网络信息开发与利用

DEVELOPMENT AND UTILIZATION OF NETWORK INFORMATION

李月明 著

國家圖書館出版社
National Library of China Publishing House

湖南省哲学社会科学基金项目(07YBA036)研究成果之一

网络信息开发与利用

李月明 著

國家圖書館出版社
National Library of China Publishing House

图书在版编目(CIP)数据

网络信息开发与利用/李月明著. --北京：
国家图书馆出版社，2015.10
ISBN 978-7-5013-5702-4

Ⅰ.①网… Ⅱ.①李… Ⅲ.①网络检索
Ⅳ.①G354.4

中国版本图书馆 CIP 数据核字(2015)第 245570 号

书　　名　网络信息开发与利用
著　　者　李月明　著
责任编辑　金丽萍　王炳乾

出　　版　国家图书馆出版社(100034　北京市西城区文津街 7 号)
　　(原书目文献出版社　北京图书馆出版社)
发　　行　010-66114536　66126153　66151313　66175620
　　66121706(传真)　66126156(门市部)
E-mail　btsfxb@nlc.gov.cn(邮购)
Website　www.nlcpress.com ——→投稿中心
经　　销　新华书店
印　　装　北京科信印刷有限公司
版　　次　2015 年 10 月第 1 版　2015 年 10 月第 1 次印刷

开　　本　880×1230(毫米)　1/32
印　　张　9.25
字　　数　250 千字

书　　号　ISBN 978-7-5013-5702-4
定　　价　38.00 元

序

国际电信联盟(ITU)发布的《衡量信息社会报告》显示,截至2014年底,全球网民约30亿,超过全球人口的40%。中国互联网络信息中心(CNNIC)发布的《中国互联网络发展情况统计报告》显示,截至2015年6月,中国网民规模达6.68亿,互联网普及率为48.8%,网络应用包括即时通信、网络搜索、新闻资讯、影音娱乐、网上购物等方方面面。可见,网络信息正深刻地影响和改变着人们的生产生活。

2012年3月,美国政府斥资两亿多美元推出《大数据研究和发展计划》,认为未来对数据信息的占有和控制将成为国家的核心资产。同年7月,联合国发布了题为《大数据促发展:挑战与机遇》的白皮书,指出各国政府可利用大数据促进社会经济发展,更好地服务和保护人民。中国共产党的十八大报告也明确把"信息化水平大幅提升"纳入2020年全面建成小康社会的目标之一。网络信息越来越上升到国家战略资源的高度。

众所周知,网络信息浩如烟海但杂乱无章,网络信息蕴含了丰富的知识,但质量良莠不齐、价值密度低,这给人们开发利用带来了极大的不便。为此,湖南图书馆李月明撰写了《网络信息开发与利用》一书。该书具有以下三个特点:

第一,全面系统。该书采用理论研究和案例分析相结合的方式,系统而全面地论述了网络信息资源产生、发展和分布现状,网络信息传播特点和趋势,网络信息组织和评价,网络信息开发的原则和形式,网络信息保存和开放存取,网络信息检索方法和技巧,网络信息服务流程和规范,以及数据库、多媒体、数据挖掘、云计算、大数据等网络信息相关技术。

第二,实践性强。该书作者长期从事网络信息开发与利用实际工作,使得该书具有较强的实践性。如第四章通过"湖南地方戏剧数据

库”的开发实践，研究了数据库开发前期如何准备，如何进行必要性与可行性分析，如何搭建数据库平台，如何设计字段、栏目结构，前台页面如何展示，以及功能如何实现；第五章通过检索“图书馆自动化系统”这一主题，再现了不同中外文信息检索系统的特色和检索流程；第六章通过《外媒看湖南》这一舆情信息服务案例，提出当前舆情信息服务存在的问题，并相应给出了对策建议。这些具体而可操作性的实例，可指导网络信息开发和利用的工作实践。

第三，资料丰富。该书具有丰富的实例和引证。如第一章和第二章引用了1997年至2015年的《中国互联网络发展情况统计报告》，分析了我国网络信息分布情况和网络信息传播情况，提出了网络信息优化配置建议和网络管理建议。第七章列举了网络环境下著作权、域名、商标权、隐私权、名誉权侵权行为和不正当竞争行为的特征和大量案例，可指导人们合理合法地开发和利用网络信息资源。

