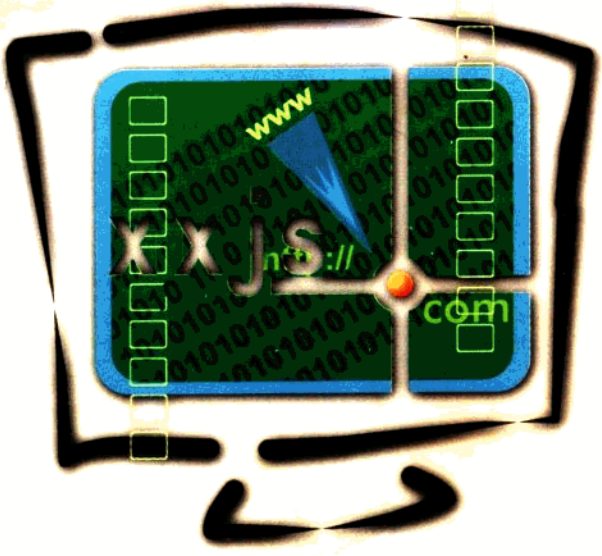
SHIYONG XINXI JIANSUO

实用信息检索

刘广普　主编



河南大学出版社

前　言

随着科学技术的迅猛发展，科技信息的数量急剧增加，种类繁多，内容分散，文种多样，要从这浩如烟海的信息中迅速准确地获得自己所需要的信息，就必须掌握一种方法。而信息检索作为一种科学的学习、研究方法，是高校大学生和教学、科研人员获取知识信息，不断改善知识结构的重要途径。本书编者根据高校大学生的实际情况和信息资源的现状，从培养学生信息利用能力入手，力图使学生了解有关信息检索的基本原理、方法和途径，学会常用检索工具的使用方法，掌握网上信息检索的原理和方法，懂得如何获得与利用信息资源，从而增强自学能力和研究能力。

本书的特点是：

1. 体现信息时代文献检索新特征。目前，信息资源的现状是，现实馆藏还不能完全数字化，数字化的网络信息资源也不能覆盖印刷型的文献，网络信息资源与文献信息资源的并存将是长期的。基于此，该书在结构体系上，既注重传统手工检索部分的内容介绍，也突出现代信息检索部分，展示现代信息检索系统的最新进展和使用方法，具有较强的时代特征，实用性强。

2. 注重网络环境。本书在介绍网络信息资源的检索方法时，着重介绍国内外已具备相当使用条件的、高价值的网络学术数据库与网络学术信息资源的查询与利用方法。高校大学生在了解并掌握具体数据库使用方法的同时，能比较全面、深入地体会当前网络环境下的学术信息交流环境与模式，从而有效地满足自己的信息需求。

3. 理论与实践结合。本书的另一特色是在实例介绍时尽量兼顾本院的粮油骨干专业。

本书共分五大部分二十六章，从信息检索原理与方法技术、信息资源的检索、信息资源的开发利用等方面进行阐述。各章节编写人如下：

第 1 章，曹庆娟；第 2 章和第 9.1、9.2 节，朱萍；第 3 章，段丽；第 4 章，吴小梅；第 5、6、7、8、章和 9.3、9.4、9.5 节，高孟霞；第 10、11、12 章，王国栓；第 13、14、15 章，张爱芳；第 17、18 章，刘广普；第 16、19、20、21、22、23 章，刘广明；第 24、25、26 章，王亚军。

全书由主编和副主编拟定大纲并统稿，由郑州工程学院图书馆张和芬教授 主审。本书在编写过程中，得到了南开大学柯平博士、中原工学院张怀涛教授的热情指导和大力支持，在此深表谢意。

由于作者学识水平有限，书中不当之处在所难免，敬请同行和读者指正。

本书编委会

2002 年 5 月于郑州

第一部分　信息检索基础

1. 信 息 概 论

1.1　信息的涵义

信息作为一个科学概念，最早是由美国贝尔电话公司的电器工程师申农所提出。1948年申农发表了他的代表作《通信的数学理论》，这被视为信息论的奠基之作。后来随着科学技术的发展，社会信息量剧增，信息概念逐步运用到各个领域，人们从不同的角度对其进行表述，由此产生了信息定义的多样化。如有的认为，信息是物质与精神的中介；有的认为，信息是收信者事先不知道的报道；还有的认为，信息是指应用文字、数据或信号等形式通过一定的传递和处理，来表现各种相互联系的客观事物，在运动中所具有的特征性内容的总称。据统计，国内外关于信息的定义已有百余种之多。我们认为，信息是生物以及具有自动控制功能的系统，通过感觉器官和相应的设备与外界进行交换的一切内容。

今天，对信息的定义仍然众说纷纭，站在不同的角度就会有不同的解释，但是，有关信息的基本内涵已取得普遍共识，就像我国“三论”专家王雨田等人论述的那样：“信息论的重要贡献之一，就是在科学史技术史上第一次提出了与质量、能量并列的概念。”即：信息虽然不是物质本身，但它是物质的一种基本属性。信息一般包括两个层面上的意义：作为一般人经常说到的信息，它是指某一具体的消息、情报、新闻、资料、事件等。作为信息论中的信息是一个具有哲学概括意味的范畴，这两个层面上的信息的关系就像一般人说的美同哲学家说的美一样，名称看似相同，其实一个是在具体的层面上使用，另一个则是在抽象的层面上使用。

信息的基本特征是：

（1）普遍性与永存性。信息既然是物质的基本属性，那么就与物质共存。物质永恒运动，无处不在，故而信息也即永恒地普遍存在。

（2）可知性与独立性。信息既然是显现出来的物质属性，因此，物质是可知的，信息也是可以被识别和认识的。信息虽然与物质有着不可分割的关系，但又具有相对的独立性，它不以什么物质作载体和消耗能量的大小为转移。

（3）传递性与转换性。信息是物质的基本属性，依附于物质，因此，物质运动，信息也即运动，物质载体被传递和交流，信息也随之被传递和交流。在传递交流过程中，信息可以从一种形态转换成另一种形态。

