基于 CNPAT 数据库数据构建上海专利指标数据库的方法研讨

代茂军　李　红
（上海大学，上海 201800）

〔摘　要〕 本文介绍了上海专利统计分析系统的模块组成以及这些模块所起的作用，详细阐述了组成系统的专利数据转换模块。

〔关键词〕 专利数据；专利数据统计；专利数据转换；专利数据库

〔Abstract〕 The paper introduces the modules that consist of Shanghai Patent Statistical and Analysis System (SPSAS) as well as the functions those modules perform, then it expounds the patent data conversion module in detail.

〔Key words〕 patent data; patent data conversion; patent data statistical analysis; patent database
〔中图分类号〕G250.74　〔文献标识码〕A　〔文章编号〕1008－0821（2004）01－0015－03

1 前　言

发达国家在利用专利数据作为技术指标分析比较国家创新情况、评估技术发展现状、预测技术发展前景等方面已有多年的理论研究和实践经验。各国科技竞争力评价指标体系普遍利用专利指标作为主要指标来衡量和分析国家技术创新状况和问题（如著名的《洛桑报告》）。但利用专利数据建立科学技术指标的工作，在我国还未得到足够的重视。鉴于专利数据统计分析对技术发展的重要作用，上海大学已于近期完成了上海市自然科学基金项目“用于创新评价工作的专利指标量化研究”的工作，建立了上海专利指标数据库，解决了专利数据自动化处理的技术难题。本篇文章详细介绍了如何利用中国专利局专利信息光盘检索系统 CNPAT（ABS）中的数据来构建上海专利统计分析数据库。

2 上海专利统计分析系统简介

上海专利统计分析系统是由上海大学开发完成的，系统前台开发工具 Delphi 7.0，后台数据库管理系统为 Access 2000。前台应用程序通过 Delphi 7.0 控件面板中所提供 ADO 数据访问对象控件建立了与后台数据库的数据传送渠道。本系统提供了单项统计和组配统计功能，同时也提供了系统所必备的其他功能，如数据的打印、复制、帮助等。这些功能主要是由下面两个程序模块实现的，分别为数据统计模块、数据转换模块，系统程序流程见图 1。

2.1 数据统计模块

通过对数据转换模块的调用，实现 CNPAT（ABS）数据库记录格式到上海专利指标数据库记录格式的转换，并添加新的专利数据记录到专利指标数据库中，从而可以随着专利数据的更新而不断更新本数据库。在这些专利数据基础上，就能够实现本系统的统计分析功能。根据用户实际工作需要，我们确定了单项统计和组配统计，其中单项统计功能有：专利权人统计、IPC 分类统计和年度统计；

