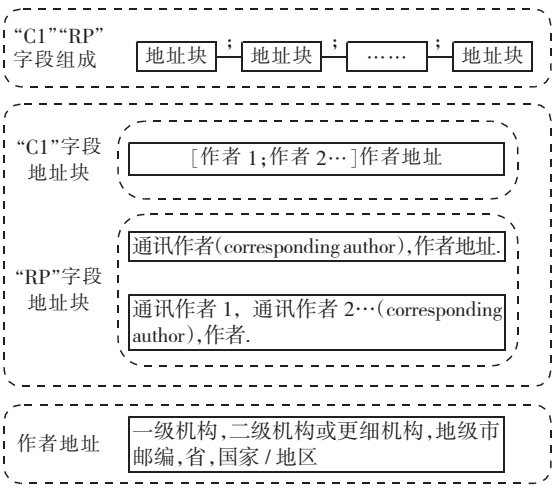

图 2　Web of Science 数据库中“C1”“RP”字段的结构

统一为姓名全称。在实际操作中，只需将“RP”字段中提取出的简写作者姓名与“AF”字段中的全称作者姓名进行匹配即可实现。

(3)引入词典机制

为将机构地址、作者、国家/地区等信息进行规范化处理，特引入词典机制，创建地址词典、作者词典、国家/地区词典。本研究中的“词典”指建立内容的不同表述方式与规范化表述方式之间的联系，可通过 Excel 等软件实

现。地址词典能够解决二级机构地址表述不一致、院系合并分解、名称变迁等问题；作者词典可以解决同一作者名称表述不同、校内部门流动等问题；国家/地区词典可以解决需要对本国特殊地区进行特别分析的问题。

(4)识别合作类型

论文一般有3种合作类型：本校合作、国内合作和国际合作。本研究判定一篇论文只能为其中一种合作类型，现定义如下3种地址类型：①地址类型A为本校地址，可通过正则表达式识别是否为本校；②地址类型B为非本校地址，地址字段中国家/地区被识别为“CHINA”；③地址类型C为国外地址，地址字段中国家/地区被识别为非“CHINA”。对3种合作类型作如下规定：①本校合作，论文中的地址只包含地址类型A；②国内合作，论文中的地址既包含地址类型A，也包含地址类型B，且不包含地址类型C；③国际合作，论文中的地址既包含地址类型A，也包含地址类型C。

识别合作类型可通过位操作中的“或运算”实现。首先，分别为3种地址类型赋值，如表1所示。其次，3种合作类型对应的计算公式和计算结果如表2所示。

表1 3种地址类型对应的二进制数值

地址类型	二进制数值
A	00
B	01
C	10

表2 3种合作类型的计算公式和计算结果

合作类型	计算公式	二进制数值
本校合作	A	00
国内合作	A|B	01
国际合作	A|C或A|B|C	10或11

(5)引入异常机制

本研究通过引入词典机制规范化处理数据，每当有新增数据时，需要对词典及时进行更新维护。为识别词典维护的契机，创建异常表，将词典中未匹配成功的数据(如使用地址词典匹配本校二级机构地址失败)存入异常表。异常表的意义在于查漏补缺，当有新增数据时，只需对异常表中的信息进行检查、识别，并将其补充到词典中。

Web of Science数据库中导出的数据可能存在个别数据信息缺失、有误等情况，需进行人工干预。为此，可创建人工处理表，将经过人工处理的有噪数据存入该表，再替换对应程序处理数据。

3.3 数据分析

(1)高产作者、高被引作者分析

从分析对象看，高产作者和高被引作者分析包括关键作者分析和所有作者分析，其中关键作者指作为第一作者或通讯作者的本校作者。对数据进行深度处理后，本研究以本校作者、二级机构为唯一标识符，去除重复信息后重新聚合数据，分别针对关键作者、所有作者按发文量、被引量重新排序，可得到高产作者、高被引作者分析结果。其中，发文量、被引量以整数计数[17]，即一篇论文针对每个作者、二级单位进行一次计数。高产作者、高被引作者分析流程如图3所示。