（4）共享性与再生性。同实物交易不同，信息的传递交流可为双方共享。人们将这

共享的信息综合概括、加工处理，又可再生出各种新的信息。

1.2 信息的种类

信息的种类很多，按照不同的划分标准可以分为不同的类型。

(1) 按照信息存在的领域划分：可划分为自然信息和社会信息两大类。

所谓自然信息，是指存在于人的意识活动之外，在自然中传递的信息，包括动物界、植物界、微生物界等方面的信息。例如，随着自然界气温、雨雪、风向、阴晴等不同信息的变化，动植物会对此做出不同的反映，呈现出不同的行为状态；另一方面，动植物的活动与生长所发出的信息，又不断影响改变着周围的环境。自然信息在未被人类认识之前，是没有什么价值的。

所谓社会信息，是指反映社会现象的信息，它包括人类社会所进行的一切活动。社会信息只存在于社会领域流通，并服务于一定的社会目的。这类信息又可分为三种：一是被人类所认识，被用来为人类自身服务的自然信息，如风雨雷电、花草树木、鱼虫鸟兽等。这部分自然信息不仅在自然界中传递，而且可以在社会中流通，达到一定的社会目的。二是反映社会历史发展过程中的具体时间、事实、数据、情况等事实性信息。三是在实践活动中产生的反映自然界和社会规律的理性信息，它呈现为系统化形式，并表现为概念、判断、推理和理论，也称为科学信息。

(2) 按信息的内容性质和学科研究对象划分：有政治信息、军事信息、经济信息、科技信息、生活信息等。

(3) 按信息的加工程度划分，有零次信息、一次信息、二次信息、三次信息。

零次信息是指未经公开发表或未交流于社会的信息。如手稿、实验记录、文章草稿、会议记录、书信、电话等。

一次信息是指未经加工或经过粗略加工的原始信息。如：图书、期刊论文、会议论文、科技报告、专刊文献等。

二次信息是在一次信息(原始信息)基础上加工整理而成的，用于检索一次信息的工具。它是查找一次信息的重要工具，主要是指文摘、索引、目录等。

三次信息是根据二次信息提供的线索，查找一次信息，并对其进行分析、研究、综合而形成的具有较强概括性的浓缩信息。如综述、年鉴、研究报告等。

(4) 按信息传递的范围划分，有公开信息、半公开信息、非公开信息。

(5) 按信息的整序特点划分，有系统化信息和非系统化信息。按一定目的或方法将信息系统地汇总、整理、贮存、保管起来，以便人们检索、利用的信息称为系统化信息。散见于各种载体之上的无序信息称为非系统化信息。

(6) 按信息的作用划分，可分为有效信息、无效信息、负效信息等。

(7) 按信息的载体形式划分，有文献信息、声像信息、多媒体信息等。

文献信息是指将文字记录在一切载体上的信息。如印刷型文献、胶片及计算机软盘、硬盘、磁带等。声像信息是指各种声音和图像的信息，如广播、电视及卫星图片等。多媒体信息是将文字、声音和图像融为一体的一种新型的信息传递形式。

1.3 信息的作用

今天，信息一词已成为家喻户晓的常用语。那么，信息究竟有什么作用？第一，信息是构成自然界的三大部分之一。就像我们前面提到的，物质、能量、信息构成了自然界的三大部分。人类社会发展有三个里程碑：第一是发现世界是物质的并改造物质世界；第二是发现物质是有能量的并利用能量；第三就是发现信息也是物质的一种基本属性，它能给我们提供物质存在、运动、相互作用和相互转化的消息。第二，按照申农的定义，信息就是能够减少或消除不定性的东西。比如，两军作战，彼此对敌方的情况如实力、装备、士气等都不很清楚，这就存在许多不定性的因素；如果通过侦察获得一些情报，使得不定性减少，那么这些情报就是信息。人们正是通过不断地获得自然界和社会的不同信息，并加以分析、归纳和处理，从而达到认识世界、改造世界的目的。所以，信息是人类意识与客观存在相互作用的媒介，是人类感知的来源，在人类认识世界改造世界的过程中，起着重要作用。

再则，信息是资源。资源是人类社会生产的对象，在工业社会中是物质资源。物质资源是有形的，如水资源、森林资源、矿藏资源、生物资源等。信息不像物质资源那样构成直接的生产手段，它是一种无形的资源，具有无限的再生性。人们掌握充分的信息可以极大地提高生产率，节省资源和能源。因此，人们视信息、物质、能源为现代社会的三大支柱。例如，日本国土狭窄，资源贫乏，但它大力推行“信息资源化政策”，充分开发利用世界信息资源，以至于战后经济高速发展，并长期在世界经济舞台上立于不败之地。目前在我国，正在形成一个以信息为基础，包括信息的生产、收集、处理、存贮、传播、使用等多环节的产业集合体——信息产业。信息产业的发展，对于推进我国的工业化进程有着巨大的意义。

1.4 与信息相关的几个概念

过去，我们常用文献、知识、情报这些概念来表述我们的认知对象。这几年，国内兴起用信息这个词汇来描述以上三者，这是因为信息这个词包含的范围更广，更能反映时代特征。但以上三者和信息的关系并不能完全等同，它们既有区别，更有联系。下面，我们做简单介绍。

知识

知识是人类对客观世界的物质形态及其运动过程和规律的认识。

自然界和人类社会中普遍存在着信息，这些信息有的已经被人感觉到，并通过人的感觉器官送到头脑，经过大脑思维加工处理，成为人对自然社会的认识，就是知识。这些知识就是序列化了的信息。也就是说，知识是人脑意识的产物，是经过人脑加工系统化了的一种信息。但并不是所有的信息都是知识。

文献

文献一词，最早见于《论语·八佾》：“子曰：夏礼吾能言之，杞不足徵也；殷礼吾能言之，宋不足徵也。文献不足故也。足，则吾能徵之矣”。这里，孔子只是把文献的足与不足作为能否讲解杞宋两国礼制的一个重要依据，而并未说明文献是什么。南宋理学大师朱熹在《四书句集注》中说：“文，典籍也；献，贤也。”按朱熹的解释，文献是指典籍和贤才。

