CRSP 数据库检索方法

彭 哲[1] 刘飞泉[2]

(1. 厦门大学王亚南经济研究院 福建厦门 361005)
(2. 华中科技大学经济学院 湖北武汉 430074)

摘 要:CRSP 数据库是美国芝加哥大学开发的金融数据库。介绍了 CRSP 包含的各子数据库及其内容,并以 CRSP 自行开发的图形化界面平台 CRSPSift 和 WRDS 网络查询界面为例,介绍了 CRSP 数据库的检索方法及个性化功能的使用技巧。

关键词:CRSP;CRSPSift;WRDS;检索技巧
中图分类号:G254.92 **文献标识码**:A doi:10.3969/j.issn.1005-8095.2013.08.021

Retrieval Methods of CRSP Database

Peng Zhe[1] Liu Feiquan[2]
(1. The Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen Fujian 361005)
(2. Shool of Economic, Huazhong University of Science and Technology, Wuhan Hubei 430074)

Abstract: CRSP Database is a financial database developed by the University of Chicago in USA. The paper introduces the component sub-databases of CRSP, and takes graphic interface platform CRSPSift and online inquiring interface WRDS for examples to expound the retrieval methods of CRSP Database and use skills of its personalized functions.

Keywords: CRSP; CRSPSift; WRDS; retrieval skill

1 CRSP 数据库概况

1.1 CRSP 数据库简介

CRSP 数据库是由美国芝加哥大学布斯商学院“证券价格研究中心”(Center for Research in Security Prices)所开发的数据库,主要涵盖了自 1925 年以来美国三大交易所的股票交易数据、上市公司相关信息、财政部债券信息等。CRSP 数据库包括 7 个子数据库(见表 1)。

1.2 CRSP 数据库在国内的使用

CRSP 数据库为国内一些金融数据库提供了有益的借鉴。北京大学中国经济研究中心(CCER)联合北京色诺芬信息有限公司推出的“CCER 中国证券市场数据库”,在基本结构和字段设计上参考了 CRSP 数据库。而北京聚源锐思数据科技有限公司联合清华大学经管学院、沃顿商学院、台湾辅仁大学所开发的 RESSET 金融数据库,在设计思想与相关算法上,也借鉴了 CRSP 的经验。此外,国泰安公司联合香港理工大学中国会计与金融研究中心所开发的“中国股票市场交易数据库”(CSMAR,China Stock Market & Accounting Research Database),在架构上也借鉴了 CRSP 数据库的一些处理技术。

目前,中国内地引进 CRSP 数据库的高校有北京大学汇丰商学院、清华大学经济管理学院、厦门大学、上海财经大学、上海高级金融学院等。

国内访问 CRSP 数据库的方式有两种:一种是通过客户端 CRSPSift 进行链接,厦门大学等高校即采用这种访问方式;另一种则是通过“沃顿研究数据服务”(Wharton Research Data Services,WRDS)进行间接访问。WRDS 是宾夕法尼亚大学沃顿商学院于 1993 年开发的数据平台。目前,WRDS 整合了 CRSP、Thomson Reuters、CSMAR 等多个数据库,国泰安公司的 CSMAR 还成为了 WRDS 在大中华地区的唯一数据来源。

2 CRSP 的基本检索方法

下面介绍 CRSP 的检索方法。为避免重复,本文用 CRSPSift 界面的查询为例,介绍股指收益率的检索方法;而用 WRDS 网络检索界面介绍单只股票收益率的检索技巧。

2.1 通过 CRSPSift 图形界面提取数据

下面以 CRSPSift Enterprise 4.2 版为例,介绍 CRSP 的基本检索技巧。CRSPSift 目前提供了获取 CRSP US Stock & Indices 子数据库、CRSP/Compustat

收稿日期:2012-12-21
作者简介:彭哲(1984-),女,2009 级经济学博士研究生,研究方向为数量经济学;刘飞泉(1980-),女,2009 级经济学博士研究生,研究方向为数量经济学。

