博士文库
BOSHI WENKU

基于WWW的学术信息检索策略

夏立新 著

华中师范大学出版社

博士文库
BOSHI WENKU

基于WWW的学术信息检索策略

夏立新 著

华中师范大学出版社
2004年6月

内 容 简 介

随着因特网的普及及其应用的不断深入，越来越多的个人、学术团体和组织将自己研究的成果和日常活动中产生的信息放在因特网上。这些信息和大量的网络数据库、网络出版物等共同构成了庞大的多学科、多语种而又异常分散的网上信息资源。因特网极大地扩展了用户获取信息的空间。然而，人们普遍感觉到网上信息检索并没有人们想象的那么有效。从用户的角度讲，检索策略研究，可以优化检索过程，以最小的花费、最短的时间获得最佳的检索效果，对于保障用户从事创新性学术研究中的信息需求和满足成型的需求具有十分重要的现实意义。本书在分析 WWW 信息检索策略与联机检索、光盘检索策略异同的基础上，全面系统地研究了检索策略所涉及的用户信息需求分析，网上学术信息资源的分布情况调查，主要检索工具或系统的分析评价，检索过程的科学控制（包括构造检索提问表达式，以及根据检索结果反馈的信息，调整和优化检索提问表达式），检索结果的管理，检索效果的分析与评价等多个方面，既有理论的归纳与总结，又有具体检索方法与步骤的阐述，试图为学术信息的检索提供理论上和技术方法上的支持。

本书可以作为大专院校信息管理与信息系统、图书、情报、档案学等相关专业的教学参考书，也可供学术研究人员、信息工作者、经济工作者、图书情报工作者参考。

新出图证（鄂）字 10 号

图书在版编目（CIP）数据

基于 WWW 的学术信息检索策略/夏立新著. —武汉：华中师范大学出版社，2004.6
ISBN 7-5622-2982-1/G·1525
Ⅰ.基… Ⅱ.夏… Ⅲ.因特网—情报检索 Ⅳ.G252.7

中国版本图书馆 CIP 数据核字（2004）第 064796 号

基于 WWW 的学术信息检索策略

责任编辑：曾 巍　　**责任校对**：罗 艺　　**封面设计**：新视点
编辑室：文字编辑室　　**电话**：027-67863220
出版发行：华中师范大学出版社
社址：湖北省武汉市珞瑜路 100 号
电话：027-67863040（发行部） 027-67861321（邮购）
传真：027-67863291
网址：http://www.ccnup.com.cn　　**电子信箱**：hscbs@public.wh.hb.cn
经销：新华书店湖北发行所
印刷：湖北恒吉印务有限公司　　**督印**：姜勇华
字数：210 千字
开本：787mm×1 092mm 1/16　　**印张**：8.75
版次：2004 年 6 月第 1 版　　**印次**：2004 年 6 月第 1 次印刷
印数：1—1 000　　**定价**：15.00 元

序

夏立新博士的著作《基于 WWW 的学术信息检索策略》出版了。我感到很高兴，并乐于为之作序。

这部著作的意义，首先在于它探索了当代科学生活中的前沿问题——在网络时代如何了解和获取学术信息资源的方法和手段。对科学研究工作者来说，因特网上数量巨大、更新及时、无地理障碍的网络信息资源是取得自己所需信息的首选。熟练地驾驭因特网上的信息，获得对这些信息利用的支配权和主动权，已经成为我们在新时代下从事科学研究工作的“基本功”。掌握这种“基本功”，无疑将大大扩展科学工作者的信息视野，提高信息搜集的本领。因此，此书的出版，将有助于提高科学工作者的研究效率。

因特网上的信息品种众多，质量参差不齐，甚至还有一些虚假的或格调低下的混杂其间。本书的主旨，在于研究网上学术信息及其获取方法。这些信息是网上最有价值、与生产科研关系最为密切的信息。这些信息产生于耗费巨大、历时长久的科学研究活动。科研活动的高成本，决定了这些信息的高价值。作者在书中分别探讨了网上的图书、期刊、会议、学位论文、专利、标准等方面的信息及其有关的数据库。这些对于从事科学活动的人员来说是非常实用的。

