·业务研究·

文献计量分析研究的分类与处理流程

化柏林

（中国科学技术信息研究所，北京 100038）

摘 要：文献计量统计分析的流程包括数据获取、数据预处理、统计计算与应用分析四大模块。数据来源分为数据库数据与网页数据，获取方式分为手工获取与自动获取。数据预处理主要是格式转换、拆分与提取，并过滤掉不符合要求的数据。统计计算按统计结果可分为 Top N 统计、奇异值统计、数量分布统计、年度增长统计、其它关联统计等。

关键词：计量分析；实现技术；数据获取；统计分类；流程

中图分类号：G35；TP311　　**文献标识码**：A　　**文章编号**：1007－7634(2007)09－1332－05

Classification and Process in Bibliometric and Analytic Research

HUA Bo－lin

(*Institute of Scientific and Technical Information of China*, *Beijing* 100038, *China*)

Abstract: Process of bibliometric analysis is made up of data acquire, data preprocess, statistical calculate and application analysis. Data comes from database or web pages, which is acquired by manual or automation. Data preprocess is charge of format transformation, split and extraction, filter. Statistic process is classified by statistic result with Top N statistic, singular value statistic, distribution statistic, annual increase statistic and other relation staticstic.

Key words: bibliometric; implement technology; data acquiring; classification of statistic; process

1 引　　言

计量分析的研究分为计量分析理论研究与计量分析应用研究。计量分析理论研究主要是研究计量分析的规律，与其数据量没有直接关系，只不过需要用适当的数据来进行验证，是数据无关的。计量分析理论研究主要包括空间分布规律、时间变化规律等。空间分布规律包括期刊分布规律、作者分布规律、词频分布规律等，如布拉德福定律、洛特卡定律、齐普夫定律等[1]；时间变化规律包括文献增长规律、文献老化规律等[1]。

应用型计量分析研究主要是针对特定的目的，利用特定的数据进行特定的研究，是数据相关的。从研究目的上应用型计量分析研究主要分为四类：评价型计量分析、主题型计量分析、预测型计量分析、资源获取型计量分析。预测型计量分析和资源获取型计量分析研究比较少。评价型计量分析以引文分析为典型，通过文献之间的引用关系对论文、作者、机构和期刊等进行评价，小规模分析是选取一种或几种期刊的某一段时间内的全部论文进行分析，如文献[2－4]。主题计量分析，选取某一时间段期刊论文进行关键词统计分析、高产作者、高产机构、区域分布、国内外对比分析等，如对知识管理[5]、网络信息计量学[6]等的统计分析，主题计量分析突出使用关键词或主题词，主要运用简单统计

收稿日期：2007－03－05

作者简介：化柏林（1977－），男，山东人，助理研究员，硕士，从事自然语言处理研究.

和关联统计技术，目前此类文章占计量分析文章的主流。这样的分析数据量大都在几百篇到几千篇，数据量较小，统计计算较容易，计算时可以有适当的人工干预。

使用清华同方、万方数据、重庆维普三大期刊全文数据库检索计量分析方面的文章，发现计量分析的文章非常多，但论述计量分析如何来实现的文章却很少，见表1。即使有，也大都直接做成管理信息系统并封装起来，把统计做成与导入、查询相并行的模块，对用户的开放性不够。这类论文（如文献[7-9]）的论述主要是关注数据库结构（字段名、字段类型等）、数据访问接口（如ODBC、ADO）、查询的实现、结果的输出等方面，而对统计实现以及计量分析技术的探讨很不充分，这样对关注文献计量的非技术人员的启迪也较少。目前的应用型文献统计分析缺乏在相应的统计软件里进行简单的编程实现多式多样的统计，把简单的工具用活用好来解决现实的复杂问题。因此小数据量（几万条之内）统计分析的量佳方案是通过VBA在excel里进行。

表1　三大期刊全文数据库相关文章检索结果(检索时间:2006/12/15)