表 1　CRSP 数据库的子数据库

子数据库名称	简　介
美国股票和指数数据库(US Stock and Index Database)	包括纽约证券交易所(NYSE)、NYSE 美国证券交易所(NYSE Amex US)、纳斯达克、纽约交易所高增长板块(NYSE ARCA)的相关数据。相关数据有日、月、年度 3 个频率。
CRSP/Compustat 合并数据(CRSP/Compustat Merged Database,CCM)	将标准普尔(Standard & Poor's)的 Compustat 数据库按照 CRSP 的数据库格式重新编排,并入 CRSP 数据库。包含有上千份月度和年度报表。
美国共同基金存活公司无偏数据库(Survivor-bias free US mutual fund database)	包括 1962-2008 年间公开交易的开放基金数据
CRSP 历史指数数据库(CRSP 1925 Historical Indexes Database)	提供了多种描述市场表现的指数。其中包括 CRSP 财政部和通胀指数系列,共包括 10 种期限的指数
CRSP 研究指数数据库(CRSP Research Indexes Database)	包括指数数据和股票数据两类。指数数据采样频率为秒,股票数据为分钟
美国财政部数据库(US Treasury databases)	包含了美国财政部债券的相关信息。月度数据始于 1961 年,日数据始于 1961 年
CRSP/Ziman 房地产数据库(CRSP/Ziman Real Estate Data Series)	由 CRSP 与 Richard S. Ziman 房地产中心联合开发。该中心隶属于加州大学洛杉矶分校的安德森管理学院。主要涵盖了在美国三大交易所赏识的美国房地产投资信托公司(REIT)的信息

注:资料来源于 CRSP 网站[1]

Merged (CCM)子数据库的图形界面。

在安装好 CRSP 客户端之后,桌面会出现 Lauch CRSPSift 的图标。双击图标,会出现图 1 界面。

我们希望提取 2000 年 1 月至 2010 年 12 月 S&P500 的指数收益率数据,为此,点击 New Query 标签(如图 1 方框部分所示)。

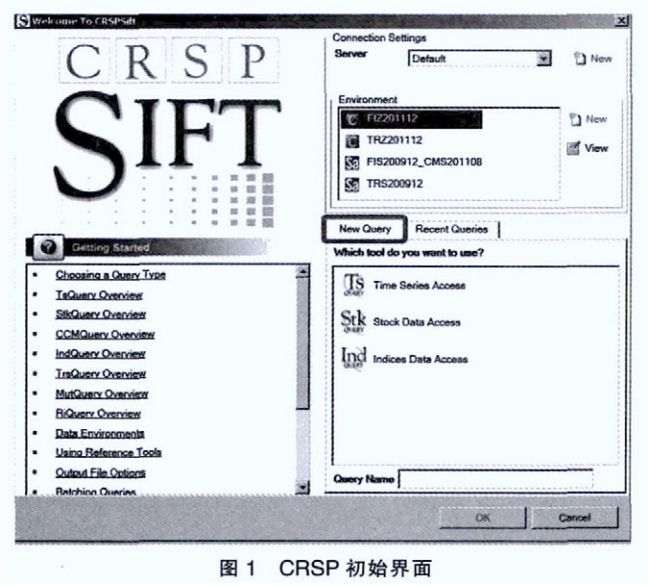


图 1　CRSP 初始界面

第一步:选中右下部分的 Indices Data Access,然后按 OK,进入数据提取页面(见图 2)。点击查询界面的左上角 File 菜单项,选择 Save As,将本次查询存储为一个 kls 文件。由于这里提取的是指数信息,因此保存后的查询文件名称为 XXX.ind.kls。

第二步:在 Database 一栏中选择数据频率,有日数据和月度数据两种选择,这里选择 Monthly。然后输入起止日期,如 20000101 To 20101231,即选取 2000 年初到 2010 年 12 月的所有月度数据。如图 3 所示。

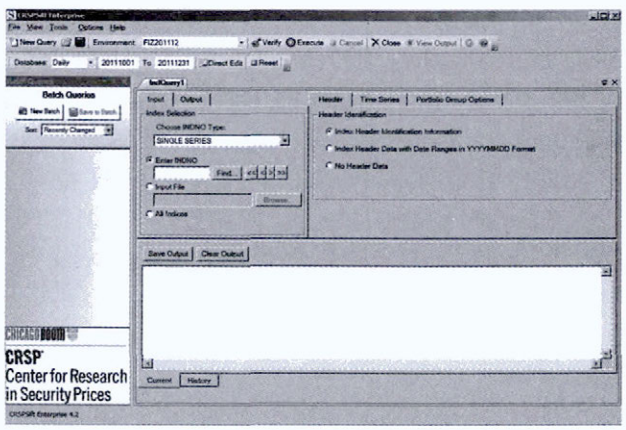


图 2　CRSP 查询界面

图 3　CRSP 菜单和关键按钮(版权归 CRSP 所有)

