基于SCIE、ESI的学科信息分析工具的设计与实现*

李 晶 师俏梅

(西北工业大学图书馆 西安 710072)

摘 要 以SCIE收录论文为分析对象,利用ESI的指标数据建立学科信息分析工具。对SCIE收录论文的学术影响力、SCIE收录论文的期刊学术影响力、以及SCIE收录论文在ESI的22个学科领域的分布情况等全面分析,采用文献计量学和情报研究的方法,为跟踪了解学科的发展态势、评价学科与科研发展,调整并优化学科建设规划提供相关依据。

关键词 科学引文索引 基本科学指标 学科服务 学科信息分析 VBA

中图分类号 G353.1 G350.7 **文献标识码** A **文章编号** 1002-1965(2012)10-0117-05

The Design and Implementation of Subject Information Analysis Tool Based on SCIE and ESI

LI Jing SHI Qiaomei

(Library, Northwest Polytechnic University, Xi'an 710072)

Abstract Taking SCIE as studying pool, we use ESI indicators data to establish the subject of information analysis tools and analyze the academic influence and the distribution of papers in 22 subject fields of Essential Science Indicators (ESI) by the bibliometrics and intelligence research method in order to track the trend of the development of subject, evaluate the subjects and R&D, adjust and optimize the subject construction plans to provide evidence.

Key words SCI Essential Science Indicators(ESI) Subject Service Subject Information Analysis VBA

0 引 言

学科是高校教学和学术研究的基本构建单元和基层机构。一所高校的学科结构状况如何,怎样正确评价学科间的科学关系,正视并解决学科发展与建设中存在的问题,对做好高校的中长期发展规划,推进学校各项事业的发展具有重要意义。论文的学科或专业属性分布直接对应于高校的科研分布体系,其产出的数量和质量能够描述高校的科研活动方向,反映不同学科文献产出能力的大小[1]。

当前,国内众多研究型图书馆纷纷设立学科馆员岗位,学科服务工作已成为热门研究课题。图书馆员借助数据库分析工具,开展科研成果的分析和评价工作,有助于我们及时把握相关学科的发展动态,了解相关学科的优势和特色,从而可以更有针对性地提供与院系师生实际需求相适应的个性化的学科信息支持。

Essential Science Indicators[2](基本科学指标)是基于Web of Science (Science Citation Index Expanded和Social Sciences Citation Index)权威数据建立的分析性数据库,能够为科技政策制定者、科研管理人员、信息分析专家和研究人员提供多角度的学术成果分析。利用ESI能够分析国家、研究机构、期刊的学术影响力;发现各学科领域的热点和前沿研究成果;获取国家、机构、期刊和论文在全球各学科中的排名信息;揭示特

收稿日期:2012-05-21 修回日期:2012-07-05

基金项目:本文系2010年度西北工业大学高等教育研究基金项目"新媒体时代基于开源软件的高校图书馆建设研究"(编号:2010GJZ05)的研究成果之一。

作者简介:李 晶(1980-),女,硕士,馆员,研究方向:数字图书馆开源软件;师俏梅(1970-),女,副研究馆员,研究方向:特色数据库建设、科技查新、用户教育、情报研究。

定学科领域中研究成果和学科整体的影响力现状。

Web of Science 数据库收录了 10000 多种世界权威的、高影响力的学术期刊,内容涵盖自然科学、工程技术、生物医学、社会科学、艺术与人文等领域,最早回溯至 1900 年。Web of Science 收录了论文中所引用的参考文献,并按照被引作者、出处和出版年代编制成独特的引文索引。

为了更好的分析 SCIE 收录论文的影响力水平和学科分布情况,了解优势学科和潜力学科的发展现状及趋势,为学校的科研绩效评估提供可靠的情报依据,深化图书信息服务的内容,西北工业大学图书馆开发设计了基于 SCIE、ESI 学科信息分析工具。该工具生成的分析结果对于及时跟踪了解高校各学科的发展态势、评价高校学科与科研发展,调整并优化学科建设规划具有可借鉴性。

