“十三五”江苏省高等学校重点教材

教育部大学计算机课程改革项目规划教材

丛书主编 卢湘鸿

经济管理信息的检索与利用（第2版）

李树青 曹 杰 主 编

蒋伟伟 郑怀丽 副主编

清华大学出版社

教育部大学计算机课程改革项目规划教材

丛书主编 卢湘鸿

经济管理信息的检索与利用（第2版）

李树青 曹 杰 主 编

蒋伟伟 郑怀丽 副主编

清华大学出版社

北京

内 容 简 介

本书主要介绍互联网各种常见经济管理类信息资源及其检索方法，重点对各种信息检索技能和方法结合实际操作演示进行详细说明，注重内容的实用性和易读性，并对互联网免费资源的获取方法专门做了必要的介绍。

本书可作为经济管理类相关专业的本科生和研究生的参考用书，同时也适合对互联网信息检索有需求的读者。

图书在版编目(CIP)数据

经济管理信息的检索与利用/李树青等主编. —2 版. —北京：清华大学出版社，2018
(教育部大学计算机课程改革项目规划教材)
ISBN 978-7-302-50876-2

Ⅰ. ①经… Ⅱ. ①李… Ⅲ. ①经济管理－情报检索－高等学校－教材 Ⅳ. ①G252.7

中国版本图书馆 CIP 数据核字(2018)第 174199 号

责任编辑：谢　琛
封面设计：常雪影
责任校对：白　蕾
责任印制：沈　露

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　编**：100084
社 总 机：010-62770175　　**邮　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 装 者：北京嘉实印刷有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　张**：25.25　　**字　数**：584 千字
版　　次：2015 年 10 月第 1 版　2018 年 10 月第 2 版　　**印　次**：2018 年 10 月第1 次印刷
定　　价：65.00 元

产品编号：080479-01

前　言

本书主要从实践操作的角度介绍各种互联网信息资源的获取方法，尤其对相关经济管理类信息资源做了重点说明。经济管理类信息资源是一种较为常见和使用频率较高的信息资源，与我们的生活和工作密切相关。掌握好相关的互联网信息检索方法和资源获取方法，具有非常重要的实用意义。

在编纂本书的过程中，我们考虑了三个主要因素：

一是加强对信息检索技能和方法的介绍，不论是互联网搜索引擎，还是各种专业的信息资源数据库工具，读者只有很好地掌握各种常见的信息检索方法，才能更好地利用这些工具来完成各种信息资源的获取。因此，本书在第 4 章专门结合搜索引擎的检索方法，对基于关键词的基本检索方法和基于 Web 目录的分类检索方法进行了专门的介绍，并结合典型案例分析对各种常见检索策略进行了详细说明。同时，对于部分特殊的高级检索方法也在后文的特定数据库介绍章节中进行了补充说明。

二是注重内容的实用性和易读性。对于理论性的信息检索内容，本书介绍的并不是很多，相反，对于具体的操作方法，本书不仅非常详细地给出具体的介绍，而且还结合实际截图对关键步骤予以说明，几乎所有的截图都取自 2015 年以后，部分内容在 2018 年修订完成，时效性很强。读者可以自行按照说明和截图演示，对相关内容予以上机实践。

三是强调免费资源的获取方法，例如对如何利用搜索引擎来获取相关信息资源做了大量篇幅的介绍，不仅有专门章节说明，而且在诸如图书、论文和其他各类信息资源介绍时，都对使用搜索引擎来获取的方法做专门介绍，同时也对互联网上的免费信息资源服务做了详细的说明，使得读者可以更容易地获取相关信息内容。

本书由李树青、曹杰任主编，蒋伟伟、郑怀丽任副主编，卢振侠、何光明、郑爱琴、王珊珊、石雅琴、许娟、俞露、凌莉、陈莉萍等也参与了本书资料整理和部分章节的校对工作。本书的写作也得益于互联网所提供的各种信息资源，现代互联网确实能够为我们提供很多有用的帮助。同时，本书还得到了南京财经大学信息工程学院同仁的大力支持，清华大学出版社也给予了很大的帮助，在此一并表示感谢！

