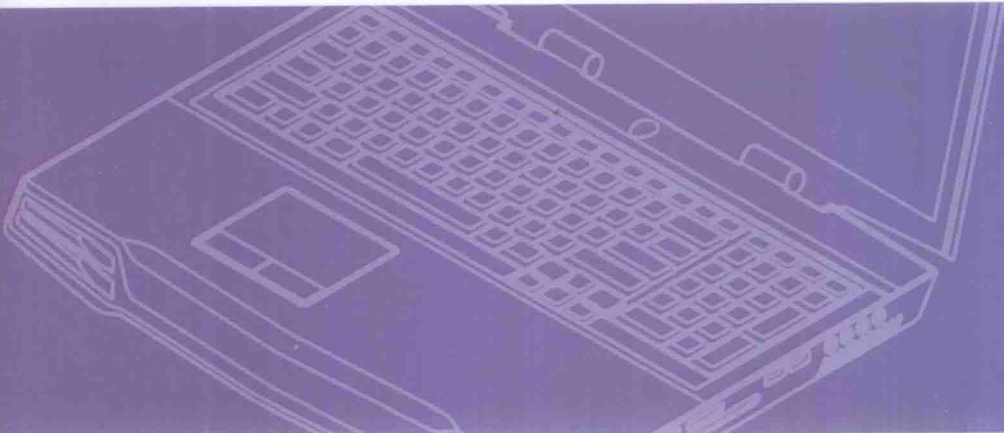




普通高等职业教育 计算机系列规划教材

计算机信息 检索技术

◆ 杨兆辉 主编
◆ 宿佳宁 杨文 王霁阳 副主编
◆ 明丽宏 主审



COMPUTER



中国工信出版集团



电子工业出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
<http://www.phei.com.cn>



配备
电子课件

普通高等职业教育计算机系列规划教材

计算机信息检索技术

杨兆辉 主 编

宿佳宁 杨 文 王霁阳 副主编

明丽宏 主 审

電子工業出版社

Publishing House of Electronics Industry

北京 • BEIJING

内 容 简 介

本书系统论述了计算机信息检索的理论和方法,全面介绍了最新、最常用的中文信息检索系统和网络信息资源。第1章绪论概述了信息检索的现状、发展趋势及意义和作用,第2章至第7章详细阐述了文献信息检索、计算机信息检索、特种文献检索、中文电子书数据库、互联网信息资源检索和多媒体信息检索。本书以计算机信息检索为主,具有内容新颖、实用性强等特点。

本书可作为高等职业院校学生的文献检索课程教材,也可作为科技工作者、教师、图书情报人员的参考用书。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有,侵权必究。

图书在版编目(CIP)数据

计算机信息检索技术/杨兆辉主编. —北京:电子工业出版社, 2015.10

ISBN 978-7-121-26997-4

I. ①计… II. ①杨… III. ①机器检索—情报检索—高等学校—教材 IV. ①G354.4

中国版本图书馆 CIP 数据核字(2015)第 195645 号

策划编辑:徐建军(xujj@phei.com.cn)

责任编辑:徐建军 特约编辑:安家宁 齐美叶

印 刷:北京京师印务有限公司

装 订:北京京师印务有限公司

出版发行:电子工业出版社

北京市海淀区万寿路 173 信箱 邮编 100036

开 本:787×1 092 1/16 印张:13.5 字数:345.6 千字

版 次:2015 年 10 月第 1 版

印 次:2015 年 10 月第 1 次印刷

印 数:3 000 册 定价:36.00 元

凡所购买电子工业出版社图书有缺损问题,请向购买书店调换。若书店售缺,请与本社发行部联系,联系及邮购电话:(010) 88254888。

质量投诉请发邮件至 zltz@phei.com.cn,盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线:(010) 88258888。

前 言

人类文明的发展史，也是一部信息文明的发展史。21 世纪，正是人类社会完成由工业文明向信息文明整体飞跃的伟大时代。

21 世纪的信息文明，是现实世界与虚拟生存交融的新环境，是信息无限增大与信息差距并存的新环境，是基于信息的创新与泛在知识的新环境。在这一新的历史条件下，社会信息资源的科学组织、建设、管理及其有效开发和利用，不仅是人类社会的共同历史使命，也是每个社会成员都无法回避的现实问题。上自国家与民族，下至不同层次的社会组织及个人，其生存和发展完全系乎于此。因此，信息管理科学的形成与发展不仅具有高度的科学价值，而且具有不可估量的社会实践价值和现实意义。

教育部决定在高等学校开设“信息检索与利用”课程，受到了学生的欢迎，通过本课程的教学和实习，在提高学生信息驾驭能力方面取得了良好的效果。而目前信息检索的最常用的载体是计算机及互联网，因为网络信息具有多样性、离散性的特点，大学生需要掌握基于现代计算机技术的信息检索方法，并具有组织信息的技能，这是知识更新、终身学习和知识再创造的需要。