1 学科信息分析工具的设计目标

为了实现分析 SCIE 收录论文的影响力水平和学科分布情况功能,基于 SCIE、ESI 的学科分析工具的主要设计目标如下:

a. 分析论文影响力分布。提供基于 ESI 提供的全球论文影响力基准值(Baselines)[3]–即 22 个学科中每年发表论文的 6 个百分位水平(0.01%,0.1%,1%,10%,20% 和 50%)的被引次数基准值,SCIE 收录论文影响力分布情况的数据列表和图表。b. 分析论文学科分布。能够生成 SCIE 收录论文在 ESI 的 22 个学科的分布情况的图表。c. 分析刊载论文期刊分布。能够生成基于中国科学院文献情报中心按年度和学科对 SCIE 期刊进行 4 个等级的分区的 SCIE 收录论文整体分区情况的图表。d. 分析论文第一作者单位分布。生成按年度统计以高校为第一作者或通讯作者单位发表的 SCIE 收录论文数据报表。

2 学科信息分析工具的模块功能设计

本分析工具通过数据导入模块、数据浏览模块、数据提取模块、数据处理、数据浏览、数据统计以及图表生成模块等功能模块实现学科信息的分析,如图 1 所示。

a. 数据导入模块负责对 SCIE 子库检索结果的全记录数据保存为纯文本格式的 txt 文件作为数据处理对象,导入到分析工具中的原始数据记录表中。b. 数据浏览模块用于将 SCIE 的记录信息从原始数据记录表中读取并将记录号、题名等相关信息显示到当前窗口,便于馆员查看和浏览原始记录信息。c. 数据提取模块用于清洗数据的格式、提取题名、作者、来源、作者单位以及记录号等字段信息,并将提取后的信息保持到以及记录年份命名的数据表中。d. 数据处理模块是本工具的核心模块,负责添加基于中国科学院文献情报中心对应的期刊的 1–4 级分区信息,增加期刊对应 ESI 的学科领域,标示以高校为第一作者或通讯作者单位发表的 SCIE 收录论文的记录信息,增加基于 ESI 提供全球论文影响力基准值的百分位水平值。e. 数据统计模块用于统计分析数据。SCIE 收录论文的 ESI 学科领域分布情况、统计按年度统计以高校为第一作者或通讯作者单位发表的论文数量、统计收录论文所在期刊的 1–4 级的分区情况以及统计基于 ESI 提供全球论文影响力基准值影响力水平。f. 图表生成模块可以将统计分析数据生成直观的示意图表。

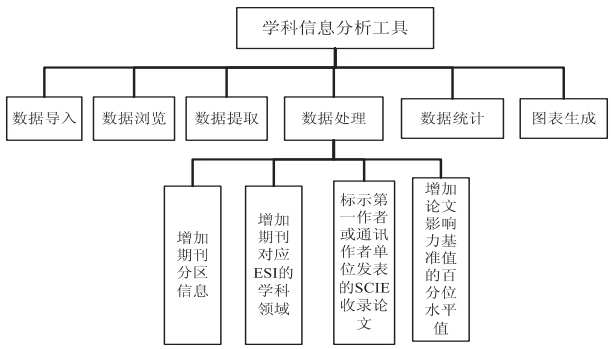

图 1 学科信息分析工具的功能模块

3 系统功能实现

考虑到系统的易用性和可维护性,系统开发采用基于 Microsoft Excel 2007 的 VBA[4,5] 语言。系统无需安装任何程序,界面设计简单,操作方便。界面如图 2 所示。

为了便于数据的统计和结果对比,建立了以年份命名的数据表用来存放 SCIE 收录论文的相关信息,字段如表 1 所示。

表 1 以年份命名的 SCIE 收录论文相关信息表

字段名称(简写)	字段名称(英文)	字段名称(中文)	说 明
AU	Authors	作者	SCIE 提取内容
AF	Author Full Name	作者全名	SCIE 提取内容
SO	Publication Name	出版物名称	SCIE 提取内容
TC	Times Cited	被引频次	SCIE 提取内容
SN	ISSN	ISSN	SCIE 提取内容
UT	Unique Article Identifier	文章唯一标识符	SCIE 提取内容
PY	Year Published	出版年	SCIE 提取内容
WC	Web of Science Category	Web of Science 类别	SCIE 提取内容
CI	Author Address	作者地址	SCIE 提取内容
期刊影响因子			参照最新的 JCR
期刊分区信息			参照中国科学院文献情报中心期刊分区信息
期刊归属大类			ESI 的 22 个学科领域
论文影响力的百分位水平			ESI 提供全球论文影响力基准值 6 个百分位水平