(2)院系贡献度分析

院系贡献度分析包括院系发文贡献度分析和院系被引贡献度分析。对数据进行深度处理后，本研究以二级机构地址为唯一标识符，去除重复信息后重新聚合数据，分别按发文占比、被引量占比重新排序，可得到院系发文贡献度、院系被引贡献度分析结果。其中，发文量、被引量以整数计数。

3.4 制作分析报告

基于本研究深度处理和分析的数据，使用Excel软件可生成发文量、论文被引量、论文学术影响力、高产作者、高被引作者、院系贡献度、国际合作等可视化分析图表。学科服务人员基于这些图表撰写分析报告，交予专家审阅，再根据专家意见进行修改，最后发布报告。

3.5 测试

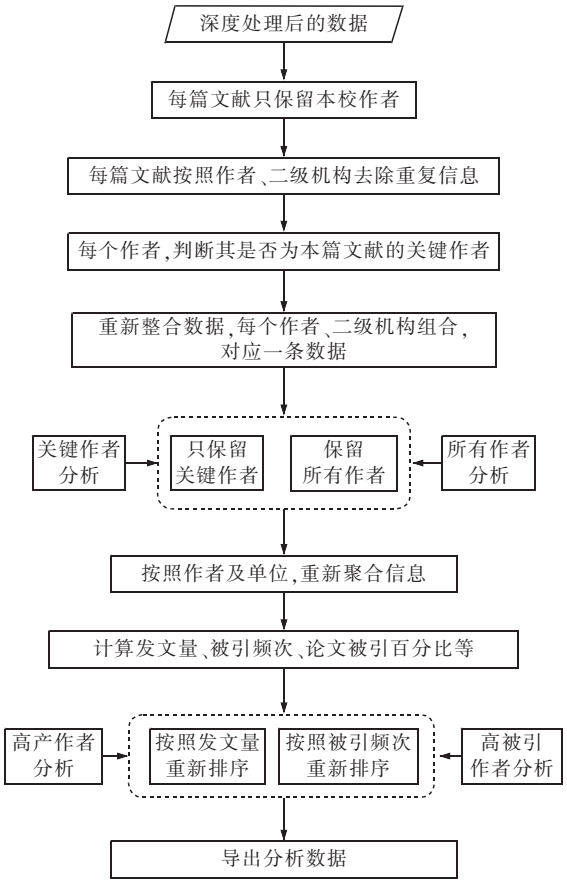

图 3　高产作者、高被引作者分析流程

本研究从两个维度进行测试。第一个维度的测试为随机抽查，即随机抽取一定比例（如 1%）的已深度处理数据，核查数据准确性。第二个维度测试为程序自动核查，前提是存在人工处理数据，本研究设计的程序能够自动核查人工处理数据是否与程序处理数据一致。这两个维度的测试可迭代多次，保证程序处理结果的准确性。

3.6　评估

本研究对框架的评估从 4 个方面进行：①时间评估，即比较程序和人工分别处理数据的时间，以评估自动化程序是否缩短了数据处理时间。②准确性评估，即评估程序处理数据的准确性。③规范性评估，即评估经过程序处理后数据内容的表述、符号、格式是否更统一、更规范。④应用效果评估，即评估通过程序制作学科分析报告的应用效果。

4　学科服务数据处理与分析框架验证：以浙江工商大学为例

4.1　框架构建背景

SCI、SSCI、A&HCI 收录了浙江工商大学 2012－2022 年发表的 5000 余篇科研成果。该校图书馆在开展学科服务时，经常需要对这些数据进行深度处理，如提取作者（第一作者、通讯作者）、作者二级机构地址、合作类型等信息。因经费有限无力购买商业分析软件，而人工处理耗时长、准确率低、不规范现象严重，所以本研究探究计算机自动处理、分析学科服务数据的框架与方法，同时在实际工作中也对其使用效果进行了验证。