随着时代的发展，“文献”一词的含义不断变化，“文”和“献”的差别逐渐消除，一般泛

指具有史料价值的图书和档案。《辞海》对文献的界说，很能反映文献概念的演变：文献“原指典籍与宿贤。……后专指具有历史价值的图书文物资料。”

当今，由于记录知识的手段和载体发生深刻的变革，文献一词的含义也赋予新的内容：1984 年 4 月 1 日实施的中华人民共和国国家标准《文献著录总则》中就明确规定了“文献”一词的定义：“文献：记录有知识的一切载体。”也就是说，文献是用文字、图形、符号、声频、视频等技术手段记录人类知识的一种载体。它包含四个因素：(1)有记录信息、知识的内容；(2)有记录信息、知识的物质载体；(3)以文字、图像、符号、声频、视频等作为记录手段；(4)以一定的形态呈现出来，如图书、期刊、档案、专利说明书、缩微品等等。

情报

情报是为解决特定问题而传递给特定对象的有用的知识。从这个定义中我们可以看到，情报具有三个基本属性：知识性、传递性、有用性。所谓知识性，就是指任何学科领域的，以任何形式出现的情报都具有一定的知识内容，情报是一种知识；传递性，就是指知识只有经过传递才能成为情报，情报是一种动态的知识；有用性，人们传递一定的知识，目的在于解决特定的问题。情报是一种动态的(传递中的)有用的知识。但由于情报一词在英语中有“间谍”之意，随着我国对外交流活动的增多，它极易引起误解，已无法适应时代的要求。所以，1992 年国家科委决定将科技情报改为科技信息，许多情报所已改为信息所，更具有时代的特点。过去所说的情报检索也就是今天的信息检索，只是检索的内容更加广泛。

由上述可见，情报是知识的一部分，文献是情报的一种载体，也是知识的一种载体。文献不仅是情报传递的主要物质形式，也是吸收利用情报的主要手段。

2. 信　息　源

2.1　信息资源的含义

信息是一种重要的资源已成为共识。对信息资源进行组织、管理、建设、开发、利用成为人们普遍关心的问题，对信息资源管理的研究也成为国内外研究的热点之一，并已发展成为一门新兴的学科。但时至今日，对什么是信息资源仍众说纷纭，国内外都未形成一致看法。有的认为信息资源就是文献资源，有的认为信息资源就是数据，有的认为信息资源就是各种媒介和形式的信息，还有的认为信息资源就是信息活动中各种要素的总和(包括信息、设备、技术和人等)。比较流行的看法达 16 种之多。

我国对信息资源概念及其有关问题的研究始于 20 世纪 80 年代中期。早期对信息资源概念的认识受国外特别是美国的影响较大。1996 年，乌家培撰文指出：“对信息资源有两种理解。一种是狭义的理解，即指信息内容本身。另一种是广义的理解，指的是除信息内容本身外，还包括与其紧密相连的信息设备、信息人员、信息系统、信息网络等。”在“96 信息资源与社会发展国际学术研讨会”上，江大川进一步解释说：“广义的信息资源是指信息和它的生产者及信息技术的集合，即广义的信息资源由 3 部分组成：(1)人类社会经济活动中的各类有用信息；(2)为某种目的而生产有用信息的信息生产者；(3)加工、处理和

传递有用信息的技术；狭义的信息资源则仅仅指人类社会经济活动中经过加工处理有序化并大量积累后的有用信息的集合，它包括科学技术信息、政策法规信息、社会发展信息、经济信息、市场信息、金融信息等多方面内容。"这两种比较流行的广狭之说，颇具代表性，实质上是对国外各种观点的综合概括。现在国内关于信息资源含义的认识正在逐渐深化，总的趋势是从广义之说向狭义之说方向转化。因此，我们对信息资源含义持狭义的理解，认为所谓"信息资源是经过人类选取、组织、序化的有用信息的集合。"其原因很简单，因为只有信息本身才是信息资源中的核心要素，而且一条信息或几条信息也构不成信息资源，只有当信息达到一定的丰度和凝聚度时，才能成为信息资源。从这个意义上说，信息资源应是多种多样信息的总和或集合。再则，有用性是一切资源的本质属性，信息资源也不能例外。第三，与非信息资源相比，信息资源最显著的特征就是有序性。对水资源、石油资源、矿产资源等自然资源来说，无所谓有序、无序，只要具备一定的丰度和凝聚度，值得人们开采、获取即可。信息资源却不然，无序的信息不仅无法利用，还会造成信息通道的"栓塞"，阻碍信息的传播、交流、开发和利用。因此，组织、序化的信息才能成为信息资源，而没有控制的、未经组织的信息将不能成为资源。

2.2 信息资源的特点

信息资源的特点，主要体现在以下几个方面：

(1) 共享性。一般的物质资源在对其利用上，总是存在着明显的竞争关系，你用多了，那我就得少用甚至不用。但信息资源不存在这种问题，举个简单的例子，大家都来读一本书，每个人从中获得的信息量因自身因素虽有不同，但不会因为其他人已经阅读或将要阅读而受到影响，人们可以同等地共享某一份信息资源。

(2) 传递性。信息资源只有被传递才能发挥作用，传递的速度越快，发挥的作用也就越大。信息资源的传递速度与载体的形式有关。由于载体的形式不同，信息资源的运动速度是不一样的，如电子型信息资源的运动速度要比印刷型信息资源的运动速度快得多。目前，由于计算机等高新技术的发展，处理信息的能力进一步增强，信息传递的速度日益提高。

