文献信息检索

主　编　武万军　冉　政　李红香
副主编　陈　娟

上海交通大學出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

目　录

项目一

文献信息检索概述

开篇导读

文献信息作为一种战略性资源，成为发展科技、经济、文化教育的重要支柱之一。信息是人们认识世界，改造世界取之不尽、用之不竭的宝贵资源。认识、获取、利用文献信息已成为每个学生的一项必备的基本素质。为了更好地掌握这项技能，提高文献信息意识和获取信息的能力，掌握文献信息的基本知识是非常必要的。本章讨论了文献信息源的基本概念以及文献检索的基本原理和效果评价，介绍了文献信息检索的基本应用。

情景导入

“信息主义”是一个较晚才出现的用语，但在此之前已有“信息社会”的概念和理论，其中就已经明确表达了信息技术决定论的思想。1962 年日本学者梅棹忠夫在《朝日放送》杂志发表题为《论信息产业》的论文中首先提出“信息社会”的概念；法国的施赖贝尔在 1980 年出版的《世界面临挑战》中也较早地明确提出了“信息社会”的概念。应该说，“信息社会”的提法和相应的理论是由 20 世纪七八十年代美国的一批社会学家和未来学家采用和推广后才产生了重大影响。其中最著名的是贝尔的“工业社会理论”、托夫勒的“第三次浪潮”和奈斯比特的“信息社会”。不同的学者采用了不同的术语（如还有“信息时代”“超工业社会”“知识社会”“网络社会”甚至“比特社会”等），所表达的都是“信息社会”正在到来的意思。

想一想

1. 什么是信息？什么是信息社会？
2. 所有的信息都可以转化为文献吗？为什么？

单元一　文献信息源概述

信息源是人们在科研活动、生产实践和其他一切活动中所产生的成果和各种原始记录，以及对这些成果和原始记录加工整理得到的成品的总称。联合国教科文组织出版的《文献术语》把其定义为：个人为满足其信息需要而获得信息的来源，称为“信息源”。

一、文献信息源的概念

1.文献的载体和媒体

文献信息主要是人类的精神产品，而精神产品又必须依附在某种物质载体上才能保存、流传。但文献信息的载体除了物理的以外，还有逻辑的载体，即用什么符号或文字表达文献信息。这两种载体都是必须同时具有的，并且都能看见。为了与真实的载体区别，人们把字符之类的信息载体称为“媒体”或“媒质”(media)。

(1)信息媒体的种类

信息媒体是信息传播的形式，它们有符号、文字、声音、图像、动画等。情报是信息和知识的活化，总是伴随着人的情报活动而产生。因此，可以把人的情报活动看作情报的媒体。

例如，蜜蜂采集花粉，尽管是为了它们生存的目的，但客观上却起着传播花粉，繁荣植物的作用。蜜蜂在采蜜过程中有“语言”交流，并且有“数学语言”。已经发现蜜蜂用一种“舞蹈语言”作为传递消息的方式，而且“语汇”非常丰富。德国的诺贝尔奖获得者奥特鲁姆(Orum)揭示了蜜蜂“舞蹈语言”的秘密。

花丛(蜜源)在何方向？距离蜂巢有多远？表达这两个问题，在数学上要用解析几何的极坐标知识，而蜜蜂却巧妙地解决了它们。蜜蜂表达这些问题所用的跳舞动作和聋哑人的手语具有相同的性质。

信息可从一种载体或媒体转移到另一种不同的载体或媒体上，其转移成本可以低，也可以高，但我们总想不出有什么信息在什么时刻可以不要载体或媒体。也就是说，信息必然始终附于某种“壳”(shell)，即使是在转移过程中也是如此。

(2)多媒体信息载体的比较

作为多媒体的信息载体，文字、符号、声音、图形、图像和动画在传递信息上有着许多截然不同的特点。

①符号

符号最不直观，或最“抽象”，你可用一个极怪异的符号代表一种只有你自己才知道的含义。但符号一般最简单，占用机器的内存少。

②文字

文字则以其“永久”构成人类文明的历史。其表达信息的能力可以“明察秋毫”到无与伦比的程度。知识和思想全靠文字得以积累和升华，以至于它们可以洞察未来，揭示从感知上得不到的东西。

文字远比符号直观并且表达丰富，又比同属于视觉信息的图像简洁和明确，而且比起与它密不可分的一体化的语音来说，文字传递信息的速度快得多，因为文字毕竟是视觉符号。这些特点使文字永远成为使用得最多的信息介质，成为计算机信息处理的“主力”。其他的媒体可以从多方面补充它、完善它，但不会完全代替它。

思政课堂

万物有时，而文化之树常青。物质的存在终将速朽，唯有文化生生不息。文化是我们自己创造的，文化又反过来塑造和影响着我们每一个人，决定着我们设定的目标能否实现。我们要坚持以人民为中心的文化价值导向，坚持为了人民、依靠人民、共建共享，注重文化熏陶和实践养成，使真、善、美的理念转化为人们的精神追求和行为习惯，不断提高文化参与感、获得感和认同感，形成积极向上、充满正能量的社会风尚。

③声音

声音表达信息的细节最为丰富，使用面最广。例如，同样一句话在不同的时间或地点、用不同的语气或声调，其含义都可能有所不同。声音大致分为语音和音乐两大类，与音乐不同，语音所传递的信息更明晰、明确得多。它还具有其他任何媒体不具有的最便于“携带”和“成本”最低的优点。

④图形和图像

图形化的“语言”给人们带来丰富多彩的感受。图形或图像传递的信息更直观、更快、细节也特别丰富。但由此也带来了信息的确定性差的问题，如你从任何一个图形上可以“一下子”看到很多东西，但它究竟代表哪一种确切的含义，如果没有对应的文字说明，你多半不能立即回答。

