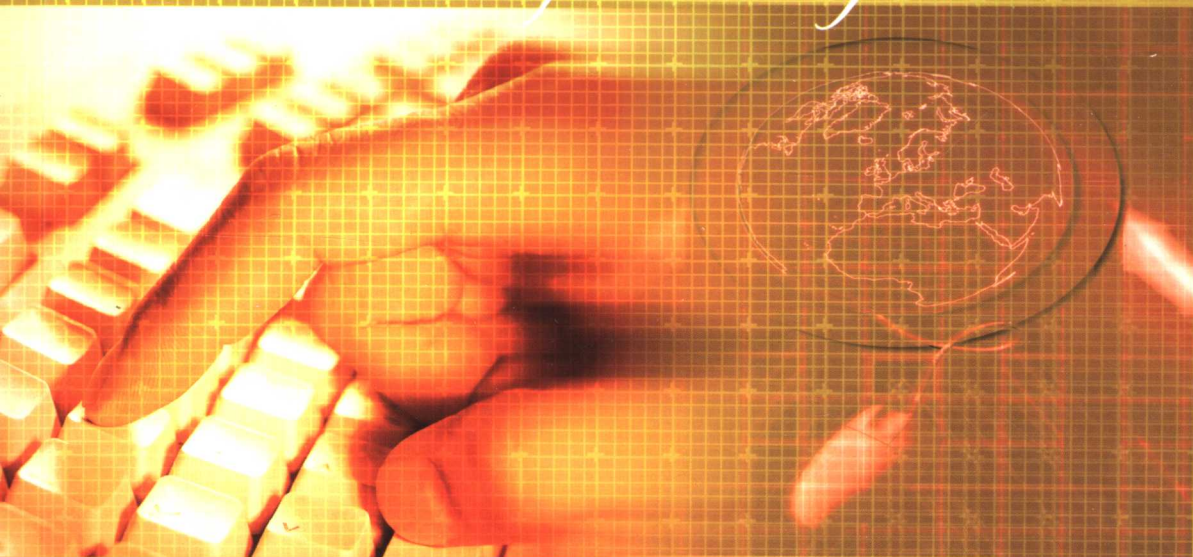


数字信息检索 教程

*Shuzi xinxi jiansuo
jiaocheng*



主 编 刘廷元

副主编 徐晓琳 黄 波 孙汝杰



华东理工大学出版社

EAST CHINA UNIVERSITY OF SCIENCE AND TECHNOLOGY PRESS



数字信息检索 教程

*Shuzi xinxi jiansuo
jiaocheng*

主 编 刘廷元

副主编 徐晓琳 黄 波 孙汝杰



华东理工大学出版社

EAST CHINA UNIVERSITY OF SCIENCE AND TECHNOLOGY PRESS

图书在版编目(CIP)数据

数字信息检索教程/刘廷元主编. —上海:华东理工大学出版社,2006.8
ISBN 7-5628-1928-9

I. 数... II. 刘... III. 计算机应用—情报检索 IV. G252.7

中国版本图书馆 CIP 数据核字(2006)第 073490 号

数字信息检索教程

.....

主 编 / 刘廷元

责任编辑 / 李国平

封面设计 / 王晓迪

责任校对 / 许 春

出版发行 / 华东理工大学出版社

地 址:上海市梅陇路 130 号,200237

电 话:(021)64250306(营销部)

传 真:(021)64252707

网 址:www.hdlgpress.com.cn

印 刷 / 常熟华顺印刷有限公司

开 本 / 787×1092 1/16

印 张 / 16.75

字 数 / 405 千字

版 次 / 2006 年 8 月第 1 版

印 次 / 2006 年 8 月第 1 次

印 数 / 1—4050 册

书 号 / ISBN 7-5628-1928-9/TP·148

定 价 / 23.50 元

(本书如有印装质量问题,请到出版社储运部调换)

序 言

在 Internet 为我们构筑的全球数字信息网络环境下,信息犹如浩瀚的海洋。早在 20 世纪 30 年代,一位美国诗人就对信息的不断增长给人类知识和智慧带来的挑战感到十分迷惑。他问道:

Where is the wisdom we have lost in knowledge?

Where is the knowledge we have lost in information?

——T. S. 艾略特(1888—1965)

面对信息的不断增长,诗人问我们:知识中我们丢失的智慧到哪里去找?信息中我们丢失的知识又到哪里去找?

虽然我们这本教材不能提供“知识中我们丢失的智慧到哪里去找”的帮助,但也许可以为读者提供“信息中我们丢失的知识又到哪里去找”的帮助。

要真正拥有知识,必须首先从内容和量这两个方面去搜寻“完全信息”(Perfect Information),而要搜寻“完全信息”则必须掌握利用信息的方法、技术和科学知识,提高信息检索的准确性和完整性,即信息的查全率或查准率。《数字信息检索教程》作为一门应用性、方法性、工具性的课程,通过讲授信息检索的理论与实践知识,能够帮助本科生、研究生和科技人员提高自己的信息素质,搜寻到对自己有用的完全信息。

近 100 年来,信息检索经历了手工检索、机械检索、计算机单机检索和计算机网络检索四个发展阶段。最早的穿孔卡片手工检索工具出现于 1904 年,1930 年后发展到边缘穿孔卡片、比孔卡片到机械穿孔卡片等。1940 年,美国试制成功第一台缩微胶片检索机。一张缩微平片可以缩摄存储几十页至几千页的文献,且存储时间较长。