“计算机信息检索技术”是一门面向多学科和专业的公共课程，学习它有助于系统地培养学生通过计算机技术和互联网来获取信息的能力、信息组织能力、信息评价能力和信息综合利用能力，提高学生的信息素养和计算机信息检索技能，满足学生快速、准确、全面、有效地获取与利用网络信息资源的需求。本书正是针对这一要求而编写的。作者基于多年信息资源检索及计算机应用课程教学的积累，编写了这本具有创新性、时代特色和实用价值的《计算机信息检索技术》，出发点是结合教学改革，通过教材的学习和教学，让教师和学生共同完成计算机信息检索任务的过程中理解专业知识，学会各种检索方式在各类检索工具中的运用，把握、运用计算机及互联网技术实现信息检索的技术和技巧，从而获取信息查询、分析与利用的能力，为学生在学习期间和未来的职业生涯中独立地获取和运用计算机信息资源、解决实际问题奠定了良好的基础。

本书由哈尔滨职业技术学院的教师组织编写，由杨兆辉担任主编，宿佳宁、杨文、王霁阳担任副主编。全书共 7 章，第 1 章、第 5 章由杨文编写，第 2 章由王霁阳编写，第 3 章、第 4 章由杨兆辉编写，第 6 章、第 7 章由宿佳宁编写。全书由杨兆辉负责统稿，明丽宏负责主审。

为了方便教师教学，本书配有电子教学课件，请有此需要的教师登录华信教育资源网 (www.hxedu.com.cn) 注册后免费进行下载，如有问题可在网站留言板留言或与电子工业出版社联系 (e-mail: hxedu@phei.com.cn)。

需要说明的是，信息检索是一个不断变化、发展迅速的领域。尽管我们努力全面系统地介绍最新的成果，但由于编者水平有限及时间仓促，内容难免有疏漏、不足之处，恳请使用本书的教师和读者予以批评指正。

编 者

目 录

Contents

第 1 章 绪论	(1)
1.1 信息检索的现状	(1)
1.1.1 信息与信息载体	(1)
1.1.2 信息的特征	(2)
1.1.3 信息的功能	(3)
1.1.4 知识、文献的概念及二者与信息的关系	(4)
1.2 信息检索的发展趋势	(5)
1.2.1 数据库进一步完善	(6)
1.2.2 网络检索简单化	(6)
1.2.3 网络检索多样化	(6)
1.2.4 网络检索工具智能化	(6)
1.2.5 信息检索可视化	(7)
1.3 信息检索的意义和作用	(7)
1.3.1 信息检索的意义	(7)
1.3.2 信息检索的作用	(8)
思考题	(11)
第 2 章 文献信息检索	(12)
2.1 信息检索的概念和基本原理	(12)
2.1.1 信息检索的概念	(12)
2.1.2 信息检索的基本原理	(12)
2.2 信息检索的类型	(13)
2.2.1 按文献载体和记录形式划分	(13)
2.2.2 按检索对象划分	(14)
2.2.3 按信息出版形式和内容划分	(15)
2.2.4 按信息资源的加工程度划分	(16)
2.3 参考工具书的特点、作用和类型	(17)
2.3.1 参考工具书的特点	(17)

2.3.2 参考工具书的作用	(17)
2.3.3 参考工具书的类型	(18)
2.4 常用事实与数据检索工具	(20)
2.4.1 万方数据资源系统	(20)
2.4.2 中国年鉴网络出版总库	(21)
2.4.3 中国大百科全书数据库	(22)
2.4.4 中国统计年鉴数据库	(23)
2.4.5 中国资讯行高校财经数据库	(24)
2.4.6 国务院发展研究中心信息网	(24)
2.4.7 中国科技统计网	(25)
2.4.8 中国宏观经济信息网	(26)
2.4.9 中国年鉴资源全文数据库	(27)
2.5 信息检索工具的内容和类型	(28)
2.5.1 信息检索工具的定义及其特点	(28)
2.5.2 信息检索工具的类型	(28)
2.6 检索工具的结构	(31)
2.6.1 编撰说明	(31)
2.6.2 正文部分	(32)
2.6.3 辅助索引	(32)
2.6.4 附表	(32)
2.7 信息检索语言的概念和类型	(32)
2.7.1 信息检索语言的概念	(32)
2.7.2 信息检索语言的类型	(33)
2.8 信息检索的内容和步骤、方法	(36)
2.8.1 信息检索的标识与原则	(36)
2.8.2 信息检索的步骤、途径与策略	(36)
思考题	(39)
第3章 计算机信息检索	(40)
3.1 计算机信息检索概述	(40)
3.1.1 计算机信息检索的含义	(40)
3.1.2 计算机信息检索的类型	(40)
3.1.3 计算机信息检索系统的构成	(43)
3.1.4 国内外计算机信息检索的发展概况	(43)
3.2 计算机信息检索的原理和技术	(45)
3.2.1 计算机信息检索原理	(45)
3.2.2 计算机信息检索技术	(46)
3.3 文本信息检索	(53)
3.3.1 顺排文档与倒排文档检索	(53)
3.3.2 加权检索	(56)
3.4 信息标引方法与技术	(58)

