

信息搜集与处理技术

xinxi souji yu chuli jishu

主编：杨和平 陈章 谢必武

 经济日报 出版社

信息搜集与处理技术

杨和平 陈 章 谢必武 主编



经济日报出版社

图书在版编目(CIP)数据

信息搜集与处理技术 / 杨和平、陈章、谢必武主编. -- 北京: 经济日报出版社, 2011.1

ISBN 978-7-80257-272-0

I. ①信… II. ①杨… ②陈… ③谢… III. ①信息—搜集—高等学校: 技术学校—教材 ②信息处理—高等学校: 技术学校—教材 IV. ①G202

中国版本图书馆 CIP 数据核字(2010)第 246639 号

《信息搜集与处理技术》

作 者	杨和平 陈 章 谢必武
责任编辑	肖小琴
责任校对	王 瑛
出版发行	经济日报出版社
社 址	北京市宣武区白纸坊东街 2 号(邮政编码: 100054)
电 话	010-63567690(编辑部) 63567683(发行部)
网 址	www.edpook.com.cn
E-mail	jjrb58@sina.com
经 销	全国新华书店
印 刷	重庆联谊印务有限责任公司
开 本	710mm*1000mm 1/16
印 张	14.5
字 数	250 千字
版 次	2011 年 1 月第 1 版第 1 次印刷
书 号	ISBN 978-7-80257-272-0
定 价	28.00 元

版权所有 盗印必究 印装有误 负责调换

编委会名单

主 编：

杨和平

陈 章

谢必武

副主编：

张 智

潘久政

文 锋

编 委：

杨和平

陈 章

谢必武

张 智

潘久政

文 锋

熊 江

周长春

袁德桢

张南宾

内 容 简 介

本书根据工学结合人才培养及信息时代对人才的要求,结合高职高专学生实际情况编写而成,本教材共分信息基础和信息搜集处理项目两大部分,信息基础主要介绍有关信息源、信息、信息检索的含义,信息检索基础知识;信息搜集处理项目分数字图书馆资源、网络信息资源检索、数据库检索、信息处理与综合利用等六个检索项目,每个项目后有学习任务。按照任务驱动、工作过程、工学结合选取教材内容,充分体现了真实性、针对性、通用性、操作性。本教材在体例上进行了全面创新,重点突出,深浅适度,实用性强。

本书适用于高职高专院校的各专业学生,亦可供广大教师、科技工作者及信息工作者学习参考。

前言

当今信息时代,全球信息化已经形成,信息、物质、能源构成了人类社会发展的三大资源,在竞争激烈的今天,信息意味着财富、机遇和发展,谁先掌握信息,谁就具有竞争优势,如何以最快的速度查找和利用各种信息,已成为提高生存、竞争和发展能力的关键,获取信息、使用信息的能力,已经是信息社会劳动者必须掌握的终身技能。

信息搜集与处理能力作为高职学生的第一方法能力,在人才培养中具有重要地位,为使高职高专院校学生掌握信息搜集与处理技术的方法与技能,了解信息的收集、整理、加工和利用在学习、工作、创新中的作用和意义,培养学生敏锐的信息意识及高度的信息素养,从浩瀚的信息资源中迅速、准确地获取各种信息,利用信息分析问题和解决问题,独立工作、开拓思维、勇于创新,提高综合竞争能力。根据教育部《关于全面提高高等职业教育教学质量的若干意见》(教高[2006]16号)、2010年9月召开的全国高职改革与发展工作会及《关于加强高职高专教育教材建设的若干意见》的精神,体现工学结合特色,我们编写了《信息搜集与处理技术》这一教材。

结合高职高专教育的实际情况,本教材在内容、体例上进行了全面创新,在编写过程中按照任务驱动选择教学内容,以职业单元设置与职业行动过程为取向,以工作过程为参照系,以此配合各专业完成对人才方法能力的培养。通过设计学习型检索任务,来组织关于信息搜集与处理所需要的基础知识的重构、使学生掌握基本检索工具的使用,由简到繁,由易至难,循序渐进地培养学生信息搜集与处理能力,在完成检索任务的过程中,实现课程教学目标。在体例上力求体现新颖、重点突出、深入浅出和实用够用的特点,突出科学性、针对性、通用性、可行性和易操作性,选取了与高职生密切相关的毕业论文写作和就业信息搜集等内容,以真实信息搜集工程为基础,对计算机信息检索、数据库检索、网络信息资源检索等内容做了较为详细的介绍,在数字图书馆、信息处理方面进行了较深入的探讨,以尽可能提高学生信息素养及信息搜集与处理技能。

