图书情报工作
第 57 卷 第 23 期 2013 年 12 月

科学计量可视化软件的对比与数据预处理研究*

周晓分 黄国彬 白雅楠

[摘 要] 从软件运行平台、数据来源、数据文件格式要求、数据导入规模与处理规模等角度对 10 款科学计量软件(Bibexcel、Bicomb、CiteSpace、HistCite、NetDraw、Pajek、SATI、SPSS、Ucinet、VOSviewer)的数据预处理要求进行比较,发现:CNKI、万方、CSSCI 和 WoS 数据库的数据可由不同的软件处理;不同的软件仅能处理相应格式的 Text、Excel、Html 和其他文件格式的文件;软件不同,所能处理的数据量也有所差别。

[关键词] 数据预处理 科学计量 可视化软件
[分类号] G353
DOI:10.7536/j.issn.0252-3116.2013.23.011

1 引 言

文献计量分析一般包含 5 个步骤:数据收集、数据预处理、数据挖掘、数据分析和报告撰写,其中数据收集和数据预处理这两个阶段在整个文献计量分析过程中所占的时间最多[1]。科学计量软件中的数据预处理是从原始数据到导入数据之间的过程,包括数据来源的确定、文件格式的要求、数据导入与处理规模等部分。数据预处理这一环节是整个数据分析的基础,预处理的目的是将收集到的数据进行格式转换,使之符合特定科学计量软件的要求,便于以后的统计分析。

然而,笔者在 CNKI 中以“知识图谱”为主题进行检索,对其中利用科学计量软件的文章进行分析后发现:在最终的论文撰写过程中,很多学者往往会忽略对数据预处理这一阶段的介绍,而是直接从“数据来源”进入“数据分析”阶段;一些学者虽然对数据规范做了一定的说明,但大都是对数据进行排序、删除、合并、分类、颜色标识、指标规范、加权处理、归一化处理等,没有对数据来源、文件格式要求、软件处理规模等方面做进一步详细的介绍。

本文在分析“知识图谱”相关论文的基础上,选取 10 款应用比较广泛的科学计量软件(Bibexcel、Bicomb、CiteSpace、HistCite、NetDraw、Pajek、SATI、SPSS、Ucinet、VOSviewer),从软件运行平台、数据来源、数据文件格式要求、数据导入规模与处理规模等角度对数据预处理过程进行详细介绍,以方便研究人员的学习和使用,从而减少软件使用过程中的重复劳动,提高研究效率。

2 可视化软件运行平台要求

可视化软件一般都会运行在使用最普遍的系统平台上,不同的软件对平台的具体要求又有所不同。下面依次对这 10 款软件的运行平台进行简单介绍。

Bibexcel[2]是瑞典科学家佩尔松(O. D. Persson)开发的文献计量学研究软件,用以帮助用户分析文献数据。其主页为 http://www8.umu.se/inforsk/Bibexcel/。该软件既可在 Windows 系统下运行,也可在 Linux 系统下运行。

书目共现分析系统 Bicomb[3]由崔雷开发,受中国卫生政策支持项目(HPSP)资助,主要功能是对生物医学文献数据库中的书目文献信息进行快速扫描、准确提取并归类存储、统计计算、矩阵分析等。软件下载地址为 http://dmi.cmu.edu.cn/dmi/research/Bicomb.php。Bicomb 软件在安装了 Windows 98/2000/NT/XP/Vista 等操作系统的电脑上均可正常运行,不建议使用 Windows ME/2003 等特殊版本。另外,因 Bicomb 的统计功能将利用 Microsoft Excel 生成报表,所以电脑中需要具备 Microsoft Office 办公软件系统;软件系统的界面包含 Flash 动画,要求操作系统中 Flash 的版本在 8 以上。在 Bicomb 的下载界面可看到需要首先运行一遍 bde-install,布置好环境,然后才能运行 Bicomb。

*本文系中央高校基本科研业务费专项资金资助项目“知识图谱软件的技术原理与评价指标体系研究”(项目编号:2012LYB02)研究成果之一。
[作者简介] 周晓分,北京师范大学政府管理学院信息管理系硕士研究生,E-mail:zhouxiaofenheb@126.com;黄国彬,北京师范大学政府管理学院副教授,硕士生导师;白雅楠,北京师范大学政府管理学院信息管理系硕士研究生。
收稿日期:2013-07-17 修回日期:2013-08-12 本文起止页码:64-72 本文责任编辑:王传清

CiteSpace[4]是陈超美博士使用 Java 平台开发的知识图谱可视化工具,该软件通过聚类和时区视图的方式展示一个领域在一定时期内的知识基础与研究前沿。目前最新版本为 2013 年 5 月 25 日更新的 3.5.R7(64 bit)版。此版本运行时,操作系统须为 Windows 64 - bit ,内存要求 2GB RAM,同时必须安装 Java 7 (64 - bit)。在其下载页面 http://cluster.ischool.drexel.edu/~cchen/CiteSpace/download.html 可看到对此的详细说明。目前 CiteSpace 可在 Linux 系统下运行,其脚本是由法国阿维尼翁大学(University of Avignon)的 E. SanJuan 提供的。CiteSpace 的一些历史版本也可在 Mac 系统上过运行,但最近几版都尚未在 Mac 系统上测试。运行 CiteSpace,除了下载到地,也可采用 Webstart 的方式。

