

第 1 章 绪 论

学习目标

- 了解信息的概念和特征。
- 了解信息源的特征。
- 掌握信息源的类型。
- 了解学习信息检索的目的和意义。



1.1 信 息

当今,社会信息化已形成巨潮席卷整个人类社会。人类各种实践活动、各个领域都与信息和信息技术息息相关、密不可分,使得整个社会的生产方式、生活方式、活动方式、行为方式、思想观念、道德规范等方面都在发生根本变化。信息促使人类大脑在时空领域得到极大扩展,从而加速了世界经济的区域集团化、一体化的进程。

信息作为一种社会资源很早就被人们所认识,而且自古以来人类就在利用这种资源,只是利用的能力和水平很低而已。例如,我国古代发明的原始感测和显示技术的指南针、原始通信技术的烽火台,以及记录、存储信息的造纸术和印刷术等。

1.1.1 信息的定义和特征

1) 信息的定义

当今世界,“信息”已成为社会生活中使用频率最高的词汇,与之相连的“信息经济”“信

息产业”也成为人们身边的现实。信息作为重要的经济、战略资源和推动社会生产力的新动力,日益受到人们的青睐,“信息时代”“信息社会”正在向我们走来,信息化成为当代社会发展的大趋势。

“信息”一词被高频率地使用,但究竟什么是信息呢?怎样认识信息的科学意义呢?我们知道世界万物都有自己的运动规律和表现特征。社会生活中人们了解到的事物间的差异都是通过一定的形式表现出来的。例如,声音、图像、文字等,这些形式所表达的内容就是信息。信息不是物质本身,而是物质的一种基本属性。

拉丁语中的“信息”一词是陈述、描述、概要的意思;英语中的“信息”是 information,有知识、资料、数据、情报、消息、报道、报告、通知等多种含义;汉语中的“信息”一词,“信”和“息”两字的意思相近,均有音信、消息的意思,但前者侧重于消息、征兆,后者强调情况、音信。

那么,信息到底是什么呢?目前,学术界还没有形成一个一致的、完整系统的定义。《辞海》中是这样定义的:信息是指客观的消息、情况、情报等,是指对消息接受者来说预先不知道的报道等。《新华字典》将信息表述为:信息是指音信、消息。信息论中是指用符号传输的报道,报道的内容是接受者预先不知道的,是事物的运动状态和关于事物运动状态的陈述。《韦伯斯特词典》将信息表述为:信息是知识和情报的通信和接受,是通过调查、研究或要求而得到的知识,是情报、消息、新闻、事实和数字等。我国科学家钱学森也对信息有过这样的阐述:信息是激活了、激化了的的知识,是为了解决一个特定的问题所需要的知识。

综上所述,信息就是消息,人们在学习、工作、日常生活中随时随地都在接受和利用信息;信息是资源,它具有使用价值和价值;信息是财富,并且是无价之财富;信息是生产力要素,更是一种不可估量的促进生产力发展的新动力。虽然学术界难以形成统一的信息定义,但从文献信息探索的角度,我们应对信息做一个基本的约定:信息是自然界和人类社会中一切事物自身的运动状态与方式,是事物的内在联系和含义的表征,信息通过一定的载体形式反映事物的本质特征及其运动变化。也就是说,信息是一种可以用来消除不定性的东西,它泛指一切事物运动的状态和运动的方式,是关于事物运动状态和运动方式的反映。正因为信息是事物运动状态和运动方式的反映,所以它才能用来消除人们认识上相应的不定性,不断提高人类认识世界和改造世界的能力。因此,人类社会的发展、科学技术的进步,都离不开信息资源的开发和利用,而且信息资源的真实状况及开发利用程度,已成为衡量一个国家经济、文化、科技,乃至综合国力的重要指标。

2) 信息的特征

信息与物质、能量并称为人类社会的三大基本资源,但是与物质和能量不同,信息具有以下特征:

(1) 信息具有客观真实性。信息既不是物质,也不是能量,信息就是信息。不论你是否认识到,信息总是存在的。信息的这一最重要的本质特征是其生命所在。

(2) 信息具有共享性。信息的共享性是它与其他资源最大的不同之处。物质和能量被专人占有和使用,进而被消耗而不复存在。信息显然不同,当某一信息被多人共享时,信息的占有者不会丧失该信息,而且,一般来说也不会因为多次使用而改变信息的自身价值。作为一种共享性的资源,信息可以通过传输和扩散的方式实现共享,在交换、传递的过程中,信息不但不会丧失,而且还可能同时获取新的信息。因此,信息还可称为一种无限再生的资源。另外,我们在实际生活中应当注意区分信息的共享性与现实共享范围,了解独占信息、

有价值信息和公共信息的特性,合理合法地充分享有信息。

(3) 信息具有可存储性。信息的可存储性指的是信息必须依附于各种形式的载体,信息和物质(事物)、能量具有密切的关系,但它仍需依附于另一种形式的载体。所以说,没有物质、没有能量,就没有事物的状态和方式,也就没有信息,而且信息的识别、转换、扩充、压缩、生成、分享等所有的过程都必须有物质和能量的支持。信息通过物质载体来存储,从远古的结绳记事、甲骨、泥板、纸张到磁、激光、大规模集成电路的广泛应用,大大丰富了信息的存储形式,也极大地增加了物质载体存储的信息量。

(4) 信息具有可转换性。信息可以被加工、整理、概括、归纳使之浓缩精练和不断扩充并变换各种形态,由原始一次信息变为二次信息、三次信息等,实现信息价值的增值。运用现代信息处理系统,将原始文件进行压缩,可达到便于传输和节省存储空间的目的。信息技术的高速发展可使信息进行声音、图形、文本、数据等多种形式的相互转换,有效地促进信息的传播和交流。例如,可将声音转换为文字或将文字转换为声音、将统计表转换为拼图等。

(5) 信息具有传输性。信息不同于物质,作为一种既无大小又无重量的非实体,它的传递速度仅取决于其承载体的性能。因此,计算机技术和光纤技术的高速发展使信息的传递超出了以往载体的限制。当信息的载体是电磁波和光波时,其传输的速度将达到每秒30万千米,而运用广播、电视、Internet等大众传媒传递信息,可在瞬间将信息迅速、广泛、全方位地扩散。

(6) 信息具有极强的时效性。信息与物质、能量的价值观不同,信息有很强的时效性,过时的信息就是马后炮。例如,天气预报信息、军情与警情信息、战略预测信息等,如果不在最需要的时间内被准确地获取、传递、使用,它将分文不值。也就是说,信息的利用价值与时间成反比,及时有效地把握信息就能实现信息的最佳价值。

(7) 信息具有增值性。信息通过人脑思维或人工技术的综合、加工和处理,不断积累丰富,提高其质量和利用价值。信息交换的结果是信息的增值。

其实,信息的特征还有许多,如知识性、识别性、生成性、动态性等,这里限于篇幅,不再一一介绍。

1.1.2 信息的类型

信息类型繁多,从不同角度划分会有不同的信息类型。

(1) 自然信息和社会信息。这是依据信息存在的领域及信息的内容划分的。自然信息是指存在于人们主观意识之外,来源于自然界,在自然界中传递的信息。社会信息则来源于人类社会,是人类对于社会现象和社会活动的认识、反映,是在人际间传播的信息及思维信息。

(2) 语言信息和非语言信息。这是依据信息的符号种类及信息的性质划分的。语言信息是由作为信息沟通最基本、最重要的工具——语言来表现的信息,它由语法信息、语文信息和语用信息构成。人们总是以信息为媒介与事物打交道,通过语法信息获得语文、语用信息,经分析比较产生决策信息,根据决策信息使事物按人们预期的目标发展。非语言信息则是指表情、体态等非语言显现的信息。

(3) 动态信息和静态信息。这是依据信息的状态划分的。动态信息是指时效性极强的瞬息多变的的信息,如军情、警情、新闻等类信息。静态信息则是指呈现稳定、相对静止状态的

信息,如历史文献、档案资料等类信息。

(4) 一次信息、二次信息、三次信息。这是依据信息的加工处理程度划分的。一次信息是在生产实践、社会实践和科学实验中产生的新成果、新知识和经验性总结等方面的信息;二次信息是依据客观需要,按照一定的科学方法,将特定分散的一次信息进行加工整理而形成的信息;三次信息是指依据二次信息提供的线索并选用一次信息的内容进行综合分析形成的信息。

(5) 单媒体、多媒体信息和文献及电子信息。这是依据信息的媒体和信息的载体划分的。用来表示信息的声音、图像、文字、数据,以及广播、电视等物质载体是信息的表示媒体。只有一种媒体表示的信息是单媒体信息,运用多种媒体综合、协调地表示的信息是多媒体信息。多媒体信息不同于多媒体技术和多媒体计算机系统的概念,而现代化的多媒体手段可以将多媒体信息带入一个崭新的应用领域。

(6) 政治信息、经济信息、文化信息、科技信息、学术信息、法律信息、军事信息、市场信息等,这是依据信息的社会属性划分的。

1. 1. 3 信息、知识、情报与数据

1) 信息与知识的关系

信息不同于知识,它是作为物质运动的反映,通过各种信息通道达到人脑成为人类意识。在人们获取、整理、利用信息之前,信息处在一种发散的状态,时刻在变化、流动、产生与消失。如果人们不断地接收信息,而且这些信息又是真实可靠的,经过人们的分析、提炼,就能准确把握客观事物的规律性,再归纳总结,这种“信息”就上升为知识、科学了。

知识是人类积累的关于自然和社会的认识和经验的总结,是人类通过收集和利用信息对自然界、人类社会及思维规律的认识与掌握,是人的大脑通过思维重新组合的系统化信息的集合。知识既包括有文字记载的认识和经验,也包括没有文字记载的认识和经验。知识是吸收了信息后凝结出来的概念,是系统化、完备化的信息,是信息中被人类总结出来的精华部分。

信息与知识往往是联系在一起,信息可以被认为是一种可传播形式的知识,信息与知识的关系如同金钱与财富的关系。只有当信息被系统地、有目的地积累起来时,它才能转变为知识。

人类既要通过信息来认识世界、改造世界,也要根据所获得的信息去组成知识。所以说,知识是信息的一部分,信息则是“流动的”知识,它是构成知识的原料,这些原料经过人脑接受、选择、处理,就能组合成新的知识。信息相对变动,具有动态属性,而知识相对稳定,具有静态属性。

2) 信息与情报的关系

信息与情报是两个最为密切的概念,以往两者的使用没有任何区别,如英文 information 表示的就是“信息”和“情报”。随着信息科学的发展,“信息”一词才逐渐取代了“情报”一词。

情报一般是指在特定的时间、特定的状态下,为特定的人提供的有用知识,情报的本质就是知识。《辞海》里是这样解释的:“战时关于敌情之报告,曰情报。”情报往往与军事、间谍相关联,带有机密的色彩,随着社会的发展,情报已不仅仅局限于军事领域,还适用于政治、

经济、军事、社会生活等诸多领域。与信息知的关系一样，信息和情报两者也是既有联系，又有区别。信息包含了情报，因为信息既包括有关军事和间谍相关的情报，又包括更多方面的“情报”“知识”，既有社会生活中人与人之间的相互沟通、思想交流、情感传递，也有人与自然的“对话”与交流，还有纯客观自然界的自然物，非自然物，动物，植物相互的信号、消息交换与传递。综上所述，信息、知识与情报的逻辑关系如图 1-1 所示。