本教材共分信息基础和信息搜集处理项目两大部分,信息基础主要介绍有关信息源、信息、信息检索的含义,信息检索基础知识;信息搜集处理项目分六个检索项目,项目一介绍传统图书馆与数字图书馆资源的利用,项目二介绍检索工具书与参考工具书,项目三介绍网络信息资源检索,项目四介绍报刊和数据库检索,项目五介绍特种文献及其检索,项目六介绍信息处理与综合利用。每个项目后有检索任务,以增强学生获取和利用信息的能力。

全书由重庆三峡职业学院杨和平教授、陈章副教授、谢必武教授担任主编,各部分编写人员如下:谢必武教授编写基础一、项目五,杨和平教授、陈章副教授编写项目六,张智教授编写项目一,潘久政教授编写基础二,熊江、周长春副教授编写项目二,张南宾副教授、袁德彪副教授、文锋副教授编写项目三、项目四。全书由谢必武教授统稿。

本书参考了前辈、同行的教材、专著和研究成果,在编写过程中得到了经济日报出版社及编者单位的大力支持,在此一并表示衷心感谢。

本教材在编写体例和内容组织上与传统的信息搜集与处理教材相比有很大的改变,这仅是一种尝试,由于编者水平有限,时间仓促,教材中错误和疏漏之处在所难免,敬请指正。

编者

2010年10月18日

目 录

I 信息基础	1
基础一 信息概述	1
[学习目标]	1
[知识要点]	1
一、信息源	1
二、信息检索	14
小 结	21
学习任务	22
思考与讨论	22
基础二 信息检索	23
[学习目标]	23
[知识要点]	23
一、信息检索语言	23
二、信息检索方法、途径及步骤	26
小 结	35
学习任务	35
思考与讨论	36
II 信息搜集与处理项目	37
项目一 传统图书馆与数字图书馆资源的利用	37
[学习目标]	37
[知识要点]	37

一、传统图书馆资源的利用	37
二、数字图书馆资源的利用	46
小 结	58
学习任务	58
思考与讨论	58
项目二 检索工具书与参考工具书	59
[学习目标]	59
[知识要点]	59
一、工具书概述	59
二、检索工具书	70
三、参考工具书	75
小 结	87
学习任务	87
思考与讨论	88
项目三 网络信息资源检索	89
[学习目标]	89
[知识要点]	89
一、Internet 基本知识	89
二、网络信息资源检索方法及策略	96
三、搜索引擎	101
小 结	116
学习任务	116
思考与讨论	116
项目四 报刊和数据库检索	117
[学习目标]	117
[知识要点]	117
一、报刊检索工具	117

二、数据库检索	129
小 结	172
学习任务	173
思考与讨论	173
项目五 特种文献及其检索	174
[学习目标]	174
[知识要点]	174
一、标准文献及其检索	174
二、专利文献及其检索	183
小 结	197
学习任务	198
思考与讨论	198
项目六 信息处理与综合利用	199
[学习目标]	199
[知识要点]	199
一、信息处理	199
二、信息的综合利用	203
小 结	217
学习任务	217
思考与讨论	217
主要参考文献	218

I 信息基础

基础一 信息概述

[学习目标]

能正确区分文献、情报、信息、知识的概念及其他它们之间的关系,明确信息源的基本类型。

[知识要点]

信息的含义、特征,信息源的含义、类型及特点,信息检索的含义、类型、意义和作用。

信息如同空气一样,跟我们生活息息相关。信息在包围着我们,人的五官、身体所感受到的可以说首先都是信息。信息普遍存在于自然界、人类社会和人的思维之中。从微观世界到宏观世界,从无机界到有机界,从植物到动物,从机器到人,都能产生信息,也能接收信息。信息联系着你和我,联系着大家,联系着世界万物。人类社会实践也在源源不断地产生出新的信息。正因为信息无所不在,无时不有,所以,它也成为许多学科的研究对象。