⑤动画

在 Flash 动画出现之前，网页动画基本上以 GF 动画为主。Flash 采用了矢量动画的形式，文件量较小，放大时也不会发生锯齿现象，并且和 Real player 格式的影片一样，支持“流媒体”播放形式，即允许用户一边下载，一边播放，因此，Flash 动画可以流畅地在窄带网络上传输。此外，最新的 Flash 动画支持导入多种声音和视频文件，并添加了丰富的多媒体交互表现手段，可以帮助用户创建更加优秀的网络动画作品。

动画的交互性强，比其他媒体更有吸引力。动画作为一种交流手段，有着特有的表现形式和优点，例如，用 10s 或 20s 就可以讲述一个人的一生。

2.文献信息源的概念

《文献情报术语国际标准（草案）》（ISO/DIS5127）认为，“为了把人类知识传播开来和继承下去，人们用文字、图形、符号、声频、视频等手段将其记录下来，或写在纸上，或晒在蓝图上，或摄制在感光片上，或录到唱片上，或存储在磁盘上。这种附着在各种载体上的记录统称为文献。”

简而言之，文献是记录知识或信息的物质载体。其中知识、信息是文献的实质内容、灵魂；物质载体是知识、信息存储、传递的主要工具和外在形式。习惯上，我们把记录科学知识的每一份物质载体称为科学文献，而把科学文献的汇总称为科学文献流。在各类信息源中，文献是最主要、最常用的基本信息源。

3.计算机化的文献信息源

（1）ASCII 码

①概念

ASCII [(American Standard Code for Information Interchange):美国信息交换标准代码]是基于拉丁字母的一套电脑编码系统,主要用于显示现代英语和其他西欧语言。它是最通用的信息交换标准,并等同于国际标准 ISO/IEC 646。ASCII 第一次以规范标准的类型发表是在 1967 年,最后一次更新则是在 1986 年,到目前为止共定义了 128 个字符。

②起源

在计算机中,所有的数据在存储和运算时都要使用二进制数表示(因为计算机用高电平和低电平分别表示 1 和 0),例如,像 a、b、c、d 这样的 52 个字母(包括大写)以及 0、1 等数字还有一些常用的符号(例如,*、#、@等)在计算机中存储时也要使用二进制数来表示,而具体用哪些二进制数字表示哪个符号,当然每个人都可以约定自己的一套(这就叫编码),而大家如果要想互相通信而不造成混乱,那么大家就必须使用相同的编码规则,于是美国有关的标准化组织就出台了 ASCII 编码,统一规定了上述常用符号用哪些二进制数来表示。

美国信息交换标准代码是由美国国家标准学会(American National Standard Institute,ANSI)制定的,是一种标准的单字节字符编码方案,用于基于文本的数据。它最初是美国国家标准,供不同计算机在相互通信时用作共同遵守的西文字符编码标准,后来它被国际标准化组织(International Organization for Standardization,ISO)定为国际标准,称为 ISO 646 标准。适用于所有拉丁文字字母。

③表达方式

ASCII 码使用指定的 7 位或 8 位二进制数组合来表示 128 或 256 种可能的字符。标准 ASCII 码也叫基础 ASCII 码,使用 7 位二进制数(剩下的 1 位二进制为 0)来表示所有的大写和小写字母,数字 0 到 9、标点符号,以及在美式英语中使用的特殊控制字符。其中:

0～31 及 127(共 33 个)是控制字符或通信专用字符(其余为可显示字符),如控制符:LF(换行)、CR(回车)、FF(换页)、DEL(删除)、BS(退格)、BEL(响铃)等;通信专用字符:SOH(文头)、EOT(文尾)、ACK(确认)等;ASCII 值为 8、9、10 和 13 分别转换为退格、制表、换行和回车字符。它们并没有特定的图形显示,但会依照不同的应用程序,而对文本显示有不同的影响。

32～126(共 95 个)是字符(32 是空格),其中 48～57 为 0 到 9 十个阿拉伯数字。

65～90 为 26 个大写英文字母,97～122 号为 26 个小写英文字母,其余为一些标点符号、运算符号等。

后 128 个称为扩展 ASCII 码。许多基于 x86 的系统都支持使用扩展(或“高”)ASCII。扩展 ASCII 码允许将每个字符的第 8 位用于确定附加的 128 个特殊符号字符、外来语字母和图形符号。

(2)汉字内码

汉字机内码,又称“汉字 ASCII 码”,简称“内码”,指计算机内部存储,处理加工和传

输汉字时所用的由 0 和 1 符号组成的代码。输入码被接受后就由汉字操作系统的“输入码转换模块”转换为机内码，与所采用的键盘输入法无关。机内码是汉字最基本的编码，不管是什么汉字系统和汉字输入方法，输入的汉字外码到机器内部都要转换成机内码，才能被存储和进行各种处理。

因为汉字处理系统要保证中西文的兼容，当系统中同时存在 ASCII 码和汉字国标码（又称“汉字交换码”）时，将会产生二义性。例如：有两个字节的内容为 30H 和 21H，它既可表示汉字“啊”的国标码，又可表示西文“0”和“!”的 ASCII 码。为此，大部分汉字系统都采用将国标码每个字节高位置 1 作为汉字机内码。这样既解决了汉字机内码与西文机内码之间的二义性，又使汉字机内码与国标码具有极简单的对应关系。

二、文献信息源的分类

1.按照文献的出版形式分类

依据文献的发行状态和出版形式，可将其划分为图书、连续出版物和特种文献三种。

（1）图书

一般来讲，图书是指内容比较成熟、资料比较系统、有完整定型装帧形式的出版物。科技图书是一种重要的科技文献源，它大多是对已发表的科技成果、生产技术知识和经验的概括论述。图书的范围较广，主要包括：学术专著、参考工具书（指对某个专业范围作广泛系统研究的手册、年鉴、百科全书、辞典、字典等）、教科书、丛书等等。要较全面、系统地了解某一专业领域的知识，参阅图书是行之有效的方法。

