

高等医药院校规划教材

(第2版)



医学信息 检索与利用

主 编 曹洪欣

副主编 彭 骏



第二军医大学出版社

Second Military Medical University Press

高等医药院校规划教材

医学信息检索与利用

(第2版)

主 编 曹洪欣

副主编 彭 骏

编 者 (按姓氏拼音字母顺序排列)

曹洪欣 范吉莲 惠朝阳

陆 敏 彭 骏 邱君瑞

王 侠 徐 维 叶 倩

张静昌

第二军医大学出版社

内 容 简 介

本书广泛吸收了国内外信息检索与利用的最新研究成果,结合当前我国信息环境的特点以及医药卫生工作者的信息需求,以信息检索的基本知识为基础,以国内外经典和重要的生物医学信息源为主线,以信息利用为重点,按照检索基础→信息资源→资源利用的思路,系统阐述了信息检索基本原理、国内外生物医学信息数据库的使用方法和技巧、医学科技论文(含学位论文)写作以及医学情报调研和循证医学等内容。全书案例丰富,资料翔实,具有较强的实用性和针对性,既可作为高等医学院校研究生、本科生教学用书,也可供广大医药卫生工作者参阅。

图书在版编目(CIP)数据

医学信息检索与利用/曹洪欣编著. —2 版. —上海: 第二军医大学出版社, 2008. 8

ISBN 978-7-81060-854-1

I. 医… II. 曹… III. 医药学-情报检索 IV. G252.7

中国版本图书馆 CIP 数据核字(2008)第 065298 号

出 版 人 石进英

责任编辑 王 楠

医学信息检索与利用(第 2 版)

主 编 曹洪欣

第二军医大学出版社出版发行

上海市翔殷路 800 号 邮政编码: 200433

发行科电话/传真: 021-65493093

全国各地新华书店经销

江苏句容排印厂印刷

开本: 787×1092 1/16 印张: 21.5 字数: 535 千字

2008 年 8 月第 2 版 2008 年 8 月第 1 次印刷

印数: 1~5 000 册

ISBN 978-7-81060-854-1/G·072

定价: 39.00 元

前 言

自 1984 年原国家教委颁发《关于在高等学校开设“文献检索与利用”课的意见》后,该课程已成为国内各高校培养大学生信息素质与能力的基础性课程,对学生吸取新知识、改善知识结构、提高自学和研究能力、发挥创造才能起到了显著的促进作用。

第二军医大学的检索课程教学始于 20 世纪 80 年代初,现有正式列编的医学信息学教研室,“医学信息检索与利用”课程已纳入全校研究生、本科生所有专业层次以及任职教育的必修课程体系。课程内容建设也经历了从简单文献检索工具利用到紧跟时代发展步伐、完整构建课程内容体系的过程。2006 年,第二军医大学出版社正式出版了由我校范吉莲等编写的《医学信息检索与利用》教材。

该教材出版发行后,得到了广大读者及专家学者的一致好评,并被许多医学院校选作信息检索课程教材,在教学实践中取得了较好效果。但随着社会、经济、科技和社会信息化程度以及院校教学改革的不断深入与发展,教学内容也需要进一步完善与充实。为此,我们重新组织有关人员在该版的基础上,以检索基础→信息资源→资源利用为编写主线,加强了对信息及信息检索原理的阐述,充实了国内外经典、常用以及新增生物医学信息源的使用方法与技巧介绍,增加了包括医学学位论文在内的医学科技论文写作等实用的信息利用内容论述,利用近一年的时间,编写了《医学信息检索与利用》第 2 版。

全书共分 10 章,第一章介绍了信息检索基础和检索基本原理以及信息素质教育相关内容,是本书的基础篇;第二章至第六章介绍了常用生物医学信息资源及互联网信息资源的使用方法和技巧,是本书的资源篇;第七章至第十章介绍了课题检索、医学科技论文(含学位论文)写作以及情报调研和循证医学有关内容,是本书的利用篇。

本书第一章由曹洪欣、惠朝阳编写,第二章由范吉莲编写,第三、六章由彭骏编写,第四章由张静昌编写,第五、十章由邱君瑞编写,第七章由徐维、陆敏编写,第八章由陆敏、徐维、王侠编写,第九章由彭骏、王侠编写。

本书在编写过程中参阅、引用了大量国内外相关研究成果,在此向有关专家、学者表示衷心感谢。本书的出版再次得到了第二军医大学出版社的鼎力支持,在此深表谢意。

“婴其鸣矣,求其友声”,期望本书的出版能够对高等医学院校的师生以及广大医药卫生工作者进行信息检索与利用的教学与实践起到抛砖引玉的作用;“初生之物,其形必丑”,尽管本书有了第一版的基础,但由于时间仓促和编写人员水平有限,必定还存在诸多不足之处,敬祈广大读者、同仁批评指正,不吝赐教。