限于个人能力，本书可能有很多不足之处，敬请各位读者和专家批评指正。

作　者

2018 年 5 月

目　录

第1章 导 论

本书主要介绍在互联网上检索各种信息资源的方法，其中对经济类信息检索有专门的说明。不过，在正式介绍之前，需要首先了解一些比较基本的概念和原理，以便后续的学习，同时也有助于那些对信息检索原理不是很了解的读者，使其掌握一些信息检索的基本理论内容，从而加深对各种实践操作的理解。本章主要介绍信息资源和信息资源检索的一些基本概念。

1.1 信息资源

随着信息技术的发展和信息技术对人类生活生产行为影响力的不断加深，信息资源本身逐渐成为和能源、材料并列的当代世界三大资源之一，而且它的重要性在很多领域甚至逐渐超过了后两者。正如美国哈佛大学研究小组著名的资源三角形观点："没有物质，什么都不存在；没有能量，什么都不会发生；没有信息，任何事物都没有意义。"

显然，这个现象与人类对信息资源价值的重视密不可分。直到20世纪80年代信息资源才作为一种独立的资源形式，然而近三十年来，特别是随着互联网的高速发展，信息资源逐渐成为一种对国家发展，对人类生活生产至关重要的战略资源。有意思的是，信息资源的可获取性也逐渐得到提高，因此不论是企业、个人还是国家政府，都在广泛充分地利用信息资源来提高驾驭物质资源和能源资源的能力。

所谓信息资源，是指人类社会活动所产生和涉及的一切文件、资料、图表和数据等信息的总称，它的存在形式包括文字、音像、印刷品、电子信息、数据库等。从广义来看，它是指信息活动中各种要素的总称，既包含信息本身，也包含与信息相关的人员、设备、技术、资金等因素；但是从狭义来看，人们通常所说的信息资源只限于信息本身，是指各种载体和形式的信息的集合。

从信息来源的角度，人们通常把信息资源分为四种形式。

1. 体载信息资源

体载信息资源指以人体为载体，通过口头语言和身体语言(体态)这些信息交流符号创造和传播并能为他人识别的信息。参与社会信息交流的每个人都是一个独立的信息源。它其实也是最为古老的一种信息资源形式，在人类的早期，大多数体现知识的信息都是以语言口授的方式得以保存，如各种远古神话传说等，例如《论语》就是记载孔子言语的一本书。

但是，口语信息资源并没有随着时代的发展而逐渐变得不再重要，相反，在现代社会，

口语信息依然是一种极为重要的信息资源，甚至有学者把这种口语信息资源称为“零次信息资源”。很多不经正式渠道流通的信息，各种存储在人类大脑中的知识，往往都只是通过口语的形式传播，也通常只保存在人们的脑海中。显然，人们有必要将其收集整理出来，以便利用。就像前段时间，有人曾经组织过对著名老科学院和老电影艺术家的人物采访，据此这些珍贵的口语信息才能保存下来。一些诸如百度知道之类的网络百科全书也正是利用这种口语信息资源来提供信息查询服务的。

2. 实物信息资源

从严格意义上讲，实物并非信息资源，但是一切物质实体蕴含着的丰富信息均可视为实物信息，它给人们提供了充分认识事物的物质条件。依据实物的人工与天然特性又可将实物信息资源分为以自然物质为载体的天然实物信息资源和人工实物为载体的人工实物信息资源。

在信息资源获取活动中，人们往往通过获取实物来间接得到信息，如产品展览会上展出的各类产品，通过了解这些产品，人们可以得到很多关于市场和竞争企业的相关信息，同时也能够了解该产品的一些具体细节信息。对于经济类信息资源而言，实物信息资源及其相应的数据信息都是一些重要的信息资源。

3. 文献信息资源

文献信息资源是用一定的记录手段将系统化的信息存储在各类载体上而形成的一类信息资源，这些载体包括印刷型载体(Printed Form)、电子型载体(Electronic Form)、缩微型载体(Micro Form)和声像型载体(Audio-Visual Form)等。文献信息资源是信息资源中的主体部分，也是信息搜集、存储、检索和利用的主要对象，也是本书主要的讨论对象。

