

信息资源管理 技术概论



汤大权 唐九阳 钱猛 刘芳 / 编著



INXI ZIYUAN GUANLI
JISHU GAILUN

国防科技大学出版社

信息资源管理技术概论

汤大权 唐九阳 编著
钱 猛 刘 芳

国防科技大学出版社
·长沙·

内 容 简 介

本书主要讲述信息资源管理的概念、基本原理和技术体系,比较全面地介绍了信息资源的描述、储存、传递、传播与利用等技术方法。全书共6章,内容分别为:信息资源管理概述,信息资源描述,信息资源存储,信息资源检索,信息资源分发,信息资源安全。

本书内容新颖、理论体系完整、可操作性强,可作为高等院校信息系统工程、计算机信息管理、管理工程等专业本科生教材,也可作为管理人员和计算机应用人员的参考书。

图书在版编目(CIP)数据

信息资源管理技术概论/汤大权等编著. —长沙:国防科技大学出版社,2009.
11

ISBN 978 - 7 - 81099 - 617 - 4

I . 信… II . 汤… III . 信息管理—概论 IV . G203

中国版本图书馆 CIP 数据核字(2009)第 197819 号

国防科技大学出版社出版发行

电话:(0731)84572640 邮政编码:410073

<http://www.gfkdcbs.com>

责任编辑:耿 筠 责任校对:王 嘉

新华书店总店北京发行所经销

国防科技大学印刷厂印装

*

开本:787×960 1/16 印张:15.75 字数:300 千
2009 年 11 月第 1 版第 1 次印刷 印数:1 - 1020 册

ISBN 978 - 7 - 81099 - 617 - 4

定价:36.00 元

前 言

信息资源管理是计算机科学、信息科学、图书馆学、情报学、社会学、管理科学与经济学等学科相结合的综合性强、交叉性显著、具有独特风格的交叉学科，同时，它又是一门实践性很强的应用科学，在实践中产生，又在实践中不断发展并完善。它的主要任务是研究信息资源管理过程内在的规律以及基于计算机等现代化手段的形式化表达和处理规律。经过人们不断探索和实践，目前已形成了信息资源管理独具特色的理论和技术体系，其应用的触角已深入到社会生活的各个方面。

信息资源管理的基本概念、原理和方法，对了解信息资源管理的一般过程，掌握信息资源管理过程的各个环节及相互间的作用，提高信息资源管理的效率具有十分重要的应用价值，对实际建设信息系统有着重要的理论指导意义。

本书主要向读者介绍了信息资源管理的基本概念、原理、方法和技术，以期帮助读者运用系统工程思想和现代计算机技术去认知和把握信息资源管理的内涵和外延，掌握信息资源管理的基础理论、方法和技能。

本书共6章，内容分别为：第1章介绍信息、信息资源、信息资源管理的基本概念，以及信息资源管理的历史和现状；第2章介绍信息资源描述模型、信息资源建模语言及建模工具；第3章介绍信息资源存储介质、存储接口以及各种存储体系结构；第4章介绍信息资源检索模型和检索方法；第5章介绍信息资源分发技术原理；

第6章介绍多层次的信息资源安全策略。

在共同讨论形成本书大纲的基础上,汤大权撰写第1、4章,唐九阳撰写第3、5章,钱猛撰写第2章,刘芳撰写6章。全书最后由汤大权负责统稿。

国防科技大学信息系统与管理学院副院长张维明教授、国防科技大学信息系统与管理学院指挥自动化系副主任肖卫东教授一直关心本书的撰写与出版,在此一并致谢。

信息资源管理技术涉及领域宽、内容多、发展快,加之编者水平有限,疏漏、不当和错误之处在所难免,恳切希望广大读者批评指正。

本书得到国家自然科学基金项目(编号:60903225)、国防科技大学“十一五”重点课程建设项目(编号:1151B018)等课题资助。

作 者

2009年9月

目 录

第1章 概 述

1.1	基本概念	(1)
1.1.1	信息	(1)
1.1.2	信息资源	(7)
1.1.3	信息资源管理	(15)
1.2	信息资源管理的历史和现状	(16)
1.2.1	信息资源的起源	(16)
1.2.2	信息资源管理的发展阶段	(18)
1.3	信息资源管理的过程	(20)
1.3.1	信息需求分析	(20)
1.3.2	信息源分析	(21)
1.3.3	信息资源采集	(22)
1.3.4	信息资源加工和存储	(23)
1.3.5	信息资源检索	(24)
1.3.6	信息资源开发	(25)
1.3.7	信息资源利用和传递	(27)
1.4	本书的组织结构	(28)
1.5	习 题	(29)