1946 年,第一台电子管计算机在美国诞生后,经过几代人改进和发展,计算机存储及其网络通信技术取得了长足进步。1958 年,具有批处理能力的晶体管计算机产生。1965 年以后,集成电路计算机及硬盘产生,数字通信技术和分组交换网得到很快发展,开始有了数据库联机的数字信息检索,著名的 DIALOG 系统以及 MEDLINE、ORBIT、BRS、JOIS 等相继开始服务。20 世纪 70 年代以后,卫星通信技术、光纤通信技术、个人计算机的产生和发展给信息检索创造了空前的发展机会,计算机检索已不受时空限制,向网络化、数字化方向发展,个人用户开始加入到数据库检索行列中来。

1978 年,美国梦想家兰开斯特就预见到“无纸的”(paperless)数字信息交流时代即将到

来。进入 20 世纪 90 年代以后,随着通信网络和信息处理技术的发展,使得基于互联网开发的数字信息资源及其检索系统有了突飞猛进的增长。到 2006 年 5 月,Google 搜索引擎已经拥有 80 多亿网页、超过 10 亿张的图片,每天接收数亿个检索。众多的搜索引擎和门户网站,给人们的信息利用带来了极大的方便,人们开始习惯于到一个固定的地方获取自己需要的信息。由于搜索引擎的检索速度非常快,而且结果通常是相关的且无需付费。目前,全世界现有无印刷副本的数字期刊大约就有 5 000 种,另有 34 427 种杂志既可以获得数字化文本也可以获得印刷型文本,其中 44% 的是科学出版物。

现在,一方面是数字信息的激增,另一方面是数字信息检索方法、技术的快速发展,给人类信息利用带来极大的便利。强大的桌面计算机连接着全球的数字信息交流网络,科学家几乎可以利用全球的数字信息资源,与任何地方的同事进行“无纸”交流。科学研究成果的数字化,从数据、文本的产生一直到科学发现的最后合法判决——作品的出版,允许包括数字图书馆在内的信息服务机构超越创造者(作者)和接收者(读者)的作品时空限制进行“无纸”信息检索。

然而,如果我们仅仅依靠初级的信息检索方法、技术,不系统地学习、掌握现代数字信息检索的高级方法、技术和科学知识,或者仅仅依靠图书馆的个别信息资源占有及其封闭性的信息服务,获得的信息内容和信息量就会十分有限。我们就会因为信息不全、不充分而落后于他人,变成一个“信息穷人”,使自己的知识链上出现“信息孤岛”或“信息鸿沟”,以致于最终不能发挥自己的聪明智慧。

面对浩瀚的数字信息,我们必须充分地了解数字信息资源及其时空分布、数字信息服务系统及其利用途径、数字信息源(数据库)及其计算机检索、搜索引擎及其开放获取、信息需求及其信息检索评价、信息利用与学术论文写作,全面掌握现代数字信息检索的方法、技术和科学知识,才能融入当今这个信息社会,成为全球信息交流共同体的智者。唯有如此,我们才不会在信息中丢失知识以及在知识中丢失智慧。

本书由国内六所大学的老师共同编写,具体分工如下:第 1 章、第 5 章、第 9 章由刘廷元(西南石油大学)完成;第 2 章、第 8 章由徐晓琳(重庆工学院)完成;第 3 章由黄波(成都信息工程学院)完成;第 4 章由程坤、刘廷元(哈尔滨理工大学、西南石油大学)完成;第 6 章由叶建华(深圳大学)完成;第 7 章由孙汝杰(上海应用技术学院)完成;全书由刘廷元主编、统稿,副主编徐晓琳、黄波、孙汝杰。参加编写的还有江华、李明扬、张明、吴宪等。

为了方便使用本书,我们编写了配套的电子课件,并由华东理工大学电子音像出版社出版。电子课件由叶建华为主编。如果需要电子课件,请与作者或出版社联系。

E-mail: tsq-qb@163.com 或 gppli@ecust.edu.cn

编 者

2006. 7

目 录

第 1 章 数字信息资源	(1)
1.1 信息及其相关概念.....	(1)
1.2 数字信息的定义与类型.....	(7)
1.3 数字信息源.....	(17)
1.4 数字信息检索评价.....	(20)
第 2 章 数字信息检索的基本原理	(24)
2.1 信息检索基本概论.....	(24)
2.2 数字信息检索工具.....	(32)
2.3 计算机检索技术.....	(34)
2.4 信息检索的基本步骤.....	(38)
第 3 章 中文数字信息资源检索	(43)
3.1 馆藏书刊目录查询.....	(43)
3.2 维普中文科技期刊数据库.....	(47)
3.3 CNKI 中国期刊全文数据库.....	(51)
3.4 万方数据资源系统.....	(57)
3.5 人大复印报刊资料全文数据库.....	(61)
3.6 电子图书数据库.....	(65)
3.7 科学引文数据库.....	(71)
第 4 章 搜索引擎及开放获取	(75)
4.1 搜索引擎概述.....	(75)
4.2 常用中英文搜索引擎.....	(77)
4.3 学科门户网站.....	(83)
4.4 开放获取.....	(93)
第 5 章 国外文摘数据库	(100)
5.1 国外文摘数据库概述.....	(100)
5.2 Ingenta 学术平台.....	(102)
5.3 ISI 三大引文数据库.....	(105)