它可以按照加工的深度分为四种：

一是零次文献信息资源。它是指最原始的文献信息资源形式，虽未公开交流，但它是生成一次文献信息的主要素材。具体形式包括未经记录或者未形成文字材料的口头交谈信息，还有未公开于社会即未经正式发表的原始文献，或没正式出版的各种文献资料。通常获取难度很大，但是在商业和军事领域，它具有特殊的利用价值，也被称为“灰色信息”。

二是一次文献信息资源。它主要是指一些具有原创性的文献信息，如各种论文、专著和新闻等。此类信息价值较大，通常也是人们最终所希望获取的信息内容。常用的一次文献主要包括图书、期刊、会议文献、学位论文、专利文献、标准文献、科技报告、政府出版物、产品样本和产品目录、档案，统称为十大文献信息源。其中，图书、期刊(报纸)被称为普通文献(白色文献)；会议文献、学位论文、专利文献、标准文献、科技报告、政府出版物、产品样本和产品目录以及档案八种类型文献被称为特种文献(灰色文献)，它是一种介于图书与期刊之间的文献类型，通常在出版发行方面或获取途径方面比较特殊，因而被称为特种文献。

三是二次文献信息资源。它主要是指对大量一次文献进行收集整理后形成的信息资源，如摘要和目录索引等，这些文献信息资源的主要目的是提供人们一种查询一次文献的途径和方法。

四是三次文献信息资源。它主要是指在对二次文献进行整理加工的基础上，按照某一个领域和学科方向编撰的带有综合性的文献信息，如百科全书和词典等。

文献信息资源也可以按照信息内容及其检索方式的不同分为三种类型：

一是全文信息资源(Full-text Information Resource)。它就是用户希望获取的最终的信息内容，通常都是一次文献信息内容。随着计算机技术的发展，今天的大多数文献信息资源都可以通过互联网来获取相应的电子全文版本。

二是书目信息资源(Bibliographic Information Resource)。书目是相对于全文而言，也就是说，此类文献信息资源通常都不是人们希望获取的最终信息内容，只是进一步检索的依据和途径，即二次文献信息资源。借助这些书目型信息资源，人们可以更方便地找到所需的全文信息资源。当然，即便没有全文信息，此类信息源依然很重要，人们可以据此来了解某一学科的发展趋势和某一机构或者个人的科研能力，它也是很多科研评价的重要参考依据之一。

三是数值信息资源(Numeric Information Resource)。它主要提供各种原始数据资料，以便学者进行科研工作时使用，如经济统计数据、人口地理数据和产品参数数据等。通常也被称为事实信息资源。

4. 网络信息资源

今天，网络信息资源往往特指以互联网为纽带连接起来的和以互联网为主要交流、传递、存储手段与形式的信息资源。具体是指所有以电子数据形式把文字、图像、声音、动画等多种形式的信息存储在光、磁等非纸介质的载体中，并通过网络通信、计算机或终端等方式再现出来的资源。网络信息资源通过网络将原本相互独立、分布于世界各地的数据库、信息中心、文献中心等连接在一起，形成一个内容与结构全新的信息载体。

相对于传统文献信息源而言，它的信息来源复杂，质量参差不齐。它既包括以网络资源形式存在的文献信息资源内容，也包括网络媒体发布的纯网络信息资源形式，同时还包括诸如大量由互联网用户发布的网页信息内容，特别是随着移动设备的广泛使用，实时通信信息资源日益成为一种重要的网络信息资源。它们数量巨大、增长迅速，传播方式也极为快捷，对现代社会的影响力逐渐增大，也成为人们日常获取信息的主要来源之一。

值得说明的是，本书介绍的信息资源检索方式主要为网络信息资源，其中包括 Web 网站资源、搜索引擎和各个传统文献数据库提供的网站检索站点等。

1.2 信息资源检索

所谓信息资源检索，有时也称为信息资源获取、信息资源查询和信息资源搜索等。它们的意思相差不大，都表示用户利用现代信息检索系统来获取所需信息资源内容的过程。虽然随着信息资源重要性的不断提升，信息资源数量的不断增多，信息资源的可获取性不断增强，人们获取所需信息资源的能力也在增强，但是在很多领域、很多时间，人们对信息资源的获取效果依然不满意，造成这种现象的原因是多方面的，除了技术和资源本身需要改进外，增强信息资源获取意识、加深信息资源形式和种类的了解、提高个人的信息资源