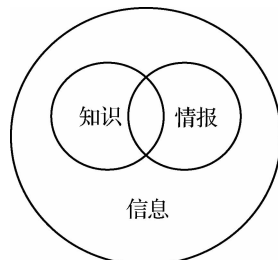


图 1-1 信息、知识与情报的逻辑关系示意图

3) 信息与数据的关系

在计算机普及的条件下，信息与数据的概念常常是混用的，实际操作中信息本身也已数据化，而数据本身也确实代表着信息。数据指的是表达事物信息的字符集合，它由特定的符号或自然数表示的数值型数据和计算机能够识别的在计算机环境下的非数值型数据构成。凡是能够用计算机进行编码和通信的数据都可以称为信息，只有被定义了的数据才是信息。信息就是加工了的数据，是数据的内容。

信息与数据通常是难以分割的，但使用中亦须加以区分。比如“2008”这样一个数据，作为一般的单纯自然数并不表示什么信息，但把这一数据放在“奥林匹克运动”上，它就可能表示“北京 2008 年奥运会”这样一则信息了。如果它出现在飞机的速度表上，表示的肯定就是飞机的速度信息了。因此，数据是一种形式，是一种表示信息的形式，它所表示的内容就是信息。但信息并不仅限于数据表示，它还可以用图形、图像等形式表示。我们还可以说数据作为一种素材、一种原始资料被有序地组织和“创作”而得到的“产品”就是信息。

1.2 信 息 源

1.2.1 信息源及其特征

1) 信息源的含义

所谓信息源，就是人们为满足信息需要而借以获得信息的来源。信息源既可以是产生和持有信息的个人和机构，也可以是负荷信息的载体。人类对于自然、社会及思维的知识均来源于实践。我们从信息的内涵知道，信息本身既不是物质，也不具有能量。因此，任何信息都离不开一定的物质载体和能量，信息总是以一定类型或形式的载体作为其存在和传递的物质基础。

信息载体又可分为信息表征载体和物质载体两类。信息表征载体是指信息内容赖以存在的文字、符号、声波、光波等用以记录信息的载体，亦称第一载体。物质载体是指信息载体赖以存在和传递的纸张、胶片、磁带、光盘等物质材料，亦称第二载体或硬载体、载体的载体。只有第一载体和第二载体有机结合，才能保证信息内容得到完整地保存和有效地传递。随着信息科学、材料科学的不断发展，信息载体也会相应地发展变化。

2) 信息源的特征

(1) 信息数量多、内容庞杂。从某种程度上来说，各种信息收藏机构收藏的信息资源包

括传统信息资源(纸质文献、各类光盘、缩微胶片等)及网络信息资源,这些信息资源数量多,内容纷繁复杂,涉及社会生活的各个方面。特别是网络信息资源包括成千上万种电子期刊、报纸,以及政府、学校和公司等机构和个人的详细信息,信息具有跨地区、分布广、多语种、高度共享的特点。网络信息资源既有最新的各类信息,也有毫无价值的大量冗余信息,数量巨大、内容庞杂。

(2) 信息的半衰期缩短,老化加快。近年来,随着科学技术的快速发展,各类文献资料及网上信息不断淘汰和更新。新知识、新理论、新技术、新产品层出不穷,加速了信息的新陈代谢。我们通常用文献的“半衰期”来描述文献的老化情况。

文献的半衰期是指某学科领域目前尚在使用全部文献中,较新的一版所出版的年限。国外有人统计不同学科文献的半衰期为:地理学 16.1 年、地质学 11.8 年、数学 10.5 年、植物学 10 年、化学 8.1 年、生理学 7.2 年、机械工程 5.2 年、社会科学 5 年、化工 4.8 年、物理学 4.6 年、冶金学 3.9 年、生物医学 3 年。由于各国科技发展水平不同,相应的文献寿命也不相同。

(3) 信息类型多、范围宽、用途广。信息资源建设得以快速发展,信息类型越来越多,包括印刷型出版物、电子期刊、图书、工具书及大量的书目数据库、联机数据库、软件资源,以及个人主页、电子邮件等种种形式。信息呈现出多种表现性,仅文本就有文字、符号、表格等,另外还有声音、动画、图像等。人们通过网络可以方便地进行人文科学、自然科学的信息交流,如发送电子邮件或召开在线视频会议等。

(4) 电子信息成为一种主要的信息资源形式并表现出鲜明的特点。随着计算机技术和情报资源相结合,产生了一种新型的信息资源——电子信息,这种信息资源以数字形式存储,通过计算机或网络检索或传输,加快了信息流通和信息资源共享,以书目、索引、文摘、全文数据库、多媒体信息、电子刊物等形式为用户提供服务。

网络信息资源是一类重要的电子信息,它通常以超文本技术链接,把各类不同的相关信息都通过节点链接起来,使得检索相关信息非常方便。但是这类信息组织特殊、控制性差(尽管从局部来说某个网站、某个数据库是有控制的、相对集中的、有序和规范的),信息质量良莠不齐,所以其优势还没有完全发挥出来。

1.2.2 信息源的类型

信息源的种类繁多,但总体上可分为文献信息源和非文献信息源两大类型。两类信息源各具特点,一般来说,文献信息源信息量大,相对稳定,易于组织、管理和检索;非文献信息源相对而言,具有生动直观、选择性和针对性强、获取速度快的特点。

1) 文献信息源

文献信息源是正规的信息源,是人类社会所特有的人工信息源。凡是以文献作为载体形式的信息源均可称为文献信息源。文献在保存和传递信息方面具有重要的作用,可以说文献信息源是人们获取信息最基本、最主要的来源。文献信息源具有特殊的优势:能够促使信息得到长久的保存和广泛的传播;所载信息固定而明确,便于进行多方面的加工和利用;是确认人类活动,进行社会规范的工具。文献信息源的主要缺陷就是在传递和交流上不够灵活和生动。

2) 非文献信息源

非文献信息源包括电子信息源、实物信息源、口头信息源。

(1) 电子信息源是指含有各种信息的电子资源。电子资源按其载体形态划分,主要有两大范畴:一是以光盘、磁盘、磁带等脱机载体出版的有形电子资源,二是在网络上出版的网上电子资源。数据库光盘和多媒体光盘曾经是图书馆电子文献的主要组成部分,但随着Internet的日益普及和飞速发展,网上资源的优势日趋显著。与光盘相比,网上资源不需要光盘塔等设备的投入和人员的日常操作维护,内容更新周期更快。因此,网上资源将成为图书馆电子资源的主体,光盘只能作为数据备份的辅助工具,这是一个不可逆转的趋势。

电子资源按发布形式不同可分为数据库、电子期刊、电子图书、电子报纸、动态性信息、软件等。其中,数据库是历史最久、影响最广的一种电子信息资源,它以先进的数据处理技术、深加工的信息,以及与网络的完美结合,成为电子信息资源的主体。数据库涉及领域广泛,按其内容可划分为以下4种:

- ① 书目数据库,包括索引、文摘等二次文献数据库,如EI、SCI等。
- ② 数值数据库,主要包含数字数据,如统计数据、科学实验数据、科学测量数据等。
- ③ 事实数据库,收录人物、机构、事务等的现象、情况、过程之类的事实性数据,如机构名录、大事记等。
- ④ 全文数据库,提供完整的原始文献数据,集检索功能与浏览原文功能于一体,是目前最便捷、最具潜力的电子资源。

电子信息源的特性如下:

① 易获取性。电子资源借助计算机与网络,向用户提供检索和传递资料的服务,Internet穿越时空的优势使读者在任何时候、任何地点均可通过网络获取电子资源。电子资源的提供使图书馆的服务在时间和空间上得以延伸,使信息的获取更为便捷。

② 共享性。与传统文献相比,电子资源使多用户可在同一时刻下共享、共用同一信息源。传统文献,如一册书或一本杂志,同一时间只能供一个读者阅读。而电子资源,如一个数据库可被多个用户同时检索,同一篇资料可供多个用户同时下载、浏览,体现了电子资源特有的实时共享性,提高了资源的利用率。

③ 资料更新快。在科学发展突飞猛进、知识更新日新月异的当今时代,信息的时效性显得尤为重要。在减少了印刷型文献印刷、装订、运输等烦冗环节之后,网上出版物和数据库通常以日或周为更新周期,其高效的动态更新和快捷的实时传递,在充分发挥信息的时效价值方面有着传统文献无可比拟的优势。

④ 形式的多样性。多媒体技术在电子资源中的利用,使文字、图形、声音、动画、三维图像等相结合,提供丰富多彩、生动逼真的信息,使读者能够闻其声、见其形,有身临其境之感。与此同时,超文本技术的运用改变了传统信息的线性组织方法,增强了人们对知识的表达能力,大大拓展了信息的获取和传播范围,使读者更容易理解和接受。

⑤ 知识的类聚性。受手工检索工具对文献揭示的不充分性和检索手段烦琐的限制,传统文献中大量有用知识被隐藏。而在网络环境下,依靠强大的检索软件可以使不同类型的相关信息聚集并有序化。目前网络数据库有大型化趋势,同一公司制作的多种数据库产品的检索界面趋向统一,并可多库操作,对检索的完整性提供了可靠保证。除各数据库之间的链接外,数据库还可与图书馆书目查询系统进行链接,这些链接方式可以帮助读者从书目、

题录、文摘等二次文献直接查阅到全文。

(2) 实物信息源是指蕴含信息的实物载体,包括自然实物信息源和人造实物信息源两大类。自然实物信息源如植物、动物、河流、山川、土地、矿物、宇宙等,作为未经加工的自然存在的能够直接反映事物的原始状态,人们可以从中获取有关自然界的的信息,如气象的变化、矿藏的蕴藏、动植物的繁衍等。人造实物信息源是经过人类有目的的加工而成的物体,如产品、样品、样机等。它是人类活动的产物,具有内隐性。其内涵为加工工艺、化学成分、物质参数、设计指标、外观状态等。对这类信息源的利用,需要经过较为复杂的分析研究,将蕴含在实物中的信息解释出来才能够实现。实物信息源的主要优势如下:

- ① 真实可靠。通过实物,人们可以直观而生动地掌握其内涵。
- ② 信息浓度大。通过分析,人们可以从中掌握生产实物的一系列活动。

③ 易于开发利用。通过实物,人们可以充分利用其信息进行仿制、改进和创新。实物信息源的主要缺陷就是传递不方便,而且蕴含的信息必须经过分析与提炼,转换成其他形式才能被人们认识和利用。

(3) 口头信息源。口头信息以大脑为载体,是人与人之间的对话、交谈、演讲等。现实社会中的每一个人,每天都在进行信息的口头交流,不断地发送信息又不断地接收信息。口头信息源有以下优势:

- ① 获取信息快,交流及时。

② 具有高度的选择性和针对性。例如,从研究同一课题的同行处获得有关的口头信息,较之查找大量散见于各种文献中的信息,既方便,又对口。

③ 信息反馈及时、准确。在直接交流时,对于所涉及的问题双方可以立即询问澄清,必要时还可以互相补充和修正,避免理解上的错误。

④ 根据人们接触时的气氛、语气、手势、暗示等,可以领会语言文字不能表达出来的东西,因而有助于全面了解对方所谈的内容,对所获得的信息进行适当评价。

⑤ 可以提供其他信息不能包含的许多信息,而这些细节往往更具启发和借鉴作用。当然,口头信息源的缺陷也是显而易见的。由于人们直接通过口头交流来获取信息的机会总是有限的,机会分布也不均衡,所获信息的满意度和完整性极不稳定;由于难于对口头交流实行有效的社会监督,故其可靠性也不易检验,加之难于积累和整理,信息会随着时间的推移而失真和泯灭。所以人们在获取、传递和利用口头信息时,常常将它转换成文献信息。