3.4.1	自动标引的基本原理	(58)
3.4.2	自动标引算法	(60)
3.4.3	统计学习标引法	(63)
3.4.4	概率标引法	(64)
3.5	汉语文献自动标引	(65)
3.5.1	汉语分词算法	(65)
3.5.2	分词算法举例	(66)
3.5.3	汉语文献自动标引	(67)
3.5.4	单汉字标引法	(68)
3.6	搜索引擎的内容和原理	(70)
3.6.1	搜索引擎的发展与分类	(70)
3.6.2	搜索引擎技术原理	(71)
3.7	信息摘要技术与方法	(72)
3.7.1	文本信息摘要的生成与实现	(72)
3.7.2	网页信息摘要的生成与实现	(76)
3.7.3	数值信息摘要的生成与实现	(79)
3.8	计算机信息检索构造及其效果评价与策略调整	(81)
3.8.1	计算机信息检索构造	(81)
3.8.2	计算机信息检索效果评价与策略调整	(82)
	思考题	(90)
第4章	特种文献检索	(91)
4.1	专利文献及其检索	(91)
4.1.1	专利基础知识	(91)
4.1.2	专利的特征	(91)
4.1.3	专利的类型	(92)
4.1.4	获得专利权的条件	(92)
4.1.5	专利文献概述	(93)
4.1.6	专利文献的特点	(93)
4.1.7	中国专利检索工具与方法	(93)
4.1.8	中国专利文献数据库	(95)
4.2	标准文献及其检索	(97)
4.2.1	标准文献基础知识	(97)
4.2.2	标准文献的特点	(97)
4.2.3	标准文献的分类	(97)
4.2.4	标准文献检索	(100)
4.3	会议文献及其检索	(106)
4.3.1	会议文献基础知识	(106)
4.3.2	国内会议文献检索	(106)
4.3.3	国外会议文献检索	(106)
4.4	学位论文及其检索	(107)

4.4.1	学位论文基础知识	(107)
4.4.2	国内学位论文检索数据库	(108)
4.4.3	中国国家图书馆学位论文数据库	(109)
4.4.4	CALIS 高校学位论文数据库	(110)
4.4.5	国外学位论文检索数据库	(111)
4.5	科技报告及其检索	(112)
4.5.1	科技报告基础知识	(112)
4.5.2	科技报告的类型	(112)
4.5.3	国内科技报告检索	(113)
4.5.4	美国四大报告检索	(114)
	思考题	(115)
第 5 章	中文电子图书数据库	(116)
5.1	概论	(116)
5.1.1	数据库及其相关概念	(116)
5.1.2	数据库的类型	(117)
5.2	中国知网	(118)
5.2.1	概况及数据库简介	(118)
5.2.2	CNKI 的检索方法及使用	(126)
5.2.3	CAJ 阅读器常用功能介绍	(136)
5.2.4	知识元搜索	(137)
5.3	维普期刊资源整合服务平台	(143)
5.3.1	系统简介	(143)
5.3.2	期刊资源整合服务系统功能模块之一——“期刊文献检索”模块	(144)
5.3.3	期刊资源整合服务系统功能模块之二——“文献引证追踪”模块	(151)
5.3.4	期刊资源整合服务系统功能模块之三——“科学指标分析”模块	(154)
5.3.5	期刊资源整合服务系统功能模块之四——“搜索引擎服务”模块	(157)
5.3.6	期刊资源整合服务系统整合服务平台功能	(158)
5.3.7	PDF 阅读器常用功能介绍	(158)
5.4	万方数据知识服务平台	(159)
5.4.1	资源概况	(159)
5.4.2	检索方法	(161)
5.4.3	检索结果处理	(165)
5.4.4	增值服务	(165)
	思考题	(167)
第 6 章	互联网信息资源检索	(168)
6.1	互联网信息资源的类型和特点	(168)
6.1.1	网络信息资源的类型	(168)
6.1.2	网络信息资源的特点	(170)
6.2	网络信息检索	(171)
6.2.1	网络信息检索概述	(171)

6.2.2	网络信息检索的方法	(171)
6.2.3	网络数据检索的方法	(172)
6.3	网络信息资源导航	(172)
6.3.1	网址导航	(172)
6.3.2	站内导航	(174)
6.3.3	搜索引擎产品资源导航	(178)
6.4	国外网络数据库	(179)
6.4.1	Ei CompendexWeb	(179)
6.4.2	Web of Science	(180)
	思考题	(182)
第 7 章	多媒体信息检索	(183)
7.1	多媒体基础知识	(183)
7.1.1	多媒体的基本概念及技术体系	(183)
7.1.2	媒体的种类及特点	(186)
7.2	多媒体信息的组织与管理	(189)
7.2.1	信息组织与管理概述	(189)
7.2.2	多媒体数据组织与管理要解决的主要问题	(189)
7.3	多媒体数据库	(191)
7.3.1	多媒体数据与数据库管理	(191)
7.3.2	多媒体数据的体系结构	(192)
7.4	检索多媒体信息	(193)
7.4.1	基于内容的声频检索	(193)
7.4.2	基于内容的视频检索	(199)
7.4.3	基于内容的图像检索	(200)
	思考题	(204)