HistCite[5],即 History of Cite,是加菲尔德博士开发的一款引文图谱分析软件。最初使用该软件需要收取一定的费用,但现在用户只需签署一份 HistCite 最终用户许可协议(HistCite End User License Agreement),并提交用户的姓名、所在机构与 Email 地址,即可进行任何非商业用途的使用。目前可用版本于 2012 年 3 月 17 日更新,可在 HistCite 主页 http://www.HistCite.com 中下载。HistCite 软件只能用于 Windows 操作系统的电脑,运行时会在本机上搭建一个服务器,并在默认浏览器(如 Internet Explore,Chrome 和 Firefox 等)中打开 HistCite 工作网址 http://127.0.0.1:1925/。

NetDraw[6]由美国肯塔基州立大学商学与经济学院管理系 S. Borgatti 教授开发,是一款免费的社会网络分析软件,由 Analytic Technologies 公司提供。NetDraw 简单易学,容易操作,且只要是 Windows 98 及以上的系统都可以安装使用。主页为:http://www.analytictech.com/NetDraw/NetDraw.htm。最新版本为 2011 年 11 月 26 日更新的 NetDraw 2.118。

Pajek[7]是 1996 年 11 月由伍拉迪米尔·巴塔格利(V. Batagelj)和安德雷·马尔瓦尔(A. Mrvar)使用 Delphi(Pascal)语言开发的大型复杂网络分析工具,其中的一些程序由 M. Zaversnik 提供。Pajek 在 Windows 95 及以上的版本、Linux(64)和 Mac 系统下运行,可免费使用,但仅限于非商业用途。目前 Pajek 软件主页已转至维基百科 http://Pajek.imfm.si/doku.php。

文献题录信息统计分析工具(Statistical Analysis Toolkit for Informetrics, SATI)[8]是国内学者刘启元等利用.NET 平台,使用 C#编程语言设计的一款免费开源的数据统计与分析辅助工具。该软件的官方网站为 http://sati.liuqiyuan.com/。在安装该软件前,需要确保电脑已经安装.NET Framework 4。因国内学者开发,所以有关软件的一切说明都是中文,方便国内学者学习和使用。目前 SATI 最新版本是 2012 年 11 月 20 日更新的 SATI 3.2。

"统计产品与服务解决方案"软件 SPSS(Statistical Product and Service Solutions)[9]是世界上最早的统计分析软件。1968 年由美国斯坦福大学的三位研究生 N. H. Nie、C. Hadlai(Tex)Hull 和 D. H. Bent 研究开发成功。2009 年,IBM 公司收购 SPSS,至今已更新至 21.0.0 版,需选择适合个人计算机的系统版本,登录后方可下载,下载地址为 http://www14.software.ibm.com/download/data/web/en_US/trialprograms/W110742E06714B29.html。SPSS 客户端支持 Windows、Linux 和 Mac OS 操作系统。

Ucinet(University of California at Irvine Network)[10]由加州大学欧文(Irvine)分校的一群网络分析者编写的一款基于菜单驱动的 Windows 程序,现由 Analytic Technologies 公司进行维护更新,该公司由 Roberta Chase 和 Steve Borgatti 共同经营。目前最新版本是官方于 2011 年 11 月 28 日更新的 6.365 版本,下载地址为 http://www.analytictech.com/Ucinet/download.htm,可免费试用 60 天。

VOSviewer[11]是雷登大学 CWTS 研究机构的研究人员 Nees Jan Van Eck 和 L. Waltman 开发的一款免费知识图谱绘制工具。该软件使用 Java 程序语言编写,运行 VOSviewer 前,需要安装 Java 环境(Java 6.0 版本或者更高)。VOSviewer 可下载安装,也可直接点击 launch 在线运行。其最新版本为 VOSviewer 1.5.4 版,下载及 launch 运行页面为 http://www.VOSviewer.com/download/。

以上 10 款软件的运行平台要求的总结见表 1。

3 可视化软件数据来源

科学计量可视化的基础是对某一主题有关的大量数据的收集。通过对"知识图谱"主题的论文进行分析,发现这些论文中的数据通常来源于中外各知名数据库。其中,中文数据主要来源于中国知网(CNKI)、万方和 CSSCI 这三个主流数据库。外文(主要为英文)数据主要来源于 Web of Science(WoS)平台中的 SCI、SSCI、A&HCI 等数据库。各数据库所支持的下载格式与各可视化软件所支持的格式各有不同,下面将以本文的检索为例,对此做详细说明。