1.2.3 文献及其特征

人们常说“文献信息”。其实,文献与信息密不可分。人们将信息、知识、情报记录在一定的载体上就形成文献。文献必须含有信息,信息也必须依附在一定的文献载体上;文献是信息的静态物质形式,信息则是文献的动态抽象形式。人们利用文献,实质上就是利用文献中的信息。因此,可以说文献是继承发扬科学知识的接力棒,信息与知识上升为科学记录于文献之中,又成为新的信息来源。如此反复循环,推动整个人类的文明进步。

在社会生活中,通过对文献信息资源的搜集、分析、整理和存储,向需要利用所藏文献信息资源的人们提供相关服务最常见的信息服务的机构就是图书馆。图书馆不仅为收藏知识积累了丰富的文化典籍,而且越来越成为大量信息的集散地,从而形成了多层次、全方位的馆藏文献信息资源。

1) 文献的定义

文献是伴随着人类对史料的记载而产生的,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,或理解为固化在一定物质载体上的知识,也可以理解为古今一切社会史料的总称。

现在通常理解为图书、期刊等各种出版物的总和。文献是记录、积累、传播和继承知识最有效的手段,是人类社会活动中获取情报最基本、最主要的来源,也是交流、传播情报最基本的手段。

随着时代的变迁,文献载体不断丰富,书籍日益增多,文献的概念也发生了变化,文献已不兼指书和人,而是专指著述。

随着文献交流的日趋频繁,学术界也力求对文献做一个统一的概定。其主要代表性的定义有:泛指一切文字、图形、音响资料和视听资料,是以特定的物质为载体的人类精神信息的固态品;用一定的技术手段将一定的信息内容记录到一定载体上形成的存储型传递媒介;记录信息与知识的一切人工附载体。国际标准化组织的《文献情报术语国际标准草案》中所指文献是:在存储、检索、利用或传递信息的过程中,可作为一个单元处理,依附载体而存储有信息或数据的载体。中华人民共和国国家标准《文献著录总则》(GB 3792.1—1983)中的文献则是记录有知识的一切载体。从以上文献的表述来看,笔者认为,文献是指人们对信息加以有序化、系统化,并以文字、图形、符号、声频、视频等信息符号记录在一定载体上的固态精神产品。

文献是重要的信息资源,是人类知识的物质保存形式。文献的定义由4个要素构成:

- (1) 文献内容,即被记录的信息、知识。
- (2) 信息符号,即记录信息、知识所用的符号,如文字、数字、声音、图像等。
- (3) 文献载体,即除人脑以外的用于记录信息、知识的物质载体,如最早的泥板文献中的泥板,以及竹简、甲骨、石头、绢帛、纸张、磁带、光盘、磁盘等。
- (4) 记录方式,即信息、知识被记录、存储到载体材料上的方式,如铸刻、书写、印刷、录制、摄制、刻录等。

其中,文献内容和文献载体为最基本的要素,两者相互依存构成文献。各类文献经过加工、整理、布局就构成文献资源。文献作为信息的宝库,是人类文明成果的总汇,是一种宝贵的智力资源和精神物质,是社会进步的重要标志。丰富的文献资源能够为人们提供前人或他人的已有成果,避免无效劳动,为社会创造效益。

2) 文献的特征

文献的特征分为外部形态特征和内部形态特征两大部分。文献特征是编制各种文献检索工具时的著录依据。

(1) 文献的外部形态特征。

① 文献名称:它是文献内容的概括。一般情况下,从一本书的名称就可以知道文献的主题与性质,然而有一些文献的名称不能反映其学科内容,如《根》,根据文献名称可判断是一本植物学内容的书,其实是叙述美国家族历史的书。

② 著者名称:它是指对文献负有一定责任的个人或机关团体。著者有本国和外国、团体和个人、古代和现代、主要和次要之分。

③ 版本:包括版次和版刻,指文献的排版次数,用来标明文献版本的重要变更。

④ 出版地:是指出版者所在地的地域名称,可反映地方性出版物的特点,往往也是用户选择文献的条件。

⑤ 出版者:包括负责整理、复印出版的机关团体或出版企业组织等。

⑥ 版期:它取决于版次,主要指某版次的出版时间。

⑦ 载体形态:指文献的页数、卷数、插图、折图、彩图、照片、开本、尺寸等。

⑧ 标准书号与刊号:指图书的 ISBN 号和期刊的 ISSN 号。ISBN 为国际标准书号的代号,后面有 10 位数字,分 4 段组成,每段之间以短横线间隔。第一段为国家、区域号,由国际标准组织(ISO)拨给;第二段为出版商代号,由国家、区域的有关标准化组织拨给;第三段为书号,由出版者自己给出;第四段为稽核号,为计算机核查数字。凡参加国际标准组织的国家,出版的书都可以根据国际标准书号编制法编制各国的标准书号。我国已正式加入国际标准书号系统,并确定我国第一位数字的组号为“7”(如“0”代表美国,“1”代表英国,“2”代表法国,“82”代表挪威,“90”代表荷兰)。一般单个数字代表较大的地域或国家,双位数字代表较小的地域或国家。例如,一本《现代信息检索与利用》的国际书号为 ISBN 7-5354-1988-8。

(2) 文献的内部形态特征。文献的内部形态特征是指可以揭示文献内容和实质的标识,有分类号和主题词两种。

① 分类号:文献分类就是根据确定好的逻辑系统(分类法),将文献信息所载知识单元及形式特征,利用号码标识进行分门别类的系统揭示与组织。在分类语言中,用来代表类目名称的标识符号称为分类号,分类号是现代文献分类法的重要组成部分。文献的分类号标识文献信息主题内容及形式特征,是提供分类排架和组织分类目录的依据。从文献分类号可以得知该文献的学科属性、内容实质。例如,文学用“I”字母来标识,中国小说用“I24”来标识。

② 主题词:它是一种以词语为检索标识,采用词语揭示和描述文献信息主题内容,按照主题词字顺组织文献的情报检索语言。主题词是文献的中心思想和所要表达的主要内容的高度概括。

3) 文献的类型

文献种类多种多样,是因为划分文献的标准多种多样。例如,按文献载体形态分为纸质型文献和机读型文献,而按文献记录手段又划分为印刷型文献和非印刷型文献。

(1) 印刷型、缩微型、机读型及声像型文献。

① 印刷型文献。印刷型文献以纸张为存储介质,以手写、印刷为记录手段而产生的一种文献形式。其优点是便于阅读和传播;其缺点是存储密度太低、体积大、占用空间多,不能实现自动检索,存取速度慢。

② 缩微型文献。缩微型文献以感光材料(胶片)为存储介质,以缩微摄影技术为记录手段而产生的一种文献形式。其优点是体积小、重量轻、信息存储的密度高,借助计算机可进行自动检索;其缺点是必须借助专门的设备方能阅读。

③ 计算机可读型文献。计算机可读型文献以磁性材料和光盘为存储介质,通过编码和程序设计,把文献信息变成计算机可识别的数字语言和机器语言,然后存储在计算机的磁带、磁盘或光盘上,检索时再由计算机输出的文献。其优点是信息存储的密度大,输入输出方便快捷,可实现自动检索;其缺点是硬件设备费用较高。

④ 声像型文献。声像型文献是一种非文字的文献,以磁性材料或感光材料为存储介质,借助专用机械装置直接记录声音信息和图像信息而产生的文献。其优点是直观、真切;

其缺点是阅读、利用时要借助专用设备。

(2) 一次、二次及三次文献。

① 一次文献。一次文献也称为原始文献,是作者直接记录生产或科研成果的总结和理论探讨的原始文章,如科技论文、专著、研究报告、专利说明书等。确定一篇文献是否为一次文献,主要看其内容是否为原始创作,而不以其存储介质和存储方式来确定。

② 二次文献。二次文献是根据一次文献的外表特征和内容特征,经过加工整理,简化组合,使之系统化并具备检索功能的文献,包括印刷型的目录、题录、文摘和用于文献检索信息的机读数据库,统称为检索工具。二次文献是检索一次文献的工具。

③ 三次文献。三次文献是指借助二次文献广泛地选用一次文献的内容,经综合分析、比较研究而编写的文献,如专题述评、动态综述、进展报告、数据手册、百科全书、文献指南等。

4) 文献的出版类型

按文献的出版形式可将文献分为三大类型。其具体类型结构如图 1-2 所示。

(1) 图书。在各种文献中,图书的品种最多、数量最大、范围最广,包括专著、教科书、参考工具书(辞书、年鉴、手册、百科全书、名录、图谱、图录、年表、历表等)和检索工具书(目录、题录、文摘等)。图书一般给人们提供系统性、完整性、连续性的知识和信息。

(2) 期刊。期刊是一种采用统一名称,定期或不定期出版的一种连续出版物。它具有出版周期短、报道速度快、数量大、信息内容新等特点,是很重要的信息源。

(3) 报纸。这是社会信息传播的一种重要文献,也是很重要的信息源。其特点是报道及时,受众面广,具有群众性和通俗性。

(4) 会议文献。会议文献是指在各种范围的学术会议上发表的论文及与会议相关的文献。其特点是反映新技术、新成果快,专业性强,能较全面地反映某一技术领域或学科的发展水平、动态和趋势。

(5) 科技报告。科技报告是指科技人员在从事某项科研工作时撰写的阶段研究报告或研究成果的正式报告。它具有单独成册、出版日期不定、内容专深、报道迅速、内容保密、控制发行等特点。

(6) 标准文献。标准文献是指在工、农业产品和工程建设的质量、规格、检验方法等方面做出的技术规定,是从事生产、建设的一个共同技术依据和准则。它是一种具有约束力的规定性、法律性文献。

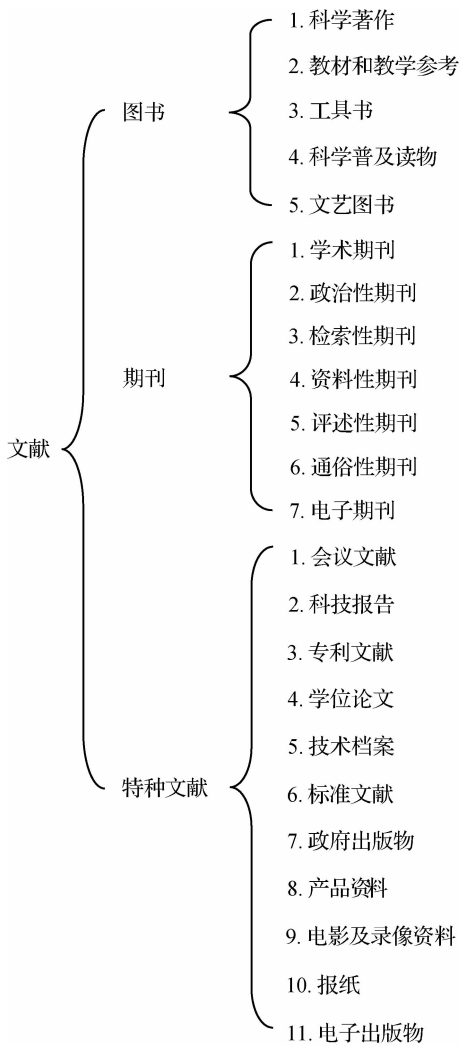


图 1-2 文献出版类型结构

(7) 专利文献。专利文献是指一切与专利制度有关的文献,主体是发明说明书和专利说明书。专利文献有统一的格式,是一种法律文件。它具有详尽、新颖、实用、报道及时等特点。