要快速、准确、全面地获取自己需要的信息，就必须要构建检索策略。检索策略是在确定检索目标的基础上，选择信息来源、检索途径、检索用词，拟订和调整检索表达式等一系列步骤的科学安排。检索策略的合理性是决定检索效果成败的关键。较之以纸质文献为基础的文献检索策略，网络信息的检索策略有了很大的发展，也存在很多新的变数。网络信息检索策略尚在探索形成中，作者根据自己的实践和有关文献，阐述了检索策略的构建原理、过程控制、效果评价及检索结果的管理等等，其中不乏新的见解。

与同领域的其他研究成果相比，本书将用户信息需求的分析、网上信息资源的了解、主要检索工具或系统的分析评价，检索过程的科学控制（包括构造检索提问表达式，以及根据检索结果反馈的信息调整和优化检索提问表达式），检索结果的管理，检索效果的分析与评价等作为一个环环相扣的系统对检索策略展开研究，视角独特，立意新颖，这项研究顺应了网络迅猛发展的需要，是情报检索界的一个研究热点，也是学术界的迫切需要。其研究的对象明确，研究目标确定，内容结构合理，文字表达流畅，归纳准确，评述得当。著作创新点主要体现在这样三个方面：第一，从确定检索用词、选择检索工具或数据集合、构造与优化检索提问表达式、检索结果的排序输出处理等方面对不同媒体（联机检索、光盘检索、网络检索）中学术信息检索系统中的检索策略进行了系统分析、对比，特别对 WWW 环境下检索策略所涉及的需求分析、检索过程的控制、检索结果的管理和再利用等进行了较深入的探讨；第二，对 WWW 环境下学术信息的分布进行了有效的归纳和总结，为用户更为有效地检索和利用网络上的学术信息提供了科学的方法和高效的检索策略的构造方法；第三，系统分析了网络环境下学术信息检索失

误的原因以及进行扩检和缩检的措施。该书阐述的观点正确，归纳分析合理，所援引的文献资料较为全面。

夏立新是我指导的博士研究生。他的这本著作就是以他的博士论文为基础修改出版的。在攻读博士学位期间，夏立新参加了我主持的国家自然科学基金资助项目、国家社会科学基金资助项目和教育部博士点基金项目的研究工作，该书的内容是与他参与这些项目的研究有关的。

在攻读博士学位期间，夏立新同志学习勤奋、科研能力强，成绩优秀。他的这部著作的出版是与他长期的学术积累分不开的。在本书出版之际，我殷切期望作者在未来漫长的学术探索道路上继续保持谦虚谨慎的作风，取得更大的成绩。

陈光祚

2003 年 9 月 16 日

目　录

1 绪　论

1.1 研究的缘起和意义

进入20世纪90年代以来，随着计算机技术和通讯技术的高速发展及其不断融合，因特网逐渐成为一项全球性信息资源共享的通信设施，并由供少数研究人员使用的深奥系统发展成为普通大众获取信息的媒介。目前，越来越多的个人、学术团体以及其他组织将自己研究领域的成果和日常活动中产生的信息放在因特网上。这些信息和大量的网络数据库、网络出版物等共同构成了庞大的多学科、多语种而又异常分散的网上信息资源。根据《自然》杂志1999年8月发布的研究结果，WWW上有近乎80亿页的可公开存取的信息①。WEB的规模仍在呈指数增长，根据另一资料估算，过去两年，WEB的规模扩大了三倍②。由此可见，网上信息资源的发展势头已远远超过当年引发"情报危机"时文献的增长，正在出现因特网环境下的"信息爆炸"。当因特网奇迹般地通过数以千万计的计算机的连接把整个地球联系在一起的时候，人们在学习、生活和工作中享受到了由此带来的便利。另一方面，网上信息的急剧增加与用户对信息的个性化需求之间的矛盾不但没有得到有效缓解，反倒愈加突出，人们似乎又碰到"情报爆炸"时代所面临的同样困境。虽然因特网使我们可以得到的信息比以前任何时候都多，但并不意味着我们一定能得到想要的信息。为此，各种网上信息的检索工具应运而生，如用于检索FTP的Archier，用于检索Gopher的Veronica、Jughead以及Gopher Jewels等等。1994年以来，WWW迅速发展，以搜索引擎为代表的基于WEB的检索工具也如雨后春笋般发展起来，WWW环境下的信息检索成为网络信息检索的主流。