图 2　学科信息分析工具界面

3.1　**记录数据浏览模块的实现**　记录数据浏览模块的实现是依据原始记录索引数据将记录的记录号、收录号、题名、作者姓名、作者单位、来源、期刊年卷期字段内容提取并显示在程序主窗口。

3.2　**数据处理模块功能的实现**　数据处理模块是本工具的核心模块,本模块主要是对 SCIE 收录论文信息表增加 5 个字段,分别是期刊影响因子、期刊分区信息、期刊归属 ESI 的学科领域、论文影响力的百分位水平、是否为第一作者或通讯作者单位发表的论文。

SCIE 收录的科学论文涉及到基础研究的各个领域,由于各学科属性与发展特点各异,以及数据库统计源的学科结构存在巨大差别,使得不同学科的影响因子和被引频次分布存在不均衡性,很难进行学科间的比较和评价。为了更科学地对学术期刊进行评价,更合理地考核科研人员的工作业绩,中国科学院文献情报中心按年度和学科对 SCIE 期刊进行 4 个等级的分区,一般而言,分区位置越高,期刊的影响力越大,命中率也相对较低。这种区分为 SCIE 期刊提供了一个较合理的评价标准,使得对某一刊物水平高低的认定维持在一个相对平衡的定义上。增加期刊分区字段是通过匹配 SCIE 记录信息的 ISSN 号与中国科学院 SC IE 期刊分区表的 ISSN 号,从而获得期刊的分区信息。增加该字段可以揭示 SCIE 收录论文的期刊水平的分布情况。

ESI 依据期刊类别对期刊文献进行分类,对收录的 1 万多种学术期刊进行分析,认为这些期刊中绝大部分是高度专业化的学术期刊,可以单独归入一个学科领域,只有约 60 多种期刊(如 Nature、Science 等)由于刊发的文章覆盖不同学科领域而被归为多学科(Multidisciplinary Sciences)[6]。显然,ESI 收录的 1000 多万篇文献依据所属期刊绝大部分按照一对一的关系被归入 22 个学科领域,对于出现在多学科类期刊中的文献 ESI 进行了二次归类。利用从汤森路透的 http://sciencewatch.com 网站下载最新的期刊列表文件[7],期刊列表文件包含了 11683 种期刊,包含期刊简名、期刊全名、期刊的 ISSN 号、期刊的所属大类四个字段信息。增加期刊对应的 ESI 学科领域字段,利用原始数据的 ISSN 号对应匹配,将所有文献依据所属期刊分别归入 ESI 的 22 个学科领域,按年统计的 SCIE 收录文献在 ESI 的 22 个学科领域的分布情况。通过该该字段的信息可以全面揭示高校强势学科和发现潜力学科。

利用 ESI 提供全球论文影响力基准值(Baselines)-即 22 个学科中每年发表论文的 6 个百分位水平(0.01%,0.1%,1%,10%,20% 和 50%)的被引次数基准值数据表,通过分析记录的期刊对应的 ESI 的学科领域、数据年份、引用次数字段信息取得论文的影响力基准值。通过该字段的信息可以了解高校 SCIE 论文的影响力水平。部分关键代码如下:

```
For i = 2 To Sheets(TextBox17.Value).Range("
```

```
A65535").End(xlUp).Row
    k = 0
    If IsEmpty(Sheets(TextBox17.Value).Cells(i,"Q").Value) Then
    Else
    sum = sum + 1
    For j = 1 To 22
    If StrComp(LCase(Sheets(TextBox17.Value).Cells(i,"Q").Value),LCase(Sheets("大类基准线").Cells(j,14).Value)) = 0 Then
    k = Sheets("大类基准线").Cells(j,15).Value
    Exit For
    End If
    Next j
    For m = 1 To 13
    If StrComp(TextBox17.Value,Sheets("大类基准线").Cells(1,m).Value) = 0 Then
    n = m
    Exit For
    End If
    Next m
    If Sheets(TextBox17.Value).Cells(i,5).Value >= Sheets("大类基准线").Cells(k,n).Value Then
    Sheets(TextBox17.Value).Cells(i,"R").Value = "0.01%"
    ElseIf Sheets(TextBox17.Value).Cells(i,5).Value >= Sheets("大类基准线").Cells(k+1,n).Value And Sheets(TextBox17.Value).Cells(i,5).Value < Sheets("大类基准线").Cells(k,n).Value Then
    Sheets(TextBox17.Value).Cells(i,"R").Value = "0.1%"
    ElseIf Sheets(TextBox17.Value).Cells(i,5).Value >= Sheets("大类基准线").Cells(k+2,n).Value And Sheets(TextBox17.Value).Cells(i,5).Value < Sheets("大类基准线").Cells(k+1,n).Value Then
    Sheets(TextBox17.Value).Cells(i,"R").Value = "1%"
    …..
```