第2章 信息资源描述

2.1 信息资源模型的分类	(30)
2.1.1 结构化模型	(31)
2.1.2 半结构化模型	(35)
2.1.3 基于语义的信息模型	(40)
2.2 Web 信息资源描述	(46)
2.2.1 从 SGML、HTML 到 XML	(46)
2.2.2 SGML	(48)
2.2.3 HTML	(49)
2.2.4 XML	(52)
2.2.5 RDF	(55)
2.3 XML 结构化网络文档	(61)
2.3.1 XML 语言	(61)
2.3.2 结构化	(65)
2.3.3 命名空间	(75)
2.3.4 XML 文档的寻址和查询	(76)
2.3.5 XML 处理	(80)
2.4 小 结	(86)
2.5 习 题	(87)

第3章 信息资源存储

3.1 概 述	(88)
3.2 存储技术基础	(89)
3.2.1 存储介质	(89)
3.2.2 存储总线接口	(97)
3.2.3 RAID 阵列技术	(106)

3.3 直连存储 DAS	(114)
3.4 网络附加存储 NAS	(115)
3.5 存储区域网 SAN	(117)
3.6 IP 存储	(123)
3.7 存储虚拟化	(125)
3.8 小 结	(129)
3.9 习 题	(129)

第 4 章 信息资源检索

4.1 概 述	(130)
4.1.1 信息检索系统	(132)
4.1.2 检索任务与文档的逻辑视图	(133)
4.1.3 信息检索的过程	(135)
4.2 检索性能评价	(137)
4.2.1 检索性能的关键度量指标	(138)
4.2.2 其他的度量	(142)
4.3 信息资源检索的模型	(144)
4.3.1 信息检索模型的概念	(144)
4.3.2 传统的信息检索模型	(146)
4.3.3 结构化文本检索模型	(148)
4.3.4 浏览模型	(151)
4.4 基于内容的多媒体信息检索方法	(154)
4.4.1 基于内容的多媒体检索系统结构	(156)
4.4.2 图像检索	(158)
4.4.3 视频检索	(163)
4.4.4 音频检索	(167)
4.5 Web 上的信息检索	(170)

4.5.1 Web 信息检索中的问题	(170)
4.5.2 Web 信息检索的方法	(171)
4.5.3 Web 搜索引擎	(172)
4.5.4 搜索引擎的性能	(179)
4.6 小 结	(180)
4.7 习 题	(181)

第 5 章 信息资源分发

5.1 概 述	(182)
5.2 内容分发网络 CDN	(185)
5.2.1 CDN 系统组成	(185)
5.2.2 CDN 工作原理	(187)
5.2.3 CDN 的关键技术	(189)
5.3 IP 组播	(192)
5.4 P2P	(196)
5.4.1 P2P 网络结构	(198)
5.4.2 P2P 网络的应用	(204)
5.5 小 结	(206)
5.6 习 题	(207)

第 6 章 信息资源安全

6.1 概 述	(208)
6.2 物理安全	(210)
6.3 数据安全	(212)
6.3.1 加密的基本概念	(213)
6.3.2 对称密码	(214)
6.3.3 公钥密码	(216)

6.3.4	HASH 函数	(219)
6.3.5	PKI 公钥体系	(221)
6.4	网络安全	(223)
6.4.1	防火墙	(223)
6.4.2	病毒防护	(226)
6.4.3	入侵检测	(229)
6.5	用户管理	(235)
6.6	安全策略	(237)
6.6.1	安全管理组织机构体系	(238)
6.6.2	法律法规管理	(238)
6.6.3	行政制度管理	(239)
6.7	小 结	(240)
6.8	习 题	(241)
参考文献		(242)