(3) 依附性。信息资源的生产与传播均离不开物质载体和将信息内容记录到载体之上的高新技术，与载体分离的信息资源是不存在的。

(4) 时效性。信息资源的时效性是指信息在一定的时间范围内具有价值或效益。一条及时的信息可以给用户带来一定的效益，过时的信息就发挥不了作用或作用相对降低。

(5) 再生性。人们在传播、利用信息资源的同时会生产出许许多多新的物质，新的物质相互作用又会产生出许多新的信息，如此循环往复。新的信息资源也就源源不断地产生。正是由于信息资源的这种再生性，自然界和人类才会不断进步、发展。

2.3 信息资源的分类

进行信息检索，必须对信息资源的类型和特点有所认识。不同的划分标准形成不同的信息资源类型。不同的信息类型，其特点和检索方法也有所不同。

2.3.1 按信息资源内容分类

(1) 按表现形式分

① 文献型信息源

把语言文字贮存在各种不同载体上以记录信息内容的信息资源是文献型信息源。它是目前信息内容最丰富、人们使用频率最高的信息资源。

② 数据型信息源

它是以数据形式出现并存储在各种不同的载体上的信息集合。据统计，1997 年有 3000 多种数据库通过因特网直接为用户提供信息检索服务，这些数据库的内容涉及不同领域、不同专业。

③ 声像型信息源

它是以声音或图像形式出现的信息资源，如广播、电视、CD 光盘以及 VCD 光盘等。这是目前正在开发的、可以在网络上向用户及时提供图像和声音的信息的传播方式。

④ 多媒体信息源

多媒体信息源是随着现代科学技术的发展而出现的一种新的信息源形式，它集文字、声音、图像于一体，多以光盘或网上形式出现。

(2) 按加工程度分

① 零次信息源

指未经公开发表或未交流于社会的文献信息。如手稿、实验记录、文章草稿、会议记录、书信等。

② 一次信息源

通常是指原始文献信息，即指作者以本人的研究成果为基本素材而创作撰写的文献信息。如期刊论文、科技报告、专利说明书、会议论文、学位论文等等。

③ 二次信息源

指将分散的、无组织的一次信息经加工、整理、编辑成目录、文摘、索引等检索工具或数据库，以便查找和利用。二次信息的重要性，在于它可以作为查找一次信息的线索，对信息具有存储、拟返和检索的作用。

④ 三次信息源

就是在一、二次文献信息的基础上，经过综合分析而编写出来的文献信息，如专题评述、动态综述、学科年度总结、百科全书、年鉴等。

总的来说，零次信息是一次信息的素材；一次信息是信息的基本形式，从一次信息到二次、三次信息的过程，是一个由博而约，由分散到集中，由无组织到系统化的过程。从信息检索来说，一次信息源是检索的主要对象，二次信息源是检索的主要手段与工具。

(3) 按传递的范围分

有公开信息源、半公开信息源、非公开信息源。

(4) 按内容性质和学科研究对象分

有政治信息、军事信息、经济信息、科技信息、生活信息等，这些方面还可以继续采用一定的标准来划分，如经济信息可以进一步区分为企业信息、商业信息、金融信息、贸易信息、股票信息、会计信息、期货信息、保险信息等。信息与其学科研究对象有着不可分割的关系，可以说，有多少学科就有多少学科信息。

2.3.2 按信息载体形式分类

2.3.2.1 印刷型信息源

(1) 图书

图书的内容大多是对已经发表的科研成果、生产技术和经验，或者某一知识领域的论述或总结。一般来说，内容比较全面、系统、成熟、可靠，是人们学习知识、从事工作不可缺少的信息来源。但图书的出版周期长，知识信息不如期刊论文和特种文献新。有资料显示，美国若干大学和英国的电气工程师们阅读的科技文献中，图书的比重分量仅占19%和14%。图书一般都有惟一标识即国际标准书号(ISBN)。由ISBN号可知该书的语种区、出版社、流水号等出版信息。ISBN号一般由10位数组成：第一位代表语区；第二位数字到第五位或第六位数字代表出版社；第六位或第七位数字至第九位数字表示该出版社出版的图书种数的流水号；最后一位数是计算机校验码。其中，0和1是英语区，2是法语区，3是德语区，4是日语区，5是俄语区……我们国家的语区代码是7。

(2) 期刊

期刊是指有固定名称、定期连续出版的刊物。它具有出版周期短、报道速度快、内容新、信息合理等特点。由于期刊的这些特点，科技工作者从期刊中获取的信息，约占整个信息来源的65%。期刊有多种类型，其学术地位和利用价值往往差别很大。据信息专家研究，许多学科专业20%的期刊包含了80%左右的有用信息，这就是所谓的核心期刊。因此，对科技工作者来说，注意掌握和利用核心期刊，是十分重要和必要的。

(3) 专利文献

广义的专利文献指专利局出版的与专利有关的各种文献，如专利公报、分类表、索引、说明书等。狭义的专利文献是指专利说明书，它是专利文献的主体。专利说明书具有内容广泛、系统详尽、格式规范、出版迅速等特点。由于有专利制度的保证，发明专利的新颖性、创造性和实用性审查，内容翔实可靠，技术、经济、法律三位一体，并且还具有地域性、时效性、发明内容单一性等特点，使用时应予注意。

(4) 标准文献

技术标准、技术规范和技术法规等统称为标准文献。它主要是对科学试验、工程设计、生产建设、技术转让、国际贸易、商品检验的质量、规格及其检验方法等方面所作的技术规定，是从事生产、建设和管理时需共同遵守的具有法律约束力的技术依据，每一件标准都是独立完整的资料。

标准按使用范围分，有国际标准、区域性标准、国家标准、部标准、专业标准和企业标准等；按内容分，有基础标准、产品标准和方法标准等；按成熟程度和约束力分，有法令标准、推荐标准、试行标准和标准草案等；按技术内容分，有计算单位、符号、术语、尺寸、形式、品种、基本参数、技术要求、试验方法、计算方法、工艺流程、包装标志、运输和保藏等标准。标准文献的特点是：①制定、审批有严格的程序；②适用范围明确专一；③编排格式、叙述方法严谨统一，措词精练准确；④有一定的可靠性和现实性；⑤对有关各方面有约束力，某些标准文献具有法律效力；⑥新陈代谢比较频繁，具有时效性。随着经济条件和技术水平的改变，标准文献常不断进行修改和补充。