4.2　数据来源

本研究以 Incites 数据库（更新于 2022 年 10 月 28 日）中的 InCites Dataset 数据集为数据来源；学科分类体系选择 Essential Science Indicators；出版年设置为 2012－2022 年；文献类型为 Article、Review；机构名称为 Zhejiang Gongshang University；数据采集日期为 2022 年 11 月 11 日；共采集得到 5321 条文献信息。此外，本研究还采集了 Web of Science 数据库中 2012－2022 年该校被 SCI、SSCI、A&HCI 收录的所有文献信息，用于补充 Incites 数据库中缺少的作者详细地址（C1 字段）、通讯作者及其地址（RP 字段）等信息。

4.3　评估和应用

（1）时间评估

时间评估指基于相同数据比较程序处理时间和人工处理时间。分别在程序的开始处及终止处获取系统时间，计算时间差，计算程序运行 10 次所需时间的平均值作为程序处理时间。程序运行环境为普通个人电脑。人工处理时间按实际情况，每天 8 小时只处理数据，提取作者（第一作者、通讯作者）、作者二级机构地址、合作类型等信息。比较程序处理和人工处理数据的时间，结果如表 3 所示。人工处理时间依据的是小样本的经验总结，具有一定

误差，但与程序处理时间相比，差值非常大。由此可见，设计学科服务数据处理和分析的自动化程序可显著缩短数据处理和分析时间。

表 3　程序处理和人工处理数据时间比较

处理论文数量(篇)	程序处理时间(秒)	人工处理时间(天)
100	4.08	≈2
5000	3 分 24	≈100

(2)准确性评估

准确性评估旨在评估数据内容处理的准确性。笔者调查分析人工处理的数据发现，受处理者熟练程度及细心程度的影响，人工处理数据会出现表述不当、信息错误、信息遗漏、信息增生 4 类违背准确性的表现，如表 4 所示。程序处理数据过程中机器处理的特性、引入的词典机制、测试阶段的随机抽样调查均保证了处理内容的准确性；程序稳定运行 3 个月后发现，偶尔会因词典未及时更新出现数据内容不准确的现象，但准确率仍保持在 99%以上。由此可见，依托自动化程序进行学科服务数据处理和分析的结果准确性更高。

表 4　人工处理数据违背准确性表现

类别	具体表现	举　例
表述不当	二级机构地址表述不当(因机构改名、合并；全称与简称混用)	“财务与会计学院”后更名为“会计学院”，“工商管理学院”简称为“管理学院”，人工处理数据中同时存在多种表述
	作者姓名表述不当(全称方式、简写方式混用)	作者姓名“张小三”，存在“Zhang，XiaoSan”“Zhang，XS”混用的情况
	国家、地区名称表述不当(对国际地址不熟悉，误使用城市名作为国家、地区名)	将国家“Portugal”误写为城市“Oporto”
信息错误	合作类型识别错误	当作者单位较多时，容易漏看；误将“国际合作”识别为“国内合作”
	关键作者识别错误	当第一作者单位或通讯作者单位较多时，容易误将“关键作者”识别为“非关键作者”
信息遗漏	遗漏通讯作者	当存在多个通讯作者时，可能遗漏一个或几个通讯作者
	遗漏通讯作者单位	当一个通讯作者存在多个单位时，可能遗漏其中一个或几个单位
	本校作者不全	当作者数较多时，容易遗漏其中几个本校作者
	作者单位不全	当作者数及作者单位数较多时，容易遗漏其中几个作者单位
信息增生	多出错误的机构地址(因 Excel 操作不当)	某篇文献的“第一作者单位”为“计算机与信息工程学院”，人工处理数据误写为“计算机与信息工程学院；信息与电子工程学院”