当然，由于该书涉及面广，有些章节的内容尚可更深入。如第三章“网络信息组织”逻辑结构不够严密，理论和实践性不够强，希望作者今后加强这方面的学习和研究，写出更具参考性和实用性的作品。

黄如花

2015年7月24日

目　录

第一章　网络信息资源概述

当今时代，以信息技术为核心的新一轮科技革命正在孕育兴起，互联网正深刻改变着人们的生产生活。在此信息环境下，信息的产生、传播、开发和利用发生了根本性的变化，网络信息呈爆炸式增长，几乎涵盖了人类整个知识领域，网络信息对人类文明和社会发展产生了极其深远的影响。网络信息虽然浩如烟海，蕴含了丰富的知识，但我们也应看到，网络信息杂乱无章、良莠不齐，给人们开发利用带来极大的不便。本书从理论和实践的角度研究了网络环境下信息的生成、特点、分类、传播、组织、开发、检索和服务，以及网络信息的合理合法利用相关问题，希望能发现网络信息传播机理和内在规律，探索科学的网络信息组织方法，开发符合用户需求的信息产品，解决海量信息精准定位问题，帮助人们快速有效地获取网络信息，进而提高网络信息的应用价值，推进我国网络信息业健康有序发展。

第一节　信息的定义和分类

信息作为与自然界和人类社会息息相关的概念，既有着源远流长的历史，又有着众说纷纭的定义，还有着各种各样的存在形式、物质载体和传播媒介。

一、信息的定义

信息是普遍存在的，一切信息来源于自然界，来源于人类社会，人们的生产、生活、科研以及社会活动等都是信息产生的源泉。

“信息”作为一个术语，最早由哈特莱（R. V. Hartley）于 1928 年在

其撰写的《信息传输》一文中使用。20 世纪 40 年代后期，随着信息论、控制论的产生，“信息”作为一个科学的概念应用于自然科学和社会科学的许多领域，成为哲学、数学、系统论、控制论、经济学、管理学等学科共同讨论和研究的重要概念。不同的学科，从不同的角度，对信息这个概念有不同的解释。据不完全统计，信息的定义有 100 多种，至今仍没有统一的、能为各界普遍认同的定义。

在经济学家眼中，信息是反映事物特征的形式，是与物质、能量相并列的客观世界的三大要素之一，是管理和决策的重要依据。哲学家则从产生信息的客体来定义信息，认为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，信息是事物本质、特征、运动规律的反映。在新闻界，信息被普遍认为是事物运动状态的陈述，是物与物、物与人、人与人之间的特征传输，新闻是具有新闻价值的信息。

在图书馆学和情报学领域，信息被认为是数据处理的最终产品，是经过收集、记录、处理，以能检索的形式存储的事实或数据。信息论创始人克劳德·艾尔伍德·香农（Claude Elwood Shannon）认为“信息是用来消除随机不确定性的东西”。美国信息管理专家霍顿（F. W. Horton）给信息下的定义是“信息是为了满足用户决策的需要而经过加工处理的数据”。我国著名的信息学专家钟义信认为“信息是事物存在方式或运动状态，以这种方式或状态直接或间接的表述”。

综合对信息概念的研究成果，目前较为全面的定义是：信息是用符号、文字、信号、数据、声音、图形、图像、动画、视频等形式，通过不同介质与不同渠道的传递和处理来表现各种相互联系的客观事物在运动变化中所具有特征内容的总称。若要获得所需信息，须具备一定的信息素养，包括判断什么时候需要信息，懂得如何去获取信息，如何去评价和有效利用所需的信息，即信息意识和认知能力，信息检索和获取能力，信息分析和评价能力。

二、信息的分类

关于信息的分类，从不同的角度划分有不同的分类方法。

1. 按信息所描述的对象来分，可分为自然信息、生物信息和社会信息等

自然信息是指自然界中处于自然状态的、客观存在的信息，它反映了事物的特征和变化以及自然界共性和特点，如自然景观、农作物长势等。生物信息是指生物体的内在联系及其交往的反映，如遗传密码，动物的声音、动作、姿势和气味等。社会信息是指人类社会在生产和交往活动中，用语言、文字、图表、数值等形式交流或交换，除人的生物信息以外的，与人类的社会活动有关的一切信息。