然而，WEB环境有其自身的特点。WEB上缺乏印本时代所广泛采用的书目控制标准，没有类似国际标准图书编号(ISBN)的符号来惟一地标识文献，没有图书情报机构所采用的分类编目这样的标准系统，没有总目录来揭示WEB上的"馆藏"。事实上，甚至有许多(也可能是绝大部分)Web文献没有作者姓名和出版日期。同时，WEB上的信息也极为分散，更迭、消亡无法预测，人们无法判断网上究竟有多少信息与自己的需求有关，这也在客观上增大了在网上进行有效检索的难度。另外，在WWW环境下的信息检索中广泛使用的搜索引擎一般都是对网页中的每一个词建立索引。这种索引方法没有考虑词的上下文关系和词的具体含义，增大

① Steve Lawrence and C. Lee Giles, "Accessibility of Information on the Web." *Nature* 400 (July 8, 1999), 107.

② http://www.oclc.org/oclc/research/projects/webstars/statistics.htm[Jan 2, 2001].

了在无关文献中查询到检索用词的可能性，它使命中文献的数量大为增加，也降低了检索结果的相关性。

同时，在 WEB 这种一个信息高度分布的网络环境中，由于现实条件的限制或人为因素的控制，不是所有站点上的所有信息都被纳入统一的索引中。搜索引擎的索引数据库中包含了成千上万的网页，但没有哪种搜索引擎能够为整个 WEB 建立索引，更不用说整个因特网。有资料显示，搜索引擎 Northern Light、AltaVista、HotBot、Yahoo 对网上信息的索引率分别为：16%、15.5%、11.3%和 7.4%①。这也进一步说明，不是所有的网上信息都能通过网络检索工具检索到。

因此，从查准率和查全率的角度来看，以搜索引擎为代表的网络检索工具，其检索效果远没有人们想象中的那么有效。除此以外，许多人抱怨搜索引擎不能提供有关课题的信息，提供了大量无关信息或重复信息，甚至形成“死链”，所提供的链接没有信息或相关信息很少，并且对所给予的检索结果没有作出解释②，专业信息的检索能力非常缺乏。“在因特网上您总能找到（甚至只能找到）您不需要的东西”，这较为形象地表达了许多人在因特网上用搜索引擎查询信息时的共同感受。