5.4	Dialog 数据库系统·····	(111)
5.5	Ei Village2 ·····	(116)
5.6	化学文摘(CA) ·····	(123)
5.7	国内外文文摘数据库检索·····	(129)
第6章	国外全文数字信息资源检索·····	(135)
6.1	国外全文电子期刊数据库概述·····	(135)
6.2	EBSCOHost ·····	(136)
6.3	Elsevier Science (ScienceDirect) ·····	(141)
6.4	John Wiley ·····	(147)
6.5	Springer LINK ·····	(152)
6.6	IEEE/IEE Electronic Library ·····	(158)
6.7	ProQuest 系列数据库 ·····	(167)
第7章	特种信息资源检索·····	(173)
7.1	专利检索·····	(173)
7.2	学位论文检索·····	(185)
7.3	会议论文与科技报告检索·····	(193)
7.4	标准信息检索·····	(201)
第8章	数据与事实信息资源检索·····	(212)
8.1	数字型工具书·····	(212)
8.2	数据和事实型数据库·····	(214)
8.3	参考工具网站选介·····	(222)
第9章	学术论文的写作与发表·····	(232)
9.1	学术论文概述·····	(232)
9.2	学术论文的撰写·····	(243)
9.3	学位论文·····	(249)
9.4	学术论文的发表·····	(252)
参考文献	·····	(259)

第 1 章 数字信息资源

1.1 信息及其相关概念

1.1.1 什么是信息

“信息”一词是汉语的固有词汇,而非“舶来品”。在中国古代早期的一些诗词歌赋、小说野史中,就有过“信息”一词使用的记载,只不过它是借助于“消息”、“音信”等词来表述其中的含义,尚未赋予“信息”一词以严格的界定罢了。例如,西晋时期陈寿的《三国志·魏书》中就曾记载:“诸葛恪围合肥新城,城中遣士刘整出围传消息”,此处的“消息”即指“信息”之意。

南唐时期,诗人李中正式使用了“信息”一词。在李中的《碧云集·暮春怀故人》中有“梦断美人沉信息,目穿长路倚楼台”的诗句,其中的“沉信息”即指杳无音信之意。明代李昌祺的《剪灯余话·至正妓人行》诗:“荡子江湖信息稀,疲兵关塞肌肤裂”,句中“信息稀”即指音信稀少之意。另外,在脍炙人口的《水浒传》第四十四回中还有这样的记载,宋江语:“只要贤弟去得快,旬日便知信息。”这里所说的“知信息”是指消息即将返回的意思。

英语中的“information”(信息)一词,其字源为拉丁文的“informatis”,意指“沟通、通讯的过程或事物”。但最早的英文名词“information”出现于何时,至今没有定论。不过,可以肯定的是,英语中的“信息”一词的出现比中文晚很多。欧洲语言中的“信息”一词最早出现在 15 世纪的法语口语词汇,且语义有各种令人费解的区别,其意群围绕的语义有调查、教育、告知或交流知识的活动、消息,等等。

根据 20 卷本《牛津英语词典》对“information”一词的解释,信息有三种含义:

(1) 作为过程的信息:当某人获取信息后,他或她的认识也会随告知行为过程而改变。从这种意义上看,信息即“告知……的行为;某种事实或情况的知识或消息的通讯;告知行为或被告知的某事的事实”。

(2) 作为知识的信息:信息也可指“作为过程的信息”(告知行为过程)所告知的知识内容:“有关某种事实、专题或事件的所通讯的知识;某人被告知的东西;情况、消息。”

(3) 作为事物的信息:信息一词也可指告知的载体,诸如数据和文献之类的载体。这些载体之所以被称为信息,是因为具有信息性,即“具有传授知识或通讯信息的特征,并具有教育性”。

然而,仅从语言角度去认识什么是信息是不够的,我们还必须从信息科学关于信息的本质角度去认识什么是信息。信息科学的重要奠基人维纳(N. Wiener)在《控制论——动物和机器中的通信与控制问题》中提出了两个经典命题:

命题 1:“信息既不是物质,也不是能量,信息就是信息,不懂得它,就不懂得唯物主义”;

命题 2:“信息是我们用于适应外部世界,并且在使这种适应为外部世界所感知的过程中,同外部世界进行交换的内容的名称。”

我们认为,信息是附属物质和能量的一切世界的第三个要素,信息科学是研究信息的所有属性及其信息、物质和能量之间的依存和转换关系。因此,为了深刻地认识数字信息检索的实质,我们有必要从信息科学的高度去了解信息的基本属性。

属性 1:本原的普遍性

宇宙的任何实体都是信息的源,世界总是在其不同时空点下形成具体结构状态的物质—能量—信息关系。不论是“自然世界”还是“人造世界”,信息所建构的都是关于世界本原的状态、结构、组织、适应性及其相互作用的总体因式。

属性 2:抽象的世界

任何抽象的对象都是信息的具体化。它源于自然世界的物质实在,或者是物质实体之间、实体和抽象对象之间的关系再现的创造。信息不但具有真实世界的属性、状态、特征、结构和动态过程,而且具有它们之间的依存和转换关系。因此,新信息能够基于现有信息及其与抽象世界的关系产生。尽管它被记录、转换、交流,信息仍然是大脑认知及其抽象世界存在的产物。

属性 3:累积性

自然世界是守恒的。根据守恒的自然定律,物质和能量既不能再生,也不能消灭(它只能被转换)。然而,信息是不守恒的,因为信息可以再生、消灭和累积,比如,人类基因信息、病毒信息(包括动物病毒和计算机病毒)、语言信息、科学知识,都遵循不守恒的自然定律,以再生、消灭和累积的方式存在。其中,累积属性是人类依靠信息进化的最重要的属性。