2. 按信息所依附的载体分，可划分为文献信息、声音信息和电子信息等

文献信息就是文献所表达的信息，以文字、符号、声音、图像为载体，是一种相对稳定的信息，一经“定格”在某种载体上，无法改变其固有属性。这种信息的优点是易识别、易保存、易传播，缺点是不能随外界的变化而变化，造成文献老化。

声音信息指人们口耳相传，或借助无线电、电话、唱片、录音机等传播的信息。声音信息具有出现早、传递快、偶发性强的特点，大部分转瞬即逝，一部分被记录或被录音成为文献而保存下来，一部分通过人类的记忆代代相传而成为口述回忆资料。作为信息留存的一种形式，声音信息无时不在、无处不有，承载着人类的知识、经验和史实，是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、通信技术和多媒体技术迅速发展的产物，一般通过计算机、手机和网络等媒介传播。电子信息在信息的存储、传播和应用方面已经从根本上打破了长期以来由纸质载体储存和传播信息的格局，是当今发展最快、最具应用价值和发展前途的新型信息，代表了信息业发展的方向。

3. 按信息的编码形式分，可划分为文字信息、图像信息、音频信

息、视频信息、数据信息

信息是物质的存在方式、形态和运动规律的表征，必须以某种编码形式储存或传输于某种介质之中。如文字信息记录在书本、纸张上，图像信息印制在画报、照片上，音频信息记录在唱片、录音带中，视频信息存储在磁带、光盘、硬盘等存储介质中，数据信息储存在计算机的信息系统里，等等。

第二节　网络信息的特点和分类

一、网络信息的特点

网络信息是将文字、图片、声音、视频等多种形式的信息，以数字化形式存储，并借助网络读取、组织、发布、传播、检索和利用的信息资源。在网络环境下，信息以计算机可识别的方式存储于网络的某一节点上，并且可以在任何需要的时候通过四通八达的全球互联网络传向任一合法的网络终端用户。

网络信息是目前世界上信息量最丰富、使用最便捷、用户最多、发展最快的信息资源，与传统信息有较大的区别。

1. 网络信息海量化

据 IDC 2012 年发布的研究报告，2011 年全球的信息量是 1.8ZB，预计到 2020 年将超过 35ZB，相当于用 400 亿个 1TB 硬盘来存储这些数据。另有统计称，全球每秒发送 2900 万封电子邮件，推特每天发布 5000 万条消息，亚马逊每天产生 630 万笔交易，全球信息量正在以每 18 个月翻一倍的惊人速度增长。网络信息之所以如此海量，最主要的原因在于：

(1)互联网作为新媒体每时每刻都在产生大量的信息。以 2014 年 3 月 8 日马来西亚航空公司飞北京的航班 MH730 失联事件为例，MH730 客机失联后，全球网络媒体连续一个多月进行全天候的报道，此事也成为论坛、微博、微信等社交网络的热门话题，产生了大量的信

息。国内媒体如网易、新浪、凤凰网、腾讯设了马航失联的专门栏目，进行了图片、文字、视频、音频的综合报道，用户还可以发表评论。2014 年 3 月 8 日至 4 月 22 日，仅针对马航失联的即时报道腾讯网就有 3500 余条，新浪网有 2800 余条。可见互联网这一新媒体产生的信息量之巨大。

（2）数据库技术和网络技术的发展和交叉应用使文献类信息数量剧增，并进一步应用到信息组织和管理、信息检索和利用、人工智能等领域。国外文献数据库种类多、资源丰富，综合性的主要有 Web of Knowledge 学术资源整合平台、Dialog 联机检索系统、Elsevier 全文数据库（SDOL）、OCLC 联合目录数据库等。我国数据库建设虽然起步较晚，但发展很迅速，据不完全统计，我国现有各类综合性数据库 3000 多个，包括中国知网学术文献（CNKI）、维普科技期刊（VIP）、万方数据知识服务平台等。

2. 存储载体数字化

网络信息不是实体信息，是经过人类加工处理的数字化的信息。显示在计算机屏幕上的文字、图形、视频只是数字化信息的输出形式，在计算机内部，无论是存储、编辑还是传输，这些信息均是以数字编码的形式存在。