标准文献的价值在于通过标准可以了解各国经济、技术政策、生产水平、资源情况和

标准化水平，可预测分析发展动向，可借鉴国外先进技术，是人们进行科研和生产不可缺少的重要文献。

（1）学位论文

学位论文是高等学校毕业生为获取学位而提交的学术论文，包括学士、硕士和博士论文三种。其中博士论文学术价值较高，具有一定的独创性，内容比较系统详尽。学位论文一般都保存在学位的授予单位，也有少数论文在期刊上发表。中国科技信息所收藏有国内外硕、博士论文部分复制品，北京首都图书馆存有我国全部博士论文。学位论文对科学研究和撰写学术论文均有参考价值。

（2）会议文献

会议文献是指在国际或国内学术会议上交流的文献。它内容新，往往反映出科学技术的最新成果和发展趋势，是了解某学科水平动态的重要信息源。会议文献按出版时间划分，可分为：①会前文献：包括会议议程、会议论文预印本和论文摘要等；②会后文献：有书本式的会议录、论文集、期刊特辑和声像资料等出版形式。

会议文献的英文标识通常有：Conference（会议）、Congress（大会）、Convention（大会）、Symposium（专题讨论会）、Workshop（专题学术讨论会）、Seminar（学术研讨会）、Colloquium（学术研讨会）、Proceedings（会议录）等，前面三种会议形式往往是国际学术行政会议，规模大，人数多；其余则是小型学术研讨会，论题较为专深。

（3）政府出版物

它是指各国政府部门及其专设机构所发表出版的文件，可分为行政性文件（国会记录、政府法令、政策、统计等）和科技文件（科技报告、技术政策文件等）。其特点是品种多、数量大，某些文件具有指令性。其作用是有助于了解国家的政策及其变化情况。

（4）科技报告

科技报告是关于某项研究成果的正式报告，或者是对研究过程中每个阶段进展情况的实际记录。其特点是内容专深具体、数据完整；有机构名称，统一编号，自成一册。科技报告有许多是保密和控制发行的，因此它的获取不如期刊容易。由于它是研究的记录和成果，代表了某一学科的科技水平，因而可以对科研工作起到直接的借鉴作用。许多最新的研究课题或尖端学科的资料，往往首先反映在科技报告中。

科技报告与书籍、期刊等不同，类型比较复杂，名称叫法往往用代码表示。因此要利用科技报告，首先要识别它们的代码含义。目前国际上常用的科技报告代码主要有：

① 机构代码：这是科技报告代码的主要部分，用来表示科技报告的名称，一般以编辑、出版、发行机构名称的首字母作报告的机构代码。例如，PB 是 Publication Board 的首字母缩略语，即美国政府出版局；AD 是 Armed Service Technical Information Agency (ASTIA) Document 的首字母缩略语，即美国军事技术情报局；NASA 是 National Aeronautical Space Administration 的缩略语，即美国国家航空航天局；DOE 是 Department of Energy 的缩略语，即美国能源部。

② 类型代码：用来表示科技报告类型。主要的类型代码有：

B——Bulletin	通报
CR——Contractor Report	合同户报告

M——Memorandum　　备忘录
N——Notes　　笔记
P——Paper　　论文
PR——Progress Report　　进展报告
SP——Special Publication　　特种出版物
TB——Technical Briefs　　技术简讯
TM——Technical Memorandum　　技术备忘录
TN——Technical Notes　　技术笔记
TP——Technical Paper　　技术论文
TT—— Technical Translation　　技术译文

③ 密级代码：用来表示科技报告的保密程度，具体有：

ARR——Advanced Restricted Report　　绝密报告
C——Classified　　保密报告
R——Restricted　　控制发行的报告
S——Secret　　机密报告
U——Unclassified　　非保密报告

④ 分类代码：用来表示报告的主题分类，如：

P——Physics　　物理学
C——Chemistry　　化学

⑤ 日期代码和序号：如："99－916"表示 1999 年的 916 号报告。

目前世界上每年写出的科技报告约有几十万件之多，其中最著名的是 AD、PB、NASA、AEC 四大报告。我国科技成果的统一登记和报道工作，从 1963 年正式开展。代表我国科技水平的科技报告是由国家科学技术研究成果管理办公室编辑、科学技术文献出版社出版的《科学技术研究成果报告》。

以上是常见的各种类型的文献。总的来说，不同类型的文献往往为不同的研究工作所需要，或为一项工作的不同阶段所需要。例如了解学科领域的背景资料，宜利用图书资料作为入门指导。进行科学研究主要利用期刊论文等。开展技术革新，新产品试制，往往参考专利文献。及时地了解和掌握国外科学技术的最新信息，使我国的科技工作建立在世界最新成就的起点上，这是科技创新的需要。因此，掌握文献信息的查阅方法，是必要的基本功。

2.3.2.2 网络型信息源

据统计，Internet 每月传递 1.5 亿包数据，可供检索的文件达 2100 万件，有着数千万台计算机存有各种各样的信息供人们利用，而且网上的信息每天都在不断增长。世界上任何地方的 Internet 用户都可以利用 Internet 的服务资源去获取 Internet 提供的所需要的信息资源。这些资源涉及人类从事社会活动的各个领域、行业，包括自然科学、技术科学、农业、医学、社会学等专业领域，社会、政治、历史、科技、医疗卫生、政府决策、金融、商业、娱乐等各个方面，各种社会公众服务领域，如体育、音乐、艺术、天气预报、旅游、消遣等，可谓五花八门，应有尽有。面对如此海量的信息资源，用户难免有眼花缭乱之感，把它