属性 4:可重复使用

由于物质和能量的守恒,所以信息能够被复制,由多个用户对象重复使用,且在数量上没有损失、在质量上也没有降低。信息能够通过通讯而扩大或倍增。因而,信息的重复利用可以导致新信息的创生。

属性 5:无量纲

信息没有空间维度。就此而言,物理实体多大或多小,它们的抽象再现,认知视觉对象,占据的都是一个相似的视界面;唯一的分别可能就是它在视界内有不同。

信息也没有时间维度。一个对象的信息模型是与时间无关的。但如果必需,时间属性又可以与对象关联。这就是为什么我们对最近的事件有一个精确的时间感觉,但对历史事件却是相当模糊的。

属性 6:无重量

基于属性 5(无量纲)的直接推论,即信息的重量总是为零。由于这个属性,我们才可以对下列现象进行解释:一个空的或满的硬盘为何具有同一重量;一个空白的或满记录的磁带、光盘在重量上没有不同;一个芯片存满了 0,1 或它们的任意组合,与一个格式化了了的空芯片相比,重量也完全相同。

属性 7:可变性

信息、物质和能量是世界的三个基本要素,它们之间是可以相互转变的。信息、物质和能量要素之间有多种可能的创生、依存、转换、变化关系,并且这些可变关系能够借助数学方

法来描述。计算机就是人类基于这种可变性计算方法的杰出产物。

属性 8:多种再现形式

此属性与属性 2(抽象的世界)相关。我们观察信息,它能够被多种形式再现,诸如等价物(电子、光子)、相似物(声音、视像)、人造物(口语、书面语言)和数字机器(电视、计算机)。在上述分类中,信息再现的数字化是最一般和基本的方法。数字化的认知基础是大脑内信息的离散性再现及其个体神经细胞单位的离散性处理;数字化的机器处理基础则是运用计算信息学方法模拟人的认知信息处理过程。因而,离散的可再现性是信息再现、存储和处理的基础。它也是现代数字信息工程和数字信息检索的科学基础。

属性 9:多种传输载体

它与属性 8(多种再现形式)平行,信息能够被多种载体(介质)传输,例如以下载体及其组合:电子的、电的、磁的、光的、机械的、书面的、口头的和符号的。值得注意的是,一个确定的媒体可能传输一种或多种形式的信息。对照而言,一个给定的信息形式可能由不同的载体传输。

属性 10:多种通讯方式

考虑信息可以被多种形式再现和多种载体传输的这两个事实,它的通讯也可以以多种方式完成。信息的通讯方式有以下四个层级:

信息碰面:一对一(1-to-1)

信息广播:一对 N(1-to-N)

信息收集:N 对一(N-to-1)

信息网络:N 对 M(N-to-M)

现代流行的数字信息检索过程就是这种通讯方式的高级表现形式。信息网络这种通讯方式是属性 5(无量纲)的推广,它表明空间性信息运动既可以相对地不考虑时间要素,也可以相对地不考虑空间要素。

1.1.2 什么是知识、数据、事实和文献

1. 什么是知识(Knowledge)

这个问题历来就有多种回答。从哲学的角度看,知识是客观事物的属性和联系的反映,是客观世界的数据、事实和信息在人脑中的主观映象。根据皮亚杰的思想和当代认知心理学的信息加工观点,知识是主体通过与其环境相互作用而获得的信息及其组织。国际经济合作组织主编的《知识经济》(The Knowledge Based Economy)中对知识的界定,采用了西方 20 世纪 60 年代以来一直流行的说法——知识就是知道什么(Know-What)、知道为什么(Know-Why)、知道怎么做(Know-How)、知道谁(Know-Who)。这样的界定可以概括为“知识是 4 个 W”。

广义地说,知识是人们通过学习、发现或感悟到的对世界的认识的总和,是人类认识的结晶。狭义地说,知识是一种有组织的经验、价值观、相关信息和洞察力的组合。利用知识所构成的框架,人们可进而评价和吸收新的经验与信息。因此,知识是建立在数据、事实和信息基础之上的。

知识之所以在数据、事实和信息之上,是因为它与人类主观目的和决策相关,更适合或

接近人类行动。信息、数据、事实虽给出了一些有一定意义的东西,但它往往会随时间流逝,效用价值开始衰减。只有通过人们对信息、数据、事实进行归纳、演绎、比较等手段进行挖掘之后,使其有价值的部分沉淀下来,并与已存在的人类知识体系相结合,这部分有价值的信息、数据、事实才会转换成知识。

2. 什么是数据(Data)

比较典型的定义有以下几种:

(1) 数据是对现实生活的理性描述,尽可能地从数量上反映现实世界,也包括汇总、排序、比例等。

(2) 数据是一系列外部环境的事实,是未经组织的数字、词语、声音和图像等。

(3) 数据是计算机程序加工的“原料”。例如,一个代数方程求解程序中所用的数据是整数和实数,而一个编译程序或文本编辑程序中使用的数据是字符串。

(4) 数据泛指对客观事物的数量、属性、位置及其相互关系的抽象表示,以适合于用人工或自然的方式进行保存、传递和处理。

其中第四个定义较符合我们对数据的理解。例如,水的温度是 100℃、礼物的重量是 500 克、木头的长度是 2 米、大楼的高度是 100 层。在这些表述中:水,温度,100℃,礼物,重量,500 克,木头,长度,2 米,大楼,高度,100 层就是数据。数据一般指单独的事实,它本身没有什么意义,除非在一定的上下文中,否则没有什么用处,如数字“10”。通过数据的描述,我们的大脑里就形成了对客观世界的清晰印象。从这些例子中我们看到,数据要通过人们约定俗成的字符和定义表现出来。我们也可以把这些字符和定义称之为数量关键词。数据就是通过对这些关键词的应用把人类认知的物质世界清晰地描述出来。