从检索者的角度讲，检索策略对于取得满意的检索结果至关重要。所谓检索策略，就是在弄清用户情报需求实质的前提下，选择检索途径、检索用词以及明确各词之间逻辑关系和查找步骤的科学安排③。对于传统的计算机情报检索系统而言，构造检索策略的步骤一般分为：分析用户的信息需求；确定检索用词；选择数据库；构造检索提问式；分析检索结果等。这样的检索策略适用于任何电子信息检索工具，包括图书馆目录和 CD-ROM 数据库。当准备选择的数据库跟 WEB 一样规模宏大、缺乏组织并且不断变化时，构造良好的检索策略显得尤为重要。建立在传统计算机情报检索基础上的检索策略的基本原理和方法有其普遍性，同样适用于 WWW 环境下信息资源的检索。但 WWW 环境下信息资源的检索策略应有自己的研究内容：在 WWW 环境下，最终用户一般直接检索，而非委托检索，即使有，也不多见。因此，最终用户需要分析自己的信息需求，弄清自己究竟需要什么；归纳总结 WWW 环境下信息资源的分布情况，帮助解决“最终用户的信息需求能否在 WEB 上得到满足”这一问题；WWW 环境下主要检索工具或系统的分析评价，主要解决“检索方式的合理选择问题”，即“如何满足最终用户的信息需求”；检索过程的科学控制，包括根据检索“命中”信息的情况，运用各种可能的检索技术，以实现扩检和缩检的需要，从而使最终用户的信息需求得到最大程度的满足；用户从因特网上检索、下载了大量数字化信息，并存放在个人电脑上，日积月累，这些数字化信息的管理成为大问题，为此，有必要探讨这些数字化信息的管理问题，为用户运用文献计量学方法在对检索结果进行统计分析的基础上采用可行的扩检和缩检来优化检索策略提供依据；检索效果的分析与评价，为检索效果的改进提供科学依据。

由此可见，WWW 环境下学术信息的检索策略研究，可以优化检索过程，以最小的花费、最短的时间获得最佳的检索效果，对于保障用户从事创新性学术研究中的信息需求和满足成型的需求具有十分重要的现实意义。同时，检索策略的研究能够丰富和完善情报检索的理论

---

① http://www.searchengineshowdown.com/strat/Summary of Lawrence and Giles Nature Article.htm[Jan 2, 2001].

② Stobart, S. and Kerridge, S. An investigation into World Wide Web search engine use from within the UK JISC-funded project undertaken by UKERNA. http://osiris.suderland.ac.uk/sst/se/results.html.

③ 陈光祚：《计算机情报检索系统导论》，书目文献出版社，1993 年，第 192 页。

体系，从而具有重要的理论意义。

本书将检索策略的范围限定在学术信息，主要有如下的考虑：首先，WEB 已发展成为学术信息交流的重要渠道。长期以来，学术信息的交流渠道处于不断的发展与完善之中。早期，学术信息主要通过以口头交流为主的非正式渠道获得，17 世纪初期出现于英国的“看不见的学院”(Invisible College)，又称“无形的集体”，最为典型。这是一种自发形成、自发解体的不稳定的非正式学术信息交流形式，如同一次科学“沙龙”。这种方式有其突出的优点，那就是：能够迅速获得最新的尚未公开的信息；由于是基于双向的交流，能够促进对真实信息需求的理解以及相关信息的交流；方式灵活；有助于不同学科领域学者之间的交流，对交叉科学的研究尤其有利；是一种人们易于并且乐于采用的方式。但是，其不足也暴露得非常充分，主要表现在：交流范围狭小；由于个人掌握的信息有时并不完整，因而容易导致一些误解；这种交流方式难以维持，因而常常不稳定。在这种情况下，以文献信息交流为主的正式交流渠道应运而生。这种交流方式有效地克服了非正式交流方式的不足，其优点主要表现在：信息能够扩散到广泛的读者群；信息更为详尽，如能够较好地表现研究方法、图表数据等；文献信息一般经过了同行评议，真实可靠；能够被引用与参考；可以作为拥有知识优先权的依据等。随着这种交流方式的发展，文献信息的类型也在不断演化，出现了零次文献、一次文献、二次文献、三次文献等多种文献形态。在这种交流方式下，学术信息需求往往需要借助于检索者这一中介代理才能满足。随着以计算机为主体的现代信息技术的发展及其在信息交流领域的应用，电子信息传递形式取得了迅猛发展。如今，因特网上电子邮件、电子会议、电子布告牌、新闻组、电子邮电列表、信息查询等多种应用形式已在学术信息交流领域得到广泛应用，因特网在学术信息交流中的地位也愈显重要。