3.3 数据统计模块的实现 为了便于数据的分析比较,将不同类型统计信息建立了独立的数据表,本工具中有基于ESI的baseline统计数据信息表、基于ESI大类分布引用次数统计数据信息表、基于ESI大类分布统计数据信息表、基于第一作者统计数据信息表以及基于中科院大类分区统计数据信息表5张统计信息表。每张信息表行是相关数值区间或者相关数值,列是年份。基于ESI的baseline统计数据信息表的结果如表2所示。

统计按年度统计第一作者或通讯作者单位发表的论文数量时需要按单位地址匹配,在这里使用like()函数,VBA内建的模式匹配功能提供了丰富的字符串比较方式,在模式表达式中可以使用通配符、字符列表(或字符区间)的任何组合来匹配字符串,支持?、*、#、[字符列表]以及[!字符列表]匹配。使用该函数可以精准匹配到第一作者的单位地址信息。部分关键代码如下:

表2 基于ESI的baseline统计数据信息表结构

Baselines	2001年	2002年	2003年	2004年	2005年	…	2010年
0.01%	论文数量						论文数量
0.10%							
1%							
10%							
20%							
50%							
低于50%	论文数量						论文数量

```
    For i = 2 To Sheets(TextBox17.Value).Range("A65535").End(xlUp).Row'匹配地址
    If (LCase(Sheets(TextBox17.Value).Cells(i,19).Value) Like "northwest * poly * univ *") Or (LCase(Sheets(TextBox17.Value).Cells(i,19).Value) Like "nw poly * univ * or nwpu") Or (LCase(Sheets(TextBox17.Value).Cells(i,19).Value) Like "710072") Or (LCase(Sheets(TextBox17.Value).Cells(i,19).Value) Like "710129") Or (LCase(Sheets(TextBox17.Value).Cells(i,19).Value) Like "n. w. poly * univ *") Or (LCase(Sheets(TextBox17.Value).Cells(i,19).Value) Like "nw * poly * univ *") Or (LCase(Sheets(TextBox17.Value).Cells(i,19).Value) Like "north-west * poly * univ") Then
    Sheets(TextBox17.Value).Rows(i).Interior.ColorIndex = 0
    j = j + 1
    End If
    Next i
    For m = 3 To 30 '得对应年份的所在数据列
    If StrComp(TextBox17.Value,Sheets("基于第一作者统计数据").Cells(1,m).Value) = 0 Then
    myyear = m
    Exit For
    End If
    Next m
    Sheets("基于第一作者统计数据").Cells(2,myyear).Value = j
    MsgBox ("完成!")
```

3.4 图表生成模块 图表生成模块用于将统计数据表的数据信息生成图表,既可以生成以学校的1年或者多年的SCIE收录论文的相关数据分析图表,同时还可以生成以学科发展的态势数据分析图表。利用该工具可以快速生成学校或学科相关信息的折线图、饼图或者柱状图等多种图表。

以我校为例从SCIE在Web of Science的SCIE数据库中检索“地址”为“northwest * poly * univ *,并且“出版年”为“2001~2010”的科技论文,统计时间

截至2011 年 10 月 28 日,按照年份下载相关全纪录数据,利用学科信息分析工具生成的 2001–2010 年的论文影响力发展态势如图3 所示,其中纵轴为 SCIE 收录的论文数量。2001–2010 年的论文期刊影响力分析如图4 所示。2001–2010 年的论文基于 ESI 的22 个学科领域论文数量前 6 名的学科如图 5 所示,可以看出材料科学、物理、工程以及化学学科是我校的强势学科。结合分析结果我们可以看到学校的 SCIE 收录论文的影响力低于 50% 所占的比重最大,SCIE 收录论文期刊分区无分区和 4 分区达到了 70% 以上,为此我校应在保持已入选学科论文数量的基础上,要注意强调论文的质量;扩大国内科研机构间以及国际间科研项目合作的广度和深度, 特别是与顶级高校和科研机构展开合作研究, 产生出高引用频次论文。