表 1　10 款可视化软件的运行平台要求

软件名称	最近更新日期	是否免费	操作系统	内存	其他
Bibexcel	未说明	是	Windows 系统,Linux 系统	无	无
Bicomb	未说明	是	Windows 98/2000/NT/XP/Visat 等,不建议使用 Windows ME/2003 等特殊版本	无	具备 Microsoft Office 办公软件系统,Flash 版本在 8 以上,运行 bde-install,布置好环境
CiteSpace 3.5.R7（64 bit）	2013 年 5 月 25 日	是	Windows 系统,Linux、Mac 系统	2GB RAM,历史版本最少为 1024MB	Java 7（64 - bit）
HistCite	2012 年 3 月 17 日	是	Windows 系统	无	有默认浏览器（如 Internet Explore、Chrome 和 Firefox 等）
NetDraw 2.118	2011 年 11 月 26 日	是	Windows 98 及以上	无	无
Pajek 3.12	2013 年 5 月 20 日	是	Windows 95 及以上,Linux(64),Mac 系统	无	无
SATI 3.2	2012 年 11 月 20 日	是	Windows 系统	无	安装. NET Framework 4
SPSS 21.0.0	未知	否	Windows 系统,Linux、Mac 系统	无	无
Ucinet 6.365	2011 年 11 月 28 日	否	Windows 系统	无	无
VOSviewer 1.5.4	未说明	是	Windows 系统	无	最新 Java 运行环境

3.1　数据下载与保存

笔者进入各数据库的下载界面,限定主题为“知识图谱”或“mapping knowledge domains”,获得各数据库相应的检索结果,检索日期为 2013 年 5 月 29 日。

3.1.1　CNKI 数据库　从 CNKI 可得 686 条数据。由于每页可显示的最大结果数为 50 条,CNKI 规定高版本浏览器(IE8.0 及以上版本)可支持导出/参考文献的最大数据量是 500 条,因此若需导出这 686 条数据,可将每页显示结果数改为 50,然后勾选本页全部 50 条数据,接着点击“下一页”,勾选“第二页”的全部数据,直至勾选满 500 条数据为止,如图 1 所示:

图 1　CNKI 导出数据限制

若要导出剩余数据,只需点击“清除”,选择所要导出的数据即可。选择好文献后点击“导出/参考文献”,勾选全部数据后继续点击“导出/参考文献”,可得到 CNKI 文献输出界面,见图 2。

从图 2 可以看出,对于数据集输出格式,CNKI 除提供. xls 和. doc 两种格式外,还提供其他 10 种引文的文本输出格式,研究者可根据需要选择相应的引文格式。其中,CAJ-CD 格式引文、CNKI 查新(引文格式)、

图 2　CNKI 导出数据格式

CNKI 查新(自定义引文格式)、Refworks、Endnote、NoteFirst、自定义(支持需输出更多文献信息的查新等用途)的数据集都输出为 text 格式文件,text 文件的编码格式为 UTF - 8;NoteExpress 可保存为. net 文件;CNKI 桌面版个人数字图书馆保存为. cnt 格式文件;CNKI e - learning 保存为. eln 格式文件。

3.1.2　万方数据库　在万方数据库的高级检索中进行检索,可得 683 条数据。万方检索结果中每页最多显示 50 条数据,但一次只允许导出 100 条数据。选择 100 条数据后,点击“导出”,可得图 3。

从图 3 可见,万方一共提供 8 种引文输出格式:导出文献列表、参考文献格式、NoteExpress、RefWorks、NoteFirst、EndNote、自定义格式、查新格式。其中,参考文献格式、RefWorks、NoteFirst 、EndNote、自定义格式、查新格式所导出的文件格式为 txt 格式,编码格式为 UTF - 8;NoteExpress 导出的文件格式为 net 格式;“导

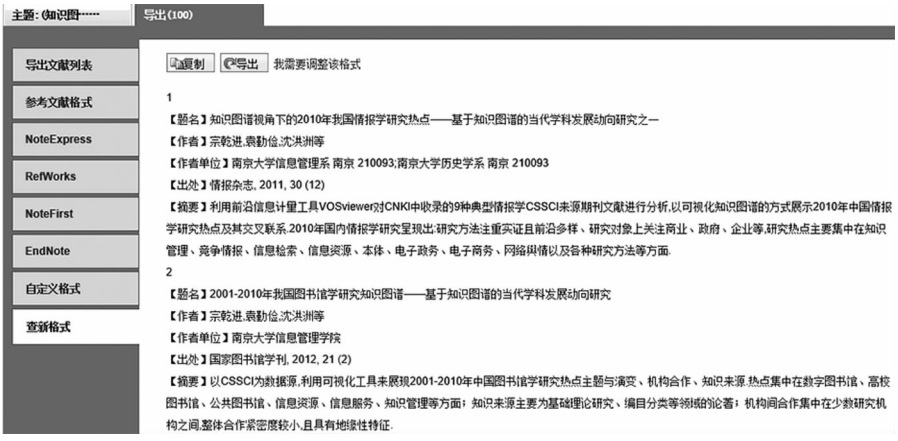

图 3 万方导出数据格式

出文献列表”需要手工将数据复制粘贴到文本(如 doc,txt 等)中。

3.1.3 CSSCI 数据库 在 CSSCI 检索页,有来源文献和被引文献两种检索途径,数据年限从 1998 年至 2012 年,目前来源数据库又增加了 CSSCI 扩展版(试运行中)2008、2010、2011 年的来源数据库和 2011 - 2012 年港澳台及海外华文期刊数据库(试验版)。以来源文献的检索为例,限定“来源篇名(词)”为“知识图谱”,可得 132 篇文献。目前 CSSCI 每页显示的最大结果数为 50 条,每次下载的数据量也仅为 50 条。这 50 条数据集的下载格式为 txt 格式,编码格式为 ANSI。

