基于 VBA 的解剖学核心期刊论文关键词词频统计

秦燕霞 王 朋

（济宁医学院解剖学教研室 日照 276826） （济宁医学院图书馆 日照 276826）

〔摘要〕 以解剖学核心期刊《解剖学杂志》为例，对其关键词词频进行计量分析，通过 VBA 编程实现对关键词的提取和对频次、高频关键词及词长的统计，认为医学期刊应在关键词选择方面进行规范化处理，数据库商在论文加工方面应对标准化问题进行完善，进而为科研提供高质量的数据储备。

〔关键词〕 VBA；解剖学；关键词；词频分析；文献计量

Keywords Frequency Statistics of Anatomy Core Journal Papers Based on VBA *QIN Yan – xia*, *Anatomy Department of Jining Medical College*, *Rizhao 276826*, *China*; *WANG Peng*, *Library of Jining Medical College*, *Rizhao 276826*, *China*

〔**Abstract**〕 Taking anatomy core journal: *Chinese Journal of Anatomy* as an example, the paper analyzes keywords frequency, through VBA programming it realizes the statistical analysis for the extracting of keywords, calculation of frequency, high – frequency words, words length. It considers that medical journals should make the keywords choosing specification, database providers should perfect their standardization issues so as to provide high quality data for scientific research.

〔**Keywords**〕 VBA; Anatomy; Keywords; Frequency analysis; Bibliometrics

1 引言

词频分析是一种情报分析研究方法，其理论基础是词频的波动与社会、情报现象之间存在内在联系，从而透过词频现象看内容本质的科学方法[1]。关键词是表达文献主题概念的自然语言词汇，能够反映文献的核心内容，因此一个学术领域在某一时期内大量学术论文的关键词的集合，可以揭示该领域学术研究的发展脉络与发展方向[2]。期刊论文关键词标引，是我国文献出版发行工作标准化、规范化建设与国际接轨的需要，也是促进我国学术研究成果走向世界的必由之路。关键词标引质量的优劣，直接影响着期刊质量和科研成果的检索效率[3]。随着信息技术的发展，医学统计分析日益受到科研工作者的重视，对医学期刊载文的统计分析能够预测医学前沿进展趋势，评价期刊质量和影响因子，为医学发展提供助力。目前国内对期刊载文关键词的统计分析以呈现统计分析结果为主[4-5]，较少阐述关键词统计分析技术实现，而医学科研工作者了解关键词统计分析实现过程对论文撰写、科研研究等都有较大现实意义。

VBA 是 Visual Basic For Application 的简称，是建立在 Office 中的一种应用程序开发工具，可以有效地自定义和扩展 Excel 的功能。用 VBA 对 Excel 进行二次开发可以自动完成许多机械重复式的工

〔修回日期〕 2010 – 04 – 13

〔作者简介〕 秦燕霞，硕士，助教，主要研究方向为医学信息计量与分析，发表论文数篇。

作，实现用户的许多个性化功能，如对行列的转换、冗余数据的过滤、有效数据的提取统计等，为科研工作带来极大便利。本文以解剖学核心期刊《解剖学杂志》为例，探讨解剖学核心期刊论文关键词词库设计，望给医学科研工作者带来便利。

2 数据来源

从中国知网的《中国期刊全文数据库》选取了1999—2008 年《解剖学杂志》论文，然后通过 VBA 编程，经过行列转换和无效数据过滤，自动滤掉会议通知、征稿简则、年度索引等非正式论文，得到可获取关键词的论文 1 550 篇，提取后的关键词部分数据，见表 1。

表 1 经过行列转换和数据过滤后的关键词示例

关键词
基底核;; 三维重建;; 可视化
核因子 - κB;; 诱导型一氧化氮合酶;; 重症急性胰腺炎;; 肺损伤;; 大鼠
…
阿尔茨海默病;; 海马;; β - 淀粉样前体蛋白;; 转基因小鼠;; 凋亡

3 关键词提取及计量分析分类

首先利用 VBA 编程对表 1 中的关键词列中的单元格进行单个关键词的提取，见表 2。

表 2 关键词提取后的数据列表示例

关键词	关键词	关键词	关键词	关键词
基底核	三维重建	可视化		
核因子 - κB	诱导型一氧化氮合酶	重症急性胰腺炎	肺损伤	大鼠
…	…	…	…	…
阿尔茨海默病	海马	β - 淀粉样前体蛋白	转基因小鼠	凋亡