图 3　CRSP 菜单和关键按钮(版权归 CRSP 所有)

第三步:找到 Input 选项卡(如图 2 右侧中部所示),在 Index Selection 框中的 Enter INDNO 项下点击 Find,查询 S&P 500 的编号。点击 Find 之后,会出现一个名为 Find INDNO 的菜单。在下拉菜单中选择 CRSP S&P 500 Indexes,然后会出现以下 4 个选项:分别对应价值加权指数、等权重加权指数、价值加权投资组合以及等权重投资组合。这里希望提取价值加权的 S&P 500 指数,故选择第一项 CRSP Value-Weighted Index of the S&P 500 Universe,点击 Select 确定。选择后,Enter INDNO 会出现序列名称 1000500。其他序列名称、编码方法及相关信息请参阅 CRSP 的说明文档[2]。

在 Header 标签页(见图 2 右侧,IndQuery1 项下右下方)中选 Index Header Data with Ranges in YYYYMMDD Format,接着在 Time Series 标签页中

选择 Pre-Defined (预先设定),然后勾选 Index returns,used counts,and values,提取收益率信息。也可以用 Individual 进行个性化选择,提取需要的信息。

第四步:点击 Output 选项卡,选择所希望导出的格式。默认的选项是 Output to Screen (直接在屏幕上输出)。这里选择 Excel 97-2003 格式,然后在 Output File Name 中输入希望保存的文件名。默认路径是存放到 My Documents 下的 CRSPSift 文件夹下。CRSP 支持的输出格式包括 txt (文本格式)、xls 与 xlsx (Excel 用)、mat (MATLAB 用)、sas7bdat (SAS 用)、dta (Stata、EViews 等软件用)、sav (SPSS 用)。

点击菜单项下的 Execute (执行;见图 3 右侧),处理完毕之后,在目标文件夹下会生成两个文件,分别名为 XXX_indhh.xls 以及 XXX_mret.xls。其中,XXX_indhh.xls 存储的是该序列的文字信息,关键字段包括 KYINDNO(序列关键标识码)、MINDCO(序列所属的序列组名)、MINDNAME (序列名)、MGROUPNAME(序列所属的组名);XXX_mret.xls 存储的是相应的数据,关键字段包括 KYINDNO、MCALDT (时间)、MTRET (指数累积收益率)、MARET(除去红利的投资组合收益率)、MIRET(指数收入回报率)、MUSDCNT(价格有效的股票支数)、MUSDVAL(价格有效的股票总价值)。以上字段中,若首字母为 M 则表示月度数据,日数据则无 M 标识。

2.2　通过 WRDS 访问 CRSP 数据库

需要说明的是,在使用 WRDS 界面之前,用户必须将个人信息提交给所在高校的 WRDS 数据库管理员。成功通过身份验证的用户将获得一组 WRDS 的用户名和密码。

下面以获取美国微软公司的股票收益率为例,说明通过 WRDS 访问 CRSP 数据库的方法。首先,输入网址:http://wrds-web.wharton.upenn.edu/wrds/,进入到 WRDS 的访问页面(如图 4 所示)。在 Sign In 标签页下,输入用户名和密码,按下 Sign In 按钮登入。

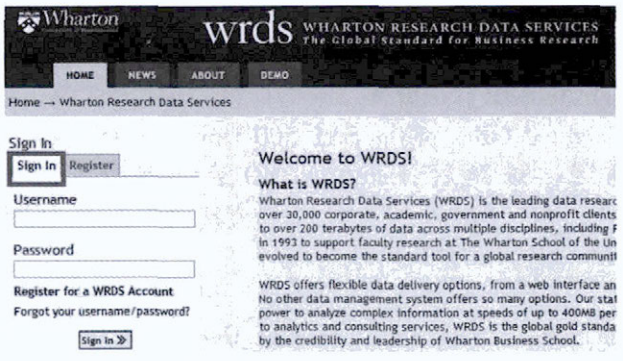


图 4　WRDS 的访问界面

登入后,屏幕左侧会出现 Select a Dataset 的下拉菜单。在下拉菜单中选择 CRSP 选项,即可进入 CRSP 数据库。进入后,屏幕右侧会显示 CRSP 数据库的简介。在屏幕左侧的 Select a Data Set 项下,会出现一个下拉菜单,并出现 Select an Available Dataset(备选的 CRSP 子数据库)的提示信息,在此菜单中,选择 CRSP Monthly Stock (月度股票数据库)。选择完毕后,屏幕右侧会显示 CRSP Monthly Stock 的字样。