检索能力，这些都能极大地改善这种情况。

1.2.1 信息检索的必要性

很多人都经常说“做功课”。例如旅游出行前，需要对旅游目的地做必要的了解，通过现有的网络搜索引擎和地图服务功能，现代人几乎可以在网上提前完成所有的行程安排及票务预订，了解注意事项。特别是随着移动设备的广泛使用，即使在出行中遇到问题，移动端的信息资源检索服务依然可以实时地提供大量有价值的资讯服务。表 1.1 给出了常见旅行安排所需的信息检索网站及其功能。

表 1.1 常见旅行安排所需的信息检索网站及其功能

旅行安排	国内常用网络信息资源站点	国外常用网络信息资源站点
行程了解、旅游攻略	百度	Google
机票预订	携程、同程	Priceline、Expedia、Orbitz
旅店预订	携程、艺龙	Booking、Airbnb
交通服务	滴滴打车	Uber
餐饮	大众点评网	Yelp

又如在生活和工作中遇到问题怎么办，过去似乎手边总是需要一本诸如“百事通”之类的手册，但今天只需打开互联网，查询一下即可。如惠普打印机出了问题，液晶屏上显示一个 Printer Mispick 的提示，就去找说明手册，其实直接在百度上输入 Printer Mispick，就可以解决问题了，如图 1.1 所示。

图 1.1 在百度中查询 Printer Mispick 的结果页面(截取于 2015-3)

在现代社会，对于个人而言，提高信息检索意识是一种重要的基本技能。有人称之为“搜商”(Search Quotient，SQ)[①]，可以把它看成是一种与智商、情商并列的人类智力因素，也就是人类通过某种手段获取新知识的能力，其本质就是查询信息和搜索信息的能力。

如果把信息检索的必要性再说大一些，人生的几件大事可能都与信息检索有关。

例如高中毕业选择高校和专业，仅仅查阅那个小小的高校专业介绍显然不够。看看大家对这些高校和专业的关注程度和相关网络信息，才可以更好地帮助他们做出选择。如百度提供的高校搜索风云榜就按照关注度对这些高校进行了排序，如图 1.2 所示。

图 1.2 百度提供的高校搜索风云榜(截取于 2015-3)

当然如果能知道专门提供此类高校专业信息的站点，则可以获取更为准确的参考信息，如教育部学位与研究生教育发展中心主办的“中国学位与研究生教育信息网”就公布有年度中国大学的学科排名信息，如图 1.3 所示。

再如就业找工作，其实就是就业信息检索。甚至连找对象这种事情也都成为现代互联网信息检索服务产业中一个很大的市场，如百合等各种婚恋介绍站点，更不要说买房买车之类的事情了。

对于企业而言，在与经济有关的各个领域，信息检索服务及其利用形式更是无处不在。例如利用搜索引擎进行广告推广已经成为一种常见的市场营销策略，用户只需在搜索引擎中输入一些查询词，搜索引擎就会把相应的广告有效地推送给用户浏览。甚至争夺自己网站在著名搜索引擎检索结果的排名位置，也成为一个专门的行当，被称为“搜索引擎优化”(Search Engine Optimization，SEO)。当然很多搜索引擎自己也开始直接在检

① “搜商”概念最早由中搜的总裁陈沛提出。

图 1.3 中国学位与研究生教育信息网公布的高校专业排名信息(截取于 2015-3)

索结果中呈现商品的具体信息，无须用户进一步单击链接去了解价格等商品信息，如图 1.4 所示。

图 1.4 在百度中查询“自行车”的结果页面(截取于 2015-3)

互联网电子商务在这几年得到了大力发展，诸如淘宝之类的网络购物站点也如雨后春笋般陆续出现并取得巨大成功。其实，网络购物的关键就在信息检索，这也是影响用户使用感受的一个最为明显的因素。如果这种系统不能很好地帮助用户找到自己所需的商品，恐怕用户就不愿意使用它们了。为此，淘宝在自己主页的显著位置上放置了一个检索框，同时也在各个商品的浏览页面中集成了各种方便用户的检索功能，甚至还提供了高级检索功能，并要求用户对此提出意见，如图 1.5 所示。