一、信息源

(一)信息源的含义

1.信息的含义、特征

(1)信息的基本含义 “信息”一词最早出自南唐诗人李中《暮春怀故人》中的“梦断美人枕信息,目穿长路倚楼台”,原意指消息。虽然人类自古以来就在利用信息,但人们开始认真研究信息,是从20世纪20年代初开始的,相关学科是到1950年后才相继出现的。从美国数学家克劳特·申农(C.E.Shannon)创立了狭义信息论以后,“信息”一词才成为一个科学的概念。但对于信息的含义,至今仍是众说纷纭,莫衷一是,各种信息定义都从不同的侧面反映了信息的某些特征,尚无一种可以为社会各界一致所接受的涵盖其全部内容的科学的定义。下面我

们介绍几种常见的信息定义的方法。

在日常生活中,人们所说的“信息”,是指音信、消息和情况,是人们在互相交流中要告诉对方的某种内容,因此,很多人长期以来就把信息看作是消息的同义语,简单地把信息定义为能够带来新内容、新知识的消息。在西方国家的文字中,信息一词来源于拉丁文“Information”,大致有以下几种解释:消息、情报、知识、见闻、通知、报、事实、数据等等。这些解释基本上都是从字面上来理解的。例如,我们说“某某杳无音信”指的是没有这个人的任何消息;利用卫星转播球赛时,是转播世界某地正在进行的球赛实况。总之,这里的信息是指交流双方要告诉对方的音信、消息和情况。

信息论的奠基人、美国数学家申农(C. E. Shannon)(1916~2001)在《在噪声中的通信》一文中从通信学的角度将信息定义为:信息是不确定性的减少或消除,是关于环境事实的可通信的知识,信息是通过各种形式,包括数据(字母、符号和数字)、代码、图形、报表、指令等等反映出来的。这一定义发现,信息的含义要比消息的含义广泛得多,不仅消息、情报是信息,指令、代码、符号语言、文字等,一切含有内容的信号都可称为信息。

控制论的创始人、美国科学家维纳(N. Wiener)从控制论的角度提出了信息定义,“信息是人们在适应外部世界并且使这种适应反作用于外部世界的过程中同外部世界进行交换的内容的名称”,“信息是有秩序的量度”,并指出“信息就是信息,不是物质,也不是能量”。根据维纳的说法,物质、能量和信息是相互有区别的,是人类社会赖以生存、发展的三大基础——世界由物质组成,能量是一切物质运动的动力,信息是人类了解自然及人类社会的凭据。信息对于物质而言具有相对独立性:信息不遵循质量守恒定律,其性质和内容与物质载体的变换无关;同样,信息在传递和转换过程中也不服从能量守恒定律,信息可以共享而能量不能共享,信息效用的大小并不由其消耗来决定。然而,信息与物质、能量又存在着密切的相互依存关系:物质、能量和信息这三者中,能量和信息皆源于物质,任何信息的产生、表述、存储和传递都要以物质为基础,也离不开能量。从另一方面来说,物质运动的状态和方式需要借助信息来表现和描述,能量的转换与驾驭也同样离不开信息。维纳把信息的定义引入了控制论,他在《人有人的用处——控制论和社会》中说:“信息这个名称的内容就是我们对外界进行调节并使我们的调节为外界所了解时而与外界交换来的东西。”

在图书馆学和情报学领域,信息被定义为事物(thing)或记录(record)。美国学者巴克兰德(M. Buckland, 1991)认为,许多事物都可以是信息,如文本、图片、录音

磁带、博物馆陈列品、自然物体、实验、事件等。但这一定义过于宽泛,未能将信息、信息载体和信息资源区分开来。另一位美国学者萨克利夫(J. Tague-Sutcliffe)则认为,信息是依赖于人类的概念化和理解能力的无形的东西,对于记录而言,它所包含的有形的字与图片等是绝对的,但它所包含的信息对于读者(或用户)则是相对的。信息是读者通过阅读或其他认知方法处理记录所理解的东西,它不能脱离外在的事物或读者而独立存在,它是与文本和读者以及记录和用户之间的交互行为相关的,是与读者大脑的认知结构相对应的东西。