数据库	检索范围	检索条件(专业检索人口)	结果
		(题名=计量分析 or 题名=文献计量 or ((题名=论文 or 题名=文献) and (题名=计量 or 题名=统计)))	2689
清华同方期刊全文库	全部期刊 1979~2006	(题名=计量分析 or 题名=文献计量 or ((题名=论文 or 题名=文献) and (题名=计量 or 题名=统计))) and (篇名=实现)	2
		(题名=计量分析 or 题名=文献计量 or ((题名=论文 or 题名=文献) and (题名=计量 or 题名=统计))) and (篇名=实现 or 篇名=系统 or 篇名=自动)	40
		(dc.title=文献 or dc.title=论文) and (dc.title=计量 or dc.title=统计)	1293
万方数据期刊全文库	全部期刊 1979~2006	(dc.title=文献 or dc.title=论文) and (dc.title=计量 or dc.title=统计) and dc.title=实现	2
		(dc.title=文献 or dc.title=论文) and (dc.title=计量 or dc.title=统计) and (dc.title=实现 or dc.title=自动 or dc.title=系统)	22
		(((T=计量)+T=统计)*((T=文献)+T=论文))	2771
重庆维普期刊全文库	全部期刊 1989~2006	((((T=计量)+T=统计)*((T=文献)+T=论文))*T=实现)	3
		((((T=计量)+T=统计)*((T=文献)+T=论文))*(((T=实现)+T=自动)+T=系统))	33

2　数据源获取

计量分析的数据来源主要有两种形式，一种是直接从数据库商获取数据库数据，这种形式需要与数据库商有很好的协商与沟通。第二种是从数据库商的网页上批量下载数据，其中下载分为手工下载与自动下载。手工下载是构造检索表达式进行检索，得到详细记录，然后复制检索结果网页上的相关内容。自动下载是通过程序构造URL，然后根据URL下载HTML网页文件，读取下载的网页文件，滤掉HTML标签，根据字段名获取数据记录。目前，较小规模的研究大都采用手工复制检索结果网页的形式。

动态网页的URL参数传递有两种方法，第一种是直接写在URL中，在文件后缀名后加问号，用等号把参数名与相应的值连起来，不同参数间用逗号隔开，这种方式显性于URL中，自动分析判别较容易。另外一种方式是通过对话Session设定参数，如JSP里的setParameter（）和getParameter（），这种方式独立于URL链接，所以识别起来较困难[10]。不同的URL构造方式有着不同的解决方案。对于第一种方式可以在程序里直接构造URL，但参数传递的交互性不好。如在java或C++里写URL类，把URL作为字符串类型的变量传进去，执行下载，其程序源代码如表2所示。

上述程序执行一次下载万方学位论文数据库某高校镜像网站的符合检索需求的学位论文全部内容。i循环控制要下载的文章数量，j循环控制论文的章节。数组变量sid存储的是学位论文在万方数据库的ID号，万方学位论文数据库使用6位数字编码。downFile是一个专门用来下载的类。

还有一种方式是构造浏览器，以URL调用相关页面，如在VB里使用WebBrowser控件，打开相

应的网站，输入检索条件，得到检索结果，用程序读取 WebBrowser 所获取的网页内容即可。检索条件的输入属于网站的内容而不是浏览器的内容，所以不需要考虑。只要检索结果就可以了。通过 WebBrowser 控件的 Document 属性可以获取网页内容。

表 2 下载万方学位论文示例程序(java)

```
1:  for (int i = 0;i < 35;i + + ){
2:    for (int j = 0;j < 10;j + + ){
3:      downFile. setURL( http://210. 44. 185. 15/CDDBN/Y +
        sid[i] + "/PDF/Y" + sid[i] + "000" + j + ". pdf");
4:      downFile. downloadFile();
5:    }
6:    for(int j = 10;j < 50;j + + ){
7:      downFile. setURL( http://210. 44. 185. 15/CDDBN/Y
        + sid[i] + "/PDF/Y" + sid[i] + "00" + j + ". pdf");
8:      downFile. downloadFile();
9:    }
10: }
```

网页分析从技术来源上讲有两种方式，一种是自己写分析代码，另一种是直接用现成的包，如 HTMLParser。网页分析从技术实现上有两种方式，一种是使用字符串处理函数进行，更常用的操作是使用正则表达式。

3 计量分析处理流程

本处理采取分步操作，一个模块一个模块地处理。既清晰地展示每一步过程，又把中间结果存储起来，留作他用。当然执行时间变长了，模块之间的关联操作使用 call 进行调用，因此也可以一次性顺序执行。文献计量统计分析的流程包括数据获取、数据预处理、统计计算与应用四大模块，详细流程如图 1 所示。