(3)规范性评估

规范性评估主要评估数据内容的表述、符号、格式是否统一、规范。笔者调查前期人工处理的数据发现，人工处理数据的规范性不高，主要为表述不统一、符号不统一、格式不统一。程序处理数据因其机器处理特性保证了符号、格式的规范性；通过引入词典机制，保证了数据内容表述的统一性。程序处理和人工处理数据结果比较如表 5 所示(表格中字体斜体部分为处理不规范之处)。由表 5 可见，程序处理数据的规范性明显高于人工处理数据。

(4)应用效果评估

基于本研究构建的学科服务数据处理和分析框架及设计的技术程序，可制作生成学科分析报告，包括院系贡献度分析报告、学科竞争力分析报告等，具体处理内容如表6所示。

表 5　程序处理和人工处理数据的规范性比较

程序处理数据结果举例		人工处理数据结果举例	
第一作者	第一作者单位	第一作者	第一作者单位
Zhang, XiaoSan	财务与会计学院	*Zhang, XS*	财务与会计学院
Zhang, XiaoSan	财务与会计学院	*Zhang XiaoSan*	财务与会计学院
Zhang, XiaoSan	财务与会计学院	Zhang, XiaoSan	*会计学院*
Zhang, XiaoSan	财务与会计学院	*Zhang Xiao-san*	*会计学院*
Li, Si	工商管理学院; Hongkong Baptist Univ(HONG KONG)	Li, Si	工商管理学院; Hongkong Baptist Univ(Hong Kong)
Li, Si	工商管理学院; Hongkong Baptist Univ(HONG KONG)	*Li, S*	*管理学院*; Hongkong Baptist Univ (*Hong Kong, China*)

其中，数据深度处理、数据分析等环节的大部分操作可由程序代替人工，极大地提升了工作效率。该程序自 2021 年应用以来，运行稳定，处理速度快、准确性高、规范性佳，效果良好。

表 6　高校学科服务数据处理和分析内容

内容	详　情	程序处理	人工处理
数据采集	下载数据		√
数据预处理	归并不同来源数据	√	
数据深度处理	提取作者(第一作者、通讯作者)、作者二级机构地址并译为中文、合作类型等信息	√	
数据分析	高产作者分析(关键作者、所有作者)	√	
	高被引作者分析(关键作者、所有作者)	√	
	院系贡献度分析(发文、被引)	√	
	其他分析(发文、被引、国际合作分析等)		√
测试	随机抽取核查		√
	程序处理数据与人工处理数据是否一致(若存在人工处理数据)	√	
评估	时间评估、准确性评估、规范性评估、应用效果		√
制作分析报告	可视化图表		√
	撰写报告		√
	专家反馈、修稿、发布		√

5　结语

本研究面向高校图书馆学科服务工作构建了完整、系统的学科服务数据处理和分析框架，基于框架设计数据分析流程，并提出相应技术实现方案。通过引入词典机制，保证了数据内容的规范性和新增数据的可扩展性；通过引入异常机制，为未匹配数据、有噪数据提供了二次处理环节，保证了数据处理的灵活性。

本研究提出的学科分析框架的优越性在于：数据处理的深度、精度、灵活性、扩展性、功能全面性均较高；可节约机构经费；处理时间、准确性、规范性方面优于人工处理方法。未来需进一步优化图形用户界面设计，推进该框架的广泛应用；在框架中增加一键可视化操作功能，实现分析结果的可视化呈现。

参考文献：

[1]李峰，张慧丽，张春红，等. 高校图书馆开展学科竞争力分析的流程与方法——以《北京大学学科竞争力分析报告》为例[J]. 图书情报工作，2020(16)：13-21.

[2]吴爱芝，肖珑，张春红，等. 基于文献计量的高校学科竞争力评估方法与体系[J]. 大学图书馆学报，2018(1)：62-67，26.