第1章

绪 论

随着信息技术的发展,互联网的应用得到广泛普及,信息环境发生了相当大的变化,信息社会给人们带来了海量的信息,供人们参考借鉴和学习。如何快速地从海量的信息中获取最有价值的东西,成了人们最棘手的问题。因此,只有掌握好信息检索的理论知识 and 技能,提高信息检索能力,才能快速合理地利用信息资源。

传统的教育模式不注意启发他们的创造能力和培养创新精神,不注意培养学生独立工作的能力,导致学生难以适应快速发展的社会。面对信息社会的挑战,现代高等教育的职能不应是单纯的知识传授,必须同时培养学生综合能力,包括创新能力、表达能力、钻研能力、自学能力和组织协调能力等。应用现代化技术手段获取各种信息、知识成为高等院校师生与广大科技工作者的一种必备知识和技能。为此,必须了解信息检索的基本知识。

1.1 信息检索的现状

1.1.1 信息与信息载体

目前,关于信息(Information)的概念解释较多,信息是信息论中的一个术语,常常把信息中有意义的内容称为信息。下面列举几个比较典型和有代表性的定义,对信息形成全面的认识,并进一步明确信息的特性和功能。

1928年,美国数学家哈特莱(L. R. Hartley)在《贝尔系统电话杂志》上发表题为“信息传输”的论文中指出:“信息是选择的方式和自由度。”从通信领域的角度将信息理解为选择通信符号的方式,并用选择的自由度来计量这种信息的大小。

1948年,美国数学家、信息论的创始人香农(C.E.Shannon)在题为“通信的数学理论”的论文中指出:“信息是用来消除随机不定性的东西。”

1948年,美国著名数学家、控制论的创始人维纳(N.Wiener)在《控制论》一书中指出:“信息就是信息,既非物质,也非能量。”他把信息看成是与物质、能量具有同等重要性的客观

世界三大要素之一。

1975年,意大利学者朗高(G. Longo)在《信息论:新的趋势与未决问题》一书的序言中指出:“信息是反映事物的形式、关系和差别的东西,它包含在事物的差异之中,而不在事物本身。”

关于信息的其他定义还有“信息就是消息”“信息就是信号”“信息就是知识”“信息是使概率分布发生变动的东西”“信息是系统的复杂性”“信息是收信者事先不知道的报道”“信息是加工知识的原材料”“信息是有序的度量”“信息是作用于人类感觉器官的东西”等。可见,信息的概念十分复杂,从不同领域、不同角度理解会产生不同的结果。

从哲学的角度来说,即在最一般的意义上将信息定义为“信息是事物存在的方式和运动状态的表现形式”。这里的“事物”泛指存在于人类社会、思维活动和自然界中一切可能的对象;“存在的方式”指事物的内部结构和外部联系;“运动”泛指一切意义上的变化,包括机械、物理、化学、思维和社会的运动;“状态”指事物在时间和空间上变化所展示的特征和规律。

信息是指一切事物存在方式和运动规律的表征。它是事物的一种普遍属性,只要事物存在,就会有表征其属性的信息。它广泛存在于自然界和人类社会中,信息无处不在,人们正是通过自然界和人类社会中产生的不同信息来区分和认识各种事物的。它是人们认识世界和改造世界,取之不尽、用之不竭的宝贵资源。在人类进入信息社会的时代,信息已成为发展科技、经济、文化、教育的重要支柱之一。

尽管信息的概念具有多样性,但我们对信息的认识,一般可以从以下几个方面理解,即信息与物质、能量既有区别又有联系;信息是事物变化和差异的表现;信息的获取,可以使主体对客体的不确定性减少。

信息载体是在信息传播中携带信息的媒介,是信息赖以附载的物质基础,即用于记录、传输、积累和保存信息的实体。如果是被人感知了,信息就会通过传导神经网络导入人的大脑。如果信息只是停留在人的大脑中,就像一个既不会说话又不能写字,没有任何表达能力的人一样,毫无价值。所以,反映到大脑中的信息只有表达出来才能发挥它的价值。如果说出来,信息就依附于声音中;写出来,信息就依附于文字上;画出来,信息将会依附于图像中。这样,声音、文字、图像等就承载了信息,成为信息的载体。包括以能源和介质为特征,运用声波、光波、电波传递信息的无形载体和以实物形态记录为特征,运用纸张、胶卷、胶片、磁带、磁盘传递和储存信息的有形载体。正是由于主体需要表达从事物中感知的信息,才要借助一定的载体。同一个信息可以依附于不同的载体表达出来。