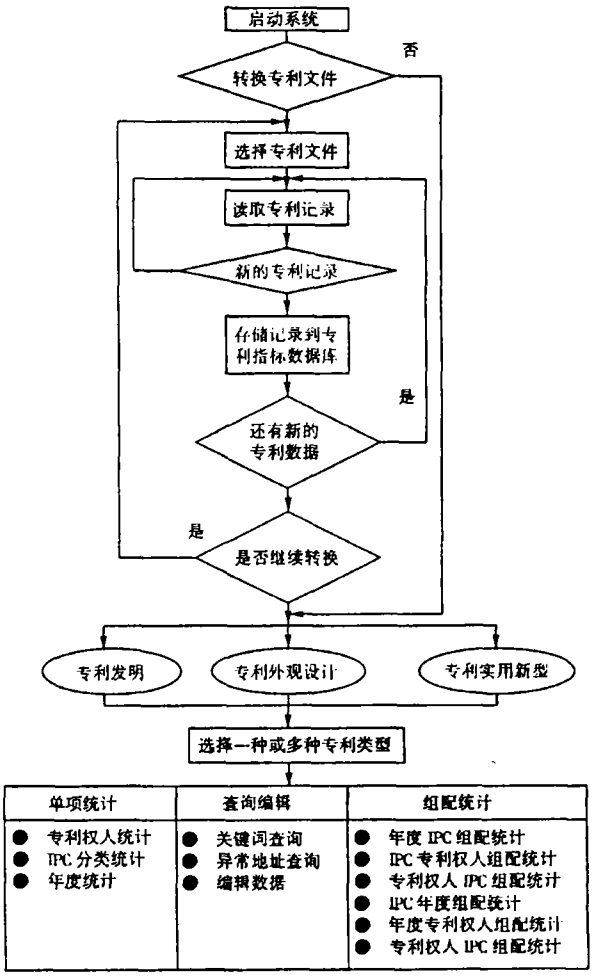

图 1　上海专利统计分析系统程序流程图

收稿日期：2003—09—15
基金项目：本项研究得到了上海市科学技术委员会的资助，项目编号：01ZH14022
作者简介：代茂军（1978—），男，上海大学国际工商管理学院硕士研究生，情报学专业。
李　红（1968—），女，华东师范大学硕士毕业，上海大学情报研究中心副教授。

组配统计功能有：专利权人和IPC组配统计、专利权人和年度组配统计、IPC和年度组配统计、IPC号和专利权人组配统计、年度和IPC号组配统计、年度和专利权人组配统计。

2.2 数据转换模块

本模块是数据统计模块功能实现的基础，实现了大量具有不同特征专利数据的自动化处理工作。本模块包括一个文本文件数据读取函数和一个缓冲区数据读取函数。该模块是本课题的技术难点之一，也是本篇文章所关注的主要方面。

2.3 专利指标数据库

PAT（ABS）有发明、实用新型和外观设计三种类型数据库，在这三种数据库中，每一个记录有以下字段组成：记录号、代理人、地址、发明（设计）人、分类号、公开（公告）号、公开（公告）日、名称、申请（专利）号、申请（专利权）人、申请日、说明书光盘号、摘要、主分类号、主权项、专利代理机构、法律状态公告日、法律状态、申请号、授权公告号、颁证日、法律状态公告日、法律状态，其中公开（公告）日字段类型是“日期类型”，摘要、主权项两字段类型为“备注类型”，其它字段为“文本类型”。考虑到所要开发数据库的实际需要，在构建上海专利指标数据库时，我们选择了其中15个字段描述一条专利数据记录，其分别为：代理人、地址、发明（设计）人、分类号、公开（公告）号、公开（公告）日、名称、申请（专利）号、申请（专利权）人、申请日、说明书光盘号、主分类号、专利代理机构、摘要、主权项。原始数据是通过中国专利知识产权局的专利信息光盘检索系统CNPAT(ABS)（1985～2001）直接检索获得的，检索表达式为：“地址=上海”，命中记录32 747个，由若干个文本文件来存储这些记录。

3 数据转换模块的程序流程和实现方法

数据转换模块的程序流程见图2（数据转换模块源程序略）。

3.1 类的方法、数据成员和重要函数说明

根据需要，我们定义了一个类Tpatten，该类继承于object，Tpatten中包括一个procedure openfn（fn3：string）过程、function getonerec：boolean和function getpos：integer函数；专有数据成员部分定义了3个变量，分别为buf、cursp、nextsp。其中第一个变量类型是数组数据类型，用来保存从转换文件中所获得的数据；另2个变量是字符型指针变量，用来存储指向一条专利记录的“记录号”字段内容的第一个字符和下一条专利记录“记录号”字段的前一个字符，实际上，nextsp和cursp两者的差值基本上就是一条专利记录的长度；公用数据成员部分主要定义了一个jltype记录数据类型的变量recp和整型变量fsize。recp变量用来存储来自数据缓冲区数组中的数据；fsize变量用来保存转换文件的长度。其中jltype记录数据类型的具体数据结构如下：

```
type
jltype = record
    ADOTDSDLR:String[100];//代理人
    ADOTDSDZ:String[100];//地址
```