[3]Mongeon P，Paul-Hus A. The journal coverage of Web of Science and Scopus：a comparative analysis[J]. Scientometrics，

2016,106(1):213-228.
[4]匡广生,郭岩,俞晓明,等.基于图的多源数据融合框架研究[J].计算机科学,2021(11):170-175.
[5]李慧,胡吉霞,佟志颖.面向多源数据的学科主题挖掘与演化分析[J].数据分析与知识发现,2022(7):1-16.
[6]俞立平,何庆光,韩钰.赋权类非线性学术评价方法伪权重及权重失灵研究——以TOPSIS评价方法为例[J].情报杂志,2022(5):190-197.
[7]Bornmann L, Williams R. An evaluation of percentile measures of citation impact, and a proposal for making them better [J]. Scientometrics, 2020, 124(2): 1457-1478.
[8]姚海燕,罗志宏.基于层次分析法的医院学科建设评价指标体系优化研究[J].中国医院,2022(2):70-72.
[9]Zhang Y, Lu J, Liu F, et al. Does deep learning help topic extraction? A kernel k-means clustering method with word embedding[J]. Journal of Informetrics, 2018, 12(4): 1099-1117.
[10]Timakum T, Kim G, Song M. A data-driven analysis of the knowledge structure of library science with full-text journal articles[J]. Journal of Librarianship and Information Science, 2020, 52(2): 345-365.
[11]俞立平,万晓云,琚春华.学术期刊影响因子过度自引的修正研究——自然影响因子[J].情报理论与实践,2019(11):62-68.
[12]Yu D, Xu Z, Wang X. Bibliometric analysis of support vector machines research trend: a case study in China[J]. International Journal of Machine Learning and Cybernetics, 2020, 11(3): 715-728.
[13]杜蕾,左昊明,李亚设.基于CiteSpace的国内智慧图书馆近十年发文热点及前沿剖析[J].图书馆理论与实践,2021(6):42-49.
[14]张毅.世界大学排名对比分析及其对"双一流"建设的启示[J].北京科技大学学报(社会科学版),2022(2):138-145.
[15]Priem J. Altmetrics[M]// Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact. Cambridge: MIT Press, 2014: 263-287.
[16]熊霞.基于PlumX的外文学术图书影响力评价实证研究[J].四川图书馆学报,2021(3):34-40.
[17]Adams J, Pendlebury D, Potter R. 数尽其用:合作世界的科研贡献管理[EB/OL]. [2022-07-14]. https://img02.ma.scrmtech.com/18476/1812/resource/1645689964/WOS_ISI_Report%2016_Jan22_%E6%95%B0%E5%B0%BD%E5%85%B6%E7%94%A8%EF%BC%9A%E5%90%88%E4%BD%9C%E4%B8%96%E7%95%8C%E7%9A%84%E7%A7%91%E7%A0%94%E8%B4%A1%E7%8C%AE%E7%AE%A1%E7%90%86.pdf.
[18]赵国荣,杨光,肖珑."双一流"高校图书馆学科竞争力分析服务调查与研究[J].图书馆工作与研究,2021(11):41-47.
[19]Pandas[EB/OL]. [2021-07-16]. https://baike.baidu.com/item/pandas/17209606?fr=aladdin.

作者简介:

林 叶(1988—),女,馆员,浙江工商大学图书馆,浙江,杭州,310018;

王丽艳(1972—),女,副研究馆员,浙江工商大学图书馆,浙江,杭州,310018;

王月苗(1989—),女,馆员,浙江工商大学图书馆,浙江,杭州,310018。

Framework Construction for Data Processing and Analysis in Subject Service of University Library

Lin Ye, Wang Liyan, Wang Yuemiao

Abstract The processing and analysis of subject service data is an important part of subject service work, and the key to optimizing subject service work lies in optimizing the processing and analysis process of subject service data. This paper constructs a data processing and analysis framework for discipline services, proposes a technical scheme to realize automatic data processing and analysis based on Python, and takes Zhejiang Gongshang University as an example to evaluate the effectiveness of the application of the framework and technology implementation. The empirical results show that the processing time of the computer program designed based on this framework for automatic processing and analysis of discipline service data is significantly shortened, the accuracy and standardization of the processing results are improved, and the application effect is good.

Keywords Subject service; Subject analysis; Subject construction; Data analysis; University library

Class Number G258.6; G250.7