(8) 学位论文。学位论文是指高等学校或科研单位的毕业生为取得某种学位资格而撰写的论文。一般有学士论文、硕士论文和博士论文之分。其内容较为系统,具有一定的深度和创造性。

(9) 技术档案。技术档案是指在生产建设和技术部门的技术活动中形成的,有一定工程对象的技术文件的总称,如任务书、计划书、审批文件、技术措施、设计图纸、工艺卡片等应入档的技术资料。

(10) 电子文献。电子文献是指以数字、代码方式,将图、文、声、像等信息存储在磁、光、电介质上,通过计算机或者类似功能的设备阅读使用,并可复制发行的一种文献,可分为磁盘型、光盘型、网络型文献。图书、期刊、报纸、检索工具等都可以制作成电子文献。电子文献既可以单机使用,也可以通过网络和联机的形式阅读。它具有存储量大,体积小,便于检索、保管和共享的特点。各图书馆非常重视电子文献的收藏和利用。电子文献有着十分广阔的发展前景。

1.3 信息社会与信息素养

随着计算机技术的快速普及,Internet 热潮席卷全球,谁也无法否认人类社会正以无法估量的速度悄然迈入知识(信息)经济时代。在这个时代,经济活动对于信息(知识)活动的依赖已达到十分突出的地步,在所有可利用的资源中信息资源的利用从总体上上升到主导地位。

1.3.1 信息社会及其特征

20 世纪 80 年代以来,由于第三次信息技术革命的推动、知识经济的勃兴,人们确信人类将逐步发展到信息社会阶段。回顾人类发展历程,我们必须清醒地看到,信息革命带给我们的是全方位的变革,正如一首歌中唱到的“这世界变化快,真让我不明白”,实际上当今时代事事时时都处于快速变化之中。奈比斯特说:“趋势就像一匹马一样,比较容易向着它已经奔跑的方向骑去。”无疑,以知识经济为显著特征的信息社会已经成为社会发展的一种无法抗拒的趋势。

1) 信息社会的根本特征

我们今天所处的时代,有人称为“信息社会”,也有人称为“后工业社会”,其中“信息社会”的概念已为大多数人所认同。马丁曾经把信息社会定义为一个生活质量、社会变化和经济活动越来越多地依赖于信息及信息开发利用的社会。在这个社会里,人类生活的标准、工作和休闲的方式都明显地被信息和知识的进步所影响。许多研究者还归纳出衡量信息社会的若干标准:一是技术标准,信息技术必须是社会关键的能动力量;二是社会标准,信息有助于提高人民的生活质量,整个社会要有广泛而强烈的信息意识;三是经济标准,信息必须成为关键的经济因素,是资源,是服务活动,是流通的商品,也是就业和财富增值的源泉;四是

政治标准,信息能够增强民主和自由,加强人们的各种参与和妥协;五是文化标准,信息具有文化价值。

如果用上述标准衡量当今社会,会得出一个无可争议的结论:我们的社会已处于信息社会发展阶段。由此可见,所谓信息社会,就是信息化了的社会,或者说信息社会最根本的特征就是社会信息化。它是对社会以材料、能源为基础的发展态势,向以信息、材料、能源并立的发展转变的一种描述和表征。社会信息化包括硬信息化和软信息化。硬信息化是指以信息基础设施建设为标志的信息技术在社会领域的全面推广;软信息化是指以信息服务发展为标志的信息资源与社会活动的深度融合。

我国政府非常重视社会信息化问题。继邓小平同志倡导“开发信息资源,服务四化建设”之后,江泽民同志又强调“四个现代化,哪一化也离不开信息化”。党和国家领导人如此重视社会信息化发展,是因为社会信息化具有强大功能:一是使社会发展空间高度拓展的激活功能;二是使社会发展因素充分组合的配置功能;三是使社会发展潜力全面发挥的诱导功能;四是使社会发展效益有效放大的倍增功能。

2) 信息化的基本特征

信息社会具有定性和定量的特征,本书主要从定性的角度来揭示信息社会的基本特征。

(1) 信息数量呈几何级数增长。据专家估计,20世纪40年代以来产生和累计的信息量已经大大超过了在此之前人类有史以来的所有信息量之和。19世纪以来,人类知识信息每50年增长1倍,20世纪中叶每10年增长1倍,20世纪70年代以后每5年增长1倍。

(2) 信息和信息技术广泛应用于社会生活的各个领域。信息产生于信息社会的各个角落,也服务于社会的各个方面。信息化手段的广泛应用和社会政治、经济、文化生活中的信息导向充分显示出信息的广泛存在和强大功能。“旧时王谢堂前燕,飞入寻常百姓家”,信息技术的普及大大缩短了信息流程,世界变成了“地球村落”。美国《时代》杂志曾破例把个人计算机选为年度风云人物,显示了现代信息技术和信息行为广泛社会化的极大影响。

(3) 人们的信息意识空前强化。人们已经充分认识到信息的重要功能和作用,全社会都在有效地开发利用各种信息资源。人们从更广泛、更快捷、更全面的途径中吸收和获取信息,主动搜寻和应用信息的意识也在不断加强,形成了一种高度的信息敏感性。人们对信息的渴望从来没有像今天这样强烈,人们的学习、工作和生活都离不开信息的参与。信息需求的自觉和自发显现,充分说明了社会信息意识——对信息功能的认同意识,对信息来源的选择和开发意识,对信息内涵的同构和再生意识,都得到了充分的加强。

(4) 从事信息职业或半信息职业的人数超过从事非信息职业的人数。在西方,人们把担任技术、管理事务的人员称为白领阶层,把直接从事生产的人员称为蓝领阶层。在发达国家,20世纪80年代以后,白领阶层人数已超过蓝领阶层人数。美国被认为有60%以上的人主要与信息打交道,而不是与物质资源直接联系。现在人们又把专门从事数据管理的人员称为灰领阶层,他们更多面对的是计算机终端的显示屏。所谓信息职业,就是创造、处理、存储和分配信息的职业。其主要包括从事市场信息工作的劳动者(信息商品生产和销售)、提供市场信息的劳动者(信息加工和服务)和从事信息基础设施的劳动者(信息技术开发和操作)。

特别是到了信息社会,新的就业机会往往就是信息职业的就业机会。因为在信息社会,无论信息职业、半信息职业或是非信息职业,信息都成为职业因素中的重要组成,任何人都

无法摆脱信息和信息技术的影响。

(5) 信息经济成为现实。按照西方经济学流行的观点,当今社会信息交流已成为经济发展的倍乘因子,即社会净产值=各部门物质生产(人力+资金)投入总和×信息流量。信息已成为经济发展的重要因素,信息经济已成为客观现实,它建立在信息资源不断积累,信息传播技术手段越来越先进、便捷的基础之上。信息经济的基本特征是:第三产业的比重不断上升,非信息职业(现在已没有纯粹的非信息职业)让位于信息职业和半信息职业;理论、知识处于经济发展的中心地位;信息产业成为多数国家的“朝阳产业”。

1.3.2 信息素养

1) 信息素养的概念

随着信息技术的飞速发展,“信息素养”越来越多地被人们提及,究竟什么是信息素养呢?

信息素养(Information Literacy)强调的是对人的内在素养的一种描述。1974年,美国信息产业协会主席保罗·泽考斯基首次提出信息素养这一概念,并把它定义为“人们在解决问题时利用信息的技术和技能”,与现在的信息素养理论相比,这个定义只是一个原始雏形。1989年,美国图书馆协会下设的“信息素养总统委员会”给信息素养下了一个这样的定义:“要成为一个有信息素养的人,就必须能够确定何时需要信息,并具有检索、评价和有效使用信息的能力。”

1989年,熊扬的《浅议企业经营者市场信息素养》是我国第一篇关于信息素养的文章。从此,我国步入信息素养概念研究的时代,出现很多不同的定义,定义虽有多种,但是其内容都表示出人们在信息社会到来后对信息行为质量和效果的高度关注和追求。引用最广泛的是美国图书馆协会提出的信息素养的概念:具有信息素养的人必须在需要时能够识别、查找、评价和有效的使用信息。

2) 信息素养的内涵

信息素养的内涵是信息素养概念的进一步延伸和细化,主要包括信息意识、信息知识、信息能力和信息道德四个方面。

(1) 信息意识。信息意识是指客观存在的信息和信息活动在人们头脑中的主观能动反映,表现为人们对所关心的事或物的相关信息敏感力、观察力和分析判断能力及对信息的创新能力。信息意识是人们产生信息需求,形成信息动机、信息兴趣,进而自觉寻求信息、利用信息的动力和源泉。通俗地讲,就是面对不懂的东西,能积极主动地去寻找答案。

(2) 信息知识。信息知识是指与信息有关的理论、知识和方法,包括信息理论知识与信息技术知识。

① 信息理论知识。包括信息的基本概念、信息处理的方法与原则、信息的社会文化特征等。有了对信息本身的认知,就能更好地辨别、获取和利用信息。信息知识是信息素养教育的基础。

② 信息技术。广义而言,信息技术是指对信息进行采集、传输、存储、加工、表达的各种技术之和。该定义强调的是人们对信息技术功能与过程的一般理解。狭义而言,信息技术是指利用计算机、网络、广播电视等各种硬件设备及软件工具与科学方法,对“文图声像”各

种信息进行获取、加工、存储、传输与使用的技术之和。

(3) 信息能力。信息能力指理解、获取、利用信息的能力和利用信息技术的能力。

① 理解信息能力,即通过对信息进行分析、评价和决策来鉴别信息质量和评价信息价值。

② 获取信息能力,就是通过各种途径和方法搜集、查找、提取、记录和存储信息的能力。

③ 利用信息能力,即将信息用于解决实际问题、学习和科学研究之中,通过已知信息来挖掘信息的潜在价值以创造新知识的能力。

④ 利用信息技术能力,即利用计算机网络以及多媒体等工具搜集信息、处理信息、传递信息、发布信息和表达信息的能力。

(4) 信息道德。信息道德是信息活动各个环节中用来规范其间产生的各种社会关系的道德意识、道德规范和道德行为的总和。它通过社会舆论、法律法规、传统习俗的约束使人们形成一定的信念、价值观和习惯,从而使人们自觉地规范自己的信息活动行为。

1.3.3 信息检索与信息素养教育

掌握信息检索的基础知识和基本技能,对于提高信息素养具有非常重要的意义,具体表现在以下方面。

1) 利于创新人才的培养

当今社会正处于知识经济时代,科学技术飞速发展,新知识层出不穷。社会对人才的需求也已发生巨变。智力劳动者的社会需求日益紧缺。培养高素质、高智商的创新人才,已是目前我国高校教育迫在眉睫的重大使命。所谓创新人才,必然是发掘和捕捉最新科技知识信息的行家里手,因为无数创新成果表明:谁能优先吸收、获取和利用最新科技知识信息,谁就能优先占领科技发展前沿阵地,谁就能在竞争中立于不败之地。为此,开设本课程就是培养学生捕捉最新科技知识信息的情报意识,学习和掌握获取最新科技知识信息的技能,为科技创新锻炼过硬的文献信息检索本领。