人类对于数据的使用和管理可以追溯到远古时代,对数据存储的载体从远古的结绳记事一直发展到纸张和现代的磁载体。传统的数据管理是数据项、记录和文件的管理,现代的数据管理是数据库管理。数据项是可以定义数据的最小单位,也叫元素、基本项、字段等。记录是由若干相关联的数据项组成,是关于一个实体的数据总和,构成该记录的数据项表示实体的若干属性。为了唯一标识每个记录,就必须有记录标识符,也叫关键字。文件是一给定类型的(逻辑)记录的全部具体值的集合。文件用文件名称标识。文件根据记录的组织方式和存取方法可以分为:顺序文件、索引文件、直接文件和倒排文件等等。数据库的内部构造是文件的集合,数据库是比文件更大的数据组织。数据库是具有特定联系的数据集合,由于这些文件之间存在某种联系因而它们不能孤立存在。

在数据项、记录和文件管理方法中,数据项、记录和文件是分散的,每个用户或每种处理都有各自的数据项、记录和文件,不同的用户或处理的数据项、记录和文件一般是没有联系的,因而就不能为多用户共享,也不能按照统一的方法来控制、维护和管理。数据库管理很好地克服了这一缺点,它通过复杂的数据模型(如层次模型、网络模型和关系模型)集中组织、控制和管理有关数据项、记录、文件和数据,以保证不同用户和应用可以共享数据项、记录、文件和数据。在信息科学领域,数据项、记录、文件和数据的价值是它可以通过编码被录入到计算机系统中。特别要注意的是,随着计算机软件、硬件技术的发展,计算机信息处理范围的扩大,数据的含义也扩大了,存储的数据也在向着“海量”方向发展。例如,当今计算机可以处理的图像、声音等,也被认为是数据的范畴。

3. 什么是事实(Fact)

我们要重温一下维特根斯坦在《逻辑哲学论》中的观点:

世界可以分解为事实(The world divides into facts)。

事实是由原子事实(Atomic facts)组成的。

一个原子事实是多个对象(Objects)的组合。

对象是简单的(The object is simple)。

即:世界—事实—原子事实—对象,这样一个从整体到局部、从抽象到具体的认识之链,它反映了客观事物的存在状态,是表现对象、载荷对象的符号。在这个层次中,对象作为最基本的模块,构成了整个认识大厦的基石。对象通过相互之间的复杂的关联构成了整个世界。这个观点也是信息科学中面向对象理论的基石。

事实的最大特点是客观性,即客观、真实。事实在一般场合下是一种未经处理过的客观事物状态,但它也可能被人改变。即使被人更改或变换,它仍然只属于客观的不以人的认识、意志为转移的事物状态。所谓“事实胜于雄辩”、“真的假不了,假的也真不了”就是指事实的这种客观性。

4. 什么是文献(Documentation)

了解什么是知识、数据、事实之后,我们还必须了解与信息概念紧密相关的“文献”。世界上最早形成“文献”这一概念的国家是中国,最早见于《论语》,以后一直沿用。在中国古代,有把“文”释为典籍、“献”释为贤人的传统,“文献”即文与人的组合概念。而在西方,直到20世纪初才开始创造“文献”一词。1903年,比利时人保罗·奥特莱特,即Institut International de Bibliographie的创建者,创造了术语“documentation”(文献),以此表示提供文件给信息寻找者的过程。

目前,关于文献的权威定义出自《文献情报术语国际标准》(ISO/DIS5127)。它将文献定义为:“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的,在载体内、载体上或依附载体内而存储有信息或数据的载体。”该定义强调了文献是可以进行加工处理的“单元”;存储(记录)“单元”的载体形式是多种多样的。简而言之,文献是记录有信息、知识、数据、事实的一切载体。文献的最主要的特征是拥有不依赖于人的“单元”载体,只要这些物质载体没有损坏或消失,所记录的信息、知识、数据、事实就可以跨越时空无限往复地为人类利用。

文献是记录有信息、知识、数据、事实的一切载体。构成文献有三个最基本的要素,即:

- (1) 构成文献内核的信息、知识、数据、事实;
- (2) 负载信息、知识、数据、事实的物质载体;
- (3) 记录信息、知识、数据、事实的符号系统。

信息、知识、数据、事实是文献的内容,离开了信息、知识、数据、事实便不能称之为文献。物质载体是符号赖以依存的寄主,从古代的甲骨、金石、竹简、树叶、羊皮、纸张到现代的磁带、磁盘、光盘,文献载体已经历了几代更替。符号系统是信息、知识、数据、事实的携带者,如语言、文字、图画、公式、声像、编码等等。物质载体和符号系统是信息、知识、数据、事实进入文献记录的前提和具体的形式、方法和过程,例如手工书写、绘画、机械印刷、光记录、电记录、磁记录,等等。文献记录的载体材料、符号系统及其具体形式、方法和过程的每一次演变、创新,都带来信息传播方式的重大变革,从而推动社会的发展。因此,文献有以下三个基

本属性:

(1) 信息性。这是文献的本质。离开了信息性(广义地),即离开信息、知识、数据、事实这个内核,文献便不复存在。

(2) 传递性。文献能帮助人们克服时间与空间上的障碍,在时空中传递人类已有的信息、知识、数据、事实,使人类因信息的交流而得到巨大的发展。

(3) 动态性。文献并非处于静止状态,其蕴含的信息随着人类社会和科技的发展在不断地有规律地运动着。

文献的信息性、传递性、动态性是社会信息学、信息计量学和信息检索研究的重要内容。

1.1.3 信息与相关概念的区别

信息是一个具有极大影响力的概念,作为一个有待阐释的术语,它与很多概念相关联,这要看理论定向的需求和迫切程度。例如,香农曾说过:“‘信息’这个词在信息论的一般领域已经由各位作者赋予不同的意义。至少,这些意义很可能在某些应用领域充分证明是有用的,需要进一步的研究和做出永久性的承认。几乎不能指望一个单一的关于信息的概念能够令人满意地对一般领域的各种可能的应用负责。”

英国科学哲学家卡尔·波普尔(K. Popper)提出的“三个世界”的理论,从哲学高度阐述了信息的概念体系。在波普尔看来,信息有“三个世界”:第一世界是包括物质实体和物质状态的物质世界,简称世界1。第二世界是精神的或心理的世界,包括意识状态、心理素质、主观经验等,简称世界2。第三世界是思想内容的世界、客观知识的世界,简称世界3。根据这个理论,信息分为三大类:第一类是有关客观物质世界的信息,即本体论意义上的本原信息,它反映事物运动的状态及其变化的方式;第二类是有关人类主观精神世界的信息,即认识论意义上的抽象信息,它反映人类能感受的事物运动状态及其变化方式,处于意识、思维状态;第三类是有关客观意义上概念世界的信息,即主体论意义上的累积信息,它反映人类所表述的事物运动状态及其变化方式,用语言、文字、图像、影视、数据等各种载体来表示,汇成一个实在的自主的“信息世界”。以“三个世界”的理论来研究信息、知识、数据、事实和文献,它们之间存在以下关系:

信息存在于全部的三个世界中(主观世界、客观的物质世界、客观的概念世界),知识存在于主观世界和客观的概念世界,但不存在于客观的物质世界中,因此知识包含于信息之中。数据、事实存在于主观世界、客观的物质世界,但不存在于客观的概念世界中,因此数据、事实包含于信息之中。文献存在于客观的概念世界,但不存在于主观世界、客观的物质世界,因此文献也包含于信息之中。在信息及其相关概念中,知识离信息最近,二者的区别最容易混淆。

概括起来,信息和知识的区别有三点:

- (1) 多样性。信息是零散的、片断的、特殊的;知识是结构的、连贯的、普遍的。
- (2) 时限性。信息是适时的、短时的、甚至是短命的;知识是持久的、无限的。
- (3) 空间性。信息是跨越空间的一种流动;知识是一种明确的存在定位,但空间上是无限的。

简言之,信息可以设想为一个过程,知识则反映一个过程;信息可以作为一个事物,知识

则反映一个事物。

由于人类获取信息的根本目的是获得客观世界的知识。所以,本书认为,信息与知识的关系可以简化为“信息—知识—信息—知识”这样一种循环转换关系之中。

本书打算从知识接受者的角度探讨如何检索并利用各种数字信息,实现知识的再创新。

1.2 数字信息的定义与类型

1.2.1 数字信息的定义与特点

从古至今,信息的再现形式、传输载体和通讯方式已经历了几代更替。今天,信息记录的载体材料、符号系统及其具体形式、方法和过程,不仅古代无法比拟,就是与50年前甚至是20年前相比也有天壤之别。直到20世纪末以前,人类还只能通过视觉、听觉和以及电报、电话、传真和电视的模拟信号传输获取现实的信息,借助于书写或排字印刷文献的阅读获取他人或前人的间接信息。现在,人类能够通过多种形式的数字信息通讯、处理工具,实现声音、图像和文字的“海量”数字信息获取和交流。

数字信息是一种处理信息的“物理符号系统”,它以冯·诺伊曼模型为基础,在信息的存储、通讯、计算和逻辑操作、控制、输入/输出上是以0和1两个数字的编码组合来实现。数字信息的通讯特点是它的电子波形幅值被限制在有限个数值之内,且不是连续的而是离散的。因此,这种离散的信号被称为“数字信号”(Digital Signal)。

信息科学中定义信源发生信息量的度量单位是“比特”(bit)。一个二进制码元所含的信息量是一个“比特”,所以信息传输速率的单位是比特/秒(bit/s)。例如一个数字通信系统,它每秒传输600个二进制码元,它的信息传输速率是600比特/秒(600 bit/s)。由于这里的码元可以是二进制的,也可以是多进制的。符号传输速率 M (单位为波特)和信息传输速率 R 有以下关系:

$$R = N \log_2 M$$

当码元为二进制时 M 为2;码元为四进制时 M 为4,……如果符号速率为600波特,在二进制时,信息传输速率为600比特/秒,在四进制时则为1200比特/秒。

由于香农的信息通讯理论避开了信息的语义,只能从信息的语法方面建立传递信息的通信系统,是一种模拟信息处理方式。而冯·诺伊曼电子计算机则实现了一种处理信息的“物理符号系统”,是一种数字信息处理方式,这是人类智能物化的新起点。