们列举分类，也难免有挂一漏万之嫌。下面仅从信息的用途上对网络资源进行简单的划分：

（1）科技类

Internet 上含有大量的科技方面的信息，许多科研机构、大学院校及个人都在网上发表他们的科研成果和学术论文。如数学、物理学、化学、天文学、航天与航空、农业、生物学、医疗卫生、环境保护、地质科学、计算机科学等等。在网上还有各种各样的科技信息数据库，供人们查找有关科技方面的信息，当然这些数据库一般都是收费的。

（2）贸易类

在互联网上，经贸信息是最活跃的，具有范围广、更新快的特点。包括政府的一些统计数字，金融、证券、股票信息，期货贸易，新产品，市场与销售，大公司排名榜，商品广告，公司名录，每日农业市场报告，房地产信息，各国或地区的经济概况等。其中政府机构建立的经济网站，信息集中，权威性强。

（3）文化类

在互联网上还有许多文化信息，包括各种绘画、雕塑、电影、音乐、文学作品，各国及地区的风土人情，网络游戏，等等，大大丰富了人们的文化生活。

（4）教育类

互联网教育资源丰富，信息量大，每个教育类网站都包含了丰富的教育资源、教学资源（如课程设置，教学课件）、教育研究资源（如教育理论知识、科研动向与成果）、教育状态信息（如学校规模、学生发展情况、教学活动状况）等，另外还有教育新闻、教育政策法规、招生与考试、教育服务、各级各类学校的在线课程、网络学校等。目前我国大大小小的专业教育类网站已达三千多个。

（5）新闻类

互联网上的新闻信息，有国家政府发布的，也有社团组织、个人发布的，因此，对信息的可靠性，需要大家在利用时加以鉴别。

（6）软件类

互联网上有许多各种各样支持 Internet 的软件供人们使用，它们和硬件一起组成的各种系统为互联网用户提供了各种各样的服务资源。正是这些服务资源把各种网络信息资源有机地结合在一起，从而构成了 Internet 的全部网际资源。这些软件有一些是免费的，有一些需要付费，一般最新版本的软件或商用软件是收费的。

网络信息资源具有数量巨大，增长迅速，内容丰富、形式多样，变化频繁、价值不一，结构复杂、分布广泛，信息商品化等特点。用户在使用时，要注意把握其特点，以便更好地获取所需信息。

2.3.2.3 数据库信息源

以特定的符号表达的信息称为数据。数据是信息的一种量化表示，数据仅反映信息，而信息依靠数据来表达。表达信息的符号可以是数字、文字或图形。计算机能接受、处理和存放的信息必须是数字化的信息，因此，必须把信息转换成可以被计算机接受的数据，也就是以二进制形式存储在计算机内并被计算机加工处理的数据。

一定数量的数据存储于特定的存储介质之上，以特定的结构相互联结于一体，就形成

了数据库。计算机数据库的建立与计算机的发展几乎是同步的。虽然 Internet 上信息资源非常丰富，但这些网站提供的信息一般娱乐性、生活性的比较多，科技含量较低。对于科技人员来说，有科研价值的信息大多数来源于大型科技文献数据库系统。

(1) 文献目录数据库

文献目录数据库是只存储有关主题领域各类文献资料的目录信息，以二次文献的形式报道文献的数据库，如题录数据库、文摘数据库、引文数据库、期刊目次数据库以及图书馆馆藏目录数据库等。目录数据库以简略的形式向用户报道文献的信息，提供查找、获取文献的线索。这类数据库信息量大，信息密度高，文献报道范围广，数据连续性、累积性强，是用户快速查找文献的有效工具。

目录数据库中的数据来源于期刊论文、会议论文、研究报告、专利文献、学位论文、图书、政府出版物、报纸等各种不同的一次文献，是经过加工、压缩的派生性数据。

目录数据库的数据结构一般比较简单，记录格式也较为固定。因此，其建库速度往往比较快，建设费用相对较低。据统计，全国现有目录数据库近 70 个，大部分记录数在 10 万条以下，大型目录数据库还很少，服务范围也比较狭小。不少数据库仅供本单位使用。数据库的商品化、标准化程度较低，有待进一步发展、提高。

(2) 全文数据库

全文数据库是存储文献全文或其中主要部分，以一次文献的形式直接提供文献的源数据库。通常将经典著作、法律条文及案例记录、重要的科技期刊、新闻报道和百科全书、手册、年鉴，以及图书馆所藏的其他重要文献的全部文字或主要文字转换成计算机可读形式，建成数据库。用户可以从中直接检索出所需的原始文献。

与其他数据库(如目录数据库，事实数据库)相比较，全文数据库有以下几方面的特点：

① 检索直接，信息客观。能直接检索出原始文献或解决问题所需的文献资料，不必进行二次检索(即根据检出的目录信息再去查找原文)。而且库中信息基本上是未经加工的原始文本，具有客观性。

② 检索彻底，报道详尽。可对其中任何字、词、句进行检索，还可表示检索词之间的复杂位置。文献的正文部分或附属部分都可以检索获得和显示，用户可以直接查看到文献正文中的每一段、每一句和每一个词，甚至还可以看到某些边缘性信息。

③ 检索语言多用自然语言，少数用受控语言。检索方法除使用布尔检索以外，位置检索占有相当突出的地位。

④ 标引全面，使用简单。绝大多数全文数据库多利用计算机进行全文自动抽词标引，能为用户提供标题、著者、关键词、摘要等多重检索途径，使用方便。

⑤ 用户接口多为菜单驱动型，或采用较简单的检索命令，易学易用。

全文数据库的发展是在 20 世纪 80～90 年代，据统计，1985 年至 1989 年的 5 年里，全球全文数据库增长了 50% 。在所有数据库中增长率最高。美国 DIALOG 情报检索系统的 300 多个数据库中，1990 年就有 100 多个库提供全文检索服务，占数据总数的三分之一。其主要原因是：a 全文数据库能提供所需的原文。b 计算机速度和容量的极大提高。c 计算机网络的发展，使全文数据库成为网上主要信息资源之一，使文献的可获得性