第1章 概述

【本章摘要】 随着信息技术在全球迅猛的发展，人们每天几乎都沉浸在信息的海洋之中，信息、材料和能源共同构成了现代社会发展的三大资源。随着技术的进步以及经济的发展，信息资源的重要性将更加突出。本章描述了信息、信息资源以及信息资源管理的基本概念，详细阐述了信息资源管理发展的历史沿革以及研究现状，信息资源管理的过程，并给出全书的总体组织轮廓。

1.1 基本概念

信息在社会生产和人类生活中起到越来越大的作用，并以其不断扩展的内涵和外延，渗透人类社会、经济和科学技术的众多领域，使人类继工业社会之后，正式迈入信息社会。信息的增长速度和利用程度，已成为现代社会文明和科技进步的重要标志。我们处于信息时代，每天都会同信息打交道，在生活、学习以及工作之中，人们几乎时刻在感受、接收、传播以及利用信息，信息已经成为人们生活中必不可少的要素。那么，什么是信息呢？这个答案可能是五花八门的，各个不同学科领域由于研究的侧重点不同，对信息的定义也不尽相同。

1.1.1 信息

信息一词古来有之，在汉语中“信”和“息”两个字都有音信和消息的含义，多将信息看作消息的同义语。清代的王之春在他所编著的《椒生随笔》中认为，“消息、信息同义，皆俗语也”，同时，还引用了《三国志》以及《二程外书》中的例子进行说明。宋朝的词人柳永也有“香闺别来无信息，云愁雨恨难忘”以及“鳞鸿阻，无信息”的词句。英语中信息所对应的单词是“Information”，它源于拉丁字“Informatio”，是解释、陈述的意思。自20世纪30~40

年代以来,随着通信以及计算机技术的发展,对人类社会的生活以及生产方式产生了深刻影响,促使越来越多的研究人员从各自的专业角度开始认真探讨信息的确切定义。

对信息这个概念进行科学阐述可以追溯到 1928 年哈特莱 (Hartley) 在《贝尔系统技术杂志》(The Bell System Technical Journal) 上发表的题为“信息传输 (Transmission of Information)”的文章,文中认为信息是选择通信符号的方式,并且采用自由度来计量这种信息的大小。通信系统的发信端总有一个符号表,发信者发出信息的过程就是按照某种方式从这个符号表中选出符合特定序列的过程。若符号表中有 S 个符号,发信者选定的符号序列的长度是 N ,也就是发信符号序列中包含 N 个符号。那么在符号表中有 S^N 种选择方式得到这个序列,即有 S^N 种长度为 N 的不同序列。发信者的发信过程就是从 S^N 个不同序列中选择某个特定序列的过程,这种选择过程是一步一步完成的,经过 N 次选择,一个符号一个符号地确定需要发送的特定序列。每当发信者选择了一个特定的符号,也就排除了一批序列,直到 N 个符号全部选定,就唯一确定了发送序列,而排除了其他 $S^N - 1$ 种序列。因此,哈特莱就推导出了最早的信息度量公式

$$H = \lg S^N$$

其中 H 为信息量, S 为符号数, N 为选择的次数。

哈特莱的信息定义是一个巨大进步,其意义在于为客观地测度信息提供了思路,从而为申农 (Shannon) 信息论的产生创造了条件。哈特莱的定义就其时代而言是个进步,但是也存在很大的局限性。首先,在定义中没有涉及信息的内容和价值,只是属于语法范畴的定义;其次,即使在语法范畴内,这个定义也没有考虑到信息的统计性质,也就是没有考虑到符号表中,符号在队列中的出现频率;第三,将信息理解为选择的方式,就必须有选择主体作为限制条件,这在定义中也没有完美地表示。

1948 年,美国贝尔实验室的申农在《贝尔系统技术杂志》上发表了“通信中的数学原理 (A Mathematical Theory of Communication)”,1949 年他又发表了“在噪声中通信 (Communication in the Presence of Noise)”,在信息的认识方面取得了重大进展。他提出了信息量的概念和信息熵的计算方法,标志着信息科学进入定量的研究阶段,并因此被视为信息论的奠基人。在文章中,申农给信息下了一个高度抽象化的定义“信息是用以消除随机不确定性的东西”。随机不确定性的东西是在传播的过程中确定的,信号从信源出发,经过信道传递到信宿。该定义考虑了信息的随机不确定性,但是也未能包含信息的内容和价值,同时,定义只是考虑了随机型的不确定性,也不能解释模糊不定性等形式的语