（2）连续出版物

顾名思义，连续出版物是指一类能够连续出版的文献信息。通常包括报纸和期刊两大类。对于科学研究而言，使用更多的是期刊类资源。

期刊（Periodicals）也称杂志（Journal 或 Magazine），是指那些定期或不定期出版、有固定名称、公开发行的连续出版物。与图书相比，期刊出版周期短，刊载速度快，数量大，内容新颖、丰富。而且，科技期刊在科技情报来源方面占有重要地位，约占整个科技信息来源的 65%～70%。

（3）特种文献

特种文献是指通过正常出版渠道不能获得的一类文献，故而也称内部资料。这类文献种类繁多，功能各异，各有用途。主要包括专利文献、学位论文、会议论文、科技报告、政府报告和技术标准等。这类资源在没有互联网的年代获取相当困难，但在网络环境下，其获取途径越来越方便，被利用率越来越高，价值也越来越大。

①专利文献

专利文献通常是指发明人或专利权人申请专利时向专利局所呈交的一份详细说明发明的目的、构成及效果的书面技术文件，经专利局审查，公开出版或授权后的文献。广义的专利文献还包括专利公报（摘要）及专利的各种检索工具。

其特点是：数量庞大、报道快、学科领域广阔、内容新颖、可靠性强。目前专利的情报价值越来越大，使用率也日益提高。

②科技报告

科技报告也称研究报告或技术报告，是科技工作者围绕某个课题研究所取得的成果的正式报告，或对某个课题研究过程中各阶段进展情况的实际记录。

其特点是：所报道的成果一般必须经过主管部门组织有关单位审核鉴定，其内容专深、可靠、详尽，而且不受篇幅限制，报道迅速。目前，世界上每年都有相当数量的科技报告产生，尤以美、英、法、德、日等国居多。

③学位论文

学位论文是高等院校和科研院所的本科生、研究生为获得学位资格（博士、硕士和学士）而撰写的学术性较强的研究论文，是在参考大量文献进行科学研究的基础上完成的。其特点是：理论性、系统性较强，内容专一，阐述详细，具有一定的独创性。

④会议文献

会议文献是指各种专业会议上所发表的论文、报告稿、讲演稿等文献。

其主要特点是：传播信息及时、论题集中、内容新颖、专业性强，往往代表某一学科或专业领域内最新学术研究成果，尤其是级别较高的专题性会议，基本上能够反映当时该学科或专业的学术水平、研究动态和发展趋势。

⑤政府出版物

政府出版物是指各国政府部门及其设立的专门机构发表、出版的文件，可分为行政性文件（如法令、方针政策、统计资料等）和科技文献（包括政府所属各部门的科技研究报告、科技成果公布、科普资料及技术政策文件等），其中科技文献占 30%～40%。其主要特点是内容可靠，与其他信息源有一定重复。美国的政府出版物数量最多，每年有几千篇公开，其他国家如英国、加拿大、法国等也出版一定数量的政府出版物。这类出版物通常需要在相应的政府网站上才能获得。

⑥标准文献

标准文献是技术标准、技术规格和技术规则等文献的总称。它们是记录人们在从事科学试验、工程设计、生产建设、商品流通、技术转让和组织管理时共同遵守的技术文件，具有严肃性、法律性、时效性和滞后性的特点。标准文献是准确了解该国社会经济领域各方面技术信息的重要参考文献。

2.按照文献被加工程度分类

信息工作的主要任务是对信息进行深层次开发和综合利用，为了有效地发掘出文献的信息内容，必须对文献进行一定的加工，因此，根据文献的产生次序和加工整理的程度不同，可将文献划分为四个层次结构。

（1）零次文献

零次文献是指未经任何加工、未经公开发表或交流的文献。特点是：客观性、零散性、不成熟性。如实验记录、草稿、私人日记、笔记、书信、草图、原始录音、原始录像、谈话记录等属于零次文献。此类文献多保留于科技人员之手。另外，科技部门、有关管理部门和计划部门也有收藏。零次文献在原始文献的保存、原始数据的核对、原始构思的核定（权利人）等方面有着重要的作用。

(2)一次文献

一次文献习惯上称作原始文献,是以作者本人的研究或研制成果为依据而创作或撰写的文献,不管创作时是否参考或引用了他人的著作,也不管该文献以何种物质形式出现,均属一次文献。大部分期刊上发表的文章和在科技会议上发表的论文均属于一次文献,如期刊论文、科学考察报告、研究报告、会议论文、学位论文、专利说明书等。一次文献具有创新性、原始性和多样性的特点。

(3)二次文献

二次文献又称检索性文献,是图书信息研究机构将大量的分散的无序的一次文献,经浓缩、整理、加工处理后,组织成系统的便于查找和利用的文献。二次文献本身具有自己的系统结构,为了方便利用,一般提供多个检索途径。所以一种好的二次文献往往由几个部分组成,具有比较固定的体系结构。书目、文摘、索引、题录、检索工具书和网上检索引擎等都是典型的二次文献。二次文献具有集中性、工具性和系统性的特点。

(4)三次文献

三次文献又称参考性文献,是在对二次文献和一次文献的内容进行分析研究的基础上通过综合概括而编写出文献,是文献的研究成果和产物。如综述、专题述评、学科年度总结、数据手册、百科全书、辞典等参考工具书属于三次文献。三次文献具有综合性、针对性和科学性的特点。一般可直接提供参考、借鉴和使用,因而普遍为科研人员和管理者所重视。

3. 按照文献的载体类型分类

(1)印刷型文献

这是以纸张为主要载体,以手写、印刷为记录手段的传统文献信息源,包括手写、油印、铅印、胶印、木版印刷等几种形式。印刷型文献信息源的优点是阅读、携带、利用方便,是迄今为止人们最普遍、最乐于接受的一种信息源;缺点是信息存储密度小、体积大、分量重、收藏和管理困难。