编 者

2008 年 4 月

目 录

第一章 绪论	(1)
第一节 信息概述	(1)
第二节 信息资源	(5)
第三节 信息检索	(9)
第四节 信息素质及其养成教育	(22)
第二章 常用医学文摘数据库	(28)
第一节 中国生物医学文献数据库	(28)
第二节 Medline	(34)
第三节 EMBASE.com	(45)
第四节 BIOSIS Preview	(52)
第五节 美国《化学文摘》	(57)
第六节 Dialog 系统生物医学资源	(60)
第三章 常用医学全文数据库	(64)
第一节 中国期刊全文数据库	(64)
第二节 维普数据库	(68)
第三节 万方数据资源系统	(78)
第四节 超星数字图书馆	(90)
第五节 Apabi 全文图书数据库	(95)
第六节 Elsevier ScienceDirect 全文数据库	(97)
第七节 ProQuest 全文数据库	(102)
第八节 Springer 电子期刊	(107)
第九节 Ovid 全文期刊库	(113)
第四章 常用专类/特种医学文献数据库	(117)
第一节 引文检索数据库	(117)
第二节 专利信息检索数据库	(124)
第三节 科技报告检索数据库	(129)
第四节 会议信息检索数据库	(133)
第五节 学位论文数据库	(135)

第六节 标准文献数据库	(137)
第五章 NCBI 及其 Entrez 检索系统	(139)
第一节 简介	(139)
第二节 生物医学文献数据库(PubMed)	(141)
第三节 GenBank 数据库	(155)
第四节 Protein 数据库	(163)
第五节 Genome 数据库	(164)
第六节 Structure 数据库	(165)
第七节 OMIM 数据库	(168)
第八节 Taxonomy 数据库	(171)
第九节 PMC 数据库	(174)
第六章 网络信息检索	(176)
第一节 网络信息资源概述	(176)
第二节 搜索引擎	(179)
第三节 医药卫生专业网站	(201)
第七章 课题检索	(242)
第一节 选题的原则、来源和方法	(242)
第二节 检索策略	(245)
第三节 检索技巧	(252)
第四节 检索实例分析	(255)
第八章 医学科技论文写作	(257)
第一节 概述	(257)
第二节 医学科研论文的撰写	(261)
第三节 医学文献综述的撰写	(266)
第四节 学位论文写作	(272)
第九章 医学情报调研	(278)
第一节 医学情报调研概述	(278)
第二节 资料的搜集、鉴别与整理	(280)
第三节 情报分析和预测	(293)
第四节 医学科技查新工作	(297)

第十章 循证医学及其信息资源	(306)
第一节 循证医学概述	(306)
第二节 Cochrane 协作网及其数据库	(309)
第三节 循证医学数据库	(316)
第四节 循证医学网络资源	(326)
主要参考文献	(334)

第一章 绪 论

第一节 信息概述

身处 21 世纪的我们,在社会生活的各个方面都可以感受到信息化的冲击和对信息的渴求。信息,同物质和能量一样,在技术革命的推动下逐渐成为人类社会发展的决定性力量之一。人类很早就注意到了能量、信息在人类生活中的作用,但是限于当时的社会观念和技术条件,其作用并未得到充分的发挥。在人类社会初期,木材、煤炭主要被人类用做烧烤食物和取暖。而从 18 世纪中叶开始,随着以蒸汽机、电力、内燃机为核心的一系列工业技术的广泛应用引发的工业革命,极大地推动了人类社会的发展。这次革命的基础之一就是对能源价值的重新审视和深入挖掘,其中就包括了将煤炭作为蒸汽机的动力燃料来使用。

20 世纪 40 年代,计算机的发明和信息论的创立为人类对信息的再认识和再开发奠定了技术和理论的基础,同时也为我们开启了迈向信息时代的大门。经过半个多世纪的发展,随着信息科学技术的飞速发展和信息科学理论的日益完善,信息观念和信息技术加速推进了社会生产、社会生活等各个方面的变革。在政治领域,它们推进着社会民主化进程,并加速了权力等级机构的扁平化进程;在经济领域,信息经济在自身成长成为一种独立经济产业的同时,还实现了对其他产业的知识化、智能化和信息化改造,知识创新成为社会持续发展的基本动力;在社会文化和生活方面,网络使得人与人之间的交流更便捷,社会价值观取向多样化,个人行为表现倾向个性化。

一、信息的概念

如同万有引力在牛顿发现以前已经存在的事实一样,信息在它被作为科学研究和人类利用对象之前早已有之。我们可以在李白的诗歌——“梦断美人沉信息,目穿长路倚楼台”中,在《孙子兵法·用间篇》中感触到信息对于古代人的意义;我们也可以在考古挖掘出土的文物,在流传至今的古代经典文献里寻求它们的现代价值。在当时,“信息”主要指的是消息、情报和知识。