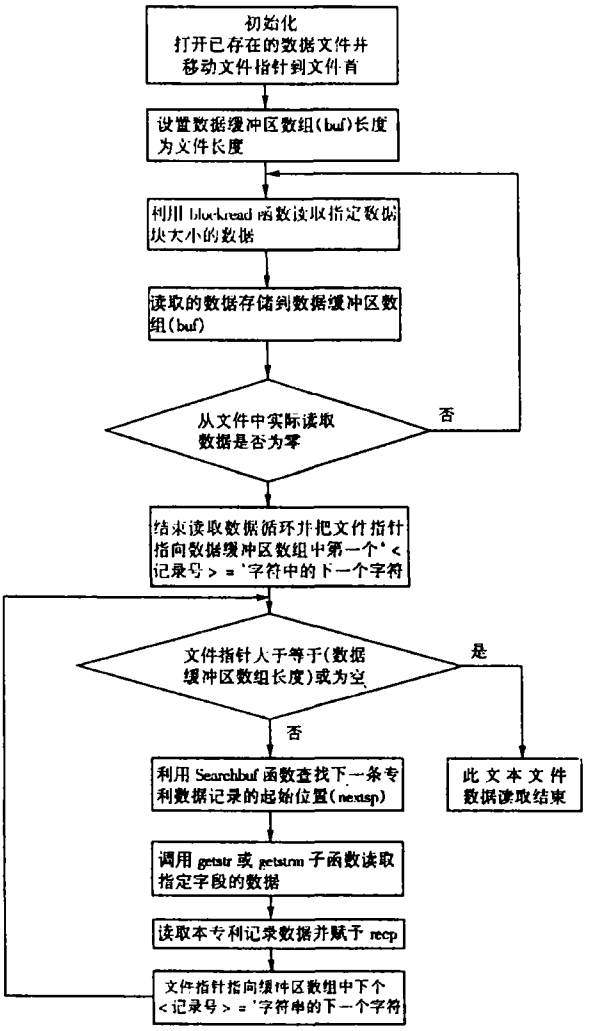

图2 数据转换块程序流程图

```
    ADOTDSFMSJR:String[100];//发明(设计)人
    ADOTDSFLH:String[50];//分类号
    ADOTDSGKH:String[50];//公开(公告)号
    ADOTDSGKR:TDateTime;/公开(公告)日
    ADOTDSMC:String[100];//名称
    ADOTDSSQH:String[50];//申请(专利)号
    ADOTDSSQR:String[100];//申请(专利权)人
    ADOTDSSQRi:TDateTime;//申请日
    ADOTDSCDH:String[50];//说明书光盘号
    ADOTDSZY:array[0..9999]of char;//摘要
    ADOTDSZFLH:String[50];//主分类号
    ADOTDSZQX:array[0..9999]of char;//主权项
    ADOTDSDLJG:String[100];//专利代理机构
end;
```

下面介绍几个在本模块中所用到的重要函数：

Procedure Blockread（Var F：File；Var Buffer；Var Count：Longint［；var Result：Longint］）：从无类型变量F读取Count个记录到变量Buffer中，一般一个记录系指包含128字节的块，但可以使用带有参数的Reset来改变从128到不同字节块大小，实际读取的字节数为Buffer*128，可选参数Result表示实际读取的记录个数，但总的字节大小不能超过

64KB。

Function SearchBuf（char * Buf，int BufLen，int SelStart，int SelLength，AnsiString SearchString，TStringSearchOptions Options = TStringSearchOptions ()≪soDown)：本函数功能是在文本数据缓冲区里查找一个指定的字符串，并返回第一个匹配字符串的首字符地址。其中参数 Buf 是查找的文本数据缓冲区；参数 BufLen 是文本数据缓冲区的长度（单位是字节）；参数 SelStart 是查找的起始位置；参数 SelLength 是指要查找的字符数量；参数 SearchString 是在 Buf 中要查找的字符串；可选参数 Options 是记录类型变量，决定查找是否区分大小写、是否完全匹配以及前向或后向查找的起始位置。

Procedure Reset(Var F:File;[;Var Result:Longint])：为读数据而打开已存在的文件，并把文件指针移到文件首，可选参数 Result 为设置所读取的每个数据块的大小。