其次，因特网商业化推动了其在全球范围内的迅猛发展和应用的不断深入，但同时也带来了诸多问题，如信息泛滥，质量良莠不一，缺乏管理，大量有用信息淹没在无用信息之中等，这给人们的检索利用带来了极大的不便。不同于娱乐信息、新闻信息等的检索，学术信息的检索对信息质量的要求较高，对查全和查准也有一定的要求。因此，我们有必要系统研究 WWW 环境下学术信息的检索策略问题。

另外，因特网的普及应用彻底改变了人们的生活与工作方式，在信息检索领域，传统检索的中介代理服务功能正逐步减弱，用户与检索者的界限已淡化。因而，面向最终用户的信息检索理论与实践的研究有待进一步深化。

1.2 研究现状

广义的信息检索包括对信息的描述、加工和有序化，即信息的存储，以及从存储的信息中查询自己所需的信息，即信息的检索。因此，长期以来，有关信息检索的研究都是围绕存储与检索两个方面展开。信息存储方面的研究主要包括数据库与文档的研究，检索系统与检索工具的研究，信息的组织与揭示方法(包括分类语言、主题语言以及基于知识单元的组织与揭示的超文本语言等)的研究等，而信息检索方面的研究主要包括检索策略、检索服务、检索效果评价等。从根本上讲，存储方面的研究是要解决信息的“可检”问题，从而为信息的检索利用提供现实保证。从检索者的角度讲，检索策略的构造与优化对于取得满意的检索结果至关重要。因而，检索策略的研究在整个信息检索研究中具有举足轻重的地位，许多研究者也对此倾注了

极大的兴趣和研究热情。通过文献调查、网上搜索、查新检索中的个案分析，笔者已广泛收集了国内外关于检索策略方面的研究文献，并进行了初步的分析研究，发现有关检索策略的研究存在诸多不足，主要表现在：

(1) 对检索策略的研究范围缺乏应有的把握，出现了简单化倾向，有的研究甚至将检索策略等同于搜索引擎为主的网络检索工具的使用技巧，或是某类信息检索方法的研究，而对检索策略所理应涵盖的 WWW 环境下用户信息检索需求分析及用概念的适当形式予以表达、网络上学术信息资源的分析、主要检索工具或系统的分析评价、检索过程的控制、检索结果的管理，以及检索效果评价等问题进行系统的研究尚未见到。

(2) 研究不平衡。一是目前的研究主要集中在检索工具的使用方法以及分析评价上，对于其他方面的研究在深度与广度上都比较欠缺；二是对于检索策略某一方面的研究，其研究内容也不平衡，如对于检索工具或系统的分析评价，又往往偏重于搜索引擎个体的分析与评价，而对于传统的文献检索系统(包括联机检索系统)在 WWW 环境下检索功能的演进，传统检索工具的 WEB 版在检索功能上的改进，以及对于 OPACs、数字图书馆、虚拟图书馆等信息的存在与组织形式所具有的检索功能等，则缺乏从检索策略的角度进行审视与研究。

(3) 有深度的研究成果尚不多见。其表现为，大多数研究成果偏重于检索方法等操作层面的阐述，缺乏理论上的概括与总结。

1.3 研究的方法、思路及内容

1.3.1 研究方法

本书采用了如下的研究方法：

(1) 文献研究。广泛收集国内外有关检索策略及相关领域的研究文献进行分析研究，并不断跟踪新信息、新动态，从理论和技术方法上全面系统地研究 WWW 环境下检索策略的相关问题。

(2) 实地调查。对国内外较著名的图书馆(主要是一些国家图书馆、科研系统图书馆、大学图书馆等)，查新机构进行实地调研和网上调查，了解他们开展网上信息检索服务的方式、方法以及用户情况。

(3) 拜访有关的专家学者和实际工作者，或通过电话或 E-mail 等方式与他们联系，就课题研究中的一些具体问题向他们征求意见，或了解有关情况。

(4) 坚持理论归纳与具体检索方法的剖析相结合，既有对因特网上学术信息资源及其检索进行较为系统的理论上的分析归纳，又有从检索方法上进行机理和例证的说明。

1.3.2 研究的思路与内容

本书的研究目的在于全面系统地研究检索策略所涉及的各个方面，既有理论上的归纳与总结，又有具体检索方法与步骤的阐述，试图为学术信息的检索提供理论上和技术方法上的支持。