情报学界引述最多的,是系统工程学家的看法。信息论的创始人申农从通信系统理论的角度把信息定义为用来减少随机不确定性(uncertainty)的东西。也就是说,信宿(信息接受方)未收到消息前不知道信源(信息产生方)发出什么信息,只有在收到消息后才能消除信源的不确定性。如果没有干扰,信宿得到的信息量与信源的不确定性相等。申农的看法,被认为是对信息认识的重大进展,因为他推导出了信息测度数学公式,标志着信息科学进入了定量研究阶段。

信息是系统的组织程度或有序程度的标记。该定义是通过与热力学中的概念“熵”进行类比推理而来的。人们常用熵来表示系统的无组织状态或无序状态,这里,信息作为与“熵”相对的概念提出来,成为“负熵”的同义词。

信息是数据处理的结果。这个定义是从信息处理的角度讲的。它把未经过加工的原始资料,无论它是数字、文字,还是符号、图像、信号,都称为数据,而把信息理解为加工原始资料后得到的、便于使用的结果。

“信息”概念的广泛应用,引起许多哲学工作者对信息本质的探讨,使“信息”从一个科学概念上升到一个哲学范畴。他们认为事物的特征通过一定的媒介或传递形式使其他事物感知,这些能被其他事物感知的、表征该事物特征的信号内容即为该事物向其他事物传递的信息。所以,信息是事物本质、特征、运动规律的反映。不同的事物有不同的本质、特征、运动规律,人们就是通过事物发出的信息来认识该事物,或区别于其他事物。

一些知名词典的编纂者也试着给出了信息的定义,如《辞海》的注释是:“信息是指对消息接受者来说预先不知道的报导。”《简明自然辩证法词典》的解释是:“信息一般泛指我们所说的消息、情报、指令、数据、信号等有关周围环境的知识。”

据不完全统计,信息的定义有 100 多种,它们分别从不同层次、不同侧面揭示了信息的特征与性质。之所以对信息有如此多的不同定义,是由于各自所强调的重点不同,反映的角度、利用的方法、用于研究的目的不同。由于信息本身的复

杂性尚未被完全认识,而且信息科学是一门新兴学科,因此它有许多分枝学科,对于它的内涵与外延,人们还在不断探索中。

就目前已有的探索,可以将其归纳为广义和狭义两类。

广义的信息指的是客观世界中各种事物的存在方式和它们的运动状态的反映,简单地说,信息就是客观世界一切事物存在和运动所能发出的各种信号和消息。

狭义的信息指的是能反映事物存在和运动差异的、能为某种目的带来有用的、可以被理解或被接受的消息、情况等。Information 一词理解为狭义的信息时,常被译为情报。

中国国家标准 GB4894-85 关于信息的定义则将两类表述合并为:信息是物质存在的一种方式、形态或运动状态,也是事物的一种普遍属性,一般指数据、消息中所包含的意义,可以使消息中所描述事件的不定性减少。

信息无所不在,可以感知,但它不是事件和物质本身,信息是事物或事物之间不确定性的量度,掌握的信息越多,所需认识的事情的不确定性就会减少;信息和物质不同,它没有物质那样的实体,而是一种抽象的又无所不在的东西;信息是客观事物存在的方式或运动状态,以及关于客观事物存在方式或运动状态的陈述;信息是原料,经过提炼加工后,能成为有用的知识;数据、资料、语言、颜色、声音、光和光谱、自然界的信号等等都可以传递信息;信息是已知的知识,又是未知知识的先导波段,或者,把这种先导性的波段称之为信息波亦未尝不可。

(2)信息的类型与特征

①信息的类型。信息广泛存在于自然界、生物界和人类社会。信息是多种多样,多方面、多层次的。了解信息的类型不仅有助于我们加深对信息内涵及其特征的认识,也有助于丰富信息检索的知识。根据不同的角度可以将信息分为不同的类型。