数据的获取有三种方式：直接从数据库商获取原始数据，这种方式最理想；从数据库商网站通过检索得到检索结果，然后复制检索结果，进行行列转换后得到二维关系数据；通过改变 URL 的传递参数来构造 URL，然后下载相关网页，从 HTML 网页中析取内容并存入数据库。

获取了二维关系数据以后要进行预处理，有一些列是需要再分的。预处理的目的是为了统计，其处理结果要符合一范式（1NF）。预处理主要包括拆分与提取。对于多值同字段要进行拆分，如作者、机构、关键词、分类号等，一篇文章有多个作者、多个关键词、多个分类号等，但这些词的属性

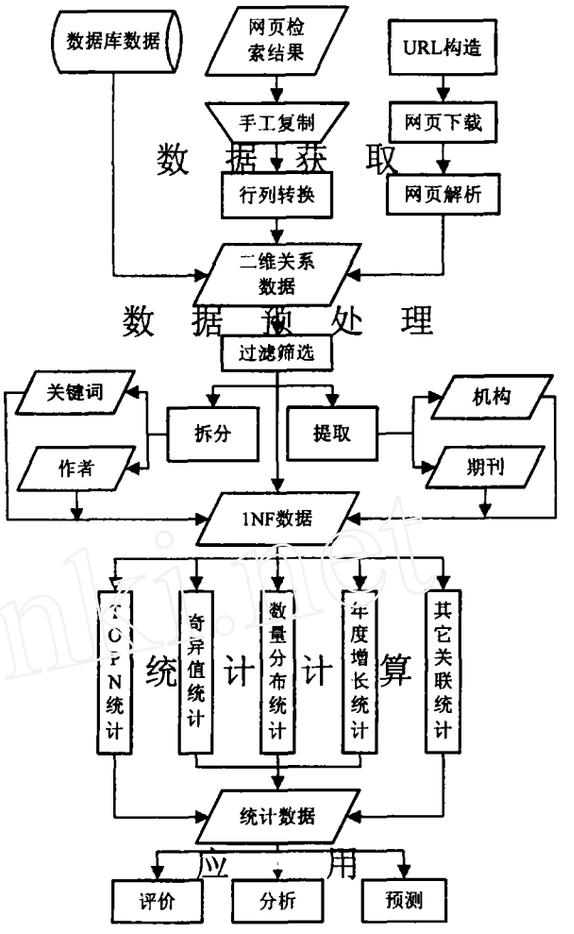

图 1 文献计量统计分析系统结构图

是同质的。对于多值异字段要进行提取（提取的概念并不准确，其实就是分割），如清华同方的单位或重庆维普的机构都含有三项内容，分别为作者所在单位、地名、邮编等信息，这些信息是异构的，数据类型、长度与取值范围都有所不同，重庆维普的刊名也含有很多信息，包括期刊名称、年、卷、期、起止页码等，需要进行拆分。

数据处理好后就可以进行统计。不同的字段有着不同的统计需求和方法。按照统计结果分为 Top N 统计、奇异值统计、数量分布统计、年度增长统计、其它关联统计等。常规的统计有按单位名、地名、作者、基金支持等字段进行简单统计，按时间进行各项目的时间统计分析。围绕关键词的统计分析一直是主题型计量分析的研究重点与核心，除了统计关键词与作者、机构、地区、期刊、分类号等的关系外，还可以统计关键词平均个数、关键词平均长度、标题的平均长度、摘要的平均长度、关键词在标题中出现的个数、关键词在文摘中出现的个数、在标题和摘要中同时出现的关键词个数。统计

按照统计方法分为为均一统计和加权统计，如作者是有位序的，而关键词基本是无位序的。

根据统计结果可以直接生成统计报告。但有时统计结果并不是最终的目标，除了统计外，还可以进行深度分析，如评价、预测与挖掘。评价是对发展历史与现状做出评判，孰多孰少、孰高孰低、孰强孰弱都要有所反映。预测需要数学模型和专门的方法，如趋势外推法、时间序列法等[11]。挖掘是要从大量的统计数据中总结出新颖的、潜在有用的知识[12]。

4 文献计量中的统计类型

文献计量中的统计按照统计对象分为作者统计、关键词、机构统计、主题统计、分类号统计、期刊统计、地区统计、参考文献统计（不同于引文分析）、基金资助统计、篇名统计、摘要统计、正文统计。统计对象的划分依据就是文献的不同字段。篇名、关键词、分类号、摘要与正文是最能反映文章主题的字段，而能反映作者观点的也许只有摘要与正文字段了。全面反映文献内容当属正文字段，而目前对于标题、摘要和正文三个字段的计量分析非常少。