1.1.2 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。通过对信息概念的分析,可以总结出信息具有如下特征。

1. 信息的传载性

信息的传载性是指信息可以传递,并且在传递中必须依附于某种载体。通常,语言、文字、声音、图像等都是信息的载体,用于承载语言、文字、声音、图像的物质也是信息的载体。例如,通过电视机看新闻,声音、图像是信息的载体,电视机也是信息的载体。

2. 信息的共享性

这是信息区别于物质和能量的一个重要特征,信息的共享不仅不会产生损耗,还可以广泛

地传播和扩散,使更多的人共享。信息的共享是通过交流与传递来实现的,正是由于信息可以共享,所以它对人类社会的进步与发展具有重要意义。但信息的共享实质上是相对的,许多信息是仅对一部分人提供,或要付出一定的代价才能共享。例如,教师上课就是一个信息共享的过程。

3. 信息的可处理性

信息可以被加工、传输、存储,特别是经过人的分析、综合和提炼等加工,可以增加它的使用价值。

4. 信息的时效性

只有既准确又及时的信息才有价值。一旦过时,就会变成无效的信息。

5. 信息价值的不定性

由于人们对信息的需求、理解及判断能力的不同,信息的价值有很大差别,同样的信息对于不同的接收者来说可能有不同的价值。同时,信息的价值也会随时间而改变。

6. 信息认识的主观性

信息的存在是普遍和客观的,但是人类对信息的认识却具有主观的能动性,同一个体对同一信息在不同的时期会有不同的认识,不同个体对同一信息也会产生不同的认识,这种信息认识的差异与个体的知识结构、信息素质、社会影响密切相关。

7. 信息的依附性

从认识论的角度来看,信息是事物运动状态和存在方式的反映。因此,信息总是依附一定的载体而存在。宇宙间的万事万物皆是信息的载体。因此,信息载体的物质形式是多种多样的。在人类社会,为了记载和保存人类对客观世界的认识,使用声音、文字、图像、符号把信息知识记录在纸张或其他特殊材料上,这是人类社会特有的承载信息的物质形式。一切生物皆具有接收和承载信息的组织和器官,人的大脑更是具有信息海量存储的功能。

1.1.3 信息的功能

信息作为维系社会发展的三大要素之一,无论是在自然界,还是在人类认识世界和改造世界的活动中都具有多方面的功能,发挥着重要作用。

1. 资源功能

信息已经成为 21 世纪人类社会的重要资源,科技的进步、社会的发展都与信息密切相关,人类通过对客观世界各种信息的认识、处理、吸收、利用和物化,促进社会的持续发展。

2. 组织管理功能

从管理的角度来讲,管理系统是一个信息输入、处理、输出与信息反馈系统,在这一系统的运作过程中,每个环节都必须以信息为依据,也必须以信息作为相互联系的条件。

3. 中介功能

信息的中介功能表现在人与客观事物及人与人之间。人与客观事物的认识和联系是以信息的存在为条件的,无论是物质系统还是精神系统,系统内部和外部的联系都必须通过相应信息的联系过程来实现。因此,信息是沟通的桥梁和纽带。

4. 消除不确定性(解惑)功能

这一功能是相对于信息接收者的状态改变而言的。在接收者收到关于某一事件的信息之前,可能对事件存在着多种估计,当接收到关于这一事件的信息之后,接收者会改变原有的估

计状态,甚至会使原有的多种估计中的某一种唯一地确定下来,消除不确定性。

5. 传播功能

信息的组织管理功能、中介功能和消除不确定性(解惑)功能都是通过其传播活动来实现的。信息的传播功能是以信息内容的可传输性为基础的。

1.1.4 知识、文献的概念及二者与信息的关系

信息前文已述及,现主要介绍一下知识文献的概念,以及信息、知识、文献的相互关系。

1. 知识

17世纪,英国哲学家洛克对“知识”的定义是:“所谓知识就是人心对任何观念之间的矛盾、排斥和联系、契合的知觉。”1997年,《韦伯斯特词典》对“知识”的定义是:“知识是通过实践研究、联系或调查获得的关于事物的事实和状态的认识,是对科学艺术或技术的理解,是人类获得关于真理和原理的认识总和。”我国学者张玲玲认为:“知识是指人们通过信息对自然界、人类社会,以及思维活动规律的认识与掌握,是人的大脑通过思维重新组合的系统化信息的集合。”卢炳惠则认为:“广义的知识是指人类认识客观世界及其自然实践经验的总结,它可以通过语言文字、各种媒体长期存储,供后人学习和借鉴;狭义的知识是指个体通过与客观外界环境相互作用所获取的各种信息及其技能。”2012年,第6版《现代汉语词典》对知识的定义是:“人们在社会实践中所获得的认识和经验的总和。”