本书属于应用性研究成果，将以理论为基础，从检索实践的视觉来研究，采用“检索策略与检索技术概述并依次以检索策略涵盖的主要内容作为专门的研究对象”这样的思路进行研究和撰写。为此，在内容上作如下安排：

第一章 绪论

第二章 信息检索技术与检索策略。信息检索技术大体经历了手工检索、联机检索、光盘检索以及因特网信息检索等阶段，本章对每一阶段的特点进行了归纳；分析了检索策略的基本原理；从确定检索用词、选择检索工具或数据集合、构造与优化检索提问表达式、检索结果的排序输出处理等几方面对联机检索、光盘检索和 WWW 检索等检索方式中的检索策略进行了分析比较。

第三章 用户的学术信息需求分析。探讨 WWW 环境下学术信息需求的类型、特点及其存在的状态；在探讨构造检索提问的方法与步骤的基础上，从学术信息需求分析的角度重点分析检索提问构造失误的原因；最后，主要探讨穆斯(Mooers)定律和齐普夫(Zipf)最小努力原则，以及检索的费用因素对 WWW 环境下学术信息需求转化为信息检索行为的影响。

第四章 WWW 环境下学术信息资源。网上信息的种类繁多，但不是所有的信息都能从网上找到。为此，本章在分析因特网信息资源现状的基础上，对文件方式、超媒体方式、数据库方式、搜索引擎、主题树、图书馆编目方式、数字图书馆与虚拟图书馆方式等常见的因特网信息的存在形式和组织形式进行了系统的分析比较，并归纳了 WWW 环境下学术信息的类型和特点，试图解决“用户的信息需求能否在 WEB 上得到满足”这一问题。

第五章 网络信息检索工具在学术信息检索中的应用。网络信息检索工具是一种对分散、无序的网络信息资源进行有效控制的工具，具有数据组织机制和信息检索机制，它对庞大的网络信息资源进行收集、记录、标引，形成索引数据库，提供检索功能，指向相关网站或其中的相关资源。因特网上这种提供信息检索服务的系统或工具，极大地拓展了用户获取信息的空间，而且检索功能强大，用户界面友好，简单易用，从而增强了用户自我信息检索服务的能力。网络用户借助于这种检索工具，可以容易、迅速和比较准确地寻获自己需要的信息。虽然使用网络信息检索工具不一定每次都能成功，但一般网络用户还是利用(甚至只利用)网络信息检索工具来检索网络资源。本章将探讨网络信息检索工具的结构和类型，从实现网上信息的虚拟组织和检索的角度分析其功能。最后，探讨了以搜索引擎为代表的网络检索工具在学术信息检索中的应用，并分析了其局限性。

第六章 基于 WWW 的联机文献检索。虽然网络信息检索工具已经成为学术研究人员获取学术信息的一种重要的途径，但其局限性也非常明显。相对于新闻、娱乐等方面的信息，学术信息的检索对于信息质量以及查全、查准的要求要高得多，因此，有必要探索、研究 WWW 环境下检索学术信息的其他途径，如基于 WWW 的联机文献检索，从而使这两种检索学术信息的途径相互补充，相得益彰，共同构成获取网上学术信息的工具体系。本章分析了 OPACs、数字图书馆、基于 WEB 的联机检索系统、网上数据库(包括传统检索工具的 WEB 版，如 CA、EI 等)等的检索功能与特点，以及在学术信息检索中的应用。

第七章 学术信息的分类检索。第四章分析了学术信息的类型、特点，第五章和第六章对网上信息的检索工具和检索系统进行系统的分析与比较，在此基础上，本章将简要地归纳图书信息、期刊信息、会议信息、学位论文、专利信息、标准信息等几种类型的学术信息的检索途径。