Function LeftbStr（const AText: string; const ACount: Integer): string 返回字符串 Atext 左边的 ACount 个字符。

3.2　专利数据由文本文件存储到数据缓冲区的读取函数(openfn)

本函数主要是把文本文件中的数据全部读取出来，并存储到数组 buf 中，这里我们暂且称之为数据缓冲区数组。具体操作过程如下：首先说明了一个和 buf 变量类型相同的并可以存储32 767个字节的数组变量 temp，blockread 函数直接把读取的数据赋值给 temp 数组，temp 数组再把所存储的值赋值到 buf 数组。然后调用本函数所传递的文件名参数赋值给一个字符串变量 fn，利用 assign 函数把文件名字符串 fn 赋给文件变量 ff，程序对文件变量 ff 的操作代替了对文件 fn 的操作，再用 reset 函数打开要读的文件（Result = 1)。这样我就已经完成了读取文件的准备工作，接下来就从已打开的文件 ff 里读取数据，所用到关键程序语句为：

```
repeat
blockread(ff,temp,32767,readbyte);
for i:=0 to readbyte-1 do buf[count+i]:=temp[i];
count:=count+readbyte;
until readbyte=0;
```

这个语句块的意思是每次从文件 ff 中读取32 767个字节数据到过渡数组变量 temp 中，再通过一个 For 循环把所读取的数据存储到数据缓冲区 buf 数组变量，重复上面的循环直到从文件 ff 中实际读取的数据为 0 止。这个循环结束后，就完成了从文本文件中提取所需要数据到数据缓冲区 buf 数组变量中。最后利用 searchbuf 函数在整个 buf 数组中查找第一个字符串“〈记录号〉=”的地址，并存储在 cursp 指针变量。我们注意到，cursp 指针变量存储的地址指向的是“〈记录号〉=”的前一个字符，而不是记录号字段的内容，所以必须改变 cursp 值以指向记录号字段值的第一个字符。

3.3　专利数据由数据缓冲区数组存储到专利指标数据库的读取函数(getonerec)

本函数主要是把数据缓冲区 buf 数组里的数据按照单条专利记录形式读取到前面所定义的记录类型变量 recp 中。getonerec 返回的是逻辑值，数据统计模块中的 N3Click 函数通过这个值来判断 buf 数组中是否还存在专利数据，如果为真继续调用 getonerec 函数提取下条专利记录数据，如果为假就说明已经没有专利数据，从而完成了文本文件转换的整个过程。下面列出判断文件转换结束的语句：

```
if(cursp>=pchar(buf)+fsize)or(cursp=nil)then
begin
getonerec:=false;
exit;end;
```

文件中每条专利数据“摘要”和“主权项”字段是备注类型，而其它字段是文本或日期类型，摘要和主权项字段内容结束的标志和文本或日期类型字段结束的标志是不一样的。在本文件中，文本或日期类型的字段内容结束标志是空格键或回车键，备注类型的字段内容结束标志不是空格键和回车键，而是一条线段，所以在 getonerec 函数中，我们定义两个子函数：function getstr(ss:string):string，function getstrm(ss:string):string。getstr 实现对文本或日期类型字段的数据读取功能；getstrm 实现对备注类型字段的数据读取功能。

在调用 getstr 和 getstrm 子函数前，getonerec 函数首先定义了 1 个字符串指针变量 tbuf 以存储当前专利记录的地址，再利用 searchbuf 函数查找下一条专利记录的起始位置，其查找的起始位置为 tbuf 变量所存储的地址，查找长度为 pchar(buf)+fsize-cursp，所要查找的字符串为‘〈记录号〉=’，并把 searchbuf 函数返回值赋予 nextsp。这样就可以调用子函数提取所要数据了，文本或日期类型字段调用语句如下：recp.ADOTDSDLR:=getstr(‘〈代理人=’)，其它字段同上；备注类型字段调用语句：strpcopy(recp.ADOTDSZY,getstrm(‘〈摘要〉’))，strpcopy(recp.adotdszqx,getstrm(‘〈主权项〉=’))。