无论人们如何去剖析、定义知识都不难看出,知识属于认识范畴。人类通过信息感知世界、认识世界,通过大脑思维加工,将获取的信息重新序化、组合形成知识,并逐步积累形成知识体系。知识是信息的一部分,是一种特定的信息。存储在人脑中的知识是主观知识,将其记录在物质载体上,就会变成可以传递的客观知识。其实,知识就是人类对客观世界的正确认识,是社会生产实践和科学研究的概括与总结,是人脑加工、整理序列化的信息。它具有规律性、实践性、渗透性、继承性和信息性等特征。以人类活动的领域为界,可将知识分为自然知识、社会知识、哲学知识。

2. 文献

“文献”在1999年版《辞海》中的定义为“记录有知识的一切载体的统称”,1983年公布的《中华人民共和国国家标准·文献著录总则》(GB3792.1—1983)把文献定义为“记录有知识的一切载体”。2012年的第6版《现代汉语词典》把文献定义为“有历史价值或参考价值的图书资料。”可见,凡是记录有知识的一切载体都可以称为“文献”。

文献由以下4个基本要素组成。

(1) 所记录的知识和信息,即文献的内容。

(2) 记录知识和信息的符号,文献中的知识和信息是借助于声音、文字、图表、图像等形式记录下来并为人们所感知的。

(3) 用于记录知识和信息的物质载体,如竹简、纸张、胶卷、胶片等,它们是文献的外在形式。

(4) 记录的方式或手段,如铸刻、书写、印刷、复制、录音、录像等,它们是知识、信息与载体的联系方式(摘自《中国大百科全书·图书馆学情报学档案学》)。

文献具有认识、存储和交流知识的作用。

3. 信息、知识、文献的关系

信息包含了知识,知识是信息被认识的部分。知识可以分为主观知识和客观知识。信息经过人脑接收、选择、处理而形成并存在于人脑中的知识称为主观知识。主观知识借助语言符号,通过各种物质载体记录下来,就会变成可以传递的客观知识,即文献。

信息、知识之间的逻辑关系是包含与被包含的关系。知识是信息的一部分,文献是信息、知识的具体体现,它不仅是信息、知识的主要物质形式,也是读者吸收利用信息、知识的主要途径。

进入 20 世纪以来,科学技术发展迅速,人类社会的信息化、网络化进程也大大加快,各类信息数量剧增,随之而来的是新学科的不断出现和学科之间相互交叉与渗透的加快,使各专业信息发布分散而无规律。主要有以下几个方面发展遇到了瓶颈。

(1) 非文本信息发展滞后。图像声频视频检索计算机技术、数字化技术、高密度存储技术为非文本信息(图像、声频、视频等多媒体信息)提供了广阔的发展空间,多媒体信息已逐渐成为网络的主流。信息检索技术正在从传统的纯文本检索向超文本支持的非线性多媒体检索发展,然而图像、声频、视频的检索技术却处于萌芽阶段,需要高新技术支持并不断创新。

(2) 搜索引擎缺陷。分类目录搜索引擎采用人工干预技术,信息分类不规范,没有一个统一的控制词表和参照标准,分类目录差别较大;搜索范围较小,数据库更新慢,查询交叉类目时容易遗漏;如果用户检索请求没有对应的分类目录,则无法进行查找;信息遗漏不可避免,查全率低。

(3) 检索知识和技能匮乏。知识检索技术信息检索过程,就是把用户请求与索引库匹配,寻找与请求关联的网页并返回排序的命中信息的过程。运用截词、词位限定、布尔逻辑运算等技术可以控制用户请求与数据库匹配的精度,但是信息检索难以避免丢失相关信息或产生大量冗余信息,即出现信息漏检与误检。信息检索效率是衡量信息检索效果的重要指标,是检验信息检索技术成熟与否的标准。知识是信息加工与序化的产物,是高浓度的有序化的信息;知识检索必然是高层次的信息检索。

1.2 信息检索的发展趋势

信息检索技术经过先组式索引检索、穿孔卡片检索、缩微胶卷检索、脱机批处理检索发展到今天的联机检索、光盘检索、网络检索等。其发展经历了由低级到高级、由传统的线性检索向超文本支持的非线性检索发展的过程。这主要依赖于信息技术的发展和信息用户的信息检索需求。这些始终是推动现代信息检索理论与技术发展的重要动力。那么,现代信息检索技术的发展趋势将来如何呢?而要对信息检索技术的发展作出成功的预测是十分困难的。但是可以在广泛调研、比较、分析的基础上把握其某些发展动态,发现一些影响现代信息检索技术发展的新变化、新技术、新理念。

手工检索方式很难全面适应现代信息的发展。信息检索由手工检索过渡到了计算机检索。计算机检索主要经历了脱机检索即批处理检索、联机检索、光盘检索和网络检索四个阶段。计算机检索的成功应用,为我们快速、高效、准确、全面、方便地获得所需信息,提供了先进的手段。