2) 利于开发、利用文献信息(智力)资源

在当今的信息化社会里,文献信息迅猛膨胀,数量惊人,类型繁杂,新旧知识更新加快,人们处于浩如烟海的文献信息当中,迅速获取广博精准的文献信息越来越困难。学习和掌握文献信息检索的科学技术和方法,即可全面、迅速、精准地获取文献信息,为进入丰盛的文献信息的智能宝库提供了金钥匙。人们利用文献信息检索技术和方法,可以遨游文献信息大海。在这个知识经济的时代里,谁能优先掌握信息资源,谁就能在竞争中站在科学发展的前沿阵地。因此,我们必须要有被时代淘汰的紧迫感,树立强烈的信息意识,积极主动地获取所需文献信息。

3) 利于继承、借鉴和创新

继承、借鉴和创新是社会发展的客观规律。科学技术的发展、人类物质和精神文明的进步也不例外。正如牛顿所说:“如果说我看得远一点,是因为我站在巨人的肩膀上。”这就是继承、创新规律的生动描述。因此,我们从事学习、科研和生产实践,必须在前人或他人的理论、技术和科研成果的基础上,深入理解知识、学精技术、吸取研究的经验教训和发现尚待解决的问题等,只有这样才能制定出可行性创新方案,对已学知识提高认识,使技术更加专深,

研究的起点更高。尤其在当今知识经济时代,继承、借鉴和创新的作用更加突出。例如,一项新的科技成果中,前入、他人的成果占 90%,而创新的成果只占 10%。为此,只有学习和掌握过硬的文献信息检索技能,才能在学习、科研时迅速准确地发现和了解前人和他人的研究成果。显而易见,只有强化文献信息意识和能力修养,才能有效地继承、借鉴和创新。

4) 利于节时、省力和高效

从事学习、科研都离不开查阅有关文献信息资料,它占据整个学习、研究的大量时间。美、日等国的调查结果表明,一个科研周期的各项工作环节所占时间为:计划思考占 7.7%,查阅文献占 50.9%,研究实验占 32.1%,撰写论文占 9.3%。

可见,为了提高查阅效率,必须选用最科学、最实用的文献信息检索方法,缩短时间,节省大量的精力,以便更好地学习。开设此课程的目的就是培养学生科学的检索技术,尤其是当今信息技术已引入文献信息检索领域,不仅要学习和掌握手工检索技能、技巧,而且要学习和掌握计算机检索技能、技巧,以实现省时、省力、高效。

5) 利于终生知识更新

世界进入知识经济时代以来,知识半衰期迅速缩短,新旧知识代谢加快。大学生迈入社会,走向新的岗位,如不及时更新知识,适应社会发展的新形势,将会成为新时代的知识盲人,落后于时代,必将被淘汰。有关专家估测:科技人员已具有的科技知识,12%是在大学获取的,87.5%是在工作岗位上通过学习与实践积累的。大学生在新的工作岗位上如果不更新知识,3~5年内已学的知识 50%陈旧失效,8~10年再不更新知识,原有的知识将失效 80%。为此,从长远出发,就要提高独立自主的学习能力、文献信息检索与利用的能力和不断更新知识的能力。

6) 有利于改善我们的生活质量

在日常生活中,小到买一本书,大到出国留学或买房买车这样的大额消费,如果掌握了信息检索的基本技能,就能及时准确地获取与生活有关的社会与市场信息,使我们的生活不仅更加丰富多彩,而且生活成本也会降低,从而为改善生活质量提供了极大的可能性。

7) 有利于缓解信息压力,有效解决信息污染

无处不在的信息虽然给我们的生活、学习、工作带来了许多便利,但不可否认,也带来了巨大的信息压力,尤其是无处不在的“垃圾”信息,往往使我们无所适从,干扰我们的思维,影响我们做出正确决策。掌握正确的信息检索技术,可以有效化解信息污染带来的风险,缓解巨大的信息量带来的精神压力。总之,学习信息检索知识是信息素养教育的重要内容,是培养高素质人才的基本要求。

1.3.4 信息检索及社会功用

鉴于信息对人类活动的重要支持和影响,以及信息激增和信息环境严峻给人们利用信息带来的不便和困难,面对文献信息的汪洋大海,要想及时、准确而系统地掌握某一专题或某一方面的具体情况(背景、现状、趋势)以促进事物的良性发展,信息检索活动显得尤为重要。

信息检索的社会功用主要表现在下面几点:

(1) 开阔视野,正确决策。通过对信息进行全面搜索,能够系统地了解前人的工作经验

与成果,及时掌握事物的最新动态及发展趋势,开阔视野,启发思路,在现有的科技开发、学术研究基础上不断创新,取得更快进展。

(2) 提高功效,事半功倍。高效开展信息检索活动,能节省人们对有用信息进行搜集利用的时间及精力,提高工作效益,做到事半功倍。

(3) 规避风险,维护权益。及时准确地开展信息检索,可以对将要开展研究的课题或项目通过全面的文献信息调研,得到其研究背景、现状及趋势方面的准确信息,使人们减少盲目开题、重复劳动等风险,从而少走弯路、提高决策的科学性,保证相关活动的正常进行,以达到预期的目的。

(4) 学习借鉴,推动创新。信息检索活动也是促进对信息资源的开发、利用及进一步丰富和繁荣的基础。它不仅促进了信息用户对现有学术及研究成果的学习和借鉴,促进了人类知识宝库作用的发挥,同时也对于全面提升人们的综合素质,创作出更优秀的文献信息是一种强大的推动力。

(5) 科学评价,知己知彼。信息检索活动一直以来就是人们开展学术研究及与相关活动的重要组成部分。它贯穿于研究项目开题立项前的文献调研、进行过程中的动态跟踪、完成研究后的成果评价等全部环节之中。同时,在国家军事、经济贸易、生产开发、企业竞争以及个人求职升迁等活动中,作为及时了解竞争对手、把握市场需求和相关情况的重要手段,发挥着不可忽视的重要作用,所谓“知己知彼,百战不殆”就是古人对信息搜集与利用价值的高度总结。

第 2 章

信息检索的基本原理与方法

学习目标

- ▶ 了解信息检索语言。
- ▶ 掌握信息检索的原理和手段。
- ▶ 了解信息检索的研究对象和范围。
- ▶ 了解信息检索程序、策略和效果评价。



2.1 信息检索语言

2.1.1 检索语言在信息检索中的作用

检索语言是根据检索的需要而发明的一种专门的人工语言,是在文献检索领域中用来描述文献特征和表达信息检索提问的一种专用语言。

信息检索是信息传递过程中的一个重要环节。随着科学技术的飞速发展,科技文献量急剧增长,内容专深且互相渗透,而科研、生产、教学及行政部门对情报的需求十分迫切,尤其是在检索课题较专深的情况下,信息检索总是变得相当复杂,为了达到较满意的检索效果,就必须做出很大的努力。

检索语言与检索效果之间有着密切的关系,它在检索过程中起着极其重要的作用。信息检索的全过程包括信息的存储过程和检索过程。当存储信息时,文献标引人员首先要对各种文献进行主题分析,通过分析选出若干能代表文献主题的概念,并用信息检索语言把这

些概念标引出来,然后纳入信息检索系统中。当检索信息时,信息检索人员也首先要对检索课题进行分析,并且通过分析明确其检索范围,选出若干能代表信息需要的概念,并把这些概念转换成信息检索语言,然后从信息检索系统中查找用该检索语言标引的文献,从而找到用户所需的信息。

由此可见,信息检索语言是信息检索系统的主要组成部分,是标引人员与检索人员之间沟通思想、取得一致意见的桥梁。

2.1.2 信息检索语言的组成、特点和功能

1) 信息检索语言的组成

实质上信息检索语言是表达、概括文献信息内容的概念及相互关系的概念标识系统。它可以是从自然语言中精选出来并加以规范化的一套分类号码,也可以是代表某类事物的某一方面特征的一套代码(如化合物的各种代码),用以对文献内容和信息需要进行主题标引、逻辑分类或特征描述。

信息检索语言是由词汇和语法组成的。在这里词汇指的是登录在分类表、词表中的全部标识,一个标识(分类号、检索词、代码)就是它的一个语词,而分类表及词表则是它的词典;语法是指如何创造和运用那些标识(单个标识或几个标识的组合)来正确表达文献内容和信息需要,以有效地实现信息检索的一整套规则。

2) 信息检索语言的特点

检索语言不同于自然语言,它所表述的概念只有一种解释,不允许一词多义、多词一义,而使概念的表述模棱两可。检索语言的这种单一性保证了表述概念的唯一性和标引与检索的一致性,从而使信息检索人员全面、准确、快速地检索到含有所需信息的文献。

3) 信息检索语言的功能

通过检索语言标引文献的主题概念,不仅能简明地提示文献包含的信息内容及其外表特征,而且还能将同一主题概念的文献集中在一起,使文献的存储集中化、系统化、组织化,便于进行有规律的检索。

2.1.3 检索语言的分类

目前世界上使用的检索语言有一两千种,但各种信息检索语言的基本原理是一致的,只是它们在表达各种概念及其相互关系和解决对它们提出的共同要求时所采用的方法不同,因而形成了不同的类型和语种。

(1) 信息检索语言按其构成原理可分为分类检索语言、主题检索语言和代码检索语言三大类。

(2) 信息检索语言按其包括的学科或专业范围可分为综合性语言和专业性语言。

(3) 信息检索语言按其适用范围可分为一种检索工具或检索系统专用的语言,某一图书、信息机构专用的语言,各类图书信息机构通用的语言。此外,信息检索语言可分为手工检索工具适用的语言和机器检索系统适用的语言。

(4) 信息检索语言按语种可划分为单语种语言和多语种语言。

(5) 信息检索语言按其标识的组合使用方法还可划分为先组式语言和后组式语言。

上述5种划分方法中,按构成原理划分的方法是最基本、最常用的方法,以下详细说明用这种方法划分的各种信息检索语言。

(1) 分类检索语言是用分类号来表达各种概念,将各种概念按学科性质进行分类和系统排列。分类检索语言包括体系分类语言(体系分类法)和组配分类语言(组配分类法)。体系分类法主要应用概念划分与概括的方法。组配分类法主要应用概念分析与综合的方法。实际上,无论是前者还是后者,都采用概念划分与概括的方法建立等级体系结构,又采用概念分析与综合的方法进行组配。

(2) 主题检索语言是用语词来表达各种概念,将各种概念按字顺排列。主题检索语言包括标题词型检索语言(标题法)、单元词型检索语言(单元词法)、叙词型检索语言(叙词法)和关键词型检索语言(关键词法)等。它们可统称为主题法系统。

(3) 代码检索语言一般只就事物的某一方面特征,用某种代码系统来加以标引和排列。下面仅就几种常用语言进行介绍。

1) 体系分类语言

(1) 体系分类语言的定义。体系分类语言是一种直接体现知识分类的等级结构的标识系统。它根据一定的观点,以科学分类为基础,以文献内容的科学性质为对象,运用概念划分与概括的方法,按照知识门类的逻辑次序,从一般到具体,从简单到复杂,进行层层划分,每划分一次就产生许多类目,逐级划分就产生许多不同级别的类目。

所有不同级别的类目,层层隶属,形成一个严格有序直线性的知识门类的等级制体系。每个类目都用分类号作为标记,每个分类号都是表达特定知识概念的词汇,这些词汇即是体系分类语言。

(2) 体系分类语言的特点。

① 体系分类法具有按学科或专业集中地、系统地提示文献信息内容的功能,有利于从学科或专业角度进行全面的检索。

② 体系分类法按结构逐级划分,具有等级结构,便于扩检和缩检。

③ 体系分类法用分类号作为检索标识,不能直接表达概念;检索时,首先要将检索的主题内容转换为分类号,然后进行检索。若不熟悉分类法,转换时就可能发生误差,但用分类号检索,可不受文种的限制。