法信息。

在同一年,维纳(Norbert Wiener)出版了他的专著《控制论——动物和机器中的通信和控制问题》(Cybernetics: or Control and Communication in the Animal and the Machine),并创立了控制论。他从控制论的角度认为“信息是人们在适应外部世界,并使这种适应反作用于外部世界的过程中,同外部世界进行互相交换内容的名称。”维纳的信息定义包含了信息的内容和价值,从动态的角度揭示了信息的功能和范围,但是,人们在与外部世界的互相作用过程中,同时也进行着能量和物质的交换,若不加以区别,将使这三者的关系混淆起来。为此,维纳的一句名言“信息就是信息,既不是物质也不是能量”,常常被人们引用为信息的定义。其实,准确地说,维纳在此只是区分了物质、能量和信息之间的界限,指出信息是一类需要单独研究的对象,并没有回答信息是什么的问题。

1975年,意大利学者朗高(G.Longo)在其出版的《信息论:新的趋势与未决问题》(Information Theory: New Trends and Open Problems)一书的序言中指出:信息是反映事务的形成、关系和差别的东西,它包含在事物的差异之中,而不是事物本身。信息就是差异,有差异就有信息。无疑这个观点是正确的,但是“没有差异就没有信息”的说法却是不够确切的。譬如,我们利用实验方法验证某个物理定律的时候,在相同的条件下得到的实验数据是一样的,这些实验数据之间没有差异,但是不能说没有得到信息,我们得出了在一定条件下,这条物理定律成立的信息。

从20世纪中后期开始,信息定义也成为图书情报领域研究的兴趣点,美国学者巴克兰德(M.Buckland)认为,许多事物可以是信息,文本固然是信息,图片、录音磁带、博物馆陈列品、自然物体、实验、事件等也是信息。因此,信息可以定义为事物或者记录。不过这个定义过于宽泛,未能对信息、信息源以及信息载体进行区别。萨克利夫(J.Tague-Sutcliffe)认为,信息是人和人所生产的记录跨越时空与他人所交流的内容,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或图片等是绝对的,但它所包含的信息对于读者而言则是相对的。信息是读者通过阅读或者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,是与文本和读者以及记录与用户之间的交互行为相关的,是与读者大脑中的认知结构相对应的东西。在萨克利夫的定义中引入了用户的概念,是认识论层面上的定义。西班牙学者库拉斯(Emilia Curras)认为,信息是所有人类活动的内驱力和能量,信息可以被传递、被感知和被理解,需要有形的载体以变为实实在在的信息。信息是一种现象和一个过程,前者是指无意识感知的信息,用来

调整我们的知识状态和态度；后者是我们需要和寻求的信息，是从文献中的数据经处理而来的。这个定义混淆了信息、数据以及知识的关系，同时将信息认为是一个过程或现象也不准确，没有体现与物质和能量的区别。

加拿大学者桑盖特（W.Thomgate）从心理学的角度研究了信息的定义，他认为，信息是存在于人类意识之外的东西，它存在于自然界、印刷品、硬盘以及空气之中；知识则是存在于我们的大脑之中，它是与不确定性相伴而生的，人们一般是用知识而不是信息来减少不确定性。心理学领域对于信息的认知无疑带有明显的心理学烙印，但从中我们也可以认识到，信息是外界的，是不依赖于人的意志而存在的；知识则不然，是人类大脑中的产物。

在计算机领域对于信息的认识是从数据处理开始，因此，在很多时候将信息和数据概念混淆在一起，但是随着技术的进步，业界越来越倾向于将这两个概念区分开来。计算机领域中最常见的表述是将数据、信息、知识表达为一个金字塔模型，如图 1.1 所示。