进入 20 世纪 90 年代以后,信息检索的发展有了大大改观,检索行为已不再是信息管理与

服务人员的专利,任何互联网用户都可以实施网络检索,信息检索步入网络化时期。信息检索的对象从相对封闭、稳定一致、由独立数据库集中管理的信息内容扩展到开放、动态、更新快、分布广、分散的 Web 内容;通过互联网来获取各种信息已成为一种普遍、日常的活动,并将成为人们检索信息、利用信息的一种主要手段和形式。信息检索的用户也由原来的情报专业人员扩展到包括教师、学生及其他各专业人士等在内的普通大众,他们对信息检索从结果到方式提出了更高、更多样化的要求。信息检索智能化成为网络环境下信息检索的发展趋势。

1.2.1 数据库进一步完善

在网络信息检索发展早期,众多搜索引擎均在数据库收录的数量和检索速度上展开激烈的竞争;而忽略了检索的针对性和准确性。现在各搜索引擎逐渐意识到只搜索出成百上千万的数据记录是远远不够的,对用户来说也不是很重要。用户更希望迅速准确地根据检索条件查询到自己想要的确切答案,而不是检索到眼花缭乱的记录集。因此,在数据库方面将发生很大的变化。在各大搜索引擎通过各种渠道不断完善自己的网站和数据库的前提下,与其网站建立合作关系,当用户搜索不到想要的确切答案后,迅速跳转到其他网站数据库继续检索,从而扩大了其检索范围,改进了检索结果。

1.2.2 网络检索简单化

现在,一般用户很少愿意应用烦琐的检索方式,大部分只是用输入几个检索词来进行检索。因而,网络检索简单化,网络检索工具界面更加通俗友好,易于操作,使用户学习和进行信息检索更加容易。例如,有些搜索引擎提供了分类导引式的网络检索;有些搜索引擎则让用户将检索范围限制在其索引或数据库的某个表的某个字段中,这样就能更直接快速地检索到结果。

1.2.3 网络检索多样化

网络检索多样化表现在网上检索工具多样化,网上检索信息多样化。现行的网上检索工具功能比较单一,网上检索工具的功能应向多元化发展:网上检索工具不仅能检索文献信息,还能检索新闻、天气、交通时刻、地图、证券、各种互动提问、百科知识、BBS 评论等;分析历史数据的变化趋势,预测未来发展方向,在数据处理中发现事物的发展规律,为科研、教学、生活提供科学依据。网络检索多样化还表现为书目情报信息系统应该具有多方面的功能,包括导向功能、评价功能和文化积累等功能。甚至网络开发出网络检索工具的回溯检索功能,不容争辩地保存有价值的网络文化。

1.2.4 网络检索工具的智能化

随着网络用户对检索精度、效率要求的不断提高,网络检索工具开始重视提高检索功能及检索服务上的智能化程度,主要是重视开发实现基于自然语言的检索形式,用户可将自己的检索提问以所习惯的短语、词组,甚至句子等自然语言的形式输入,智能化的检索软件将能够进行自动分析,而后形成检索策略进行检索。用户只须告诉网络检索工具想要查找什么,而无须考虑烦琐的检索规则、句法等,使检索过程变得轻松、随意。较有代表性的如 FSA、Eloise 和

FAQ Finder, 它们是通过模拟传统检索服务中的咨询来获取有关的检索数据。

1.2.5 信息检索可视化

据研究,人们获取的信息 70% 来自视觉, 20% 来自听觉, 10% 来自触觉。因此,人类是非常适应可视信息的。可视的图像较容易记忆,而且在传达某种信息时比任何方式都快且更有效。因而,可视化检索具有自己独特的优势。可视化信息检索包含 2 个方面:一个是检索过程的可视化,另一个是检索结果的可视化。检索过程的可视化是指用户在检索过程中各检索对象之间的关系以可视化的形式展现在用户面前。用户顺着可视化的检索画面一步一步地发现检索结果。斯坦福大学的高线出版社(High Wire Press)推出的主题地图(Topic Map)就是一种可视化检索环境,它总共包括 54149 个主题。按学科分类,并按学科之间的关系构成主题树。用户可以可视化地看到学科与学科之间的关系,以及每一个学科下子学科的数目,用户通过拖曳鼠标来选择自己感兴趣的主体,最后显示每一等级主题下的论文数目。

相对于其他信息检索技术,可视化信息检索技术的好处有,对文献或检索式内部语义关系的理解,有助于用户判断一个检索中的相关文献;一个透明的检索过程使检索更容易、更有效,即视觉过程寓于检索过程中是可视化的核心;一个可视化的环境为用户展示更丰富、更直观的信息。相关性在传统的检索中指检索结果与检索式相关,而在可视化检索中指检索结果之间的相关度。它具有开发新的检索系统的潜力;它引入了人的认知能力,使得用户可以进行交互式输入,允许在信息空间里进行动态移动,允许用户修改数据的展示方式,使他们理解数据的个人偏好可视化;减少了理解检索结果的时间;可对相关的信息进行聚类分析,而聚类分析可帮助人们发现新的学科点,也可作为反馈的工具,提供对检索的有效的反馈机制;提高检索的准确度;帮助和提高检索系统与人之间的交互性;检索结果可以模仿网络环境形成拓扑图,在拓扑图中所有的相似文献或其他类型资源将被归为同类,以等高线来表达与其他信息内容在概念上的远近,CATIA 的 Theme Scape 便是基于概念图谱运用的例子之一。