数字化的信息最大的特点就是压缩性，即通过物理的手段将信息“压缩”，以“0”和“1”保存和传递。这样既可以增加信息存储的容量，又可以提高信息的传输速度。如一个普通的 1TB 硬盘可以容纳 5200 余亿个汉字，相当于一个拥有 100 余万册字数为 50 万字图书的中小型图书馆，而利用网络传递一套 33 卷的《大不列颠百科全书》的图文信息，仅需要 4.7 秒钟的时间。所以，网络信息的数字化特性使存储介质可以存储海量的信息，使其可以快速地在网络上传播。这大大提高了人们获取、处理、存储、管理、传播和利用信息的能力。

3. 信息类型多样化

在互联网发展初期，网络信息的表现形式一般以文本信息为主，图片信息为辅，信息类型多为 txt、jpg、gif 等格式。随着网络技术的发

展,传输速率的提高,网络资费的下调,需要较高带宽的图像、声音、视频等多媒体信息传输成为可能,目前网页的表现形式综合了文本信息和多媒体信息,信息类型包括文本(pdf、txt、doc)、图形图像(jpg、gif)、声音(wav、mp3、midi)、动画和视频(swf、rm、wmv、mpg、avi)等格式。在我国多媒体网络信息中,jpg 信息占比最大(31.5%),pdf 信息次之(31.4%),doc 信息第三(24.2%)。

4. 检索利用便捷化

信息以网络媒介为基础,在全球范围内实现信息资源共享,极大地推动了网络信息检索的发展与普及。网络信息检索指的是通过计算机网络系统连接世界上各种公用数据库和商用数据库,利用超文本的检索工具与多媒体的传输能力来访问、获取和利用信息资源。一个完善的网络信息检索系统不但具有内容新、数据量大、覆盖面广、传输速率快、界面友好、操作简便、不受时空限制等特点,而且还能提供多种信息服务功能,如文件传输、电子邮件、电子论坛、数字图书馆等。

5. 使用成本低廉化

在互联网上,大部分信息资源都可以免费使用,用户只需支付网络使用费用。虽然也有一些有偿的网络信息资源,但是与获取其他形式的信息资源相比,网络信息资源在满足用户信息需求的情况下,节省了大量的人力和时间成本。相对来说,利用网络信息比利用其他信息花费成本少。

6. 信息质量良莠不齐

互联网上大量的信息分散在不同层次和节点上,没有一个中心点。用户在互联网上存储和发布信息时可以匿名,有很大的自由度,且没有系统性和组织性,审核机制也不健全,这就使得网络信息在蕴含了海量丰富的知识的同时,也催生了大量冗余、粗制滥造、虚假甚至有害的信息。信息质量良莠不齐,给用户的利用带来诸多不便,尤其是某些宣扬血腥暴力、侮辱诽谤、色情的低俗信息,容易诱发人们特别是青少年的不良思想行为,干扰青少年正常学习生活,严重者还会扰乱社会治安,造成违法犯罪。

二、网络信息的分类

网络信息包罗万象，是除传统实体型信息资源之外的又一资源宝库。它类型繁杂、形式多样，人们从多种维度对网络信息资源进行了划分。

1. 按出版形式可分为非正式出版信息、半正式出版信息和正式出版信息

非正式出版的信息是指信息量大、流动性强，而质量难以保证和控制的动态信息，如电子邮件、专题讨论小组和论坛、博客、微博、微信、即时通讯等信息。这类信息多由终端用户贡献，可信度不够高。随着 Web2.0 的普及，以及更加个性化的 Web3.0 的到来，这类信息在互联网中占据着越来越大的比重。

半正式出版的信息是指比较正规，但是又没有纳入正式出版信息系统的“灰色”信息。如各政府机构、团体、企业、商业部门、行业协会、大型活动和国际组织等网站所提供的描述性信息，这些信息一般由官方机构发出，具有较高的可信度。

正式出版的信息是指有版权的直接在网络上出版的知识性和分析性信息，包括在网络上正式发行的电子出版物（如电子期刊、电子图书等）、各种数据库（如书目数据库、联机数据库、全文数据库等）、新闻网站发布的新闻（如人民网、新华网、新浪网等）、权威部门发布的报告（如 CNNIC 的《中国互联网络发展情况报告》）以及其他动态信息（如交通、股市行情等）。这些信息等同于正式出版物，可信度很高。需要说明的是，当前，新闻网站越来越强调用户的参与和评论，用户贡献的信息都属于非正式出版信息。