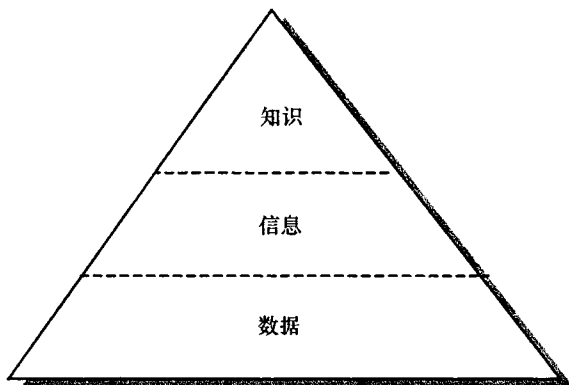


图 1.1 数据、信息、知识的关系

图中最下层是数据，数据（data）可以是文本、图像、视频、音频等基本数据元素。数据是按某一规格化方式对事实和概念的一种表示，适于人或自动装置进行通信、解释或处理。它是任何有意义或可以赋予含义的表达形式，例如字符或数字。在计算机领域，数据是输入给计算机程序或例行程序的内容，它们可以经过算术或逻辑运算的处理，求得所处理后的结果。数据是信息的来源和原始资料，用数据可以表达信息。信息是经过处理、组织和表现出来的数据。或者也可以说，信息是构成一定含义的一组数据。读、听、看、理解的信息，经过归纳和总结得出知识（knowledge）。广义地说，知识有经验知识和理论知识之分，知识是人类认识的成果和结晶，是有助于解决问题的有格式可重

用的信息。图 1.1 中用金字塔形状，下宽上窄，表示层次越高，抽象级别就越高；另一方面，随着层次的上升，要求的表达数据量就越少。还有的学者用“一个事实的生命周期”来解释数据、信息和知识的关系。原始数据总是与新生事物联系在一起的，对原始数据的评价产生了信息，成熟的信息构成了知识，而事实的最终“死亡”形成了相关的知识库。相对而言，信息要比数据重要，知识则比信息重要。

我国学者钟义信教授在《信息科学原理》一书中，比较了多种有关信息的定义，认为信息是事物运动的状态和（状态改变的）方式，并由此推导出了完整的信息概念体系。宇宙间的一切事物都在运动，都有一定的运动状态和状态的改变方式，因而一切事物都在产生信息，这是信息的绝对性和普遍性；同时，由于一切不同的事物都具有不同的运动状态与方式，信息又具有相对性和特殊性。这就是本体论意义上的信息定义，它不受任何条件约束，所表征的事物实际运动状态与方式也不受主体意志的影响，具有最为广泛的适应性。

如果引入认识主体这一约束条件，本体论意义上的信息定义就转化为认识论意义上的信息定义，也就是认识主体所感知或所表述的事物运动的状态与方式。其中，认识主体所感知的事物运动状态与方式，是外部世界向主体输入的信息，称为第一类认识论意义上的信息或感知信息；由认识主体思考（也是一种事物）产生的运动状态及其变化方式（代表认识主体的意志），是主体向外部世界（包括向其他主体）输出的信息，称为第二类认识论意义上的信息或再生信息。认识论层次的信息受认识主体约束，也就是说，没有主体就没有认识论信息。一般而言，在人类所及的有限时空中，本体论信息与认识论信息是可以互相转化的，其转化过程大致与人类认识和改造世界的过程相统一。作为认识主体的人为了认识世界而从事物的本体论信息中提出所需要的信息，如事物的功能结构方面的信息，然后通过分析、综合等思考过程优化了所获得的信息，再作用于事物，使事物产生认识主体所期望的改变。这样，事物就产生了新的本体论信息，新一轮转换过程又开始了……这种周而复始螺旋式上升的过程其实就是人类认识世界和改造世界的过程，这个过程不仅造就了大量的物质资源，而且积累了丰富的信息资源。

信息是一个大家族，要全面系统地认识信息，就必须对信息进行分类。但由于信息本身是一类独立的研究对象，有必要选择多种标准对其进行分类，这样既有助于人们全面地认识信息，同时又有助于人们弄清哪些信息属于信息资源的范畴。