图 1.5 淘宝商品的“高级检索”界面(截取于 2015-3)

从总体来看，今天互联网上的信息量已经呈现出一种爆炸性增长的态势。据报道，由中国互联网信息中心(CNNIC)发布的《第 41 次中国互联网络发展状况统计报告》显示，截至 2017 年 12 月，我国域名总数为 2085 万个，中国网站总数为 533 万，年增长 10.6%，而国际出口带宽为 7 320 180Mb/s①，年增长 10.2%②。面对这个海量的信息资源，用户使用网络信息检索的能力现状如何呢？事实上，人们难以有效地获取所需知识，主要原因在于这种信息资源的增长速度远远超出了人们能够处理它们的能力。约翰·奈斯比特(John Naisbitt)在《大趋势》一书中这样形容人们目前所处的困境：“信息是丰富的，而我们正在渴求知识(Rich Data But Poor Information)。”③

搜狗实验室在 2007 年曾经发表过一篇研究论文④。文中指出，在对搜狗搜索引擎一

① b/s 是指每秒传输数位数(Bits Per Second)，它是衡量网络带宽的重要指标，今天人们使用的网络宽带普遍可以达到 100Mb/s 左右。

② 第 41 次中国互联网络发展状况统计报告. http://www.cnnic.net.cn/hlwfzyj/hlwtjbg/201801/P020180131509544165973.pdf.

③ [美]J. 奈斯比特. 大趋势[M]. 北京：新华出版社，1998.

④ 余慧佳，刘奕群，张敏，茹立云，马少平. 基于大规模日志分析的搜索引擎用户行为分析[J]. 中文信息学报，2007，21(1).

个月内近5000万条查询日志进行分析处理后，发现长度不超过3个词的查询占了总查询数的93.15%，平均长度为1.85个词，这说明用户输入的查询通常都比较短。实际上，查询词语越短就越难以有效地表达用户的准确信息需求。更为吃惊的现象是，只有约0.73%的查询含有用于高级查询功能的符号，即目前大多数中文检索用户只是通过输入很少的几个关键词[①]就开始查询。其实，这些包括布尔查询在内的高级查询往往都能更为有效地表达用户的信息需求。

这充分说明，在更多的情况下，用户检索技能方面的改进空间更大一些，甚至可以说，如果用户不能很好地掌握信息检索方法，再好的信息检索系统也难以发挥它们的威力。

到此，可以对信息检索的必要性有一个感性的认识。在现代社会中，信息检索已经成为一种重要的用户行为，和我们日常工作生活密切相关。所以，我们有必要学习如何更好地使用各种诸如搜索引擎在内的信息查询系统，同时也应该了解不同领域中常见有用的信息资源站点，知道从哪些站点可以更为方便地获取哪些高质量信息，从而为我们提供更多的便利。

1.2.2 信息检索的基本过程

如果把信息资源抽象成一个巨大的人类知识体，那么信息检索活动就是一种人类认识知识和获取知识的基本活动过程。在这种场景中，这种巨大的知识体既可以包括互联网网页信息资源，甚至也可以将图书、报纸等各种传统媒体资源包含进来。因此，用户必须掌握与这种知识体交互的方法，即信息检索方法，才能更好地使用它们。

其实，这包括两个重要条件：一是要存在这样的一个知识体，不管是图书和报纸等传统纸质文献，还是互联网上存储的电子资源，它们都是一种知识体的具体存在形式，因此知识体是客观存在的。这里主要探讨如何在这个信息资源知识体中获取所需信息；二是用户能够表达需要什么样的知识，相对于第一个条件而言，似乎这个条件更为简单，然而对于用户来说，这才是需要着力掌握的技能之一。这其实也就是一种信息检索的能力，越能有效地掌握检索知识，用户就越有可能在今天海量的信息世界中找到自己所需的内容。

不管用户使用搜索引擎还是任何其他信息检索系统来查询知识体，一般而言，主要分为两个步骤：一是用户发出对信息的检索请求；二是信息检索系统响应用户，返回请求的检索结果。不过，这种理解过于简单，用户和知识体并不能直接交流。中间存在两个主要的转换环节：