按照统计结果又分为 Top N 统计、奇异值统计、数量分布统计、年度增长统计、其它关联统计。Top N 是最常用最基本的统计，如高产作者统计、高被引作者（或文章或机构）统计、高频关键词统计等，以分析核心作者、核心期刊、核心研究机构等，Top N 的输出以表格式形式所列，一般不进行图形显示。

奇异值统计包括最长、最短、最多、最少等端点值的统计，它不同于 Top N 统计。Top N 统计某一特征的前 N 项，奇异值统计的是某一特征的端点值，而且有些特征本身就比较特殊，返回的是一个值，这种特征有时是一些很特殊的需求，所反映的是个别现象或特殊情况，如字符数最多的关键词、不含英文字符与标点符号的最长的关键词是什么，有多长，篇含关键词最多的个数，最短标题的长度，用等值统计和加权统计差别最大的作者（前者是不管第几作者都按一篇计算，后者按位序乘以相应的权重，一篇文章所有的和为 1，分析是否有挂名现象等）。这些统计不是没有意义，例如找出最长的关键词可以确定在使用关键词构成的词库对标题、摘要等字段进行向量分词时确定最大向量长度。奇异值统计不适合以任何图形形式展现。

数量分布统计主要统计数量分布关系，如实验中对图书情报学核心期刊的 42，989 篇文章进行统计分析，发现篇含关键词数量为三到八个的占到 95%，这也与大多数编辑部要求提供三到八个关键词有关，反过来也可以对一些规定进行验证其合理性。再者统计出四字关键词占关键词总数的 41%，而且这些这键词大都由两字名词加两字动词构成，如“数据挖掘”、“信息检索”等。这种统计常以曲线图、柱状图、饼状图等形式展现。

年度增长统计主要进行和时间有关的统计，如作者发文量的增长、关键词年度增长情况等。按年度统计可以分析新的生力军、新的研究热点等，按关键词统计年度分布可以分析某项研究的生命周期等，作者与关键词及年度的关系可以反映作者的研究轨迹。比较是年度增长统计的主要分析手段，无论是增长量还是增长率，都是双目运算。在年度增长的统计图中，必然要有年度作为一个时间维，这种统计常以曲线图或双柱状图形式展现，不适合以饼图形式展现。

关联统计主要统计其它分布情况，如关键词与机构的关系反映机构的研究重点，关键词与期刊的关系可以反映期刊的侧重，如统计发现《图书情报工作》是 17 种图书情报核心期刊中发文献计量方面文章最多的期刊。这种关联统计最适合饼图形式展现，当然柱状图也可以。

5 结 语

除上述按照统计对象与统计结果的分类外，还可以按照语言的处理层面把统计分为形态统计、结构统计、语义统计、语用统计等。常规的作者、关键词、引文分析等都属于形态统计。主题统计、观点统计等应该属于语义统计，如数据挖掘与数据库知识发现的关系有这样几种观点，“数据挖掘也就是知识发现”，“数据挖掘是知识发现的一个步骤”，“数据挖掘与知识发现无论从研究对象、方法、结果等各方面都是不同的”，每一种观点有多少人支持，这样的统计属于语义范畴。而像图表结构的统计、行文结构的统计、概念定义的句子复杂度等属于结构统计。目前的统计分析主要停留在形态统计，而结构统计与语义统计属于内容分析的研究范畴，目前进行自动化统计的难度还很大。而语用分析会越过了内容分析，比如同一观点有哪几种阐述

方式，统计不同专业背景的人撰写同类型文章时有何不同。实现语用的统计分析还有很长的路要走。

主题型计量分析正处主流状态，但这种论文（如文献【2-6】）很快就会变少。其实这种计量分析文章如同网络日志分析一样，完全可以由计算机来实现，像 Google 新推出的搜索趋势，中国知网（清华同方）新推出的学术趋势。现在的这些文章主要是计算机进行计算、统计、生成图形，然后人工对图形作以简单说明，这种说明的方式大有规律可寻，如对奇异值的说明、Top N 的说明、符合某种图形关系的说明、符合文献规律的验证说明等。