第一步:在屏幕右侧的"Step 1:What data range do you want to use? "(您希望使用的数据区间是什么)项下,设置数据的时间范围。

第二步:在屏幕右侧的"Step 2:How would you like to search this dataset?"项下,选择公司代码的编码方式。CRSP 提供了 7 种方式——TICKER(交易所分配给股票的代码,数据始于 1962 年,根据上市的交易所不同而长度不等);PERMNO(CRSP 分配的永久 5 位股票代码,一个 PERMNO 可以对应若干个 CUSIP);PERMCO (CRSP 分配的永久公司代码);CUSIP (8 位股票历史代码,数据始于 1968 年);NCUSIP(按照股票名称命名的代码,若同一只股票历史上使用过若干个名称,则对应数个代码);SICCD(标准产业分类代码);HSICCD(证券标准产业分类代码的最后一位非零数,如果 CRSP 数据库中没有某公司的 SICCD,那么该公司的 HSICCD 为零)。这里选择 PERMNO,通过 Code Lookup(代码查询)可知,微软公司的 PERMNO 代码为 10107。关于公司代码的编码信息可参阅可丹麦安胡斯大学编写的说明。

第三步:在"Step 3:What variables do you want in your query"项下,选择查询变量。CRSP 股票月度数据库提供了 Identifying Information (识别信息,有 20 项可选),Time Series Information(时间序列信息,有 11 项可选),Share Information(股份信息,有 3 项可选),Delisting Information (退市信息,有 8 项可选),Distribution Information (分配信息,有 11 项可选),Nasdaq Information (纳斯达克信息,有 4 项可选),Market Information(市场信息,有 4 项可选)。

为了查询微软公司的股价和收益率信息,在识别信息中选择公司名称,在时间序列信息中选择 Price(股价)和 Holding Period Return(持股期间收益率)。

第四步:在"Step 4:How would you like query output?"项下,选择查询数据的输出格式并提交查询。输出的文件格式有 10 种,这里选择 xls 格式;文

(下转第 87 页)

在文献检索中,主题词、同义词、缩写词、简称的检索在科技查新检索中起着重要的作用。如果选择的检索词不合理,直接影响科技查新的质量。在使用上述检索词检索时应注意以下几个方面。

(1)规范检索用语。由于医学文献中,同一事物或概念可以用很多不同的词来表述及不同的书写方式,有规范的和不规范的,有学名和俗名,有同义词、近义词、多义词、缩写词、全称、简称和英文缩写等[2]。如果仅选择上述检索词中的一种或选择用户提出的检索词来进行检索,检出的文献不够全面,漏检率高,往往会降低科技查新的质量。检索员应与用户共同探讨,分析课题主题概念的内在含义,扩展概念组面的检索项,使用规范的检索用语,同时还应考虑主题词、关键词、同义词、全称、简称等的混合检索途径,避免漏检,提高科技查新的检索质量。

(2)合理使用逻辑组配。在文献检索中,要重视检索策略的制定,应根据课题的新颖性,备有"一宽""一窄"两套检索策略[3]。如检出的文献不能满足需要时,注意是否检索词选择不合理,主题范围是否限制过宽或过窄,以及逻辑提问、概念组配是否恰当等,应及时调整检索策略,避免漏检。

(3)提炼最精炼的检索词。在实际查新检索工作中,要充分利用专业词表、辞海、术语标准、词典、检索工具书等,提炼最精炼的检索词。同时还应考虑该检索词的同义词、近义词、全称、简称、缩写词、外文名称中文简写等不同的书写方式,以及数据库默认的检索表达式,多方位、多角度、多途径进行广泛检索,尽量避免漏检,获得所需要的医学文献。

参考文献

[1] 国家科技部. 科技查新规范 [EB/OL]. [2012-09-01]. http://Lib.zjfc.edu.cn/Articlo/kjcx/201111/560.html.