3.1.4 Web of Science 数据库 外文(主要为英文)数据主要来源于 Web of Science 平台中的 SCI、SSCI、A&HCI 等数据库,可得检索结果有 2 490 条数据。WoS 每页可显示 50 条记录,支持下载的数据集为 500 条。但在下载前需要对记录数进行标记,以防记录重复下载。

WoS 共提供 7 种导出格式:保存为纯文本格式;保存到其他参考文献软件,保存为制表符分隔的格式(win,utf - 8)和(mac,utf - 8)的 txt 文件编码都为 utf - 8;保存为制表符分隔的格式(win)和(mac) 格式的 txt 文件编码均为 Unicode;保存到 bibtex 的文件格式是 .bib,可用 ultraedit 打开;保存为 HTML 的文件格式为 .html。

3.1.5 其他数据库 Bicomb 还可以处理生物医学文献数据库 PubMed 中的数据。

CiteSpace 还可处理 arXiv、Derwent 专利数据库(http://www.pencils.co.uk/)、美国国家科学基金会 NSF Awards、Project DX、斯高帕斯数据库 Scopus 和 SDSS 数据库中的数据。

Pajek 向以下网络提供分析和可视化操作工具:合著网、化学有机分子、蛋白质受体交互网、家谱、因特网、引文网、传播网(AIDS、新闻、创新)、数据挖掘(2 - mode 网)等[12]。

3.1.6 其他来源 NetDraw、SPSS、Ucinet 和 VOSviewer 还支持其他社会网络分析软件的相应格式文件。VOSviewer 也可接受其他软件(如 SPSS、Pajek、NetDraw 等)绘制的图谱。详见 4.4 小节。

3.2 可视化软件对数据的要求

在数据预处理过程中,首先需要了解的是不同软件对数据来源、数据格式、数据量等的要求,这样才能进行后续的操作。下面以数据来源为依据,将其对数据的要求做一整理。

3.2.1 支持 CNKI 数据 包括 Bicomb、CiteSpace、SATI 和 Ucinet。

Bicomb 能够处理 CNKI 的数据,但需要进行一些编码转换,详见 4.1 小节。CiteSpace 能够处理中文数据,但进行处理前需要选择中文编码(Preferences——Chinese encoding)。需要注意的是,CNKI 数据需要先导出 Refworks 格式,且 CNKI 数据不包含参考文献,做图时选择 Cited Reference 是不能得到可视化图谱的。SATI 是国内学者刘启元等开发的,所以支持较多的中文数据。软件开发者建议:使用 CNKI 的数据时,数据导出为 EndNote/ EndNote2(知网旧版)格式,因为 CNKI 提供的 EndNote 格式题录数据较为完整[8]。Ucinet 能够处理中文数据,但原始数据须为矩阵格式。所以,CNKI 数据需要先转换为矩阵形式。矩阵的获得,既可以使用 SATI 3.2,也可以使用 Excel 的数据透视表功能[13]。VOSviewer 能够处理转换后的 CNKI 的数据,详见 4.4 小节。

3.2.2 支持万方数据 包括 Bicomb 和 SATI。

在这 10 款软件中,只有 Bicomb 和 SATI 明确支持万方数据的处理。Bicomb 对万方数据库中的数据没有额外的要求;在 SATI 中,因为万方提供的 NoteExpress/NoteFirst 格式题录数据较为完整,所以 SATI 开发者推荐使用万方数据(WF)提供的 NoteExpress/NoteFirst 格式题录数据。

3.2.3 支持 CSSCI 数据 包括 Bibexcel、CiteSpace、SATI、Ucinet、VOSviewer。

使用 Bibexcel 和 CiteSpace 处理 CSSCI 的数据前，都需要利用刘盛博开发的软件 CSSCIREC 进行格式转换。CSSCIREC 也需要 Java 运行环境，只要 CiteSpace 能正常运行，CSSCIREC 就可以正常运行。CSSCIREC（可从陈超美的科学网博客介绍中下载）程序运行界面会显示如何下载数据和保存数据。SATI 支持 CSSCI 数据库新旧版平台导出的来源文献。Ucinet 对 CSSCI 数据的要求同对 CNKI 数据的要求。VOSviewer 能够处理格式转换后的 CSSCI 数据，详见 4.4 小节。

3.2.4　支持 WoS 数据　包括 Bibexcel、Bicomb、CiteSpace、HistCite、SATI、Ucinet 和 VOSviewer。

Bibexcel、Bicomb、CiteSpace 和 HistCite 都要求将从 WoS 下载的数据以纯文本的形式保存。SATI 推荐 Web of Science 数据库平台导出的题录数据为 HTML/text 格式[8]。Ucinet 只能处理矩阵形式的 WoS 数据，转换方式同 3.2.1 小节。VOSviewer 能处理格式转化后的 WoS 数据，详见 4.4 小节。