大大提高。d 封装型的全文光盘系统大量问世。

我国的全文数据库起源于语言学研究。20 世纪 80 年代中期以后，开始向其他领域拓展。1992 年初，北京大学与美国 3M 公司合作生产了英文版的“中国对外经济贸易法律全文数据库”光盘，这是我国第一个 CD-ROM 全文数据库。从此，我国的全文数据库开始与 CD-ROM 这种载体结合起来。目前已开发出了一系列的 CD-ROM 全文数据库。

(3) 事实数据库和数值数据库

事实数据库是存储有关事物(如人物、机构、事件等)的一般指示性描述的参考数据库。因此又称之为“指示性数据库”或“指南数据库”。其主要用途是供用户查询特定事物的发生时间、地点、过程等简要情况。

事实数据库的类型很多。若按信息类型划分，主要有以下几种：①人物传记数据库；②公司名录数据库；③基金指南库；④技术标准指南库；⑤软件数据库；⑥产品指南库；⑦术语数据库。

事实数据库具有以下特点：①提供实体信息，每个款目都是对事实的描述；②数据库记录长，字段多，许多字段都可以按数字或字母顺序排列；③以名称检索为主要检索手段，并提供与事实信息有关的较多的检索点；④查准率高，查全率低。

数值数据库是一种以自然数据形式表示、计算机可读的数据集合。它所记录的数值数据是人们从文献资料中分析提炼出来的，或是从实验、观测和统计工作中直接得到的。数据库生产者将这些数据收集起来，经过核实、检验和加工整理，按一定的方式组织起来，利用计算机进行存储和检索，就是数值数据库。主要记录和提供特定事物的性能、数量特征等信息。其信息报道范围常覆盖某大类的专业领域。如在军事上、部队的实力、装备、武器的性能等信息；在商业和经济领域中，特定产品的性能特征、价格趋势、国家经济增长率等数值信息；在科技领域，物质的物理化学性质、结构、频谱及实验数据、计算公式等。

数值数据库对所收集的数据的可靠性要求比较高，有时还需要列出数据的误差估计、数据来源和实验条件等，以减少用户使用中的误差。数据通常成组排列，本身并不被检索，但它们与字母或数字形式表达的可检索的关键词或叙词相联结，成为可检索的数据。数值数据库检索的结果是一个或一组特定的数值。

为了满足用户的不同需要，许多数值数据库还常附有特定的检索软件包，以提供数值的运算、统计分析、分类、排序和重组等功能。

(4) 多媒体数据库

多媒体数据库是相对于传统的仅支持单一媒体的数据库而言，将图像、图形、文字、动画、声音等多种媒体数据合为一体，并统一地进行存取、管理和应用的数据库。多媒体数据库可以对客观世界存在的对象所表示出来的信息进行更生动、更全面的保存和查询。用户除了能浏览相应对象的文字描述，还可以观看其对应的图形或动态的图像，听到对应的声响。提供多媒体数据存储和检索功能的系统将是一个更直观、更真实的科学数据库系统。

3. 信息检索知识

3.1 信息检索的概念及原理

3.1.1 信息检索的概念

信息检索(Information Retrieval)是指将信息按一定的方式组织和存贮起来,并根据信息用户的信息需求查找所需信息的过程和技术。所以,它的全称又叫"信息存贮与检索"(Information Storage and Retrieval)。信息检索的全过程应包括以下两个主要方面:

一方面是信息标引和存贮过程,就是对大量无序的信息资源进行标引处理,使之有序化,并按科学的方法存贮,组成检索工具或检索文档,即组织检索系统的过程;另一方面是信息的需求分析和检索过程,就是分析用户的信息需求,利用已组织好的检索系统,按照系统提供的方法与途径检索有关信息,即信息检索的应用过程。

以上所说的信息检索是广义的信息检索,主要是对信息工作者而言。

狭义的信息检索则仅指该过程的后半部分,即用户根据需求,借助于检索工具,从信息集合中找出所需要信息的过程,也就是利用信息系统、检索工具或数据库查找所需信息的过程。相当于人们通常所说的信息查询(Information Search)。本书所介绍的信息检索就属狭义信息检索。

作为信息检索对象的信息资源,它有着不同的表现形式,有的以文献的形式出现,有的以数据和事实的形式出现。因此,按检索对象形式的不同,信息检索可分为文献检索、事实检索和数据检索。

(1) 文献检索(Document Retrieval),凡是以文献(包括文摘、题录和全文)为检索对象的称为文献检索。要查找某一主题、某一时代、某一地区、某一著者、某一文种的有关文献,以及回答这些文献的出处和收藏处所等,这些均属文献检索的范畴。从性质上说,文献检索是一种相关检索,检索结果只是文献线索,还需进一步找到这些文献,阅读后,才能决定取舍。例如"关于纳米技术有什么文献?"这就属于文献检索范畴的问题。

(2) 数据检索(Date Retrieval),是以数据为检索内容的信息检索,这里的数据不仅包括从检索系统存储的大量原始调查数据和其他统计数据中查出专门的数字资料的数值数据,如参数、系数等,还包括非数值数据,如计算公式、化合物分子式、图表、工业技术产品名称与规格等等。例如,2000 年我国城市人口人均住房面积是多少平方米? 2001 年我国大学生在校人数有多少? 食品厂、粮油加工厂的生产产量等,这些都是数据检索。数据检索是一种确定性检索,数据是事先经过专家精心测试、评价、筛选过的,用户检索出来后可以直接使用,无须再查原始文献。

(3) 事实检索(Fat Retrieval),是以具体事实为检索内容的信息检索,这里所说的事实,是指从检索系统存储的各种原始信息资源中查出某一事物的存在情况、发生的时间、地点和过程等,是从信息资源中找出用户所需的事实过程,或者是对信息资源中已有的基本事实进行逻辑推理,然后输出新的事实过程。例如:各种厂商名录、世界名人录以及我国"九五规划"执行的结果如何等均为事实检索。