④ 分类法不宜经常修订,列举在分类表中的类目不能随时改变,因而分类法不能及时反映新的科学技术。遇到检索主题属于新兴学科或主题狭窄的课题时,有可能找不到相应的类目和分类号。

⑤ 分类法是按直线或序列编制类目,遇到跨学科的课题时,与课题相关的文献只能标引在一门学科的类目之下,因此检索时可能漏检。

⑥ 体系分类语言是一种先组式检索语言,因此,不能按检索课题的需要进行组配检索。体系分类语言除了以上特点以外,还有一个独特的优越功能,就是组织图书资料的排架,这是其他检索语言不可比拟的。

2) 标题词语言

(1) 标题词语言的定义。标题词语言是规范化了的自然语言。它以经过标准化处理的名词术语作为标识,来直接表达文献所论及或涉及的事物之间的相互关系(这种关系是借助

于参照系统来间接显示的)。

(2) 标题词语言的结构及类型。

① 标题词语言结构。标题词语言的三大基本构成要素是主标题、副标题和说明语。

- 主标题。这是表达文献核心内容的主题概念,一般由经过控制的自然语言中的词、词组和短语来充当。
- 副标题。它是从一特定方面对主标题进行说明、限定,并经过规范化的词、词组进行限定,说明主标题所表达的文献主题的某个部分、某一应用方面、研究对象和方法、性质、场所以及文献类型。
- 说明语。它是用来详细说明和限定主标题、副标题的内容与方面的词、词组或短语。它由未经控制的自然语言表达,但不纳入词表正文,而是排在主、副标题之下。

② 标题词语言类型。标题词语言是按事物来集中相关文献的。因而一个标题下的副标题就可能涉及相当于分类表中多个类目的范畴。一般都将其某些标题下的内容进行细分。细分方法多种多样,相应地形成了不同类型的标题形式,主要包括以下几种:

- 单级标题。仅由一个名词术语构成的标题,它可以是一个单词,称单词标题或元词标题;也可以是一个词组,称词组标题或复词标题,但都只有一级,故称单级标题。这种标题系统的优点是比较简单,缺点是专指性差,在一个专题下往往集中资料太多,不便于检索。
- 带说明语的单级标题。例如,“起重机,用于建筑施工的”“期刊,建筑工程的”等,这实际相当于一个复词标题。由于用说明语表达复杂概念比较自由,既能表达事物的特征,也能表达事物的方面,因而可以达到较高的专指度。但其缺点也是显而易见的,那就是比较冗长,排列次序不够明确。
- 多级标题。即在标题下再加标题。汉语中常用破折号将两级标题分开,破折号后的标题称为子标题,用以表示该标题所表达事物的某一“方面”。子标题之下还可以有子标题,称为次子标题;次子标题下还可以有子标题,称为再次子标题。不过,实际应用中超过三级的多级标题形式很少见,绝大多数是采用二级或三级标题,各级子标题可统称为副标题,第一级标题则称为主标题。
- 倒置标题。将复词标题中起限定作用的词倒置于被限定词之后,例如:

钢结构桥梁	桥梁,钢结构
木结构桥梁	桥梁,木结构
水泥结构桥梁	桥梁,水泥结构
石结构桥梁	桥梁,石结构

由此可见,倒置排列的结果可以将某一方面的资料集中在一起,这是其主要优点,但它不能单独构成标题系统。

- 带限定词的标题。例如,“建筑学(辞典)”“古建筑(维修)”“苹果(水果)”等。限定词一般都置于括号中,用于做形式复分,或区分两个字面相同但无法用其他方式区分的标题。显而易见,这种标题形式也不能单独构成标题系统。
- 混合标题。同时采用上述几种标题形式中的若干种构成。混合标题系统的优点是能提高检索效率,查找比较方便,但其结构复杂,编制难度较大。

(3) 标题词的书写格式。标题词若是一个词组,则有以下3种书写格式:

① 正序式。按词组的正常书写顺序书写。

② 倒序式。把词组的正常书写顺序颠倒过来,并加逗号表示。利用倒序式可将某些同族事物集中显示其种属关系,例如,“COMPUTERS, ANALOG”(模拟计算机)和“COMPUTERS, DIGITAL”(数字计算机)。

③ 并列式。用“与”的形式把两种概念组合在一起,如“METAL AND ALLOYS”。

一般来说,标题词间的相互关系,即它们所表达的各种概念之间的相互关系,可分为等同关系、等级关系、相关关系 3 种类型。我们可以用“见”和“见自”这一对参照来显示等同关系,用“参见”和“参见自”这一对参照来显示等级关系和相关关系。具体说明如下:

- 等同关系一般有标题词的不同写法、同义词、学名与俗称、新称与旧称、全称与简称、不同译名、其他等同词等。在每组等同关系的几个词(词组)中,都只能选用其中一个作为“标题词”,其余的则作为“非标题词”保留在标题系统中,并用“见”和“见自”这一对参照来标引它们与标题词之间的关系,例如:

电脑 见 电子计算机

这样一来,进行检索时,就不一定要判断清楚哪一个词是被采用的,哪一个词是不被采用的。可以从所想到的任何一个词或任何一种标题形式入手检索。如果入手检索的词或标题形式正好是不被采用的,则“见”参照就会把检索人引到正式被采用的标题词中。

- 等级关系一般包括属种关系、学科与其分支的关系、整体与部分的关系等。

等级关系实际上是广义词(泛指词)与狭义词(专指词)之间的关系。广义词是上位词,即上级标题词;狭义词是下位词,即下级标题词。在标题法中,一般只做上级标题词“参见”下级标题词的参照,而不做下级标题词“参见”上级标题词的参照。例如:

电机 参见 电动机 发电机

海洋学 参见 海洋地质学

消化系统 参见 肠胃

- 相关关系一般有交错关系,矛盾关系,因果关系,某种原理、方法、材料、设备与其某个应用方面的关系等。由于具有相关关系的两个标题词下的资料是可以互相参考、互相补充的,所以应用相互“参见”,例如:

图书馆学 参见 情报学

情报学 参见 图书馆学

(4) 标题词表。标题词表是把标题词按一定方式组织与展示的词汇表。它由编制说明、主表、副表 3 部分组成。

下面以美国“工程索引”使用的《工程标题词表》(简称 SHE)为例加以说明。《工程标题词表》是《工程索引》(EI)文摘条目编制和检索的依据(1993 年前),EI 的月刊和年刊都是按此标题表编排的。此表由美国工程情报公司编制。

SHE 由主标题词、副标题词、“见”或“参见”项及说明语 4 部分组成。词表的后面有 5 个附录,它们是:附录 1,副标题索引;附录 2,此部分附录中包括计算机辅助分类代码,主题指南和机构索引;附录 3,缩略语表;附录 4,希腊字母及发音;附录 5,美英不同拼写对照。

SHE 的正文是按主标题词的字母顺序排列的。主标题词全文大写黑体印刷,其下含数

量不等的若干副标题词,副标题词的第一个字母用大写。

主、副标题词共计 14 000 多个。其具体结构举例如图 2-1 所示。

BULLDINGS ^①	402,(404) ^②
(Use for general subject ,applying code 404 if for civil defense or military purposes . Use geographical location as subheading . If significant, use heading for specific type of building . For specific types represented by main headings, see BULLDINGS/REFFRECE LIST ^③)	
Acoustics	See ARCHITECTURAL ACOUSTICS
Air Supported Structures	See ^④ INFLATABLESTRUCTURES
Aluminum Applications ^⑤	415
Arctic	
Brick	
Ceilings	
Climate Control	(Beginning 01/83)
Cion Systems	(See also ^⑥ Signal Systems)
Concrete	718

说明:①标题词;②分类代码;③说明语;④“见”项;⑤副标题词;⑥“参见”项。

图 2-1 主、副标题结构示意图

SHE 是用计算机编制的,可提供计算机检索磁带,词表中标题词右侧的 3 位数字,便是该标题相对应的计算机分类代码。

“见”(see)项可指引读者选用更切题的标题词,“参见”(see also)项用以指引读者进一步选用含义相关的其他标题词,以扩大检索范围。

说明语位于主标题词或副标题词下,置于括号内,用斜体字排印,用于说明该标题词的定义、适用范围以及使用方法等。

在 SHE 的副标题索引中,每个副标题词的右边分别标有一个符号,它们是关于该副标题的简略范围注释,具体符号及其含义如下:

- S 表示它可以作为普通副标题使用,而且是直接从主标题转过来的。
- # 表示它只能做副标题使用。
- @ 表示它与主标题词的形式不同。
- + 表示它是属于地理位置、材料或船只的名称。

SHE 是查找 EI 的主题索引和相同的辅助索引的工具,通过 SHE,可以准确了解 EI 采用的标题。

(5) 标题词语言的特点。

- ① 标题词语言标引文献,是以事物为中心,从有关该事物出发便于进行特性检索。
- ② 标题词在词表中按字顺排列,无字级性。
- ③ 标题词语言只能通过参照系统进行扩检和缩检,因此,查全率较低。

3) 关键词语言

(1) 关键词语言的定义。关键词语言是为适应主题目录、主题索引编制自动化的需要而产生的一种主题语言型检索语言,是文献量迅猛增长、信息传播速度加快的必然产物。

关键词是指在文献的标题、摘要或正文中出现的,对表达文献内容具有实质意义,能作

为检索入口的,起关键性描述作用的词汇。

关键词语言,就是针对文献中的关键词选定或抽出,不做或极少做词汇控制,按字顺排列,从而提供检索途径的一种检索语言。

(2) 关键词语言的特点。关键词语言可实现检索工具编制过程的计算机化,因而降低对编目人员知识和业务水平的要求,降低标引难度,节约人力。利用计算机自动标引文献,编制关键词索引非常迅速,从而大大缩短检索刊物的报道时差。由于这些优点,关键词语言在国外被普遍应用。关键词语言直接使用自然语言,因此用它作为信息检索语言,质量较差。

4) 单元词语言

单元词又称为元词,它是从文献中抽取出来并经过控制处理的,能表达文献主题的最小、最基本的词汇单位。它可以是一个单纯词,也可以是一个合成词。这些词具有一个共同的特点:它们在概念上不能再进一步分解,如再分解,就再也不能表达原来所代表的特定概念,从而失去检索本意。

单元词语言就是经过词汇控制的,能表达文献主题的元词集合。

元词的灵活组配性使它处理文献的能力很强,但在检索中也产生了相当的误检,因此,这种语言现在很少用,它被叙词语言所代替。

5) 叙词语言

(1) 叙词语言的定义。叙词是经过词汇控制的,在标引中用来显示文献主题,在检索中用来构成提问式的一种检索词汇。

叙词这一术语是由美国人穆尔斯在 20 世纪 40 年代末提出的。叙词语言是以受控的自然语言为语词基础,以字顺和分类系统为词汇显示的基本手段,以语词的概念组配为重要特征的一种标引和检索文献的理论方法。

叙词语言采用科技术语作为标识文献主题概念的标目。叙词语言可谓是博采各种信息检索语言之长,吸取了多种信息检索语言的原理和方法。叙词语言是一种采用规范化的单词或词组,由标引人员或读者自行组配,来表达文献(或课题)主题概念的一种后组式索引语言,或者叫做后组式的检索语言。