3.3　自建数据

可视化软件不仅能够处理数据库中的数据，还可以自建数据以方便研究人员的使用。提供此类功能的软件有：NetDraw、SPSS、Ucinet 和 Pajek。

NetDraw 使用记事本文件创建数据，需要按照 NetDraw 所要求的数据描述格式来描述节点信息。总体来说，要描述的内容共分为三个部分：Node Data、Node Properties 和 Tie Data，但研究者可以根据需要来对这三部分做选择性的描述，不必每个文件都包含三部分。Node Data 主要包含用于描述网络中节点所代表的研究对象的属性；Node Properties 部分和 Tie Data 部分很相似，不同的是前者所包含的变量一般是用来描述节点的坐标、大小、颜色和形状等，而 Tie Data 主要用于描述节点与节点之间的关系属性[14]。

SPSS 在数据编辑窗口建立数据文件。进入 SPSS 数据编辑窗，可看到 SPSS 数据文件是按个案（行）和变量（列）组织的。在此数据文件中，个案表示各个调查对象。变量表示对调查中提出的每个问题的回答。SPSS 数据文件的文件扩展名为. sav，包含所保存的数据。

Ucinet 自建数据文件有两种方法：①利用写字板或 Microsoft word 等任一种文字处理程序，在一个文本文件中输入矩阵数据。或者，利用 Ucinet 自带的文本编辑器“file——text editor”将矩阵数据保存为纯文本文件，再利用“Data——import text file——DL/Raw”可将文件转换为 Ucinet 格式的数据。②使用 Ucinet 的数据编辑器，见图 4。可在其中直接输入或粘贴 Excel 矩阵数据[15]。

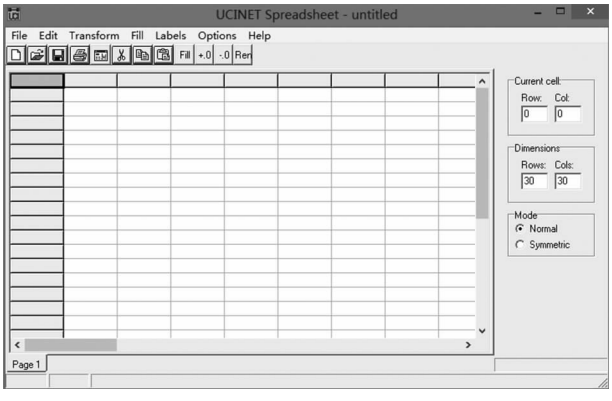

图 4　Ucinet 数据编辑器

Pajek 自建数据有 4 种途径：采用手工方式、利用专用辅助软件、使用字处理软件和应用关系型数据库软件。因篇幅限制，在这里不再一一详细介绍，可参阅相关文献[12]。

表 2 清楚地展现了这 10 款可视化软件对数据来源的要求。

表 2　10 款可视化软件数据来源要求

数据来源	单次下载数据量（条）	可导出文件格式	支持该数据库的软件
CNKI	500	. xls，. doc，. txt（编码格式为 UTF-8），. net，. cnt，. eln	Bicomb、CiteSpace、SATI、Ucinet
万方	100	. doc，. txt（编码格式为 UTF-8），. net	Bicomb、SATI
CSSCI	50	. txt（编码格式为 ANSI）	Bibexcel、CiteSpace、SATI、Ucinet、VOSviewer
WoS	500	. txt（编码格式为 UTF-8），. txt（编码格式为 Unicode），. bib，. html	Bibexcel、Bicomb、CiteSpace、HistCite、SATI、Ucinet、VOSviewer
其他数据库	–	–	Bicomb、CiteSpace、Pajek
其他软件产生的文件	–	–	NetDraw、SPSS、Ucinet、VOSviewer
自建数据	–	–	NetDraw、SPSS、Ucinet、Pajek

4　数据文件格式要求

最常见的文件格式是 text、word 和 excel 文件。从第三部分内容可看出，不是每种文件都可以由可视化软件来处理，且每种数据库都可提供多种文件格式，那么具体每种文件格式都可由哪种软件进行处理呢？下面将主要介绍 text、excel、html 和其他文件格式的数据

预处理要求。

4.1 text 格式

Bibexcel、Bicomb 、CiteSpace、HistCite、NetDraw 和 Ucinet 都可处理 text 格式的文件。其中,Bibexcel 所能处理的最初文件格式为 txt 格式,之后经过 Bibexcel 自身的转换,可生成. doc,. out,. cit,. oux,. coc,. ma2,. xls 格式的文件,具体过程可参考相关文献[16]。但有一点需要注意,在转换 txt 数据时,需先用“Editpad Lite”转换成 Windows 格式,否则所得的 doc 文件的内容为空。NetDraw 和 Ucinet 支持使用记事本创建的数据,具体方法可见 3.3 小节。Bicomb 、CiteSpace 和 HistCite 对 text 文件格式的要求可如表 3 所示:

表 3 10 款可视化软件所能处理的文件格式

	text 格式	Excel 格式	html 格式	其他软件格式
Bibexcel	用“Editpad Lite”转换成 Windows 格式	–	–	–
Bicomb	从数据库下载的记录,都应以 txt 的形式储存。对于从 CNKI 等中文文献库中下载的记录,需要打开. txt 文件,将编码选项从 UTF – 8 改为 ANSI 格式后另存	–	–	–
CiteSpace 3. 5. R7 (64 bit)	从 WoS 和 CNKI 下载的数据须以纯文本的形式保存,并以 download * . txt 的形式命名;对于 Project DX 的数据,包含 node 信息的文件命名形式为 * nodes. txt,包含 edge 信息的文件命名形式为 * edges. txt;从 Scopus 数据库中下载的数据保存为完整的 CSV 格式,然后把 CSV 格式保存为制表分隔符的 txt 文件,并将文件以 download * . txt 的形式命名	从 research. gov 网站下载获奖记录时,需将数据保存为 xlsx 格式	处理 NSF Awards 数据时,若是从 www. nsf. gov 网站下载的获奖记录,则需保存为 xml 格式	–
HistCite	text 格式	–	–	–
NetDraw 2. 118	使用记事本创建的数据			Ucinet 系统文件(. ##h),Ucinet DL(Ucinet 以 DL 语言描述数据格式的文件),Pajek 文件和软件自身的格式文件(. vna)
Pajek 3. 12	–	–	–	蜘蛛网络格式. net(Pajek networks),蜘蛛矩阵格式. mat(Pajek matrices),Vega 格式,世系谱数据的标准数据格式(GEDCOM 格式,Ucinet DL Files),三种文本格式(Ball And Stick,Mac Molecule 和 MDL MOL)用于化学专业
SATI 3. 2	text(WoS)格式、CSSCI 数据格式,但在分析前都需转换为 xml 格式	–	其他格式的文件自动转化为 xml 格式 SATI 专用数据文件。目前,其提供转换的其他文件格式有:html(WoS)格式、EndNote 格式、NoteExpress 格式、NoteFirst 格式、zotero(开源文献管理插件)	–
SPSS 21. 0. 0	文本编辑器软件生成的 ASCII 数据文件 * . txt	Excel 文件(* . xls, * . xlsx),无其他特殊要求	–	SPSS 文件(* . sav),由 dBASE、FoxBASE、FoxPRO 产生的 * . dbf 文件,文本编辑器软件生成的 ASCII 数据文件(* . dat),Access(* . mdb),SAS(* . sd7, * . sd2)等数据文件
Ucinet 6. 365	使用记事本创建的 text 文件	Excel 格式	–	csv, ntf, dl, net 等格式的文件,数据语言形式的文本文件(Data Language,DL),多元数据文件(Multiple Data Files),VNA 软件使用的文件(VNA),Pajek 软件使用的文件(Pajek),Krackplot 软件形式的文件(Krackplot),Negopy 软件使用的文件,ASCII 型的初始文件(Raw)
VOSviewer 1. 5. 4	–	–	–	map 文件和 network 文件

4.2 Excel 格式

CiteSpace 在处理从 research. gov 网站下载的获奖记录时,需将数据保存为 xlsx 格式;Ucinet 能够处理 Excel 格式的数据,详见 3.3 小节;SPSS 也可处理 Excel 格式的数据,详见 4.4 小节。

4.3 html 格式

CiteSpace 处理 NSF Awards 数据时,若是从 www. nsf. gov 网站下载的获奖记录,则需保存为 xml 格式。SATI[8]开发者规定:为方便后期题录数据的存储、交换和分析,要将其他格式的文件自动转化为 xml 格式 SATI 专用数据文件。目前,其提供转换的文件格式有:html(WoS)格式、text(WoS)格式、EndNote 格式、NoteExpress 格式、NoteFirst 格式、zotero(开源文献管理插件)和 CSSCI 数据格式。

4.4 其他文件格式

NetDraw、Pajek、SPSS、Ucinet 和 VOSviewer 还能够处理自身及其他可视化软件的导出结果。这里详细介绍 SPSS 和 VOSviewer 所处理的其他软件格式的预处理要求,NetDraw、Pajek 和 Ucinet 的预处理要求见表 3。

SPSS 所处理的数据文件有两种来源:①SPSS 环境下建立的数据文件,即在 SPSS 数据编辑窗口建立数据文件,见 3.3 小节;②调用其他软件建立的数据文件。SPSS 可以调用的多种文件格式见表 3。另外,通过使用 ODBC(Open Database Capture)的数据接口,可以直接访问以结构化查询语言(SQL)为数据访问标准的数据库管理系统,通过数据库导出向导功能可以方便地将数据写入到数据库中等[17]。

VOSviewer 能够处理两种主要的文件类型:map 文

件和 network 文件,这两种文件都是简单的文本文件。因此 VOSviewer 不能直接处理 WoS 的输出文件,需要中间步骤将 WoS 文件转换为 map 文件或 network 文件(如 Pajek 等)。在 VOSviewer 使用手册中,开发者推荐使用免费的 SAINT Toolkit[18]。但是,我们也可以利用 Pajek 将 WoS 文件转换为. net 文件。经过试验,发现也可以使用 VOSviewer 处理中文数据。具体方法为:将 CSSCI/CNKI 数据导入 SATI 3.2 进行中文共现分析,再依次通过 Ucinet、NetDraw 软件将格式转换为. net 文件,即可导入 VOSviewer 进行相应的分析。