第八章 检索过程的控制与检索结果的管理。在选定检索用词和检索系统(或工具)之后，在开始实际检索的过程中，一般还会遇到如下情况：检索提问表达式的构造不当或检索入口的选择不当而引起的检索失误，导致没有达到预期的检索效果；用户在检索结果的基础上有新的发现，并进而产生进一步检索的要求。这两种情况都要求在原来检索结果的基础上采取扩检或缩检的措施，使检索结果尽可能达到预期的检索效果，这就是检索过程控制的实质。所谓检索过程的控制是指根据检索过程反馈的信息，用户可能需要重新选择数据库文档甚至检

索系统或工具，以及检索用词，运用系统所提供的各种可能的检索技术优化检索提问表达式。本章第一节将在分析检索失误原因的基础上，探讨在检索策略的实施过程中可以采取的扩检和缩检措施。

用户通常将从因特网检索、下载的信息资料和文件以文件夹的形式分门别类地存放在个人电脑中，以便记忆和使用。然而，时间久了，用户存储的文件会越来越多，硬盘中会积累一大堆文件，面对着这么多杂乱无章的文件，用户很难知道自己所需的文件有哪些，存放在何处。为此，本章第二节建议采用数字化专题文库的形式对下载的检索结果进行管理，并结合具体的管理软件探讨了专题文库的建立、管理与维护的方法。

第九章　检索效果评价与改进。检索效果直接反映了检索系统的性能，影响系统在用户市场上的竞争力。长期以来，人们将收录范围、响应时间、输出形式、用户负担、查全率、查准率等作为评价检索系统检索效果的标准。应该说，上述六个评价标准在 WEB 环境下仍然有现实意义。但需要指出的是，上述评价标准主要是针对相对封闭的检索系统，其查全率和查准率是建立在可控状态下的评价实验基础上，它要求有含有确切文献数的测试平台和标准检索提问集，每个检索提问所“命中”文献中也有相关的和不相关的。在 WEB 环境下，满足这种情况几乎不可能。每个检索系统或工具（主要是以搜索引擎为代表的网络信息检索工具）有自己的索引，索引的范围各不相同，较之整个因特网，索引的覆盖率非常有限。迄今为止，还没有网络信息检索工具能够为整个 WEB 建立索引。利用网络信息检索工具从事信息检索，人们担心的往往不是检索不到信息，而是信息太多，以至人们难以从检出的信息中筛选出满足自己需要的信息。所以，我们认为，在 WEB 环境下，查全率没有实际意义，人们更关心的是“查准”问题。正因为如此，本章对上述六条评价标准做了一些修订与补充，并从检索工具的角度探讨了检索效果的影响因素。最后，本章还展望了 XML 在改善检索效果方面的应用前景。

2 信息检索技术与检索策略

2.1 信息检索的基本概念与一般原理

2.1.1 信息检索的基本概念

(1) 信息检索 信息检索是一个发展迅速的交叉学科领域，它涉及计算机科学、概率论、信息论、统计论、图论、数理逻辑、现代语言学等相关学科知识，需要将情报学专业知识与上述相关学科知识有机地融为一体。正因为如此，不同学科领域的研究者对“信息检索”这一概念有不同的认识和理解，迄今尚未形成公认一致的定义①。基于情报学的研究视角，我们认为，信息检索是从信息集合中识别和获取信息的过程②。信息检索最初是信息工作者的基本术语，广泛应用于图书情报界，但在因特网不断普及和应用不断深入的今天，信息检索已成为现代人适应社会生存的信息能力的一个重要组成部分，并广泛存在于社会各个领域，如今“信息检索”、“信息搜索”已成为人们耳熟能详的词汇并经常混用。但从严格的意义上讲，“信息搜索”和“信息检索”是有区别的③。