(2)缩微型文献

缩微技术起源于英国,是一种涉及多学科、多部门、综合性强且技术成熟的现代化信息处理技术。这类信息源一般以感光材料为载体,利用照相设备和其他缩微设备将印刷型文献源按照一定的缩小比例摄录在胶卷或胶片上,其产品称缩微品或缩微复制品,包括缩微胶卷、缩微胶片(平片)、缩微卡片等几种形式。其特色在于:存储密度大、寿命长、易于还原拷贝和多功能使用、可作为法律凭证。但存在信息衰减、不能直接阅读(需要配备专用的显示还原设备)等缺点。

(3)机读型文献

这是以磁性材料、光学材料或网络为载体的信息源,其特点是在存储时要将相关信息转换成计算机可以识别、理解和处理的二进制代码,输出时需要还原这些代码的原貌,即还原成“人读信息”。其优点是信息存储密度高、存取速度快,可借助于高速信息网络实现远距离传输。

(4)声像型文献

这是运用录音、录像、摄影等技术将声音和图像直接记录在磁性或光学材料上的信息源，如唱片、录音带、录像带、电影拷贝、幻灯片等。其记录的对象主要不是文字，而是富有动感的声音和图像。这类信息源能给人以直观形象的感觉，因而用途广泛。其优点是可以逼真地再现事物和现象，在某些难以用文字描述和反映的场合有着独特的作用。

小贴士

上述四种类型的文献信息源各具特色。印刷型文献信息源是最基本、最广泛采用的信息源；机读型文献信息源所占的比例正在逐年增加，并在某种程度上代表着文献信息源的演变方向；缩微型和声像型文献信息源在一定时期的某些特殊场合内仍发挥着难以替代的作用。

4. 按照发售途径和获取难易程度分类

根据发售途径和获取难易程度的不同，文献信息源可分为三种类型：白色文献、黑色文献和灰色文献。

(1)白色文献

白色文献是通过正式渠道出版发行的文献，具备内容的公开性、发行范围的广泛性等特点。如图书一般经由出版社出版并通过新华书店系统发行，期刊常由杂志社出版并通过各地报刊发行部门向国内外公开发行。其优越性在于信息获取容易，用户可以很方便地通过书店、书摊、邮局购得，因而是一种非常重要的信息源。但因信息的经济价值通常只在非对称的信息环境里显现出来，而白色文献很容易获取，所以在很多情况下其经济价值要大打折扣。但这并不意味着白色文献没有利用价值。事实上，从白色文献中萃取有价值情报的活动屡见不鲜。

(2)黑色文献

黑色文献是指不正式出版、发行范围狭窄、内容保密的文献，如军事情报资料、技术机密资料、个人隐私材料等。这些黑色文献受法律保护，只对负有保密义务的人开放，一般不允许复制。其缺点是保密程度高，非负有保密义务的人无法获取。所以竞争情报活动一般不以黑色文献为信息源。但是，按照各国法律规定，黑色文献迟早会被解密。从实践上看，有不少黑色文献虽然被解密了，但仍然可以作为竞争情报活动的信息源。

(3)灰色文献

灰色文献是介于白色文献和黑色文献之间的一类文献，其名称来自英语的“grey literature”，出现于20世纪70年代。1997年，在卢森堡召开的第三次国际灰色文献会议提出，灰色文献是指不经营利出版者控制，而由各级政府、学术单位、工商业界所产制的各类印刷与电子形式的资料。

灰色文献的特点是不正式出版，也非秘密文献，常见的类型有研究报告、学位论文、会议录、技术规范与标准、企业内部出版物(厂报、厂刊等)、经济函件和商务通信、非官方公布的统计资料以及工商行会、学会、协会、政治和贸易团体的出版物等。另外，一部分黑色文献在解密之后也可以转化为灰色文献。按照《科学引文索引》(SCI)所进行的研究

以及诸如美国国家航空航天局(NASA)、德国卡尔斯鲁厄专业信息中心和意大利高级卫生研究所编辑部等组织所作的估计,灰色文献的比例现在可能已超过文献信息源的20%。

灰色文献所含信息通常是非常珍贵的原始信息,而且往往具备新颖性,因而是竞争情报活动很有价值的信息源。但这类文献复本少,且不公开发行,因而获取较难,一般除在专门的图书馆或信息中心可以查到外,只有与作者本人联系才有可能获取。与白色文献相比,竞争情报人员获取灰色文献通常要付出更大的代价。

单元二　文献信息检索原理

信息检索的本质是信息用户的需求和信息集合的比较与选择,即匹配(Match)的过程。从用户需求出发,对一定的信息集合或信息系统采用一定的技术手段,根据一定的线索与准则指出命中(Locate Hit)的相关信息。

一、信息检索的基本原理

1. 信息检索的基本原理

信息检索的基本原理是:通过对大量的、分散无序的信息进行收集、加工、组织、存储,建立各种各样的检索系统,并通过一定的方法和手段使存储和检索这两个过程所采用的特征标识达成一致,以便有效地获得和利用信息源。其中存储是为了检索,而检索必须先进行存储。

信息存储过程包括:

(1)文献信息采集

即根据一定的原则收集文献。

(2)文献标引、著录

即对文献的信息特征,包括内容特征和形式(外部)特征进行揭示和描述。其中,对文献内容特征的揭示,是按照系统所采用的信息检索语言(分类表、主题词表等)对文献主题进行标引,为文献的内容特征加上标识。另外,根据需要,对文献形式特征中有检索意义的项目,如著者、文献题名等也可以做出标引,用作标识。