信息作为科学的概念,最早是在通信科学中被专门研究的。1928 年,哈特莱指出,信息的数量取决于通信中传递的符号。他在分离出信息意义的前提下将信息作为符号加以定量研究,也是第一次将信息作为相对脱离人类价值需要的客观对象加以研究。1948 年,申农在自己的论文里将信息定义为“不确定性的消除”,认为信息是指“有新内容的消息”。他认为通信的直接目的就在于消除接收端(信宿)对于发射端(信源)发出消息的不确定性,强调了信息在通信过程中的作用。申农对通信系统的信息度量和传递规律的系统研究——信息论,为现代通信技术的发展作出了重大贡献。与此同时,维纳在自己对控制论的研究里指出:“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的东西。”在这里,维纳试图从信息自身意义的层面上探索信息的本质,认为信息的价

值在于它的存在使得我们有了和外界交流的可能。

我国学者钟义信在深入研究前人对信息不同理解的基础上,从一般意义上指出:“信息是该事物运动的状态和状态变化方式。”这表明信息是一种客观的存在,不以主体的存在、感知与否为转移。但这样的信息对于我们人类又有什么用呢?钟义信随后为上述定义引入了一个约束条件——人类认识,认为信息是指“主体所感知或表述的关于该事物运动的状态和状态变化方式,包括状态及其变化方式的形式、含义和效用”。从主体的角度来认识信息,信息的形式可以被人类所感知,信息的含义可以为人类所理解,信息的效用可以允许人类加以利用。依此,信息可以分为语法信息、语义信息和语用信息。

我们认为信息在被作为同物质、能量等具有普遍意义的并列概念时,是指事物运动的状态和状态变化方式。在具体研究信息或利用信息的时候,我们则可以从不同的角度对上述定义加以各种约束,进而创造适用于自己需要的定义。本书中信息的概念非经特殊说明一般是指普遍意义上的信息。

二、信息的类型和特征

(一) 信息的类型

根据不同的划分标准,可将信息区分为不同的类型。

1. 按载体形式划分

(1) 印刷型信息:一般是指以纸张为载体、以印刷为记录手段而产生的一种传统信息形式,如报纸、期刊、图书、海报、广告等。其优点是便于阅读和流通;缺点是存储密度低,收藏、管理成本高。

(2) 缩微型信息:为弥补印刷型信息单位容量小的不足,缩微型信息应运而生。它是以缩微胶片或平片为载体,利用缩微摄影技术记录信息,如一张全息胶片可存储 20 万页信息。但其缺点则是必须借助缩微阅读机才能阅读。

(3) 声像型信息:常见形式有录音带、录像带、电影胶片等。它们都是以磁性和感光材料为存储介质,以模拟信号的方式记录客观世界。其优点是直观、亲切、表现力强;缺点是较新出现的数字型信息成本过高,且阅读方式难以统一到同种设备上。

(4) 数字型信息:指以数字信号的方式将图、文、声、像等信息存储到磁、光、电等介质上。其阅读方式已经从计算机发展到各类数字阅读终端上,比如手机、数字电视等等。生活中常见的数字型信息有网络数据库(以网络访问的形式提供服务)、计算机软件、电子出版物等,常见的载体形式主要是光盘、磁带、磁盘阵列等。

2. 按出版形式划分

(1) 图书:是目前仍被人类广泛使用的记载、传播信息的古老途径之一,一般是指页码不少于 49 页的非定期出版物,常见的有小说、教科书、专著、百科全书等。其特点是符合人类阅读习惯,可以用于对某一问题进行系统、全面、集中的论述,但同时存在着出版周期长、报道速度慢等缺点。

(2) 报纸、期刊:是指名称、出版周期、内容范围相对固定的连续出版物。其特点是内容新颖、信息含量大、传播速度快,可以快速反映社会进展和最新科研成果。

(3) 会议论文:是进行学术交流的一种重要方式。在国际或国内重要的学术或专业性会议上发表的论文,学术性强,代表了某一领域内的最新成就,反映了行业科技发展水平和

趋势,是获取科技信息的重要途径之一。

(4) 科技报告:是描述一项科学技术研究的结果、进展或一项技术研制试验和评价的结果;或是论述一项科学技术问题的现状和发展的文件。在内容方面,科技报告比期刊论文更为专深、详尽与可靠。目前国际上较著名的是美国政府四大报告,即 PB(美国商务部出版局)、AD(国防科技)、NASA(航天航空)、DOE(能源类)报告。

(5) 专利文献:是实行专利制度的国家及国际性专利组织在审批过程中产生的官方文件及出版物的总称,通常以专利说明书、公报、文摘、索引、权利说明书等形式出版。

(6) 政府、机构出版物:政府出版物是指政府部门在国家管理过程中形成的以文件形式为主体的信息,常见的有政府法令、方针政策、统计资料等。机构出版物常见的有标准文献和行业出版物,比如国际标准化组织(ISO)和世界卫生组织(WHO)发布的各类信息。