图 5　SATI 导入数据时“XML 格式保存”的对话框

5　数据语种、导入与处理规模

每种软件的运行速度都会受到数据量的影响。数据预处理时,也需要了解可视化软件相应的数据处理能力,才能保证软件的正常运行。数据导入规模和处理规模是反映可视化软件数据预处理能力的两个重要方面,而语种则是导入数据前要首先考虑的内容。

5.1　语种要求

除了 HistCite 只支持英文数据,其余 9 款软件都支持中文和英文数据。其中,Bibexcel 还支持瑞典语;SPSS 还支持德、法、日、意等语言,可在“菜单 - 编辑—选项—常规”界面,在输出和用户界面中的语言下拉列表中更改语言选项。

5.2　数据导入规模

5.2.1　一次只能导入一个文件　可视化软件 Bibexcel、Bicomb、HistCite、NetDraw、Pajek、SATI、SPSS 和 VOSviewer 一次只能导入一个文件。需要注意的是,使用 SATI 导入数据时,需按照下述步骤进行:选择转换格式——单文件/文件夹——去重——转换,且必须弹出“XML 格式保存”的对话框(见图 5),保存在原始数据所在文件夹,数据才是真正导入成功。使用 Pajek 时,将第一次打开的所有文件保存为. paj 的文件,下次只要打开其中一个 paj 文件,就可打开所有 paj 文件。Ucinet 大部分程序仅可导入一个文件,小部分可导入多个文件。

5.2.2　一次导入一个文件夹或多个文件　CiteSpace、Bicomb 和 SATI 支持一次导入单个文件夹。HistCite 则可导入多个文件。

5.3　数据处理规模

Bibexcel、Bicomb 和 HistCite 并未对数据处理规模做详细说明。

CiteSpace、SATI 和 VOSviewer 未限制数据处理规模。但考虑到软件运行速度及其他因素(如计算机配置高低)的影响,对数据规模还是做相应限制较好。

NetDraw 软件在处理数据时,若数据文件为 vna 格式,那么 NetDraw 可处理的数据规模可以很大。如对于非常稀松的数据,计算机的 RAM 为 1G 时,NetDraw 可处理 3 500 个节点,为 2G 时可处理 10 000 个节点[6]。

Pajek 可以处理多达 999 999 997 个顶点的网络。一般情况下,绘制网络图不应超过几千个顶点。因为对庞大的网络进行绘图将会非常费时,而且画出来的图也往往不美观。在默认情况下,Pajek 不对超过 5 000个顶点的网络进行绘图。用户可在”options—read - write“菜单中修改这个限制。当然,计算机本身的配置性能也可能造成其他方面的限制[12]。

SPSS 对数据处理规模虽然没有明显限制,但为了保证运行速度,多数情况下建议处理 100 × 100 以内的矩阵。

Ucinet 能处理 32 767 个网络节点。当然,从实际操作来看,当节点数在 5 000 - 10 000 之间时,一些程序的运行就会很慢[19]。

对这 10 款软件的数据语种、导入和处理规模的概

括,如表 4 所示:

表 4 10 款可视化软件数据语种、导入和处理规模

软件名称	导入规模	处理规模	语种
Bibexcel	导入一个文件	未说明	中、英、瑞典语
Bicomb	导入一个文件或一个文件夹	未说明	中、英
CiteSpace 3.5.R7(64 bit)	导入一个文件夹	无限制,但若数据量过多,会处于无反应的状态	中、英
HistCite	导入一个或多个文件	未说明	英
NetDraw 2.118	导入一个文件	文件为 vna 格式时,可处理的数据规模可以很大	中、英
Pajek 3.12	将第一次打开的所有文件保存为. paj 的文件,下次只要打开其中一个 paj 文件,就可打开所有 paj 文件	可以处理多达 999 999 997 个顶点的网络,在默认情况下,不对超过 5 000 个顶点的网络进行绘图	中、英
SATI 3.2	导入一个文件或一个文件夹	无限制	中、英
SPSS 21.0.0	导入一个文件	多数情况下建议处理 100×100 矩阵	中、英、德、法、日、意等
Ucinet 6.365	大部分程序仅可导入一个文件,小部分可导入多个文件	处理 32 767 个网络节点;当节点数在 5 000 – 10 000 之间时,一些程序的运行会很慢	中、英
VOSviewer 1.5.4	导入一个文件	无限制,可处理数以万计的数据	中、英

6 结 语

掌握数据预处理的方法对熟练使用科学计量软件,节约科研时间,加速科研进度有很大帮助。从以上 10 种科学计量软件的分析中可看出,这 10 款软件大都能够处理主流数据库(CNKI、万方、CSSCI、WoS)的数据,但在数据文件格式以及数据导入等方面有较大的限制,且不同软件对数据预处理的要求有很大不同。以同是科学知识图谱绘制工具的 CiteSpace 和 VOSviewer 为例,两者都可处理来自 CNKI、CSSCI 和 WoS 数据库中的数据,但两者在文件命名、数据导入规模等方面存在不同。若对这些细节有所忽略,就会造成使用上的障碍。希望本文能对读者了解这 10 款软件的数据预处理方法有所帮助。

参考文献:

[1] 虞飞华. 基于 Google Scholar 的文献计量分析研究的数据预处理技术[J]. 情报杂志,2008(12):48 – 50.