(3)建立检索系统或编制检索工具

即按标识引用语的顺序,将著录的大量文献款目有机地组织成一个排检系统,形成有序的、系统化的检索工具或数据库检索系统。

2. 信息检索的前提

信息检索的前提是信息存储有序化。无论是手工检索工具、机械检索工具,还是计算机检索工具都会根据自身系统特性,在一定专业范围内进行信息收集和选择,对收集的信息进行分析、选择、标引、描述以及组织加工转化,形成信息数据库,以便用户检索使用。检索的过程是信息存储的逆过程,即用户对检索课题进行分析,形成检索提问信息,

选取合适的检索用语，利用检索工具或检索系统查出相关信息。

简单地说，检索就是查找，查找的过程实际上是一个逻辑匹配的过程，即确定检索用语并将检索用语与标引用语作相符性比较，检索用语与标引用语一致，即找到了符合要求的文献信息。

二、信息检索的效果评价

1.检索效果的评价指标体系

检索效果指的是利用检索系统或工具开展检索服务时的有效性。它直接反映着检索系统的性能，影响系统在信息市场上的竞争能力和用户的利益。根据 F. W. Lancaster 的阐述，判定一个检索系统的优劣，主要从质量、费用和时间三个方面来衡量。

其中，质量标准主要通过查全率与查准率进行评价；费用标准即检索费用是指用户为检索课题所投入的费用；时间标准是指花费时间，包括检索准备时间、检索过程时间、获取文献时间等。查全率和查准率是判定检索效果的主要标准，而后两者相对来说要次要些。

(1)查全率

查全率是指系统在进行某一检索时，检出的相关文献量与系统文献库中相关文献总量的比率，它反映该系统文献库中实有的相关文献量在多大程度上被检索出来。其数学表达式如下：

查全率＝(检出相关文献量/文献库内相关文献总量)×100％

(2)查准率

查准率是指系统在进行某一检索时，检出的相关文献量与检出文献总量的比率，它反映每次从该系统文献库中实际检出的全部文献中有多少是相关的。其数学表达式如下：

查准率＝(检出相关文献量/检出文献总量)×100％

查全率和查准率与文献的存储和信息检索两个方面是直接相关的，也就是说，与系统的收录范围、索引语言、标引工作和检索工作等有着非常密切的关系。

一般来说，查全率和查准率之间存在互逆关系，即当某一系统的查全率和查准率处于最佳比例关系时，继续提高查全率，检出的相关文献量会增加，但同时由于检出文献中不相关文献的增加会导致查准率降低；继续提高查准率，由于增加检索限定，就会造成查全率降低。

2.影响查全率和查准率的主要因素

(1)从存储角度看

从存储角度看，影响查全率的因素主要有：文献库收录文献不全；索引词汇缺乏控制和专指性；词表结构不完整；词间关系模糊或不正确；标引不详；标引前后不一致；标引人员遗漏了原文的重要概念或用词不当等。

(2)从检索角度来看

从检索角度来看，影响查全率的因素主要有：检索策略过于简单；选词和进行逻辑组

配不当，使用逻辑 AND 太多，或不适当地使用了 NOT；检索途径和方法太少；检索人员业务不熟练和缺乏耐心；检索系统不具备截词功能和反馈功能，检索时不能全面地描述检索要求等。

(3)从存储角度来看

从存储角度来看，影响查准率的因素主要有：索引词缺乏专指性，不能准确描述文献主题和检索要求；组配规则不严密；选词及词间关系不正确；标引过于详尽；组配错误。

(4)从检索角度看

从检索角度看，影响查全率的因素主要有：检索时所用检索词(或检索式)专指度不够，检索面宽于检索要求；检索系统不具备逻辑 NOT 功能和反馈功能；检索式中允许容纳的词数量有限；截词部位不当；检索式中使用逻辑 OR 不当等。

实际上，影响检索效果的因素是非常复杂的。根据国外有关专家所做的实验表明，查全率与查准率是成反比关系的。要想做到查全，势必要对检索范围和限制逐步放宽，则结果是会把很多不相关的文献也带进来，影响了查准率。企图使查全率和查准率都同时提高，不是很容易的。强调一方面，忽视另一方面，也是不妥当的。应当根据具体课题的要求，合理调节查全率和查准率，保证检索效果。

3.用户对检索效果的评价

从用户的角度考虑，可以从检索到文献的相关性、适用性及新颖性三个方面判断检索效果是否令人满意。

(1)相关性

相关性即用户判断检索到的文献信息与实际信息需求之间的关联度。现实的信息系统是回答用检索式表达后的信息提问。虽然检出的是与信息提问相关的信息，但不一定是真正切题的信息，用户只有在阅读后才能对其切题性做出判断。

(2)适用性

适用性即检索到的文献对用户的实际需要的满足程度，或能够给用户带来的效果和产生的效益。

(3)新颖性

新颖性即对用户而言，从检索系统中检出来的、含有新颖信息的文献占文档中总相关文献数的比例。

思政课堂

锐意进取、改革创新，对新生事物秉持开放、理性、包容的态度，是当今的社会共识。对于符合事物发展规律、具有强大生命力和远大前途的新生事物，人们总是抱有更高期待。时代的荣耀属于创新者。大学生要放开“思维缰绳”，打破思维定势，以宽广的眼光看待新生事物，以宽容的态度对待新生事物，以进取的精神培育新生事物，那些“才露尖尖角”的“小荷”就能得到滋养、向阳生长，终成“接天莲叶无穷碧”的壮美景象。

4.影响用户检索满意度的原因

由于用户的因素导致检索不满意或者检索效果不佳的主要原因有以下几点。

(1)信息检索思维方法的缺陷

①在自身信息需求的分析上,从始至终都是一个需求,不懂得通过变换检索词、采用相关词或同义词来提出更多需求。

②在具体检索时,以短句为单位,不能恰当切分关键词,找出关键词之间的关系。

(2)对检索工具的选择缺陷

①在检索工具的选择上更倾向于搜索引擎,网络上的资源与专业网络数据库相比,无论是质量还是数量上都有很大的差距,仅用搜索引擎来查找文献是远远不够的。

②忽略文摘数据库,过度依赖全文数据库。虽然检索文摘数据库不能马上得到全文,但是文摘数据库的数据量大、范围广,是查找文献线索极好的检索工具。且文摘数据库收录的文献是经过筛选的,质量更有保证。