2. 按加工程度可分为一次信息、二次信息和三次信息

一次信息指没有经过加工处理的原始信息，如网络论坛、网络新闻组、电子公告板、博客等实时产生的信息，以及各类网站发布的信息。这类信息在互联网上呈现出数量多、内容杂的特点，而且它没有经过加工整理并且表现为无序状。可以这么说，用户在互联网上搜寻到的大多数

信息都属于“原始信息”。我们必须将存在于数量庞大、排列无序的“原始”信息资源中,将有价值的信息提取出来,才能加以利用。

二次信息是在对大量分散的原始信息进行收集整理、内容浓缩的基础上,按照一定的规则组织而成的可供利用的一种信息资源,如书目数据库、索引数据库、网络文摘、学科导航库、电子图书、电子报刊等。这类信息一般是同类信息的汇编,是引导和使用一次信息必不可少的工具。

三次信息是指对一次信息资源或二次信息资源进行系统分析、综合研究、编辑加工而生成的信息资源。这类信息具有系统性、综合性、知识性和概括性等特点,有较大的参考价值,如综述、专题报告、百科全书等。

3. 按行业可分为政府信息、企业信息、商务信息、教育科研信息、舆情信息和个人信息

政府信息是指党政机关发布的政府新闻、政府职能、统计资料、政策法规、办事指南、通知公告等信息;企业信息是指企业网站提供的企业介绍、产品介绍、企业动态、售后服务、技术支持、行业新闻、招聘信息等信息;商务信息指各类网上购物、电子商务、酒店预定、网上订票、股票交易信息;教育科研信息是指各教育机构、科研机构、学术团体发布的学校/机构介绍、课题申报、学术成果、学术会议、学术动态等信息,以及各类学术数据库,如电子图书、电子报刊等;舆情信息是指各传统媒体网站、各新闻门户网站发布的新闻,现在也将论坛、博客等社交性网站发布的信息称舆情信息;个人信息主要指个人网站发布的信息,以及网络论坛、网络新闻组、电子公告板、博客、即时通讯等媒介中个人贡献的信息。

网络信息还可以按文件类型分为文本信息和多媒体信息,按组织方式分为文件、数据库、主题目录和超媒体四种类型,按数据类型分为结构化信息和非结构化信息,按网络信息资源所产生的功用分为价值信息与非价值信息,按照传输协议划分成 Web 信息资源、网络论坛、FTP、Gopher 等。

第三节　我国网络信息资源特点与优化配置

我国网络信息资源数量巨大、内容丰富、结构多元，分布广泛但并不均匀，不同省份网络信息的质和量有差别。只有深入了解网络信息资源的地域分布情况，才能科学地对网络信息资源进行深层组织、合理开发，从而达到优化结构、高效利用的目的。

一、我国网络信息资源发展特点

从1997年开始，中国互联网络信息中心（CNNIC）开始对我国网民规模、结构特征、接入方式和网络应用等情况进行连续的调查研究，每年1月和7月定期发布《中国互联网络发展状况统计报告》，到2015年1月，该统计报告累计发布35次。笔者抽出1997年、2000年、2005年、2010年和2014年的数据进行了对比分析（如表1-1所示），希望通过对每4—5年网络发展的基础数据的对比，找出我国网络信息资源发展特点。

表1-1　我国网络发展基础数据表

	1997年	2000年	2005年	2010年	2014年
互联网普及率（%）	-	1.6	8.5	34.3	47.9
网民总数（万）	62	2250	11 100	45 700	64 900
国际出口带宽（M）	25.4	2799	136 106	1 098 957	4 118 663
域名总数（万个）	0.4（CN）	12.2	259.2	866	2060
网站总数（万个）	0.15	26.5	69.4	190.8	335
网页个数（亿个）	-	1.5	24	600	1899

从上表可以看出，我国网络信息资源发展速度相当快，具体来说：

1. 互联网普及率

2000年我国互联网普及率仅1.6%，2014年达到了47.9%，年均