VBA 程序代码如下：

```
Sub 关键词提取（）
Dim i, irowcount, icol, ifind, iflaglen, j As Integer
Dim stemp, sflag As String  '定义变量
i = 1  '第一篇期刊论文记录的所在行数
irowcount = 1550  '待进行关键词拆分的期刊论文记录总行数
icol = 3  '关键词字段所在列
j = 1  '关键词拆分后工作表的列初始值
sflag = ";;" '关键词字段中的分割标记符
iflaglen = 2 '关键词字段中的分割标记符所占字符长度
For i = 1 To irowcount '遍历行循环
    stemp = Worksheets（" sheet1"）. Cells（i, icol）
  '表 1 所示表的名称为 sheet1，提取后的表名称为 keywords。
    For j = 1 To 8'假定关键词数量最多为 8 个，最多遍历 8 次。
        ifind = InStr（1, stemp, sflag）'返回标记符在字符串中位置
        If ifind > 0 Then'如果含标记符就进行提取，并调整待分析字符串。
            Worksheets（" keywords"）. Cells（i, j） = Mid（stemp, 1, ifind - 1）
            stemp = Mid（stemp, ifind + iflaglen）
        Else  '如果标记符在字符串中位置为 0，将最后一个值提取，然后跳出循环。
            Worksheets（" keywords"）. Cells（i, j） = stemp
            Exit For
        End If  'if 语句结束
    Next  'for 循环控制符
Next  'for 循环控制符
End Sub  '过程结束
```

表 2 中的关键词需要归列，然后进行词频统计，最后对关键词去重，得到高频关键词列表，统计 1 550 篇期刊论文中共出现关键词 6 115 个，高频关键词，见表 3。

表 3 高频关键词列表示例

关键词	频次	关键词	频次
大鼠	259	脑	32
海马	81	应用解剖	32
小鼠	64	视网膜	30
免疫组织化学	54	发育	28
神经元	51	神经干细胞	28
脑缺血	44	细胞分化	27
脊髓	38	心	26
凋亡	37	三维重建	23
一氧化氮合酶	36	胚胎	22
细胞凋亡	33	生长抑素	21

关键词归列、词频统计及去重的 VBA 程序代码如下：

```
Sub 关键词归列（）
Dim i, j, k As Integer
Dim stemp As String  '定义变量
j = 2    '关键词归列后工作表的列初始值
For i = 1 To 1550   '遍历行循环
    For k = 1 To 8'假定关键词数量最多为 8 个，最多遍历 8 次。
        stemp = Worksheets (" keywords"). Cells (i, k)
        If stemp = "" Then  '如果关键词为空，则跳出循环。
            Exit For
        End If
        If InStr (1, stemp, " ") < > 0 And Asc (Mid (stemp, InStr (1, stemp, " ") + 1, 1)) = 63 Then '过滤字符“?”的代码
            stemp = Mid (stemp, 1, InStr (1, stemp, " ") - 1)
        End If
            Worksheets (" count"). Cells (j, 1) = stemp  '关键词归列后工作表名称为 count
            j = j + 1
      Next   'for 循环控制符
    Next     'for 循环控制符
End Sub      '过程结束
Sub 关键词词频统计（）
Dim i As Integer
Dim stemp As String  '定义变量
For i = 2 To 6116   '遍历行循环
    stemp = Worksheets (" count"). Cells (i, 1)
    Worksheets (" count"). Cells (i, 2) = WorksheetFunction. CountIf (〔a2：a6116〕, stemp)
Next  'for 循环控制符
End Sub  '过程结束
```

在对关键词词长统计过程中，需要验证词长的频数之和是否等于关键词的个数 6 115，如果不一致，则需要调整测试关键词最大长度代码，表 4 为关键词长度值与对应的频数列表。

表 4 关键词长度频数

长度值	频数	长度值	频数
长度为 1	204	长度为 12	31
长度为 2	1531	长度为 13	18
长度为 3	1074	长度为 14	14
长度为 4	1395	长度为 15	8
长度为 5	657	长度为 16	3
长度为 6	539	长度为 17	1
长度为 7	236	长度为 18	1
长度为 8	162	长度为 19	0
长度为 9	120	长度为 20	0
长度为 10	66	长度为 21	1
长度为 11	50	长度为 22	4

关键词词长统计的 VBA 代码如下所示：

```
Sub 关键词词长统计（）
Dim i As Integer
Dim stemp As String  '定义变量
For i = 2 To 6116   '遍历行循环
    stemp = Worksheets (" count"). Cells (i, 1)
    Worksheets (" count"). Cells (i, 4) = Len (stemp)
    If Len (stemp) > 22 Or Len (stemp) < 1 Then  '测试关键词最大长度代码
        Worksheets (" count"). Cells (1, 1) = i
    End If
Next  'for 循环控制符
Worksheets (" count"). Cells (7, 11) = WorksheetFunction. CountIf (〔d2：d6116〕, " 1")
Worksheets (" count"). Cells (8, 11) = WorksheetFunction. CountIf (〔d2：d6116〕, " 2")
……
Worksheets (" count"). Cells (27, 11) = WorksheetFunction. CountIf (〔d2：d6116〕, " 21")
Worksheets (" count"). Cells (28, 11) = WorksheetFunction. CountIf (〔d2：d6116〕, " 22")
End Sub  '过程结束
```

（下转第 54 页）