(3)信息检索操作方法的缺陷

①在使用具体检索平台时,只使用初级检索功能,而没有发现其高级检索功能和限制检索选项。

②在检索结果的利用上,不善于利用检索系统对检索结果的分析、分类功能对检索策略进行优化调整。

③没有对检索词进行限制,包括字段限制、时间限制及分类限制等;提炼概念不够具体或概念具有多义性导致误检;对所选的检索词截词过短。

小贴士

信息检索依据其检索对象的不同,可分为三种类型。

第一种是文献检索,即所检索到的是关于文献的信息或文献全文,它所回答的是诸如“关于铁路大桥有哪些文献?”之类的问题。

第二种是事项或事实检索,它所回答的是诸如“世界上最长的铁路大桥是哪一条?”之类的问题。

第三种是数据或数值检索,它所回答的是诸如“世界上最长的铁路大桥有多长?”或“世界上有多少条铁路大桥?”之类的问题。

第二、三种检索所得到的是能够确切解答问题的信息,或者说是文献中的具体信息。

单元三　文献信息检索应用

随着信息技术的发展,互联网的应用得到广泛普及,信息环境发生了相当大的变化,信息社会给人们带来了海量的信息,供人们参考、借鉴和学习。如何快速地从海量的信息中获取最有价值的东西,成了人们最棘手的问题。因此,只有掌握好信息检索的理论

知识和技能，提高信息检索能力，才能快速、合理地利用信息资源。

一、文献信息检索工具

检索工具是指人们用来存储、报道和查找信息的工具，具体地说，就是汇集各种信息并按照特定方法编排，以供查考的工具或系统。作为检索工具，它具有存储和检索两方面的基本功能。存储功能是指检索工具把汇集的有关信息按照其特征记录下来，使之成为条条有序化的信息线索，这就是所谓的信息存储过程。检索功能是指检索工具提供一定的检索入口，让人们能够按照一定的检索方法来查找所需信息，也就是信息的检索。

按照加工处理信息的手段不同，可以将信息检索工具分为印刷型检索工具和计算机检索工具。

1.印刷型检索工具

印刷型检索工具，是计算机检索系统普及之前应用最为广泛的一类检索工具，主要包括参考工具书和检索工具书两种类型。

(1)参考工具书

参考工具书是能为读者提供所需具体资料的工具书，供人们查阅数据、结论、定义、公式、分子式、人物简介等数据或事实信息。一般包括百科全书、年鉴、手册、类书、年表、历表、图录、字典、辞典等，属于三次文献范畴。

(2)检索工具书

检索工具书是在一次文献的基础上，整理、编制出的提供文献信息线索的二次文献。一般包括目录、题录、索引、文摘、文献指南等，主要用于查找国内外书刊资料的线索，借此获得索取原文的途径。

随着计算机的广泛应用，现在人们在出版印刷版参考工具书和检索工具书的同时，也会同步发行其数字版，同时，相关的出版机构或情报部门还会对一些印刷版检索工具做数字化回溯工作。检索工具书的数字化将给用户的文献信息查找带来极大的便利，如著名的二次文献《科学引文索引》、美国《化学文摘》等电子版或网络版，因其更能满足快捷检索的需要而成为人们利用的首选检索工具。但是，印刷版的字典、词典、年鉴、百科全书、手册等，因更符合人们长期形成的阅读习惯，他们依然是人们查找信息乐于使用的重要信息源。典型检索工具简介如下：

①目录

它也叫书目，是揭示和报道文献外部特征的检索工具，是有序的文献目录如国家书目、再版目录、期刊目录、联合目录和专题书目等。

②题录

它是以单篇文献作为报道单位，揭示文献外部特征的检索工具。它对信息的报道深度比目录大，检索功能比目录更强。其著录项目通常有篇名、作者、原文出处等。

③文摘

它是除描述文献外部特征之外，还用100～300字简练的语言揭示文献的主要内容，是向读者报道最新研究成果的一种检索工具。它是检索工具的主体，是二次文献的

核心。

④索引

索引是指将特定范围内的某些重要文献中的有关各种事物的名称，如书名、人名、地名、篇名、术语等有价值的知识单元，分别摘录并注明页码，为读者提供文献线索的检索工具。有的检索工具本身全由索引组成，如《科学引文索引》。

⑤辞典

它是字典、词典的总称。如现代《新华字典》、东汉许慎编的《说文解字》等。

⑥年鉴

它是以全面、系统、准确地记述上年度事物运动、发展状况为主要内容的资料性工具书。年鉴汇辑一年内的重要时事、文献和统计资料，按年度连续出版，它博采众长，具有资料权威、反映及时、连续出版、功能齐全的特点，属信息密集型工具书。

⑦百科全书

百科全书是记载人类一切门类或某一门类的知识，以词典形式编纂的系统完备的检索工具。目前世界上的大型百科全书一般超过 30 卷，也有超过 100 卷的超大型百科全书。比较著名的百科全书有《大不列颠百科全书》《美国百科全书》《中国大百科全书》等。除以上几种外，常见的工具书还有手册、名录、表谱、图谱、类书、政书等，在此不做详细介绍。

2.计算机检索工具

计算机检索工具又称计算机检索系统，是利用电子技术、计算机技术、网络通信技术等构成的用于存储和查找信息的检索系统。包括各类文献信息数据库、联机数据库、网络搜索引擎以及各类网站分类目录等。计算机检索系统具有更新速度快、检索途径多、检索效率高、检索结果输出灵活等特点。