常见的信息分类主要有：

- (1) 以信息的性质为依据，信息可分为语法信息、语义信息和语用信息。
- (2) 以认识主体为依据，信息可分为客观信息（关于认识对象的信息）和主观信息（经过认识主体思维加工的信息）。
- (3) 以主体的认识能力和观察过程为依据，信息可分为实在信息、先验信息和实得信息。
- (4) 以信息的逻辑意义为依据，信息可分为真实信息、虚假信息 and 不定信息。
- (5) 以信息的生成领域为依据，信息可分为自然信息、社会信息和思维信息。
- (6) 以信息的应用部门为依据，信息可分为工业信息、农业信息、军事信息、政治信息、科技信息、文化信息、经济信息等。
- (7) 以信息的记录符号为依据，信息可分为语音信息、图像信息、文字信息、数据信息等。
- (8) 以信息的载体性质为依据，信息可分为文献信息、光电信息、生物信息等。
- (9) 以信息的运动状态为依据，信息可分为连续信息、离散信息、半连续信息等。

对于信息的认识，经历了从简单地抽象，到全面系统概括的历程。而迄今为止，对于信息的定义还没有形成较为一致的认识。这是因为，一方面信息科学本身还是一门年轻的科学，人们对它的认识还有待于进一步深化；另一方面，信息本身的普遍存在、形式变化多端，难于概括。随着信息科学的发展，对信息本质的认识会不断深化，通过近年来对信息定义的变化，可以看到一个趋势，研究者都试图为信息下一个通用于各学科领域的定义，同时也给出带有各自学科特点的定义，使得对信息的认识在抽象和具体两个方向上共同发展，信息概念的内涵和外延都更加完善，也越来越体系化了。

但是对于信息资源管理学而言，它有着不同于其他学科的自身特点。本书认为，在认识论的意义上，信息是客观世界的表征，是关于事物的运动状态和运动方式的反映，能消除或减小主观对相关认知的不确定状态，通过实现与客观世界的交换与互动，适应客观世界，并最终实现与客观世界的统一。对于人类社会而言，信息是区别于物质和能量的另一种基础性社会资源，它能够用来减少或消除人的认知不确定性，以减少人类适应自然和改造客观世界的成本。

1.1.2 信息资源

信息资源是本书的研究对象,弄清它的内涵与外延是信息资源管理理论研究的逻辑起点。从字面和语法上看,“信息资源”是一个偏正词组,“资源”是中心词,而“信息”则是限定语。“信息资源”的含义是指“信息性的资源”或“与信息有关的资源”。因此,理解“资源”是认识“信息资源”的基础。

1. 资源

资源是人类活动的对象和条件。认识、开发、管理和利用资源贯穿于人类活动的始终,资源的存在为人类的社会经济活动提供了最基本的条件。考察一下人类社会发展史就可以发现,资源的范围是伴随着认识的深入和实践的扩展而不断扩大的。现在已很难考证资源一词最早出现于何时,一般认为把它首先作为科学概念来使用是在经济学领域,在经济学中资源是指为创造物质财富而投入生产活动中的一切要素。但经济学最初并不研究资源本身,只研究稀缺资源的有效配置问题,因而它没有给出严格的资源定义。在实际使用中,不同的学科、不同的研究者都有适合自己需要的资源定义。这些定义可归纳为狭义资源和广义资源两类。

狭义的资源观认为,资源就是自然资源,包括自然物质以及自然环境。

广义的资源观认为,对资源的理解并不局限于自然科学,而主要是从社会科学,特别是从经济学(它把资源作为投入要素)和管理学(它把资源视为管理活动的客体)的角度来考察。认为除了自然资源之外,资源还包括人力、财力、智力、文化、时间等。不难看出,这两类资源观之间存在着不少交叉重叠之处,但它们又都从不同的侧面揭示了资源的部分实质。

2. 信息资源的概念

可以说从人类社会有了文献开始,就已经有了信息资源,但是,作为一个专业术语,信息资源则是一个比较新的概念。随着科学技术的进步,特别是以计算机和通信为代表的信息技术的高速发展,从20世纪50年代末、60年代初以来,美国信息经济学家马尔夏克(J. Marschak)、斯蒂格勒(G. Stigler)等人在率先从微观角度入手对信息价值、信息成本、信息价格、信息效率、信息在社会稀缺资源配置中的地位、作用及其机制进行了研究;另一方面马克卢普(F. Machlup)和波拉特(M. U. Porat)等人又从宏观角度入手,对知识产业和信息产业展开了全面的研究。这些研究从理论上证实了信息是一种与物质资源、能量资源同等重要的现实资源。