除以上介绍的信息类型外,信息的出版形式还包括产品说明书、技术档案、学位论文、电子出版物等。其中电子出版物是目前流行的一种出版形式,除与原有出版形式载体不同外,还可以提供交互功能,以利于读者使用。

3. 按发生领域划分

(1) 物理信息:是指无生命世界的信息,诸如自然风景、春夏秋冬四季等。

(2) 生物信息:是指生命世界的信息,诸如动物种群内部的交流协作、人类生命的遗传信息。

(3) 社会信息:是指人类社会活动产生的各类信息,诸如政治、经济、科技、文化、军事等信息。

(二) 信息的特征

所谓信息的特征,就是指信息区别于其他事物的本质属性,主要表现为以下几个方面:

1. 客观普遍性

信息不是虚无飘渺的事物,它的存在可以被人们感知并加以利用。信息源于“物质的运动及其变化”。但信息不等同于物质,它可以脱离物质而寄生于载体,相对独立地存在。物质及其运动的普遍性决定了信息是无处不在的,也就是说,只要有物质运动的地方就有信息存在。

2. 动态性

由于客观事物本身的不断变化和人类认识的发展变化,因而表征事物运动状态和状态变化方式的信息以及人类所能认识到的信息也必然会随之改变,昨天的新闻在今天就可能已经成为旧说。

3. 共享性

信息不同于物质、能量的一大特征就是信息可以被无限复制,可以被同步或异步共享。这里的复制、共享是指信息的量和质并不因被复制、被提供而减少或失去。正如大家熟悉的一个例子中所说的那样:苹果不会因为交换而增多,观念(信息)却能因为交换而增加。网络技术实现了信息的大范围共享,进而充分发挥了信息的效应。

4. 可塑性

信息的可塑性是指信息可以被无限再生产。人类知识的发展史是信息再生产的典型表现。就我们所熟知的知识而言,在已有知识的前提下,我们可以通过新的实践和理论研究来创造出新的知识。“科学技术是第一生产力”这句格言随着信息时代的发展不断得到印证。

信息和知识(科学技术知识)将逐步成为我们这个时代最主要的资源和财富。信息和知识的再创造也将成为经济增长的主要推动力。

三、知识、文献、情报和信息

我们经常听到的知识、文献和情报等概念,同信息有着千丝万缕的联系,那么它们之间具体的联系和区别在哪里呢?

(一) 知识

在日常生活中,我们很容易就能区分什么是知识、什么不是知识。比如医学教材上的内容被当作知识看待,《人民日报》的评论员文章则被当作新闻看待。知识一般被认为是人类在认识和改造世界的实践中形成的对事物本质和规律的认识,是人类文明的集中体现。知识的这一定义包括4方面的内容:其一,知识来源于人类的实践,来源于外界事物和人类自身所具有的信息;其二,知识是人类已经认识到的信息;其三,知识是对人类已经认识到的信息的体系化、系统化的总结;其四,知识的发达程度基本反映了人类文明的高度。知识作为人类对信息加工改造的产物,在分享信息所有特性的同时,还拥有系统性、实用性等特点。知识的获得来之不易,而知识只有得到有效的传授和继承才能发挥累积效应,实现无限增殖。

(二) 文献

人类在漫长的实践中产生了知识。最初人们可以通过示范和口头交流的形式传授知识,但随着知识数量的激增或是知识在较大范围内传播的需要,人们开始在龟甲、石板、青铜器上记录知识用以传递。这些记录有知识的载体就被认作是文献。科学技术的发展,拓展了知识的表达手段、丰富了记录知识的载体材料。目前我们常见的知识表达手段除了文字、符号、图形以外,还有照片、声频、视频、计算机代码,常见的载体材料有纸张、磁带、光盘、硬盘等。仅仅是在最近的几十年,文献这一概念才被信息逐步替代。因为在信息的现代概念和价值被人类发现以前,一般信息对于人类发展的作用极为有限,而只有系统化、体系化的信息——知识才会被人类广泛重视并加以利用,用以记录传承知识的文献就成为人类在那个时代的首选表达方式。因此直到今天,文献这个词也还在被广泛使用。

根据文献内容性质和加工程度的不同,常将其划分为一次文献、二次文献、三次文献和零次文献4个级次。一次文献是指人们以自己的生产、科研、社会活动等实践经验为依据公开发表和出版的原始创作,如专著、期刊论文、科技报告、学位论文、会议录等。二次文献是指为了方便用户查找文献,由信息工作者对无序、分散的一次文献进行再加工,通过整理、提炼和浓缩,依照其外部特征(如著者、题名等)或内容特征(如主题、学科等)将其有序化,形成的新的文献形式,如目录、索引、文摘等。三次文献是指人们围绕某一专题,在利用二次文献的基础上,经过系统地阅读、分析、研究、综合、整理撰写而成的,如综述、述评、进展、趋势等,用户借此可以系统、快速地了解某一专题当前发展的水平和动向。零次文献包括两类:一是未经正式发表的文献,如书信、手稿、笔记等;二是未经加工的信息,如实验数据、观测记录、调查资料等。