文献信息数据库是计算机检索系统的核心，主要有事实或数值数据库、引文数据库文摘数据库和全文数据库等。文献信息数据库通常存储着若干文档(File)，文档是数据库中部分记录的有序集合，通常依据数据库所属的学科领域、收录范围和时间范围给文档归类，因而文档又被称为子数据库，每个文档包含有若干条记录(record)。记录是文献信息数据库的信息单元，记录用于描述原始信息的主要特征，每一条文献记录通常由包含文献的题名、著者、出处等特征的字段组成。

二、文献信息检索语言

语言是人类用来传递和交流思维以及信息的表达方式。同样，信息检索语言也是用于信息交流和转换时的表述，它在信息检索过程中以规则的标识符号来反映信息内容及特征，也就是按照特定语言将信息表述为特定的标识符号，并将这些符号进行整理、序化、存储于检索系统中。另外，用户查找信息时，根据所要查找的问题内容及特征，通过规范的检索语言从检索系统中获取需要的信息。因此，检索语言实际上就是人与信息系统、信息专业人员与用户之间交流的一种语言。

检索语言主要包括描述信息外部特征的语言(如名称语言、序号语言、引文语言等)

和描述信息内容特征的语言(如分类语言、主题语言等)。

1.名称语言

以人名(著者、编者、译者等)、地名、书名、刊名、篇名等代表信息外在特征的名称为检索标识来标引信息,形成作者索引、题名索引等,检索时可直接通过相应的作者途径、题名途径等来查找信息,这种语言标引直观,使用简单。

2.序号语言

以文献独有的顺序号,如标准号、专利号、化学分子式等为标识符号,标引文献信息作为检索的途径。

3.引文语言

利用文献间的相互引证关系对文献进行标引和检索。例如,现有 A 文献引用了 B 文献,则或以 A 为标引词描述 B,或以 B 为标引词描述 A,即为引文语言。引文语言能够表现出文献之间的层级关系,尤其是在网络信息时代,它以超链接的技术和形式广泛应用于数据库等网络信息检索系统中。

4.主题语言

主题语言使用自然语言的词语作为检索标识来揭示主题内容的检索语言。主题语言可分为标题词语言、单元词语言、关键词语言和叙词语言,使用时可参照有关的主题词表。

(1)标题词语言

标题词是经过规范处理的名词术语,是一种先组式主题词,这些标题词以固定的组合方式组织在主题表中,形成标题,检索时按既定组配执行。标题法比较直观、查找速度快,但查全一门学科或具体某一属性事物的文献比较困难。

(2)单元词语言

单元词又称元词,它是经过规范的、最小的、最基本的词汇单位,能表达一个独立的概念。单元词是一种后组式的检索语言,在检索的时候才组配起来,每个元词都可作为检索入口,根据需求增加或减少元词的数量,来改变检索范围。

(3)关键词语言

关键词语言直接从文献的题名、摘要、正文中选取不经人工规范化处理,直接抽自文献的标题、文摘或全文当中的具有实质性独立检索意义的专业名词术语。

关键词选用除冠词、介词、副词和连词之外的实词,关键词因为使用自然语言,使用简单,处理方便。检索效果与关键词的选取有关,这就取决于关键词与文献主题的偏离度,如果偏离度大,则会导致误检、漏检等情况,影响检索质量。

(4)叙词语言

叙词语言结合了标题语言、单元词语言、关键词语言的优缺点,选用能够概括文献主题内容的基本概念并经过严格规范的名词或术语,它也属于后组式检索语言,使用时参照相关叙词表,可按需要灵活组配,形成多种检索标识,并在词与词之间建立起参照系统,具有良好的检索功能。

5.分类语言

分类语言是用分类号和相应的分类名称表达信息内容，并按学科体系的逻辑关系划分和组织的检索语言，它能反映信息的学科系统性以及相关、从属、派生等关系，便于按学科门类进行族性检索。

分类语言受到严格规范，专业性强，用它进行标引和检索时要依靠相关的分类法则。

三、文献信息检索程序

文献信息检索全过程，大致可分为：分析研究课题，明确检索目的；制定检索策略；查找文献线索和索取原始文献 4 个步骤。

1.分析研究课题，明确检索目的

(1)明确检索目的

不同的目的决定着对检索结果的不同要求。因此，用户在开始检索之前，必须明确自己的检索目的，是需要该课题系统而详尽的信息，还是需要该课题最新的信息，或者只是需要该课题的片段信息解决一些具体问题。

(2)明确课题的主要内容

找出课题需要解决的关键，从而归纳出几个既能代表信息需求又具有检索意义的主题概念，包括哪些是主要概念，哪些是次要概念，概念之间的相互关系等。

(3)明确课题涉及的学科范围

搞清楚课题所涉及的学科领域，是否属于多学科或交叉学科，以便按学科选择检索工具(系统)。

2.制定检索策略

(1)选择相关检索工具(系统)

明确了课题的检索范围和要求后，就要据此来选择检索工具(系统)。首先应从课题的需求出发，考虑该课题所属学科及其相关学科领域内有哪些检索工具(系统)，然后根据检索工具(系统)的质量、性质、检索人员以往的经验等，初步选定符合要求的检索工具(系统)。检索工具(系统)的选择还需考虑文献类型、时间范围。

(2)确定检索途径、检索方法

通过对检索要求的分析并同时考虑所选用的检索工具(系统)的特点，才能较合理地确定所要查找课题的检索途径、检索标志。

采用哪种检索途径主要受检索工具(系统)所能提供的检索途径限制。有的检索工具(系统)本身检索途径就有限，无法进行选择。如果某一检索工具(系统)检索途径比较齐全，有多种检索途径可供选择，则要从课题检索要求出发。如果课题检索要求资料范围较广即泛指性较强的课题，则最好选用分类途径。