getstr 子函数获取字段值的实现方法为，首先定义了 2 个字符型指针变量 ts1、ts2 记录每个字段的起始位置和结束位置，然后再利用 searchbuf 函数查找第一个 ss 字符串的开始位置，并赋值给 ts1 变量，实现语句 ts1:=searchbuf(tbuf,buflen,0,0,ss)；这里的 buflen 等于 1 条专利记录数据的长度.ts2:=ts1+len(ss)。知道 ts1，ts2 值后，就可以获取字段值，实现语句为：

```
while((ts2[0]=#13)or(ts2[0]=#10))do ts2:=ts2+1;
ts1:=ts2;
while not((ts2[0]=#13)or(ts2[0]=#10)or(ts2[0]="<'))
do ts2:=ts2+1;
i:=ts2-ts1;getstr:=leftbstr(ts1,i);
```

每个字段值的最初几个字符可能是空格或回车键，就需要改变 ts2 的值以去掉这些不需要的空格或回车键。当 ts2 存储地址所指的内容为空格键或回车键或‘<’时结束 while 循环，说明 ts2 指针已指向字段内容末尾，那么 ts2-ts1 就是该字段内容长度，利用 leftbstr 提取这些数据并赋值给函数名 getstr。以上是提取某一字段内容的程序实现方法，其它文本或日期类型字段同上。这里我们要注意判断条件中 ts2 [0] 的含义是 ts2 指针变量中的指针值所指的内容。

getstrm 子函数读取备注类型字段值的方法和 getstr 基本相同，也是定义 2 个字符型指针变量 ts1、（下转第 21 页）

方法，但是对于查询中文期刊来说，本地计算机必须要安装或运行“网络实名”之类的工具软件（这类软件网上可以下载，其中北京因特国风网络软件科技开发有限公司开发的 3721 就是一种很好的网络实名软件），利用这种方法进行查找，即使网上没有准确的所想查询的信息、或输入的中文名称有误差，也可以检索出一些相关或类似的信息。

3.1.2　利用网络搜索引擎进行搜索。这是一种比较科学的方法，目前可用的搜索引擎许多，比较有名的中文搜索引擎有 google、搜狐、网易、新浪、雅虎中文、搜索客、天网、悠游等；外文搜索引擎有 yahoo、infoseek、altavista、lycos、excite、hotbot 等。

3.1.3　通过网上学术期刊数据库来查询。这是一种可以同时查得许多学术期刊信息的便捷方法。现在的网上学术期刊数据库，或由出版社/公司自己将其所出版的纸本刊物提供上网销售发行；或由专门的网络信息提供商（数据库开发公司），通过版权协议把众多的纸本刊物进行数字化转换，形成自成特色的数据库产品公开销售。这些数据库所收刊物，少则几十种，多则数千种；有专科性的，也有综合性的。在不同级层的查询页面，可以通过逐层浏览的方法，了解到有关刊物的刊物特色、办刊方向、来稿要求、编辑部及主编/编委的通讯地址、联系电话、E－mail 等信息。有价值的网络学术数据库，可以通过高校图书馆网页来了解，外文数据库还可以通过“全国文献保障系统”的“文理中心”、“工程中心”等了解。

3.1.4　通过图书馆、文献信息中心等文化、科研、教育机构网页上的相关链接来查找。现在，许多信息服务机构为了充分挖掘和有效利用网络信息资源，都在积极开展网络信息资源学科导航工作，许多图书馆特别是高校图书馆的网页上都设有“学科导航”栏目，通过此类栏目也可以了解到一些分学科、分主题的学术期刊网站地址或链接，从而查询非常有针对性的期刊投稿信息。

3.2　网上投稿信息利用中需注意的几个问题

3.2.1　不能盲目相信网上投稿信息

一般来说，网上投稿信息修改、补充方便，更新比较及时，能够保证最高的准确性。但也不能排除有些期刊在迁址、更换电话、变换主编/编委后，没有主动告知有关的期刊报道网站或检索工具的编制者，或在公共媒体公示有关的变更信息，致使其相应的网上信息过时。因此，科研人员首次利用或再次利用长期未用的网上投稿信息投稿之前，最好先通过其电话或 E－mail 进行联系以验证其准确性，以免延误了科研成果的及时发表。