(三) 情报

情报,简而言之,就是指有用的信息,有用性是情报的第一属性。在此基础上,情报学专家维索斯基提出了情报四原则——准确的信息,合适的时间,对口的用户和合适的成本,很好地归纳了情报的内涵。情报同信息、知识在内容上的关系如图1-1-1所示。情报不同

于信息,因为情报是根据用户特定需求(内容、时间、成本)主动采集,并经过去粗取精、去伪存真加工后传递给用户的信息,只是信息的一小部分。情报和知识在内容上则是相互交叉的。有些情报内容本身就是知识,比如科技情报搜集的对象之一——高科技知识;有些情报内容可能是非知识的信息,比如战争中搜集到的敌方兵力投放信息。按情报的应用范围可以将情报划分为军事情报、科技情报、经济情报、政治情报等。国内常见的情报机构有军队情报部门、国家安全部门、省(市)科技情报所、大型企业(机构)中的情报机构和图书馆的情报(信息)服务部门。

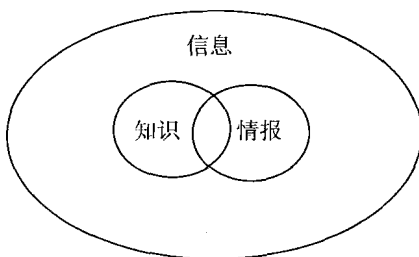


图 1-1-1 情报、信息、知识关系示意图

第二节 信息资源

一、信息资源

信息无处不在,但大多数信息是不能轻而易举地为人类所利用的,因而普遍自然存在的信息对于人类的直接作用比较有限。资源是人类对可利用的事物的总称,因而信息只有成为资源才能为社会所广泛利用。资源,一般被认为需要具备 3 个方面特征:其一,人类生产活动所需;其二,具有一定的丰度和凝聚度;其三,自然资源的获取需要人的参与。比照上述标准,信息资源首先应该是根据人类需要而生产的信息;其次,信息资源应该是有序信息的集合;再次,信息资源的获得需要通过人类劳动的参与。据此,我们认为只有那些经过人类开发与组织的信息的集合才是信息资源。

知识和情报都符合信息资源的属性要求,但信息资源的范围却远不止于知识和情报。Google 提供的高清晰度卫星照片、门户网站提供的分类信息、进行过用户细分的手机号码乃至超市收集的顾客购买偏好信息等都属于信息资源的范畴。信息资源不同于其他资源的特点主要表现在它的智能性和整体性。不同于物质、能量资源,信息资源是人类脑力劳动的直接产物,人类的智能在很大程度上决定信息资源的质与量。信息资源的整体性体现在其只有通过大范围的共建和共享才能实现最大的效益,不完整的信息资源不但会降低其利用的效率,甚至有时候还可能对用户产生误导。

知识、情报很早就得到了人类的重视和利用,但信息资源得到人类广泛关注却晚于 20 世纪中叶。一般认为,只有当信息资源的数量积累到一定程度后,人类才有可能将信息资源视为一种潜在资源。但仅有量的积累也还是不够的,只有等到现代信息技术的出现才真正使得信息资源在现实中为人类广泛地加以利用。最终,技术上的成熟和认识上的完善互相影响并共同推动着人类对信息资源的积极开发和广泛利用。大型工商企业是较早注意到信

息资源的价值并加以利用的机构。早在 20 世纪中期,企业规模的快速膨胀使得它们已经无法单凭人力处理诸如财务数据等大规模数据,大型企业开始引入计算机进行自动化数据处理。之后,计算机处理信息的强大功能进一步推动管理者将其应用范围扩大到企业内部人员信息、产品信息和企业外部的用户信息、竞争对手信息的收集管理上来。通过引入信息技术开发和利用信息资源给企业带来了可观收益,进而推动了企业在信息技术研究和信息资源开发上的投入,并最终推动了全社会对信息资源开发和利用的重视。

二、信息价值链

当信息被作为一种资源来对待的时候,人们逐渐开始关注信息的生产和消费。信息是如何成为信息资源进而为人类所利用的呢?这要从人类的信息需求说起。一般认为,信息需求从被提出到被满足需经历如图 1-2-1 所示的 6 个阶段。

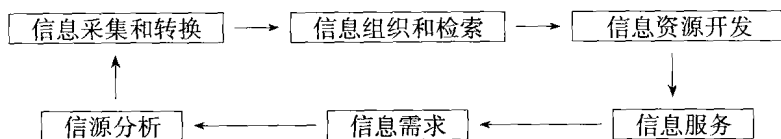


图 1-2-1 信息价值链 6 个阶段示意图

(一) 信息需求

信息需求,是指人们在从事社会实践的过程中,为解决各类问题而产生的对信息的欲望和要求。关心体育比赛的人,会希望尽早知道比赛的结果;追赶时尚的人,总希望第一时间获取潮流变化的信息。信息需求在每个人身上都会或多或少存在。