以上三种检索类型中以文献检索为主，它是信息检索中最重要的一部分。文献的检索方式可以分为手工式检索（简称“手检”）和计算机检索（简称“机检”）。

3.1.2 信息检索的作用

信息检索是查找信息的重要方法和手段，它能使人们在浩如烟海的信息海洋中迅速、准确、全面地查找到自己所需要的信息。可以说信息检索在人们的工作、学习和生活等各方面具有很重要的作用。

（1）信息检索能够充分开发和利用信息资源，避免重复劳动。

科学研究具有连续性和继承性的特点，这就要求科研人员在探索未知或从事研究工作之前，应尽可能地占有与之相关的信息资料，即利用信息检索工具，查阅大量信息资料，然后进行深入地分析、研究和综合，充分了解他人在此探索或研究的问题中已做过的工作，取得的成就，以及发展动向等，这样才能做到心中有数，防止重复劳动，避免或少走弯路，将有限的时间和精力用于创造性的研究中去。

据有关信息报道，科学研究中出现的许多问题，几乎有 95%～99%需要而且可以通过科技文献查找得到启发、帮助和解决，仅有 1%～5%的问题要靠自己的创造性劳动来解决。在这方面曾经有两件很典型的事例。一件是：某国家 50 个企业曾联合进行了一项有关电路设计的研究工作，历时 5 年，耗费 50 万美元而没有结果。后来发现另一个国家已经研究过这方面的问题，并已获得成功，而且有论文发表。另一件是：美国阿波罗宇宙飞船登月计划中，有一项钛合金空舱压力试验。他们用了 20 个钛合金空舱充甲醇做试验，结果因出现穿孔而报废，经济损失高达 150 万美元。事后才知道只要事先查一查美国的《化学文摘》就可以完全避免这次损失。因为早在 10 年前的文献中，就已经发现了解决这个问题的办法，只要在甲醇中加 2%的水就行了。

（2）信息检索能为人们更新知识、实现终生学习提供途径。

在当今社会中，科学技术迅速发展，科技成果大量涌现，知识日新月异，人们需要终生学习，不断更新知识，才能适应社会发展的需求。美国工程教育协会曾估计，学校教育只能赋予人们所需知识的 20%～25%，而 75%～80%的知识是走出学校后，在研究实践和生产实践中根据需要，不断学习而获得的。因此，掌握信息检索的方法与技能，是形成合理知识和更新知识的重要手段，是做到无师自通、不断进取的主要途径。

（3）缩短查找文献时间，提高工作效率。

科研人员为完成科研课题查找所需要的文献资料是需要花相当多的时间的，据调查统计，一般科研人员进行一项研究，其查找资料所需时间约占科研总时间的 40%～50%。所以如果不掌握科学实用的方法，不仅会浪费大量时间，而且最终仍找不全甚至找不到你所需要的信息资料。只有掌握了信息检索方法，才能迅速、准确、全面地找到你所需要的信息资料，减少人力或投资方面的费用。

3.1.3 信息检索原理

3.1.3.1 信息检索方法

信息检索有两种方法。一是直接从信息源和信息文献载体中获取信息，称为直接检索；二是通过利用信息检索工具获取所需的信息，称为间接检索。

直接检索是信息用户惯用的方法。它的优点是可以直接明确地判定所检索到的信息

是否符合需要，且对于非文献载体的信息检索来说比较快速、方便，但是，对于现代大规模的信息检索，却很难广、快、精、准地查到所需的全部信息资料。例如，美国《化学文摘》总编 Bernler 说过这样一段话："假如一个化学家懂得 30 国语言，每小时有读 4 种杂志的速度，一周之内用 40 小时来阅读有关化学专业的论文，要读完当年化学文献，需要 10 年以上的时间。"

间接检索克服了直接检索的缺点。它把信息资料"贴上"分类号或主题词等多种检索标志，并按照这些标志把信息有效地组织起来，在信息检索时利用检索标记可以查到所需信息线索或信息本身。

直接检索和间接检索是人们查找文献的两种基本方法，各有所长。人们进行文献检索时，应该注意把这两者科学地结合起来。

3.1.3.2 信息检索系统

信息检索系统是拥有一定的存储、检索技术装备，存储有经过加工的各类信息，并能为信息用户提供检索服务的工作系统。

信息检索系统包括存贮系统和检索系统两大部分。在信息存贮时，需按照信息的外表特征和内容特征进行标引，形成信息的特征标识。在信息检索时，将要寻找的信息进行检索标引，形成检索提问的特征标识，再与存贮在信息检索工具(或信息库)中的信息的特征标识进行比较。当两者的特征标识一致，或存贮的信息特征标识包含有想要找的信息的特征标识时，就能通过存贮信息的特征标识从存贮信息库的文档中输出所需信息的线索或直接输出所需的数据、资料、图形、图像、文献等。这个过程就是信息检索过程。

由此可见，信息的存贮与检索是信息检索系统不可分割的统一的两个存贮检索分过程。即：首先运用检索语言对采集来的信息进行著录、标引，然后对信息需求用一定的检索语言加以陈述，并据此从信息存贮的数据库中进行检索。见信息检索系统原理示意图 3.1-1：

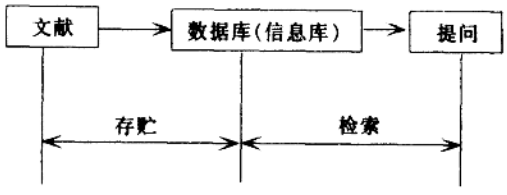


图 3.1-1 信息检索系统原理图

目前，信息检索系统按使用的技术手段可分为手工检索系统和计算机检索系统。

(1) 手工检索系统

手工检索系统又称传统检索系统，是通过人工查找信息的检索系统。其主要类型有各种书本式的目录、题录、文摘和各种参考工具书等。检索人员可与之直接"对话"，具有方便、灵活、判断准确，可随时根据需求修改检索策略，查准率高的特点。但由于全凭手工操作，检索速度受到限制，也不便于实现多元概念的检索。

(2) 计算机检索系统

计算机检索系统又称现代化检索系统，是用计算机技术、电子技术、远程通信技术、光