3.2.2　网上投稿信息不全是免费的午餐

网上投稿信息有些是公开免费的，有些是商业性收费的。免费信息在任何地方上网都可以查询，如专门的报刊目录网、各高校学报、科研院所的出版物等；收费信息则必须在特定的网域内上网才能查询，如一些外文数据库，一般是利用 IP 地址来控制使用权限，因而只有通过购买单位预先申报了 IP 地址的校园网或局域网上网才可查询。查询时应根据当时所处的网络环境来决定访问、查询哪种网上投稿信息源。

3.2.3　共享与共建并举

对于图书馆来说，不仅要有效共享他人的成果，更要积极参与网上投稿信息资源的共建。图书馆可以根据本校、本部门/系统科研人员的专业特点和科研方向，努力收集各种类型的相关信息资料，筹划建立符合本馆服务对象需要的综合性或专题性的投稿信息文档或数据库，或将本馆所藏的有关检索工具书进行数字化转换，并将其推送上网供人们查询。科研人员也应该把自己了解的投稿信息及时反馈给图书馆，以供他人共享。在建立自己的信息文档或数据库过程中，要注意对他人信息产权的尊重，凡是禁止随意转载的网页或资料，未经他人许可绝不进行转链接或原版式嵌入、照抄利用，以免引起不必要的法律纠纷。

（上接第 17 页）

ts2 记录每个字段的起始位置和结束位置，然后再利用 searchbuf 函数查找第一个 ss 字符串的起始位置，并赋值给 ts1 变量，实现语句 ts1：＝searchbuf(tbuf,buflen,0,0,ss)；参数含义同上。ts2：＝ts1＋len(ss)．2 个函数的不同主要是在获取数据循环结束的判断条件上。getstrm 程序实现语句为：

```
while not((ts2[0] = ' < ')or((ts2[0] = '_')and(ts2[1] = '_')))
do ts2: = ts2 + 1;
i: = ts2 - ts1 - len;getstrm: = leftbstr(ts1 + len,i);
```

getonerec 函数调用语句全部执行后就完成了对 1 条专利数据记录的读取，并已存储在 recp 变量中，在数据统计模块的 N3Click 函数中使用 Appendrecord 方法把 recp 变量中存储的数据添加到专利指标数据库。为了实现对其它专利记录的提取，cursp 需要指向下条专利的记录号字段值的第一个字符，再继续程序调用直到 cursp 值大于等于 pchar(buf)＋fsize 或为空时结束数据转换，并结束 N3Click 函数中的 while 循环。至此就完成了对一个文本文件专利数据的读取过程，其它文件数据的操作过程相同。

4　总　结

综上所述，首先文本文件数据读取函数（openfn）从所需要转换的文本文件中读取数据到数据缓冲区数组中，并存储专利数据记录的起始位置。缓冲区数据数组读取函数（getonerec）中的 2 个子函数 getstr 和 getstrm 从专利数据记录的起始位置读取专利数据记录到 recp 变量，每次读取到 recp 变量中新的专利数据记录都将在数据统计模块中添加到专利指标数据库，然后再读取下条专利数据记录，一直循环到文件结尾。这样就实现了所需要转换的文本文件数据的读取、存储和添加整个全过程。

参 考 文 献

[1] 刘海涛．学用 Delphi 4［M］．北京：清华大学出版社，2001．

[2] 国家计委高技术产业发展司编．中国高新技术产业发展报告［M］．北京：中国计划出版社，2001．

[3] 徐冠华．当代科技发展趋势［J］．中国信息导报，2002，(6)．

[4] 赵英莉．我国专利技术评价与预测［J］．情报理论与实践，2001，(1)．

[5] Information Products Division/TAF Branch，U.S.PTO：All Technologies Report（1963－2001）．

[6] Information Products Division/TAF Branch，U.S.PTO：Patenting by Organizations，2001．