从产生信息的客体的性质来分,可分为自然信息(瞬时发生的声、光、热、电、形形色色的天气变化、缓慢的地壳运动、天体演化……)、生物信息(生物为繁衍生存而表现出来的各种形态和行为,如遗传信息、生物体内信息交流、动物种群内的信息交流)、机器信息(自动控制系统)和(人类)社会信息。一切存在都在进行着某种形式的表达,只不过人类的表达要丰富得多,因为他们的存在内容更丰富。社会信息就是指人与人之间交流的信息,既包括通过手势、身体、眼神所传达的非语义信息,也包括用语言、文字、图表等语义信息所传达的一切对人类社会运动变化状态的描述。按照人类活动领域,社会信息又可分为科技信息、经济信息、

政治信息、军事信息、文化信息等。

以信息所依附的载体为依据,可分为文献信息、口头信息、电子信息、生物信息等。文献信息,就是文献所表达的内载信息,以文字、符号、声像信息为编码的人类精神信息,也是经人们筛选、归纳和整理后记录下来的信息(recorded information),它与人工符号本身没有必然的联系,但要通过符号系统实现其传递。文献信息也是一种相对固化的信息,一经“定格”在某种载体上就不能随外界的变化而变化。这种性质的优点是,易识别、易保存、易传播,使人类精神信息能传于异地、留于异时,缺点是不能随外界的变化而变化,固态化是文献信息老化的主要原因之一。

口头信息指存在于人脑记忆中,通过交谈、讨论、报告等方式交流传播的信息。它反映了人们的思考、见解、看法和观点,是推动研究的最初起源。口头信息具有出现早、传递快、偶发性强的特点,但缺乏完整性和系统性,大部分转瞬即逝,一部分通过文献保存,一部分留存在人类的记忆中,代代相传而称为口述回忆或口碑资料(oral tradition)。作为信息留存的一种形式,口头信息无时不在,无处不有,承载着人类的知识、经验和史实,是一种需要重视和开发的极为丰富的资源。

电子信息是计算机技术、通信技术、多媒体技术和高密度存储技术迅速发展的产物。在本书,电子信息特指通过电视、计算机、数据库、网络等传播的瞬时信息,以区别于相对固化的电子文献(光盘等)。这是当今发展最快、最具应用价值和发展前途的新型信息源。

信息分类还有其他划分方法,如以信息的记录符号为依据,可分为语音信息、图象信息、文字信息、数据信息等;以信息的运动状态为依据,可分为连续信息、离散信息;以信息的加工层次而论,可分为初始信息(或“感知信息”、“原生信息”)和再生信息(或“二次信息”、“三次信息”),后者是对初始信息进行加工并输出其结果的形式,也是信息检索的主要对象。

②信息的特征。信息虽然是一种资源,但它是不同于物质和能量的一种特殊的资源。这种资源具有以下主要特征:

客观真实性:客观、真实是信息的最重要的本质特征,这也是信息的生命所在。

传递性:传递是信息的基本要素和明显特征。信息只有借助于一定的载体(媒介),经过传递才能为人们所感知和接受。没有传递就没有信息,便谈不上信息的效用。

信息的可传递性,一方面是指信息可以通过一定的形式进行传递,另一方面,获得信息必须经过传递。电子计算机技术、现代通信技术和人造卫星技术相结合,使人们可以利用计算机网络在极短的时间内将大量的信息传递到地球上的每一个角落。因信息的传递性与物质的传递性不同,信息的占有者不会因传递信息而失去信息。一般来说,也不会因为多次使用而改变信息的价值。

依附性:信息的价值在于其知识性和技术性。人类所交流的都是各种有意义、有价值的信息,动植物、矿物等都是具体有形的,而信息却是无形的,必须借助物质载体才能得以存在和传递,其中,声波、光波、电磁波这些无形物质载体是信息的第一载体;纸张、胶卷、磁带、磁盘等有形的物质载体是第二载体。没有载体的信息是不存在的。