[2] Persson O D, Danell R, Schneider J W. How to use Bibexcel for various types of bibliometric analysis[M]//Celebrating Scholarly Communication Studies: A Festschrift for Olle Persson at His 60th Birthday. Leuven: International Society for Scientometrics and Informetrics,2009:9 – 24.

[3] 书目共现分析系统 Bicomb 用户操作使用说明书[EB/OL]. [2013 – 05 – 12]. http://dmi. cmu. edu. cn/dmi/resource/20120724. pdf.

[4] 陈超美. CiteSpace Ⅱ:科学文献中新趋势与新动态的识别与可视化[J]. 情报学报,2009(3):401 – 421.

[5] Histcite[EB/OL]. [2013 – 05 – 28]. http://www. histcite. com.

[6] NetDraw[EB/OL]. [2013 – 05 – 11]. http://www. analytictech. com/Netdraw/netdraw. htm.

[7] Mrvar A, Batagelj V. Pajek and Pajek – XXL programs for analysis and visualization of very large networks reference manual: List of commands with short explanation version 3.12[EB/OL]. [2013 – 05 – 29]. http://pajek. imfm. si/lib/exe/fetch. php? media = dl:pajekman. pdf.

[8] 刘启元,叶鹰. 文献题录信息挖掘技术方法及其软件 SATI 的实现——以中外图书情报学为例[J]. 信息资源管理学报,2012(1):50 – 58.

[9] SPSS 中国[EB/OL]. [2013 – 05 – 25]. http://www. spss. com. cn/index. aspx.

[10] Ucinet[EB/OL]. [2013 – 05 – 13]. http://www. analytictech. com/Ucinet/.

[11] VOSviewer[EB/OL]. [2013 – 05 – 22]. http://www. vosviewer. com/.

[12] 诺伊,姆尔瓦,巴塔盖尔吉. 蜘蛛:社会网络分析技术[M]. 2 版. 林枫,译,李葆嘉,审订. 北京:世界图书出版公司北京公司,2012.

[13] 魏瑞斌. 社会网络分析在关键词网络分析中的实证研究[J]. 情报杂志,2009(9):46 – 49.

[14] 王运锋,夏德宏,颜尧妹. 社会网络分析与可视化工具 NetDraw 的应用案例分析[J]. 现代教育技术,2008(4):85 – 89.

[15] 刘军. 整体网讲义——Ucinet 软件实用指南[M]. 上海:格致出版社,2009.

[16] 姜春林,陈玉光. CSSCI 数据导入 Bibexcel 实现共现矩阵的方法及实证研究[J]. 图书馆杂志,2010(4):58 – 63,42.

[17] 数据管理[EB/OL]. [2013 – 05 – 12]. http://zhibao. swu. edu. cn/epcl/spss/spss_edit/spss2. html.

[18] Eck N J v, Waltman L. VOSviewer manual[EB/OL]. [2013 – 05 – 10]. http://wenku. baidu. com/view/84cc2fd33186bceb19e8bb0a. html.

[19] Ucinet[EB/OL]. [2013 – 05 – 09]. http://baike. baidu. com/view/2343008. htm.

Comparison Between Scientific Visualization Metrology Software and the Data Pretreatment

Zhou Xiaofen　Huang Guobin　Bai Yanan
School of Government, Beijing Normal University, Beijing 100875

[**Abstract**] The paper selects 10 most widely used scientific metrology software, which are Bibexcel, Bicomb, CiteSpace, HistCite, NetDraw, Pajek, SATI, SPSS, Ucinet and VOSviewer, and gives a review of data pretreatment requirements in detail from the aspects of software platform, data source, data file format requirements, data import scale and the processing size of data. This paper finds that different software can process different databases like CNKI, Wanfang, CSSCI and WoS; different software can only handle the appropriate file formats, such as Text, Excel, Html, and others; different software can deal with different amount of datum.

[**Keywords**] data pretreatment　scientometrics　metrology software

(上接第 63 页)

A Study on Library Users' Self-service Use Behavior: The Case of Self-return Service

Yang Tao
Library of South China Normal University, Guangzhou 510631

[**Abstract**] This paper studies library users' self-service use behavior through self-return data of South China Normal University Library between 2009 and 2012 school year. It analyses users' self-service usage behavior in terms of self-service usage, time distribution, time spending, quantity distribution, and user types. The results show that library users prefer to use self-return, and prefer to use self-service during opening hours. Self-service cannot save users time. Self-service may not be able to increase collection circulations. There are differences between different kinds of user. User distribution is broadly in line with the thirty-seventy rule. Library should make a correct understanding of the significance of self-service and actively promote self-service, plan well in self-service, train users to improve the efficiency of self-service, and guarantee service for key users.

[**Keywords**] self-service　user behavior　self-return