从信息是否被发现、被表达和被满足 3 个方面,人类的信息需求可按图 1-2-2 进行细分。

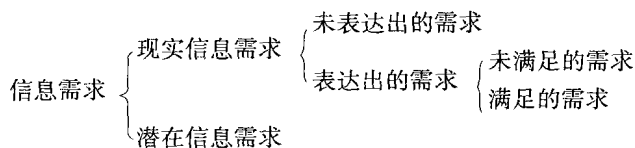


图 1-2-2 信息需求示意图

潜在信息需求是指那些我们没有明确意识到的而实际上对我们有用的信息需求。现实信息需求是指我们明确意识到的信息需求。现实信息需求依据是否能够被明确表述可分为表达出的需求和未表达出来的需求。一般只有那些明确表达出的信息需求才有可能从外界信息源中得到满足。从社会总体而言,由于人类认识和社会生产力发展的局限,明确表达的需求也只有部分能够得到满足。

从社会学角度来看,用户的信息需求具有 3 个特征:个人特征、组织特征和社会特征。个人特征因人而异,既有先天因素又有自致因素。组织特征则反映了用户所从事工作所在行业对用户信息需求的影响。社会特征是由用户所处时代背景和现实处境所决定的。无论是用户或是信息提供者都应该注重从这几个方面分析并提出或满足信息需求。

(二) 信源分析

信息源简称信源,是指信息的来源。因为信息是指事物的运动状态和状态变化方式,而

且是针对一切事物而言的,因而我们所能接触到或认识到的事物都可以成为信息的来源。

从信息采集的角度出发,信息源通常包括个人信息源、实物信息源、文献信息源、数据库信息源、组织机构信息源。根据信息源的运动方式,信息源可以分为静态信息源和动态信息源。

如同物质、能量资源自然分布的不平衡一样,信息源的分布也很不平衡。信息源的不平衡分布主要表现在地区差异、机构差异和个人差异 3 个方面。地区差异具体表现在发达国家和发展中国家之间、城市和农村之间,这是因为前者集中了更多的信息资源生产者和消费者,同时在信息资源的技术保障和管理、传播上更有优势。机构差异主要是指信息生产、储存、传播等机构同信息利用机构之间的差异。科研机构、信息中心较之于工业生产、医疗服务等机构而言拥有更为丰富的信息量。媒体、科研工作者同工、农业生产者之间信息量的差异则是信息源个人差异的表现之一。

各类信息源自身不同的特点以及信息源分布的不平衡性要求我们在信息采集和利用过程中对不同的信息源要区别对待,以提高信息采集利用效率。

(三) 信息采集和转换

信息采集是指选择信息的过程。我们在信息的选择过程中应该注意以下 2 个方面:其一,不同信息源的特点和信息源分布的不平衡性要求我们在信息采集过程中强调针对性;其二,面对信息技术的飞速发展、信息流动性的不断加快、信息数量的爆炸式增长、信息污染不断扩展等问题,我们在信息采集工作中还应该坚持下面 5 项原则:一是应当在科学预见基础上做到计划性和灵活性的统一;二是注重信息采集的适用性和经济性,做到有的放矢,提高效率;三是强调信息的真实、可靠;四是注重信息采集的主动性和及时性;五是坚持信息采集的系统性和连续性。

信息转换主要包括信息资源所有权或使用权的转换、信息资源符号的转换和信息资源载体的转换三方面的内容。信息资源的搜集需要耗费人力、物力和财力,使用信息资源也就必须承担相应的费用。当然,这种费用的支付可以是直接或间接的。信息资源符号的转换可以有多种形式,比如记者采访到的口语信息资源会通过文字、音频等方式保留下来,电影、电视剧则是把剧本中的文字信息资源转换成视频信息资源。信息资源载体转换就更为常见了,比如为了阅读方便我们经常会把电子文档打印到纸张上来。

(四) 信息组织和检索

信息组织是将处于无序状态的特定信息,根据一定的原则和方法,使其成为有序状态的过程。我们常见的汉语字典是依照拼音字母顺序将所有汉字汇集在一起,目的是为了更方便我们查找。门户网站的新闻栏目、体育栏目等则是以大家生活中习惯的、感兴趣的题材来组织信息。信息组织的方法多种多样,但他们的共同点在于都是依赖于事物属性来进行序化排列的。我们能在事物上发现多少种属性,就可以用多少种方法序化信息。就目前人类对事物属性的认识而言,事物属性可以被归纳为形式、内容和效用 3 种类型。这三种信息组织方法分别对应着前面我们提到过的语法信息、语义信息和语用信息,因而信息组织方法可以分为语法信息组织、语义信息组织和语用信息组织。