为了迅速、准确地查找文献，还必须根据用户的检索目的，以及有关主题的学科发展状况，选择合适的检索方法。

(3)制定检索式

只有在计算机检索中才需要编制检索式。检索式就是采用计算机信息检索系统规

定的算符(如布尔逻辑算符、位置算符、截词符、限制符等),将检索词进行组配,确定检索词之间的关系,准确地表达课题需求内容的式子。

3.查找文献线索

根据确定的检索途径,查找某种索引或把检索式输入检索工具中自动进行查找,如果发现检索出的文献不符合检索课题的要求,可以对检索策略及时调整。

(1)对检索数量比较少的结果,可以扩大检索范围,如可使用检索词的上位词、相关词来进行检索;增加或修改检索途径;去掉一些次要主题词;合理使用计算机算符,如截词符等。

(2)对检索数量过多的检索结果,进一步缩小检索范围,如使用下位词来限定检索;增加“AND”组配算符等。

4.索取原始文献

有一些检索工具(系统)是直接提供原文的,而有些检索工具(系统)只是提供文献信息的线索,对这一部分文献我们可以通过馆藏目录,馆际互借等渠道获取。

四、文献信息检索方法

进行文献信息检索时,应遵循一定的查检方法,采取一种快速、准确、有效、省时间的检索方式去查检所需信息。常用的检索方法有以下几种。

1.常规法

常规法是以主题、分类、著者等为检索点,通过检索工具获得文献信息的一种方法。这是一种常规的科学检索方式。使用这种方法首先要明确检索目的和检索范围,熟悉主要的检索工具的编排体例和作用。根据不同的检索要求,常规法又分为三种,即顺查法、逆查法、抽查法。

(1)顺查法

顺查法是根据检索课题的时间范围,按从远及近的时间顺序查找文献的方法。这种方法比较全面、系统,查全率、查准率高,但检索的工作量大,费时、费力。适用于重大课题和各学科发展史以及新兴学科等方面的研究课题的全面检索。

(2)逆查法

逆查法是按照课题的时间范围,利用一定的检索工具,逆时间顺序由近及远地回溯查找文献信息的一种检索方法,目的是获取近期的最新文献信息。这种方法多用于一些新课题、新观点、新理论、新技术的检索。此方法省时,查得的信息有较高的新颖性,但查全率不高,漏查的可能性比较大。

(3)抽查法

抽查法是针对某一学科领域内的课题,重点对某一时间段内文献信息进行检索,以便获得这一客体在特定时间范围内的大量信息。这种方法多用于写专题调查报告。

2.追溯法

追溯法也称引文法,这是一种跟踪查找的方式。根据文献著者在文献末尾所附的“参考文献”的指引,追查到那些参考文献的原文,若有必要,还可以根据以追查到的原文

后面的参考文献再继续追查下去。像《科学引文索引》就是基于文献间的引证关系而编制的。在检索工具不全的情况下，可采用此法查到一批相关的文献，但查全率往往不高。

3.循环法

循环法也称交替法、综合法，即交替使用“追溯法”和“常规法”来进行文献信息检索的综合查检方法。利用检索工具从分类、主题、著者、题名等入手，查找到一批文献信息，再利用文献信息后面所附的参考文献追溯查找，不断扩大检索线索，这样分段交替进行，循环下去，直到满足检索要求为止。运用此法，不会造成漏查，检索的效率高。

知识拓展

知识、文献与信息

信息包含了知识，知识是信息被认识的部分。知识可以分为主观知识和客观知识。信息经过人脑接收、选择、处理而形成并存在于人脑中的知识称为主观知识。主观知识借助语言符号，通过各种物质载体记录下来，就会变成可以传递的客观知识，即文献。

信息、知识之间的逻辑关系是包含与被包含的关系。知识是信息的一部分，文献是信息、知识的具体体现，它不仅是信息、知识的主要物质形式，也是读者吸收利用信息、知识的主要途径。

进入 21 世纪以来，科学技术发展迅速，人类社会的信息化、网络化进程也大大加快，各类信息数量剧增，随之而来的是新学科的不断出现和学科之间相互交叉与渗透的加快，使各专业信息发布分散而无规律。主要有以下几个方面的发展遇到了瓶颈。

1.非文本信息发展滞后

非文本信息（图像、音频、视频等多媒体信息）的检索技术、数字化技术、高密度存储技术为非文本信息提供了广阔的发展空间，多媒体信息已逐渐成为网络的主流。信息检索技术正在从传统的纯文本检索向超文本支持的非线性多媒体检索技术发展，然而图像、音频、视频的检索技术却处于萌芽阶段，需要高新技术支持并不断创新。

2.搜索引擎缺陷

分类目录搜索引擎采用人工干预技术，信息分类不规范，没有一个统一的控制词表和参照标准，分类目录差别较大；搜索范围较小，数据库更新慢，查询交叉类目时容易遗漏；如果用户检索请求没有对应的分类目录，则无法进行查找；信息遗漏不可避免，查全率低。

3.知识和技能匮乏

知识检索是一种全新的信息检索方式，是把用户请求与索引库匹配，寻找与请求关联的网页并返回排序的命中信息的过程。运用截词、词位限定、布尔逻辑运算等技术可以控制用户请求与数据库匹配的精度，但是信息检索难以避免丢失相关信息或产生大量冗余信息，即出现信息漏检与误检。信息检索效率是衡量信息检索效果的重要指标，是

检验信息检索技术成熟与否的标准。知识是信息加工与序化的产物，是高浓度的有序化的信息；知识检索必然是高层次的信息检索。

课后思考

1. 信息检索的基本原理是什么？
2. 什么是“元搜索引擎”？
3. 评价信息检索效果的指标有哪些？
4. 什么是文献信息源？
5. 信息媒体的种类有哪些？