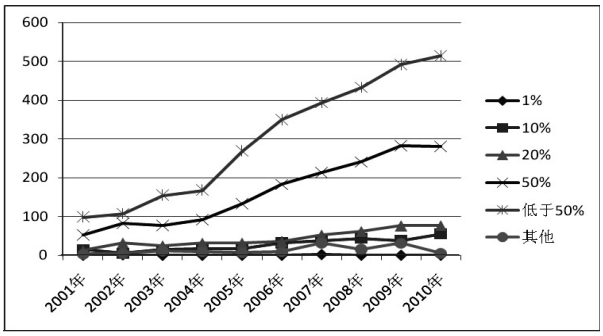

图 3　2001–2010 年西北工业大学 SCIE 收录论文的影响力态势图

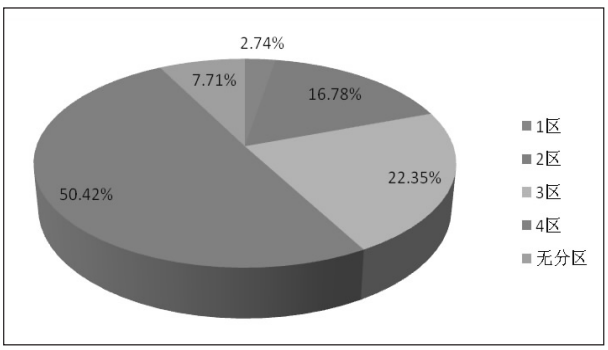

图 4　2001–2010 年西北工业大学 SCIE 收录论文基于中国科学院文献情报中心期刊分区分布图

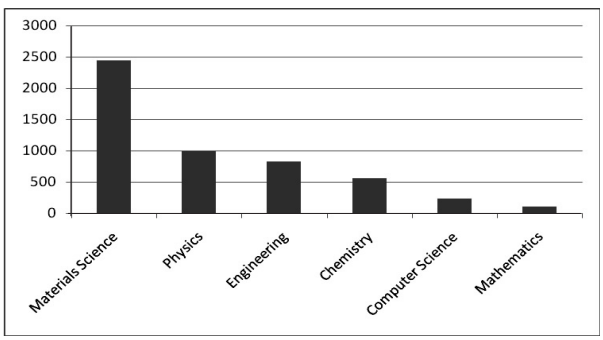

图 5　2001–2010 年西北工业大学 SCIE 收录论文基于 ESI 的 22 个学科领域前 6 名学科

4　结束语

通过构建基于 SCIE、ESI 的学科信息分析工具,可以快速、精准的掌握学校或学科发展态势分析,SCIE 收录论文的影响力水平、SCIE 收录论文期刊水平、以及论文第一作者或通讯作者单位发表的 SCIE 收录论文数量和比例是了解学科发展的重要量化数据,学科馆员利用学科信息分析工具结合对口院系的优势和定位提供与其实际需求相适应的个性化的学科信息支持, 更好地服务于学校的教学和科研工作。从而定量揭示高校学科的优势和弱势,便于高校建设发展的决策者发挥自身的比较优势、找出问题和差距,从而有针对性地提高国际竞争力和影响力。

参 考 文 献

[1]　毛　莉,陈惠兰. 从 JCR 期刊分区看高校学科与科研发展[J]. 科技管理研究,2010(17):101–105

[2]　http://esi.webofknowledge.com/home.cgi

[3]　Percentiles for papers published by field[EB/OL]. [2012–03–20]. http://esi.webofknowledge.com/percentilespage.cgi

[4]　Excel Home. Excel VBA 实战技巧精粹[M]. 北京:人民邮电出版社,2008

[5]　李　政,梁海英,李　昊. VBA 应用基础与实力教程[M]. 北京:国防工业出版社 ,2005

[6]　Classification Of Papers In Multidisciplinary Journals[EB/OL]. [2011–12–20]. http://sciencewatch.com/about/met/classpapmultijour/

[7]　Journal List[EB/OL]. [2011–12–20]. http://sciencewatch.com/about/met/journallist/

(责编:王平军)