信息检索是指从大量信息中获取所需信息的过程。未经过开发和组织的信息不易为人类所利用,因而我们通常所说的信息检索是针对那些经过开发组织的信息——信息资源而言的。在信息组织的过程中,一般会用一套标识系统来表示事物的属性。信息检索是以信

息组织所依据的事物属性为基础,针对表示事物属性的标识系统进行相符性匹配的过程。比如,图书馆在组织图书的过程中,会产生一套以作者排序的卡片目录或数据库文档,读者在需要检索特定作者的著作中,就可以通过使用这套目录或数据库文档实现高效的检索。

(五) 信息资源开发

信息资源开发是指对已有的信息资源或经过检索获取的信息资源进行深层次的加工和开发,然后提供给用户。信息资源开发的最终目的是形成针对市场需要的信息产品。信息产品可以分为生产型信息产品和再生型信息产品(图 1-2-3)。

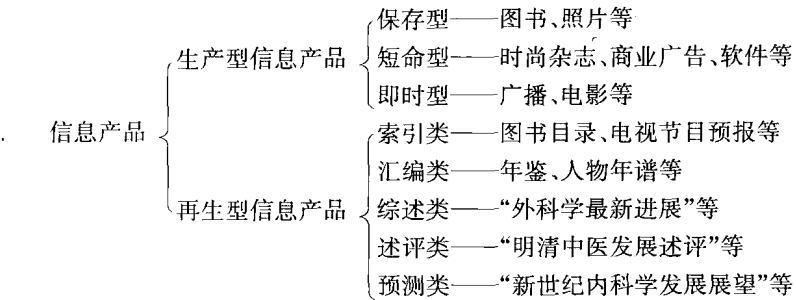


图 1-2-3 信息产品分类示意图

生产型信息产品一般由传媒产业和软件企业生产制造,大多可以作为最终产品为人类所消费。再生型信息产品则是由科研院所等科研机构或是图书情报机构来生产,一般是作为中间产品,为产品生产、科学研究、决策判断等做前导。

(六) 信息服务

信息服务是指以信息为内容的服务业务。图书馆、新闻媒体、门户网站都是我们生活中常见的信息服务提供者,它们和信息用户分别构成了信息服务的主体和客体。

信息服务者在提供信息时,一般要考虑 3 个方面的因素:信息传递方式;信息利用效果的分析;信息法律、法规。

信息服务的模式按照传递的方向和方式可以分为 4 类:多向主动传递(广告、门户网站);单向主动传递(决策支持服务、定题服务);多向被动传递(档案馆、公共图书馆);单向被动传递(口头问询、科技查新)。

信息服务的目的是为信息用户创造价值,而用户对信息的吸收因受到用户自身知识结构、用户情绪等多种因素的影响,大多难以达到信息服务者初期设想效果,因而信息服务者还应该积极获取用户反馈并加以分析,为进一步改进信息服务提供参考。

信息法律法规的目的在于维持正常的信息服务秩序和保障各方面权益,因而信息服务提供者必须在信息法律法规允许的范围内从事信息活动。知识产权法、保守国家秘密法、计算机信息系统安全保护条例是我们在信息活动中应该予以特别注意的信息法律法规。

三、信息化

信息化是相对于工业化而言的。社会信息化被认为是从有形的物质产品创造价值的社会(工业化社会)向无形的信息创造价值的社会(信息化社会)转变的过程。具体而言,信息化是指在人类社会活动中,通过广泛地采用信息技术,大力开发、利用信息资源,推动经济发展和进步的过程。从信息化发展的实践来看,信息化可划分为信息产业的形成和传统

产业的信息化、经济和社会信息化两个阶段。

（一）信息产业的形成和传统产业的信息化

这一阶段可以细分为信息设备生产、信息网络建设、信息内容产业融合、传统产业信息化 4 个部分。

以计算机、通讯设备和软件的制造与维护为代表的信息设备生产是信息化的起点,传统产业对信息设备的强大需求推动了信息设备生产的规模化。

孤立的信息设备和信息系统的功用都还很有限,只有等到信息网络建成,它们才可能发挥出最大的效能。信息网络的建设最初只是在机构、企业的内部加以建设和利用。从 1993 年美国出台 NII(国家信息基础设施)建设以来,包括我国在内的世界各国相继推出了自己的 NII 计划,加速抢占信息网络建设的高地。

信息内容产业融合是指以市场需求为导向,将过去分散在传统产业中的与信息生产、利用等直接相关的单位和资源进行优化整合,从而形成相对独立的经济实体。信息内容产业同信息设备制造业、信息网络建设事业一起构成一个具有独立地位的产业——信息产业。