独立性:信息无论以何种方式记载在何种载体上,也无论这种载体在空间和时间上如何传播,它的内容是不变的。文本、图像、数字、语言等各种形态的信息均可互相转换。

不灭性:信息的不灭性与物体和能量的不灭性不一样。一个杯子被打碎了,构成杯子的陶瓷其原子、分子没有变,但已不成为一个杯子。又如能量,我们可以把电能变成热能,但变成热能后电能已经没有了。而信息的不灭性是一条信息产生后,其载体可以变换,可以被毁掉,如一本书、一张光盘,但信息本身并没有被消灭。

共享性:同一信息可以在同一时间由两个或两个以上的使用者使用,便是信息的共享性。这也是信息区别于物质和能量的主要特性,信息不属于特定的占有对象,可以为众多的人们共同享用。一件物品转赠成为他人的所有物时,就不再属于原主了。而双方交流信息,两者都有得无失,一个理论、一项发明公诸于世,谁都可以使用。特别是在计算机网络技术有了迅速发展以来,信息的共享性问题就日益突出了。信息的这个特点也带来了一些消极影响,如信息的侵权行为也较其他行为更为复杂和特殊。

时效性:信息的功能、作用、效益都是随着时间的延续而改变的,这种性能即信息的时效性。时效性是时间与效能的统一性,它既表明信息的时间价值,也表明了信息的经济价值。一个信息如果超过了其价值的实用期就会贬值,甚至毫无用处。例如国际金融市场信息的时滞一般不应超过6小时,否则,过了时限的信息,再详尽也只是昨日的黄花,非但无用,反而会使人作出错误的决定。大多数经济活动过程都很短暂,如国际股票交易市场的变化仅仅几秒钟;又如战争时的信息,敌方的信息在某一时刻有非常重要的价值,可以决定战争或战役的胜负,但

过了这一时刻,这一信息就变得毫无用处。所以说,信息有非常强的时效性。

但某些信息的时效性却表现在越古老的信息越有价值,如考古研究等。

可处理性:信息的可处理性包括多方面内容,如信息的可拓展、可引伸、可浓缩等。这一特征使信息得以增值或便于传递、利用。同时信息可以廉价复制,可以广泛传播。信息的复制不像物体的复制,一条信息复制成百万条信息费用十分低廉。尽管信息的创造可能需要很大的投入,但复制只需要载体的成本,可以大量地复制,广泛地传播。

可转化性:信息犹如物质、能源,在一定条件下,经过开发利用,可以转化成巨大的物质财富。当今信息时代,谁掌握的信息多,谁就是强者,发达国家不仅出口产品,更重要的是出口信息。正是这一性质,构成了本课程开设价值的基点。

2.信息源的含义

信息源,顾名思义,就是信息的来源,信息源是产生信息的源头。由于信息的含义十分宽泛,信息源的定义也因学科领域的不同而有不同的解释。从传播学的角度看,“传播的来源是指生成、制作和发送信息的人,也可以是群体——指发生信息的部门或机构。”

在图书情报领域,信息源被解释为“人们在科研活动、生产经营活动和其它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”,又可进一步分为文献信息源和非文献信息源(包括口头信息源、实物信息源)。

依据信息源的层次及其加工和集约程度,信息源可分四次:所有物质均为一次信息源,也称本体论信息源,从一次信息源中提取信息是信息资源生产者的任务;二次信息源也称感知信息源,主要储存于人的大脑中,传播、信息咨询、决策等领域所研究的也主要是二次信息源;三次信息源又称再生信息源,主要包括口头信息源、体语信息源、文献信息源和实物信息源四大类型,其中又以文献信息源(包括印刷型和电子型文献信息源)最为重要;四次信息源也称集约信息源,是文献信息源和实物信息源的集约化和系统化,前者如档案馆、图书馆、数据库,后者如博物馆、样品室、展览馆、标本室等。

各种定义都是从特定的角度来解释的,强调的侧重点不同而已。简而言之,信息源可看作是产生、持有和传递信息的一切物体、人员和机构。是指信息传递过程中的信息发送端或生成端的总称,它是信息产生的源头。

(二)信息源的类型及特点

广泛存在着的多种类型的信息源,既是信息工作的基础,又是信息工作的对