(2) 叙词的组配原则。叙词所用的词类主要是名词和动名词。叙词的词序都采用自然序列,尽量采用单词作为叙词,利用单词的组配来表达复合概念,使叙词语言的优越性能(多途径检索和自行扩检、缩检)得以发挥。直接采用专指性的词组是受到适当控制的,只是在用单词组配会造成错误和混乱时才采用词组。语词的组配方式有概念组配和字面组配两种。叙词的组配是概念的组配,应当严守概念组配的原则,只有使之符合概念逻辑,才具有检索意义。

概念组配有以下 3 种:

① 概念相交。时钟和收音机是同级词,谁也不属于谁,组配后得时钟收音机。两个或多个同级词组配后,由于不同概念相交,就会形成一个新概念,这个新概念是组配前各概念的下位概念。若用 A 和 B 表示两个不同概念,则概念相交可用图 2-2(a)表示,其中阴影部分表示组配后形成的新概念。

② 概念并列。两个或多个同级词组配后,由于不同概念并列,也会形成一个新概念,例

如,长篇小说和历史小说这两个概念并列后,得到长篇历史小说。和概念相交不同,概念并列后的新概念是组配前各概念的上位概念,如图 2-2(b)所示。

③ 概念限定。刀具和热处理可以组配成刀具热处理。刀具和热处理不是同级词,刀具是事物,热处理是该事物的某一方面,称方面词,它对事物起限定作用。因此将表示事物的叙词和它的方面词组配后,由于方面词概念的限定作用,就会形成一个新概念,它能说明事物的某一方面。概念限定也可用图 2-2(a)表示。

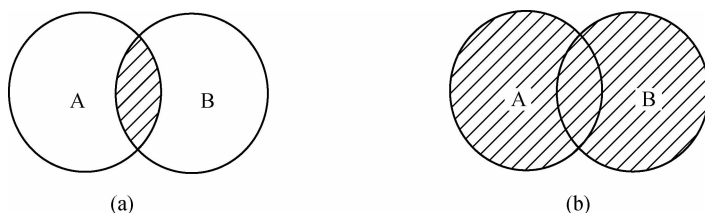


图 2-2 概念组配示意图

在各方面概念组配中,常见的是第一、第三种,特别是概念限定。因此,叙词经过概念组配后,往往能够表达更加复杂的概念,从而提高检索标识的专指性[把标识限制在图 2-2(a)中的阴影部分],以满足特性检索的需要,提高查准率。

(3) 叙词表。叙词语言所对应的叙词表是在吸收了各种词表优点的基础上发展起来的具有现代意义的词表。按其选词的学科范围,可以分为专业性叙词表和综合性叙词表。我国的《汉语主题词表》就是一部综合性叙词表。《冶金专业叙词手册》《机械工程主题表》等属于专业性词表。

叙词表一般都由一个主表和几个辅表组成。主表是叙词表的主体,包括的内容最全,对每个叙词的著录也比较完备,所以可独立存在。辅表则是为方便主表使用而编制的各种辅助索引,不能单独使用。

① 主表。主表即叙词字顺表,包括全部的叙词和非叙词;叙词条目下列有与该词有关的所有上位词、下位词、被代替词和相关词,从而可以详细了解到这类词的组成情况,同时也可以得到相关文献的“参见”项,以便扩大检索,提高查全率。

② 辅表。辅表包括叙词等级索引、叙词分类索引、叙词轮排索引。叙词等级索引按概念的等级关系排列,即将有上下从属关系的词按等级集中在一起,这样的一组词称为词族,最上位的一级词称为族首词,全部款目按族首词的字顺排列。叙词分类索引是按学科类别划分叙词,然后按类汇集,同类下的叙词按字顺排列,因此从学科角度可以查到该类的全部叙词,也可以从已知的类号转换成相应的叙词,以便从主题途径检索文献。叙词轮排索引就是将叙词中的复合词,按其包含的单词做轮番排列。一个叙词词组由几个单词构成,便可轮排几次,在几处都能查到它。

以下是由《工程索引》EI 的《工程叙词表》中摘录下来的主表的著录格式及几种受控主题词表中的参照符号对照。

《工程叙词表》的主表是由 9 400 个叙词和 9 000 多个非叙词按词的字顺排列而成,在叙词款目下列出一系列参照符号,在非叙词款目下列出 EI 的叙词。其著录格式如下。

叙词款目: Acoustic imaging

SN: Production of read-time images of the internal structure of opaque objects

DT:Ptrdates 1975
 UF:Acoustic lenses
 BT:Imaging technigues
 NT:Ultrasonic imaging
 RT:Acoustic holography
 Acoustics
 Diagnosis
 Lenses

非叙词款目:Acoustic Lenses

Use:Acoustic imaging

其中一些符号的含义见表 2-1。

表 2-1 《工程叙词表》中符号的含义

符 号	含 义	说 明
Use	用	非正式叙词转为正式叙词
UF	代	只用于正式叙词,指出正式叙词所代替的非正式叙词
NT	分	指出所列叙词下分得更专指的词,这些词均为正式叙词
BT	属	指出所列叙词的上位概念的词
SN	注释	对所列叙词的词义进行说明
DT	日期	所列叙词起用时间
RT	参见	指出与所列叙词相关的叙词

(4) 叙词语言的特点。叙词语言是在体系分类语言、组面分类语言、标题词语言、单元词语言和关键词语言的基础上,扬长避短而发展起来的一种检索语言,它具有很多优异的性能。

① 叙词语言是后组式检索语言。概念组配是叙词语言的核心。叙词经过自由的、广泛的组配后,能形成概念网罗性高的检索标识,表达主题所含的复杂概念,从而能进行特性检索,提高查准率;能形成若干检索标识,索引深度高,从而可以进行多途径检索。

② 叙词表的等级结构,便于增删、修订叙词。科学技术有了新的进展,增添相应的叙词后,就能标引和检索这些新科技的文献。

③ 叙词语言直接用叙词作为检索标识,不必转换,具有语言的直接性。

④ 叙词语言有一套较完整的参照系统,能显示叙词之间的相互关系,可供扩检或缩检用。

⑤ 利用叙词的词族索引和范畴索引,还可以按学科分类进行族性检索。

总之,与现有其他检索语言比较,叙词语言是一种多功能的较完善的检索语言。科技人员都熟悉自己的专业,从叙词表中找到切题的叙词后,只要组配得当,就能大大提高检索效率。

2.2 信息检索原理和手段

2.2.1 信息检索原理与类型

1) 信息检索原理

人类的信息检索行为总是从特定的信息需求开始,并在特定环境和信息检索系统中完成,这里所说的环境包括产生需求的环境、信息检索系统的运行环境和其他制约因素。特定的检索系统包括完成检索过程所需的一定设施和工具,它可以是图书馆、信息中心或信息经济人,也可以是某种工具书(如文摘索引、目录、资料集、手册、词典等)或机读信息源(如各种机读数据库)。

广义的信息检索包括信息的存储和检索两个过程。信息的存储就是将搜索到的一次信息,经过著录其特征(如题名、著者、主题词、分类号等)而形成款目,并将这些款目组织起来成为二次信息的过程。信息的检索是针对已存储好的二次信息库进行的,是存储的逆过程。存储是为了检索,而为了快速而有效的检索,就必须存储。没有存储,检索就无从谈起。这是存储与检索相辅相成、相互依存的辩证关系。

然而,由于职业、知识水平、个人素质,甚至习惯等因素的差异,信息存储人员(标引者)与信息检索用户(检索者)对同一信息的分析、理解也会存在不同。例如,《计算机在生物化学中的应用》一文,标引者可能将其归为“生物化学”类,而检索者则可能在“计算机”类查找该文。这样,标引者与检索者之间发生了标引错位,存储的信息就无法检索到。

怎样才能保证信息能存储且能取出呢?那就是存储与检索所依据的规则必须一致,也就是说,标引者与检索者必须遵守相同的标引规则。

信息存储与检索共同遵循的规则称为信息检索语言。只要标引者和检索者用同一种检索语言来标引要存入的信息特征和要查找的检索提问,使它们变成一致的标识形式,信息的存储过程与检索过程就具备了相符性。相应地,存入的文献也就可以通过信息检索工具(系统)检索出来。如果检索失败了,就要分析检索提问是否确切地描述了待查课题的主题概念,在利用检索语言标引时是否出了差错,从而导致检索提问标识错误。只有检索提问标识和信息特征标识一致时,相关的文献才能被检索出来。

信息检索正是以信息的存储与检索之间的相符性为基础的,如图 2-3 所示。如果两个过程不能相符,信息检索就失去了基础。检索不到所需的信息,存储也就失去了意义。

2) 信息检索类型

掌握了信息检索的原理,就有必要了解信息检索的类型。按照不同的标准,信息检索可划分为不同的类型,其特点也各自不同。

(1) 依检索内容划分。

① 文献信息检索。凡是利用目录、文摘或索引等二次信息查找某一课题、某一著者、某

一地域、某一结构、某一事物的有关信息以及这些信息的出处和收藏单位等,都属于文献信息检索范畴。其检索的结果是文献信息。例如,“设计人行天桥的参考文献有哪些?”便属于该类命题。

② 数据信息检索。凡是利用参考工具书、数据库等检索工具检索包含在文献中的某一数据、参数、公式或化学分子式等,都统称为数据信息检索。其检索结果为数据信息。例如,“某一新型载货汽车的载重量是多少?百公里油耗是多少?”就属于数据检索。

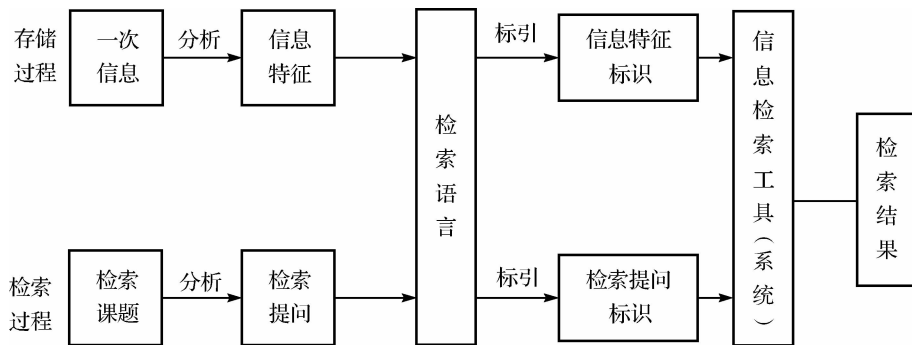


图 2-3 信息检索基本原理示意图

③ 事实信息检索。凡是利用百科全书等检索工具从存储事实的信息系统中查找出特定事实的过程,都称为事实信息检索。其检索结果是基本事实。例如,“世界上最长的斜拉桥是哪座?”“该桥位于什么地方?”“何时建成?”等。

文献信息检索是一种相关性的检索,检索的结果是文献线索,还必须进一步查找才能检索到有关的一次信息;数据与事实信息检索是一种确定性检索,检索的结果是可供用户直接利用的信息。一般情况下,文献信息检索通过二次信息来实现,而数据与事实信息检索则通过 3 次信息来完成。

(2) 按照信息存储和检索的方式和技术划分(详见 2.2.2)。

① 手工检索。手工检索是指人们通过手工的方式来存储和检索信息。

② 计算机检索。计算机检索是指人们利用数据库、计算机软件技术、计算机网络以及通信系统进行的信息存储和检索,其检索过程是在人机协同作用下完成的。计算机从其存储的大量数据中自动分拣出与用户提问相匹配的信息,用户则是整个检索方案的设计者和操纵者。其检索的本质没有发生变化,发生变化的只是信息的载体形式、检索手段、存储方式和匹配方法。