一切以冯·诺伊曼模型为基础而生产、制作的数字形式都是数字信息。而所谓数字形式,简单地讲就是指能被电子计算机识别的,不同序列的“0”和“1”代码构成形式。由于可以使用数字代码形式来再现、存储和传输一切知识、数据、事实和文献,包括文字、图片、声音、动态图像等数字信息,这就彻底改变了人类的信息传播与利用方式,为实现数字信息的广泛共享创造了条件。

数字信息不同于传统的印刷型文献和各类模拟视听资料,与之相比,其特点主要有:

1. 存储介质和传输形式发生变化

利用数字信息技术进行制作,将信息存储在磁盘、光盘、硬盘等介质上,并可以通过光纤、卫星、微波等传输形式实现远距离的“海量”信息传输。数字存储介质便于信息传输、交换和携带,数字传输形式能够将传统印刷本上的文字、图像及各类音像资料中的声音、动态图像结合在一起。一张普通 CD 光盘的最大存储量为 700 MB(兆),而一张蓝光(Blu-ray) DVD 光盘的最大存储量可以达到 27 GB(千兆),后者足够存储多套《不列颠百科全书》(国际中文版,共 20 卷,计 4 300 余万字);而一个 80 GB 的普通硬盘则可以存储多达 1 万册的数字图书,相当于一个小型的图书馆。现在,数字信息的存储容量单位已经从“兆”变为“千兆”(GB)甚至“兆兆”(TB)。一个数据库的容量通常是以 GB 或 TB 为单位计算的。在正常传输速度下,从世界上任何一个地方下载一篇几千字的数字信息资源只需要 1 分钟左右的时间。

2. 资源以多媒体作为内容特征且类型多种多样

多媒体数字资源集文本、图片、动态图像、声音、超链接等多种形式为一体,具体、生动、全方位地向信息寻找者(用户)展示主题,用户可以因此更加深入细致地了解所需信息的内容及其特征。

数字资源的类型多种多样,各种文件格式可以相互转换、交叉和连接。它既包括数据库、电子期刊、电子图书、电子报纸等正式出版物,以及会议论文、专利、标准、学位论文、教学课件等特种文献,也涵盖了 HTML、新闻组(Usenet 或 NewsGroup)、电子公告板(BBS)等非正式出版的数字信息。数字信息交流的途径因此不再是单一化的,而是多层次、全方位的。

3. 信息服务的功能更加方便和快捷

数字信息最初的服务功能只有检索、浏览、下载等简单功能。发展到今天,已经产生了一系列的更加方便和快捷的新功能,例如:(1)多种检索方式和保存检索档案;(2)信息推送、主动报道(如期刊目次推送服务);(3)读者文稿处理和网上投稿;(4)内部引文链接,外部资源链接;(5)文件传递,如 FTP 服务;(6)信息发现,如网络学术资源导航、分类主题指南等;(7)网上讨论,如 BBS、新闻组等;(8)个性化设置,自己最喜爱的书刊、文件存放等。这些服务功能扩展了传统出版物的职能,使数字信息资源得到更大程度、更深入的利用。

与此同时,数字信息的更新速度加快、时效性增强。传统的印刷型出版物需要数月的出版、发行周期,且一旦出版后,信息的内容就无法更改,必须要修订后出版新版本。而数字信息资源的出版、更新和发布就容易得多,只要有人负责不断跟踪各个领域的最新发展变化,就可以随时修改内容,每月、每周、每日甚至实时更新,及时发布给用户。

4. 信息传播不受时间、地域限制

即没有收藏地点(如图书馆)、收藏时间(使用时间)的局限,可以随时随地存取。主要体现在两个方面:一是信息传播的时间、地域界限不复存在。也就是说任何一个地区的读者只要拥有一台与网络相连的计算机并遵循数字信息传播的相关规定,就可以随时浏览、检索、复制、下载网上的数字信息,而不会因地域界限而造成信息传播的障碍。二是阅读数字信息时的独占性不复存在。人们通过网络阅读数字信息时不再需要同时占有该信息的介质和载体,使得不同读者在同一时空中可以阅读同一数字信息,从而消除了印刷型文献所固有的文献载体与文献所包含信息不可分割的矛盾。

数字信息资源的这些特点使人类的信息获取、交流方式都发生了深刻而巨大的变化。

今天,强大的个人桌面计算机连接着全球的数字信息资源网络,通过我们称之为的因特网和WWW许可,几乎可以与任何地方的同事进行即时交流。网络计算机连接着知识发现的工具,支持由基于全球科学家组成的分散合作,实时地收集、交换和分析他们的研究发现。使用电子邮件,遥远的科学家之间也能很容易地进行合作,无须面对面地会见;书稿也能够有效地合著,以最小的延迟考虑提交编辑。科学研究记录的数字化,从数据、文本的产生一直到科学发现的最后合法判决——作品的出版,允许包括数字图书馆在内的网络机构超越创造者(作者)和接受者(读者)的作品时空限制进行正式或非正式的信息学交流。信息技术完全支持科学的“无纸”交流系统,并且变化还在继续发展。科学家从自己的研究成果印刷到电子交流的完成,如果本人愿意在线传递的话,它的时间、空间界限都会不受到任何限制。

1.2.2 不同记录形式的数字信息资源

数字信息资源的类型可以从数字信息的记录、时序、出版形式、制作形式四个方面进行划分。根据记录信息的形式及其符号编码系统的不同,数字信息资源有文本类、图形(图像)类和声(视)频多媒体类三种。