一是用户需要将自己的信息需求通过查询表达出来，例如在搜索引擎中就是用户输入的关键词等，这既需要用户掌握一些信息检索技能，同时也需要检索系统提供一个良好的界面以方便用户表达信息需求和使用信息检索功能，由此也能看出很多信息检索系统的界面差别正是体现了它们对用户检索感受的不同理解和对当前检索任务的特点的考虑。

---

① 我们通常把用户输入的查询词语也称为“关键词”(Key Word)、“查询词”(Query Term)或者“搜索词”(Search Word)等。它们的含义基本相同。

二是信息检索系统要能够在知识体中找到用户所需的信息内容，这就需要信息检索系统对这些知识体的信息内容做必要的处理，以保证在较短的时间内找到最为相关的信息。然而对于如此巨大的互联网信息检索来说，这并非一件很简单的事情。所以，大多数搜索引擎都是由一些技术先进的大公司来运作和维护的。即使是传统文献检索系统，高质量的文献标注和索引也是提高检索效果的重要基础性工作。

用户在信息检索系统中检索信息的完整过程如图 1.6 所示。

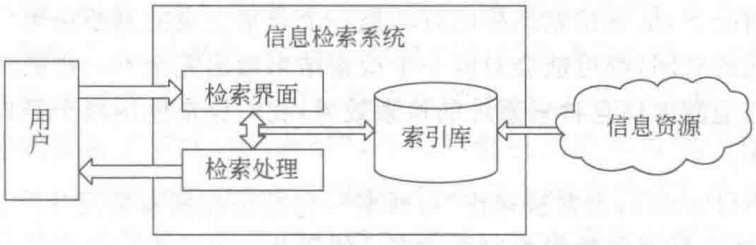

图 1.6 信息检索系统中信息检索过程示意图

从图 1.6 中可以看出，信息检索的基本过程如下：

(1) 信息检索系统为实现快速响应用户检索需求，必须事先对信息资源进行收集、整理和索引，对于搜索引擎而言，就是通过爬虫程序来下载互联网上的各种可以收集到的网页信息，对于文献数据库而言，就是录入诸如文摘或者全文等文献的内容信息。这个过程对于用户而言不可见，但是它是保证信息检索系统整体有效性的数据基础。

(2) 用户通过信息检索系统的检索界面发出检索请求，这需要用户通过系统提供的检索界面表达出自己的信息需求，具体形式可以是在搜索引擎搜索框中输入查询关键词，也可以是在地图上单击所希望查看的地点，甚至是通过语音方式实现输入。准确的信息检索结果往往依赖于用户有能力正确表达自己真实的信息需求，同时也依赖于检索系统可以提供良好的界面供用户来表达这些需求。

(3) 在信息检索系统中，用户的检索请求与索引库的记录按照已有的匹配方法进行计算，最终可以获取到相关结果及其每个结果的相关度，并以友好的方式呈现出来。这种呈现方式直接受限于用户和检索环境的需求，对于海量数据规模的网页检索而言，有效的结果排序方式显得非常重要，而对于文献检索而言，精确匹配并将所有命中结果展示出来则更为重要。

这种信息检索活动是一种常见的行为，尤其在互联网上。一般而言，使用搜索引擎就是一种典型的信息检索行为。那么其他一些查询操作算不算信息检索呢？例如会有人说：我通常上网并不是使用搜索引擎，只是看看网页，这也是信息检索行为吗？

举个例子，用户打开网易主页，看到了主页上的体育新闻，很快单击该超链接，在弹出的新页面中看到了更多的体育新闻。由于该用户是个篮球迷，于是在这个网页中又连续单击看到很多关于 NBA 联赛的消息。

这种操作看起来并不像是信息检索，其实它具有信息检索活动的全部特点，即用户有比较明确的信息需求，同时也在不停地获取满足这种需求的各类信息资源。具体来看，用户为什么要单击关于体育的新闻呢？又为什么继续单击 NBA 的消息呢？这些都能反映出用户的一种个性化的信息需求，也就是说，正是因为该用户想了解这方面的信息，才会