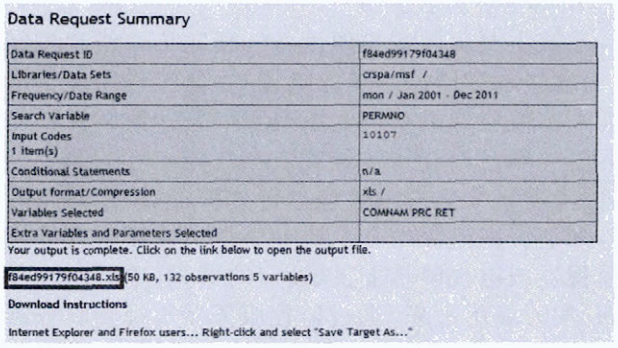
[2] 朱康玲. 同义词的获取对医学科技查新查全率和查准率的影响[J]. 中华医学图书情报杂志,2012,21(3):78-80.

[3] 孙淑萍,等. 医学科技项目查新检索的查全范围及检索技巧[J]. 医学情报工作,1995,16(1):35-37.

(上接第 84 页)

件的压缩类型有 3 种,默认选项为不压缩;数据的时间格式有 6 种,选择 YYMMDDn8,即按照 8 位年月日格式显示。

点击"SUBMIT QUERY"后,得到数据查询摘要界面 (如图 5 所示)。得到的数据文件会 Data Request ID(数据查询编号)命名。


Data Request Summary

Data Request ID	f84ed99179f04348
Libraries/Data Sets	crspa/msf /
Frequency/Date Range	mon / Jan 2001 - Dec 2011
Search Variable	PERMNO
Input Codes 1 item(s)	10107
Conditional Statements	n/a
Output format/Compression	xls /
Variables Selected	COMNAM PRC RET
Extra Variables and Parameters Selected	

Your output is complete. Click on the link below to open the output file.

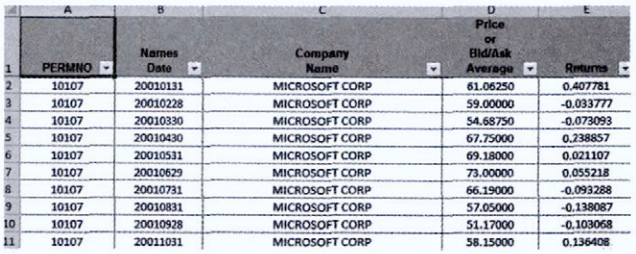
f84ed99179f04348.xls (50 KB, 132 observations 5 variables)

Download Instructions

Internet Explorer and Firefox users... Right-click and select "Save Target As..."

图 5　CRSP 股票月度数据查询结果摘要

点击查询摘要中 xls 文件的链接(如图 5 画框部分所示),将 xls 文件保存到本地硬盘。用 EXCEL 打开,查询到的微软公司的股价和收益率的数据信息如图 6 所示。



	A	B	C	D	E
1	PERMNO	Names Date	Company Name	Price or Bid/Ask Average	Returns
2	10107	20010131	MICROSOFT CORP	61.06250	0.407781
3	10107	20010228	MICROSOFT CORP	59.00000	-0.033777
4	10107	20010330	MICROSOFT CORP	54.68750	-0.073093
5	10107	20010430	MICROSOFT CORP	67.75000	0.238857
6	10107	20010531	MICROSOFT CORP	69.18000	0.021107
7	10107	20010629	MICROSOFT CORP	73.00000	0.055218
8	10107	20010731	MICROSOFT CORP	66.19000	-0.093288
9	10107	20010831	MICROSOFT CORP	57.05000	-0.138087
10	10107	20010928	MICROSOFT CORP	51.17000	-0.103068
11	10107	20011031	MICROSOFT CORP	58.15000	0.136408

图 6　xls 格式查询结果

综上所述,CRSP 金融数据库设计科学,信息丰富,为国内外学者提供了有益的参考,对国内一些金融类特色数据库的建立,亦有极大的参考作用。随着国内金融研究逐渐与国际接轨,将会有越来越多的国内机构购买 CRSP 数据库,CRSP 数据库在金融研究中的应用也将越发广泛。

参考文献

[1] CRSP Product Documentation[EB/OL]. [2012-10-21]. http://www.crsp.com/documentation/index.html.

[2] Data Description Guide for CRSP US Stock and CRSP US Indices Databases [EB/OL]. [2012-10-21]. http://www.crsp.com/documentation/pdfs/stock_indices_data_descriptions.pdf.