总之,可视化技术缩短了用户理解信息的时间,提供了感觉与思考之间的有效反馈机制,它代表着信息检索的未来,将会取代以布尔逻辑为基础的、传统的信息存取系统,如 OPAC 和搜索引擎。尽管目前成熟的、商业化的可视化信息存取系统还未问世,但随着网络技术的发展,以及 XML、RDF、Ontology、Grid 在信息组织、构架中的应用,可视化检索的优势将越来越突出,成为现代检索技术的发展方向。

1.3 信息检索的意义和作用

1.3.1 信息检索的意义

1. 信息检索是获取知识的有效途径

在文献和知识急剧增长的信息时代,如何在浩如烟海的文献信息流中迅速地找到所需的文献信息,并加以合理有效地利用,是摆在科研工作者面前的一道重要课题。据估算,在 1950 年前后,人类知识总量翻一番大约需要了 50 年;20 世纪 90 年代前后,人类知识总量翻一番需要 5~7 年;到 2020—2030 年,人类知识总量翻一番可能只需要 73 天。一个人如果不能随

时更新知识,就会成为 21 世纪的新文盲。而且,文献信息的分布已极不平衡,同一学科的论文分散在几十种,甚至几百种期刊上,导致文献的查找越来越困难。因此,如何以最少的精力、最短的时间充分占有文献信息,成了人们亟待解决的实际问题。信息检索正是有效解决这一问题的最好途径,它可以帮助人们快、准、全地获取所需信息,最大限度地节省查找时间,使文献信息得以充分利用。

2. 信息检索是创新型人才应具备的基本技能

创新型人才首先要有创新思维,创新思维是指人们在创造性活动中所特有的思维过程,它是以独特的思维方式发现、提出、解决疑难问题,创造出新观点、新理论、新知识、新方法的一系列心理过程。只有掌握大量的信息资料,在自由想象中创造灵感,才能在前人不曾涉及的领域有所建树和突破。创新型人才必须具有自立和创新精神,而自立和创新精神的培养离不开对信息的搜集、整理、分析与利用。只有掌握信息检索技术与方法,才能高效获取、正确评价和充分利用信息。所以说,信息检索是创新型人才应具备的基本技能。

3. 信息检索是继承借鉴前人成果的主要手段

马克思指出,每一项科学劳动都必须在前人劳动成果的基础上进行科学继承和发展。对于科研工作人员来说,能否掌握信息检索技能已成为衡量其科研能力大小的一个重要标志。一个人知识的多少就是其拥有的信息量的多少,一个人的能力大小就是指其处理信息能力的大小。

1.3.2 信息检索的作用

1. 信息检索有助于知识更新

随着科学技术的飞速发展,知识老化现象也不断加重。据我国有关部门的典型调查,20 世纪 70 年代的大学毕业生,5 年后有 45% 的知识老化,10 年后有 75% 的知识老化。由于科技发展越来越快,所以当今毕业生的知识老化速度也在加快。只有不断自学、进修,才能适应迅速发展变化的信息时代的要求。

知识更新,重要的是拓宽知识面。近几年美国学术界认为,有必要建立一种通才学,就是要求学生有较宽的知识面,不仅要掌握理工科渗透的知识,还要了解文科渗透的知识。这是因为,科技发展使得人类社会生产的产业结构正处在急剧变化之中,大批知识密集型工业相继涌现,边缘科学大量出现,现在的高校毕业生知识面过于狭窄,已不适应飞速发展的新形势。

及早掌握信息检索的本领,就会在未来的竞争中取得更大的主动权。

2. 信息检索有助于发展教育

(1) 教育的重要性越来越突出。在过去的农业社会和工业社会中,教育的作用虽然显著,但没有文化或文化不高也能够从事劳动生产。对于日本经济腾飞的经验,日本前首相三木武夫曾归结为“日本拥有世界上最好的教育”。

随着信息社会的到来,教育的重要性也越来越突出。信息社会是知识密集的社会,各种现代化技术将会渗透到所有行业和部门,不经过系统教育就不能胜任现代化设备仪器的操作。从这个意义上说,新技术革命也可称作“知识革命”,由此带来的信息时代也应称为“知识时代”,这就对当代教育提出了更高的要求。

(2) 必须把教育放在优先位置。国家要强大,必须科技领先,而要重视科技事业,必须优先发展教育。当今世界,谁要充当“领头羊”,就必须把加强教育放在最优先的位置。

美国前总统克林顿多次强调,每个美国人必须终身学习。在美国政府 1996 年提出的“教