清华同方、万方数据、重庆维普等全文数据库商目前只提供检索功能，随着把全文检索系统改成中国知网、知识链接门户等更大的工程，检索不再是唯一的功能，紧随其后的应该就是统计分析功能。不久的将来，这三大数据库商会陆续推出统计功能，也就是文献计量自动分析查询系统，届时大多数计量分析方面的文章将由计算机来实现并提供，而不是人来写。大多数编辑部将不再接受此类文章。作为管理信息系统的三大常规模块之一的统计模块，会成为未来几年数据库商的一大竞争点。

参考文献

1 庞景安.科学计量研究方法论[M].北京:科学技术文献出版社,2002:121-142.

2 苏新宁.图书馆、情报与文献学学术影响力研究报告(2000-2004)——基于 CSSCI 的分析[J].情报学报,2006,(2):131-153.

3 邱均平,王宏鑫,冯新霞.《情报学报》与我国情报学发展(Ⅰ)——《情报学报》创刊 20 年来引用文献的计量分析[J].情报学报,2002,(5):514-523.

4 邱均平,王宏鑫,冯新霞《情报学报》与我国情报学发展(Ⅱ)——《情报学报》创刊 20 年来引用文献的计量分析[J].情报学报,2002,(6):642-655.

5 马费成,张 勤.国内外基于知识管理研究热点—基于词频的统计分析[J].情报学报,2006,(2):163-171.

6 李长玲,化柏林.我国网络计量学研究的文献计量分析[J].图书情报工作,2006,(9):46-50.

7 陈 涛.武警学院学术论文统计系统开发及功能实现[J].武警学院学报,2005,(3):94-96.

8 张守胜.基于 Web 的科技论文统计信息系统的应用研究[J].安徽理工大学学报(自然科学版),2004,(1):59-63.

9 袁遥路.基于 ASP 的学术论文信息检索统计系统[J].微机发展,2004,(2):57-60.

10 化柏林.从 IPO 分析未来的搜索引擎[J].情报学报,2007,待发.

11 蔡筱英,金新政,陈 氢.信息方法概论[M].北京:科学出版社,2004:231-239.

12 粟 湘.数据挖掘在科技论文分析中的应用研究[D].中国科学技术信息研究所,2003.

(责任编辑:孙晓明)

(上接第 1324 页)

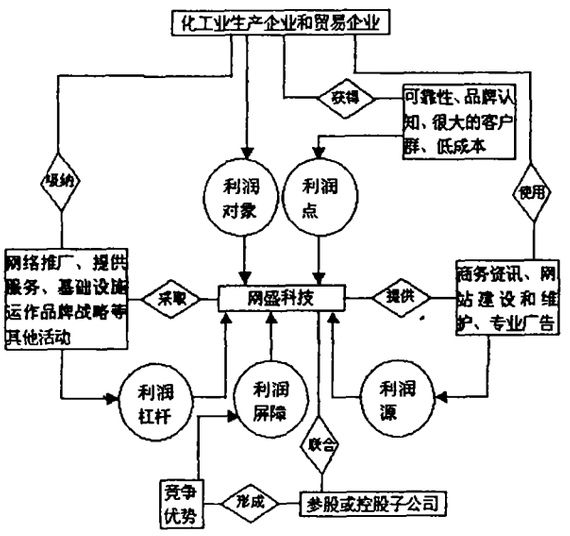

图 3 网盛科技盈利模式

参考文献

1 李京文.经济学也要与时俱进[N].经济日报,2002-07-08.

2 杨晓燕.值得借鉴的跨国公司盈利模式[J].企业经济,2003,(7):35-37.

3 Nicolas. The Economics of Networks[J]. International Journal of Industrial Organization. 1994,(2):673-699.

4 Gillen, David, Ashish Lall. The Economics of the Internet, the New Economy and Opportunities for Airports[J]. Journal of Air Transport Management. 2002,(4):49-62.

5 张小蒂.网络经济概论[M].重庆:重庆大学出版社,2005:103-104.

6 《第十九次中国互联网络发展报告》[EB/OL]. www.cnnic.gov.cn. 2007-01-09.

7 Dubsson, Tobay Magali, Alexander. Osterwalder, Yves Pigneur. e-Business Model Design: Classification and Measurements[J]. Thunderbird International Business Review, 2001,(2):1-22.

(责任编辑:孙晓明)