传统产业的信息化是同信息产业的形成同步进行,互相推动的。产业信息化主要表现为企业的信息化,即企业通过对管理过程和产销过程的信息化改造来提升企业的管理水平和产销能力,进而增强企业的竞争力。

（二）经济信息化和社会信息化

经济信息化是指通过对整个社会生产力系统实施信息化改造,在社会经济生活中逐步实现信息化的过程。经济信息化是信息产业化和传统产业信息化在经济领域发展的最终结果。一般认为由信息产业创造的价值在国民生产总值中所占的比重超过传统产业时就可以说基本实现了经济信息化。经济信息化最终将会促使信息资源成为经济发展的重要基础和战略资源、信息技术成为推动社会进步的主导技术、信息人员成为推动社会进步的中坚力量。

社会信息化是以信息产业化和传统产业信息化发展为基础、以经济信息化为核心向人类社会活动的各个领域逐步扩展的工程,其最终结果是人类生活的全面信息化,比如文化娱乐、教育和医疗卫生的信息化。社会信息化带给我们的不止是生活水平的提高,更重要的是,它还将会为个人创造力的发挥提供一个更为宽广的舞台,使得每一个作为独特个体的人有机会从集体的幕后走上前台向世界展示自我。

第三节 信息检索

一、信息检索的概念

通常情况下我们所针对的信息检索是狭义范畴的概念。狭义的信息检索(Information Retrieval)是指根据特定的信息需求,运用科学的方法,采用专门的工具,从大量信息源中获取所需信息的过程。而广义的信息检索还要包括信息的加工处理与存储过程。所以完整意义的信息检索又称信息存储与检索(Information Storage and Retrieval)。本书所阐述的有关信息检索除特别说明外均指狭义范畴概念。

二、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。

(一) 根据检索对象划分

1. 文献检索

凡是利用目录、文摘或索引等二次文献查找某一课题、某一著者、某一地域、某一机构或某一事物的有关信息以及这些信息的出处和收藏单位等,都属于文献检索(Document Retrieval)范畴。其检索的结果是文献信息。如针对“某篇论文的参考文献”的检索就属于文献检索。

2. 数据检索

凡是利用参考工具书、数据库等检索工具检索包含在信息中的某一数据、参数、公式或化学分子式等,统称为数据检索(Data Retrieval)。其检索结果为数据信息。如针对“某药物的血药浓度半衰期”、“某药物成分的化学分子式”等的检索均属于数据检索。

3. 事实检索

凡是利用诸如百科全书、年鉴等存储事实的信息系统查找出特定事实的过程称为事实检索(Fact Retrieval)。其检索结果是基本事实。如针对“我国第一颗人造卫星成功发射的年度”、“某国某年的 GDP”等的检索就属于事实检索。

4. 图像检索

图像检索(Image Retrieval)即以图形、图像或图文信息为检索内容的信息检索。目前主要有基于文本、基于内容和基于语义 3 种图像检索技术。基于文本的图像检索系统目前已经得到了广泛的应用,但它对图像的含义却是不准确的,与单纯的文字检索系统相比,检索的准确性还是有相当大的差距;基于内容的图像检索主要用图像的视觉特征来表示图像信息,并根据这些信息来进行检索,此在某些特定领域得到了广泛应用,如用于指纹识别、商标检索和医学图像检索等;基于语义的图像检索是最人性化的图像检索方式,在网络中的应用前景十分广阔。基于语义的图像检索技术将基于视觉特征和基于文本的检索技术结合在一起,通过建立以人的兴趣、爱好为模型的知识库,实现了提取图像语义和图像特征的功能,而且充分考虑了用户的反馈信息,从而使得下一轮的检索结果更贴近用户的需求。这种检索技术是未来建立高效、实用、快速的图像检索系统的主要发展方向。

图像检索是视频检索的基础,实现高效的图像检索对实现视频检索有着重要的意义。

5. 多媒体检索

多媒体检索(Multimedia Retrieval)是指以图形、图像、文本、声音、动画等多媒体信息为检索内容的信息检索。目前主要有基于文本、基于内容的两种多媒体检索技术。基于文本的多媒体检索主要是运用文本检索的方法体系,即通过分析多媒体信息的内容,给其以关键字或者其他文本形式的表述,然后通过检索这些描述信息达到检索多媒体信息的目的;基于内容的多媒体检索,简单地说就是依据多媒体信息的内容,在无须过多人工参与的情况下自动识别或理解图像重要特征的检索方法,是一种相似度检索。目前,基于内容的多媒体信息检索的主要工作集中在识别和描述图像的颜色、纹理、形状和空间关系上,对于视频数据,还有视频分割、关键帧提取、场景变换探测以及故事情节重构等问题。基于内容的多媒体数据检索涉及到的技术比较多,主要有人-机交互、图像特征的描述、提取与匹配技术、数据库的组织与管理以及检索评价技术等。随着多媒体技术的发展及其应用范围越来越广泛,多媒