2.2.2 信息检索手段

根据信息检索的基本原理,实现信息检索的基本方式可分为传统信息检索和现代信息检索。传统信息检索简称手检;现代信息检索简称机检。

1) 传统信息检索

传统信息检索是检索人员利用手工检索工具手翻、眼看、大脑思维判别,索取原始文献的一种方式,其优点是:

① 检索条件简单,成本低。

- ② 在检索过程中可以随时获取反馈信息,及时调整检索策略。
- ③ 可对不同的检索工具同时进行对比,从而提高检索质量。
- ④ 可以参阅检索工具中的附图。

其缺点是:

- ① 速度慢、效率低,检出的文献款目必须抄录。
- ② 手工检索工具提供的检索点有限,很难进行多元检索。
- ③ 难以找到涉及几个概念组合的多主题的文献。

手检的方法有以下两种:

(1) 直接检索。这是常用的检索方法,即利用一次文献进行检索。人们常常在需要信息资料时,直接到图书馆查阅各种图书、期刊及其他资料,从中找到所需的信息。

(2) 间接检索。间接检索是利用各种检索工具(目录、题录、文摘)查找信息的方法。这是文献检索最常用的方法。

2) 现代信息检索

现代信息检索是检索人员利用计算机检索系统查找文献的一种检索方式。计算机检索系统包括数据库技术、计算机技术和通信技术。机检可以克服手检的缺点,但机检对设备条件的要求比较高,所需的投资比较大。计算机检索已从单机检索、联机检索发展到今天的网络检索,并正向智能化的方向发展。

(1) 单机检索。单机检索是计算机检索的最初形式,随着计算机存储介质的发展变化,也在不断发生变化。目前主要有计算机磁盘检索、计算机磁带检索和计算机光盘检索 3 种形式。

(2) 网络检索。网络检索是近十几年发展起来的,它目前主要有图书馆局域网的检索、联机检索及 Internet 信息检索 3 种形式。

3) 传统信息检索与现代信息检索的比较

现代信息检索是在传统信息检索的基础上发展起来的,因此两者的检索原理相同,但在网上的信息检索技术与以往的计算机检索技术又有很大的差别,主要有以下几方面的异同点:

(1) 检索语言方面。检索语言就是在检索信息时使用的专门的语言词汇,手工检索与计算机检索使用的检索语言基本相同,不同的是手工检索使用规范语言较多,而计算机检索则可以使用较多的自然语言,Internet 检索一般使用自然语言。

(2) 概念的组配方面。手工检索无检索运算符,检索时各种概念之间用人脑进行组配,用计算机检索时,则是利用各种逻辑运算符进行概念的组配,网上检索概念组配方式还不成熟,目前一般只用布尔逻辑运算符进行组配。其中计算机检索的概念组配最严谨,手工检索的概念组配最灵活。

(3) 检索入口方面。检索入口就是手工检索工具中的检索途径,是根据信息的某种特征编制的索引进行的检索,如题名、著者等。手工检索的检索入口较少,而计算机检索和网络检索的检索入口较多而且非常灵活,计算机检索过程中一个信息的所有特征都能作为检索入口,如年代、语种、文献类型、各种代号等,检索时还可以多途径同时进行检索。

(4) 检索结果方面。手工检索的检索结果查准率最高,误检率最低,而查全率较低。计算机检索的结果查准率要低于手工检索,误检率要高于手工检索,查全率则较高。Internet 检索结果误检率和查全率也较高。

2.3 信息检索研究范围与对象

信息检索作为一门学科,有它自己的研究范围和对象,也有自己的理论、方法和技术。从总体上看,信息检索的研究对象是比较明确具体的,研究范围广泛而边界有些模糊,理论和方法已经逐渐形成体系。

2.3.1 信息检索的研究范围

信息检索的研究范围包括一切与信息存储与检索有关的系统、过程、理论和方法。一切可供存储和检索利用的信息类型,如文献、数据、事实、知识、声音、图形等;各种信息检索系统及其运行过程,如信息采集、标引、组织、存储、处理、匹配、传送等各种过程中使用的方法;以及在信息检索实践和研究的基础上形成的各种理论和假设,均包括在这个范围内。信息类型侧重于文献,其次是数据和事实。

2.3.2 信息检索的研究对象

信息检索的研究对象主要有以下几方面:

(1) 信息检索理论。它主要包括检索语言与标引理论、信息检索的数学模型、知识表示理论、相关性理论,以及有关的哲学问题。

(2) 信息检索系统。它是实现信息检索的物质基础,是现实的研究对象,主要研究其结构、功能、演变,其设计开发技术、管理维护技术和评价技术,还研究它与其他信息系统乃至整个外部世界的关系。其中,数据库是信息检索系统的核心部分之一。数据库的建造和维护是非常重要的信息技术。数据库的质量直接影响信息检索系统的功能和效率。

(3) 计算机信息检索。它涉及许多计算机设备、软件技术、存储技术革新、检索技术、系统管理和经营知识、市场营销技术等,是一门综合性很强的技术。

(4) 检索策略与方法。它是用户从信息检索系统中获取有关信息所必需的。好的检索策略是检索成功的必要前提,计算机的应用为检索策略和方法的发展提供了有力的支持。近30年来,涌现了许多新颖而有效的检索技术和方法,如布尔检索、位置检索、截词检索、加权检索、聚类检索等,人们利用、研究、评价和完善现有的各种检索策略和方法,研究开发新的更有效的策略和方法。

(5) 用户研究与培训。用户是信息检索系统的生命,无论是系统的研制开发、管理维护、功能和服务的扩展,还是系统评价,都离不开用户研究工作。用户培训是用户研究的继续,是与用户建立紧密联系和发展新用户的一种非常有效的措施。此外,还有自动标引、自动分类和自动摘录,以及相关设备等研究领域。

2.4 信息检索方法、程序、策略及效果评价

信息检索是根据课题的要求,利用各种有关的检索工具,按照一定的检索程序和方法,

采用各种检索策略查找信息的过程。

2.4.1 信息检索方法

常用的检索方法如下：

1) 追溯法

追溯法是指在没有检索工具的情况下，人们利用书、刊之后的引用或参考文献查找资料的方法。

2) 常用法

常用法这是指利用各种类型检索工具，通过技巧查找资料的方法。常用法又可分为顺查和倒查两种。

顺查法：即由远而近的查找方法。如已知某研究课题的起始年代，需要了解它的全面发展情况时，可从最初研究的年代开始，由远及近地进行查找。

倒查法：也称为叫逆查法。和顺查法相反，倒查法是由近及远的查找方法，这种方法常用于正在研究课题的过程中。

3) 循环法

循环法又称为分段法，这是前两种方法的结合，即首先利用检索工具查找一部分资料以后，为了了解其前因后果，再利用这部分资料的引用文继续追查的方法，利用这种方法能够保证所查的资料较全面、系统。

2.4.2 信息检索程序

查找信息的过程如下：

1) 分析课题

信息检索之际，首先要对课题进行充分的了解，掌握课题有关的基本知识，详细摸清课题所需资料范围及要求，以及弄清课题有关的名词、术语等。

2) 选择检索方法

根据检索人员已掌握的资料情况，确定采用何种检索方法。如果是新立的课题，一般采用顺查法和追溯法结合查找，正在进行中的课题一般用倒查法。

3) 确定检索途径

在检索过程中，根据所掌握的有关课题的资料，确定检索途径。一般情况下，所研究的课题较专深、单一时，常用主题途径；课题所涉及的内容较复杂时，常用分类法。

4) 查找信息线索

在确定检索途径之后，根据检索标识，通过有关索引进行信息检索，查找信息线索。

5) 索取原始信息

通过检索获得信息线索后，确认有一定的参考价值，需要详细查阅信息的全文时，即可根据文摘指出的信息出处，向有关部门索取原始信息。

2.4.3 检索策略

信息检索总是根据一定的课题进行,但并不是有关该课题内容的全部信息都需要。要想在数量庞大、类型复杂、内容专深、文种多样的信息海洋中迅速、准确地查找到满足要求的信息,必须制定查找信息的策略。

下面仅介绍手工检索策略的制定方法。

1) 先国内后国外

首先利用国内检索工具或用中文编辑的报道外文的检索工具入手检索。这样,检索者从自己熟悉的文字可以很容易地掌握课题的概况,并可以从中选择出切题的关键词、主题词,为进一步查找外文检索工具提供方便。

2) 检索时多选几个同义词、近义词

信息语言复杂多变,作者使用的语词不尽相同,为了避免漏检,检索时尽量多选几个同义词、近义词作为检索词。

3) 巧妙地利用上、下位词的关系

当要求查全率高而不计查准率时,可利用上、下位词进行检索。

4) 尽量避免从字面出发选择检索词

在检索时,尽量从内容上选词,以避免有关问题因选词不当被漏检。

5) 变换检索词的词序

某些课题,在检索中反复查找而没有结论时,应考虑倒置词序的可能。检索词序的准确与否是检索成败的关键。

6) 利用参见系统,扩大检索范围

检索工具中的参见(see also)项,把内容有关的主题联为一体,以便族性检索。例如,CA中每一大类末尾均列有与该类有关的全部相关大类和具体相关文摘。绝大多数检索工具中的主题词表、主题索引或文摘部分中均没有参见项,检索时分别查找各有关的标题,以获取某一课题的全部信息。

7) 各种索引配合使用

要想把各种索引配合使用,首先熟悉各种索引的结构、特点、款目内容及其功能。例如,CA中的化学物质索引与分子式索引、登记号索引、环系索引、杂原子索引关系十分密切。如果从化合物登记号索引入手检索,只能获得化合物的名称与分子式而查不出文摘号,必须转查化学物质索引才能获得信息线索。又例如,对商品名称与俗称,必须首先查索引指南。类似的情况在检索时常常遇到,只有熟练掌握各种检索工具的结构体例、编排特征、各索引款目结构及其相互关系,检索起来才能得心应手,取得满意的效果。

2.4.4 检索结果的评价方法

利用检索工具或检索系统查找信息时,检索结果并不是在任何情况下都能如愿以偿的。其主要原因是:从目前的检索系统看,它只能处理信息标识,而信息的标识只能表示信息的中心主题,不表示信息的次要主题。即系统只能回答信息提问而不一定能满足信息需求。

从检索者来看,一般只能大概地表述自己的信息提问,而不能确切地衡量信息与提问的切题程度。再加上,检索课题的性质、范围、深度不断变化,查找信息的方法尚未形成固定模式,补充词表的出版不及时,人们不可避免地受到知识面的局限等,致使信息标引与检索失误在所难免。

在目前情况下,对检索结果的分析评价方法有以下几种:

1) 查全率

查全率是对所需信息被检出程度的量度。

2) 查准率

查准率是衡量检索系统拒绝非相关信息的能力。

查全率和查准率均可以用公式来表示:

$$\text{查全率}(R) = \frac{a}{a+c} \times 100\%$$

$$\text{查准率}(P) = \frac{a}{a+b} \times 100\%$$

式中, a 表示被检出的相关信息, b 表示被检出的非相关信息, c 表示未检出的相关信息。

除此之外, $c/(a+c) \times 100\%$ 为漏检率,是查全率的补数; $b/(a+b) \times 100\%$ 为误检率,是查准率的补数。

一般来说,查准率与查全率之间存在反变关系。当某一课题的查全率与查准率处于某一比例关系时,继续提高查全率,查准率会降低。

3) 相关性

相关性是用户判断信息与信息需求之间关系的标准。信息系统回答的不是用户提问本身,而是检索式表达的信息提问。因此,检出的是与信息提问相关的信息,但不一定是对用户切题的信息。