1. 文本类数字信息资源

它是用语言、文字记录的信息资源,是最重要、最主要的数字信息资源,也是狭义的文献。从自然语言角度,它可再分为中文、英文、法文、德文、日文、俄文等。从人工语言及其符号编码系统角度,它是使用计算机文字处理软件产生的文件保存格式,其主要格式有:

(1)TXT 文本文件格式;(2)DOC 文件格式;(3)HTML 超文本格式;(4)PDF 格式。

虽然 DOC、HTML 和 PDF 格式中都可以包含图形,甚至声音等多媒体信息,但由于它主要是以文字、文本信息为主,图形、声音是辅助性的,因而我们仍然把它看成是文本类数字信息。

2. 图形(图像)类数字信息资源

分别包括各种照片、绘画、图形、图谱、图片、图纸、图画、图表、乐谱等。目前,常见的图形(图像)格式大致可以分为两大类:一类为位图;另一类称为描绘类、矢量类或面向对象的图形(图像)。前者是以点阵形式描述图形(图像)的,后者是以数学方法描述的一种由几何元素组成的图形(图像)。一般说来,后者对图像的表达细致、真实,缩放后图形(图像)的分辨率不变,在专业级的图形(图像)处理中运用较多。

图形(图像)类数字信息资源的文件保存格式主要有:

(1)BMP 格式;(2)JPG 格式;(3)GIF 格式;(4)TIFF 格式;(5)PSD 格式;(6)CDR 格式;(7)AI 格式;(8)DXF 格式。

3. 声(视)频多媒体类数字信息资源

“多媒体”一词译自英文“multimedia”,而该词又是由 mutiple 和 media 复合而成的。“媒体”(Medium)一词原有两重含义,一是指存储信息的实体,如磁盘、光盘、磁带、半导体存储器等媒质;二是指传递信息的载体,如数字、文字、声音、图形等媒介。与多媒体对应的一词是单媒体(Monomedia),从字面上看,多媒体就是由单媒体复合而成。多媒体系统需要将不同的媒体数据表示成统一的结构码流,然后对其进行变换、重组和分析处理,以进行进一步的存储、传送、输出和交互控制。它既用文字、图、表、符号,也用声音、影像等记录,是集文字、声

音、图形(图像)、影像于一体的数字信息资源,目前最为流行的声(视)频多媒体文件格式有:

(1)WAVE,扩展名为 WAV。(2)MOD,扩展名为 MOD、ST3、XT、S3M、FAR、669 等。
(3)MPEG-3 和 MPEG-4。MPEG-3 扩展名为 MP3。MPEG-4 的扩展名包括 MP4、
asf、mov 和 DivX、AVI 等。(4)Real Audio,扩展名为 RA。等等。

1.2.3 不同文献时序形式的数字信息资源

根据文献时序形式及其有序化程度的不同,数字信息资源有以下四种类型:

1. 零次文献(Zeroth Literature)

它也称零次信息。指未经正式发表或不宜公开和大范围内交流的比较原始的素材、底稿、手稿、书信、工作文稿、工程图纸、考察记录、实验记录、调查稿、原始统计数字,以及各种口头交流的知识、经验、意见、论点等,此类更为原始性的文献多保留于科技人员之手。它是一次文献信息资源形成前对作者学术思想、研究成果、技术发明、文艺创作等的最原始的记录。

此类文献的印刷形式为手抄件、打字件、油印件、内部铅件、复印件、描图等;电子形式为内部录音、录像、E-mail、BBS 帖子、CHATROOM 神聊、电子记事、电子文档等。

零次文献的特点是信息来源直接、真实,内容新颖。这类文献在较小的范围内交流、使用、参考,其传播渠道少,或常保密,或限制使用以及因珍稀的原因,不为人们知晓而多被埋没。非常重要的零次信息资源以档案或作数据备份方式长期保存。

2. 一次文献(Primary Literature)

它也称原始信息,习惯上称作原始文献。它是作者依据本人的科研和工作成果而形成的文献,是对作者学术思想、研究成果、技术发明、文艺创作等的正式记载,是科研成果的一种主要表述方式,组织了可供交流的系统性信息、知识、数据和事实。

此类文献的印刷形式主要包括图书、期刊和报纸、科学考察报告、研究报告、会议论文、学位论文、专利说明书、技术标准、政府出版物、产品样本等;电子形式包括事实数据库、电子期刊、电子图书、电子预印本和发布一次文献的正式学术网站(如作者自建档)等。

一次文献有如下特点:

(1)创造性:一次文献是人们根据自己在生产和科学研究中的成果撰写的,是创造性劳动的结晶,作者对自己的作品享有著作权。它报道新成果、新技术、新发明、新创造。例如专利,它具有新颖性和创造性,反映了发明创造、技术革新与改进的创造性劳动成果。正由于一次文献的创造性,故受到人们重视。

(2)原始性:一次文献是初次发表的一种原始的创作,一般是作者根据自己所积累的原始素材、原始数据创作而成,故其既有真实可靠的一面,又有特定原始创作的不成熟一面。

(3)多样性:一次性文献是每个作者的不同成果,故在内容上多样化;另外,在表现形式上也呈现出多样性,有文学读物、期刊论文,也有研究报告、学位论文等等。

3. 二次文献(Secondary Literature)

它也称二次信息,习惯上有时又称检索工具(Retrieval Tools),是指根据实际需要,按照一定的科学方法,将特定范围内的分散的一次文献进行筛选、加工、整理使之有序化而形成的文献。由于它能较为全面地、系统地反映某学科、某专业的文献线索,因而是检索和评价