“信息检索”这一概念的理解通常有广义和狭义之分，信息检索的广义理解包括两部分，即信息的存储与检索。信息的存储主要包括在对某一专业或领域范围内进行信息选择的基础上对信息的内外特征进行描述、加工并使其有序化，形成信息集合。信息的检索是指借助一定的设备与工具，采用一系列方法与策略从信息集合中查询所需的信息。狭义的信息检索仅指该过程的后一部分。存储是检索的基础，检索是存储的反过程。对于信息用户而言，狭义的信息检索更为重要，因为，用户需要的是方便、快捷、高效地获取所需的信息内容，而不必了解和掌握信息的搜集、加工、存储等事项。

信息检索的本质是用户的信息需求和一定的信息集合的匹配。

(2) 信息检索方式 信息总是以一定的载体形式存在，载体形式不同，其检索方式也存在重大差异。信息的载体形式经过了一个长期的发展、演进过程。目前，信息的载体存在形式主要有：一是印刷型，它以传统的纸张为存储介质，通过印刷术将文字或图表的原稿制成印刷品，如图书、期刊等；二是缩微型，以感光材料为存储介质，以缩微照相为记录手段而

---

① 焦玉英，符绍宏，何绍华：《信息检索》，武汉大学出版社，2001 年，第 14～16 页。

② 陈光祚：《计算机情报检索系统导论》，书目文献出版社，1993 年，第 1 页。

③ 夏立新：《情报检索的理论和方法在改善搜索引擎搜索性能中的应用初探》，载《情报科学》2001 年第 7 期。

产生的一种文献形式，如缩微胶片、缩微平片等；三是声像型，它是以磁性材料或感光材料为存储介质，借助特殊的机械装置，将语言、声音和文字、图像记录下来，通过视听设备存储与播放信息的一种动态文献形式，如录像磁带、录音带、电影胶片等；四是电子型，以计算机进行存储、管理和利用的现代信息产品，如电子期刊、电子图书、电子名录，以及各种联机数据库和光盘数据库等。

根据信息的存储手段和载体形式，信息检索可分为手工检索和计算机检索两种形式。手工检索方式简称“手检”，主要使用印刷型信息检索工具，其检索过程就是大脑分析、思考和手工操作的配合过程。计算机检索方式简称“机检”，主要使用计算机信息检索系统（包括各种数据库、应用软件、通讯设施等），检索过程就是人的设计操作和计算机自动化处理相结合的过程。

（3）信息检索系统（IRS-Information Retrieval System）　为了方便实现用户的信息需求和一定的信息集合的匹配过程，必须建立一个旨在满足用户信息需求的信息检索系统。如同一般意义上的系统一样，这个系统是目标、信息资源、设备、方法和功能的有机统一体。具体而言，信息检索系统是指为了满足用户的信息需求而建立的经过加工的信息集合，拥有一定的输入、匹配、输出的技术装备，提供一定的检索服务功能的一种相对独立的实体。与手工信息检索系统相比，以计算机为主要设备的机检系统的信息资源是数据库，因而检索是针对数据库进行的，数据库是计算机可读数据的集合。检索过程是在人和计算机的协同作用下完成的，从计算机存储的大量数据中自动分检出用户所需要的信息，即与用户提问相关的信息。匹配是由计算机执行的，而人则是整个检索方案的设计者和操作者。

（4）信息检索入口　信息检索入口又称检索点或检索标识，是指用以标识信息的外部特征和内部特征的属性值的集合。检索标识是用户作为检索的出发点和依据，用户和信息检索系统之间的交流必须有一定的检索标识，否则会导致检索失败。检索标识包括主题词、分类号、著者、标题、机构、代码等，它们通常由人工赋予，或由计算机自动生成。

2.1.2　信息检索的一般原理

信息检索包括信息的存储与检索两个过程，人们试图用流程图的形式概括信息检索的一般原理①。在这种情形下，信息的存储与检索是建立在系统词表（如分类表、主题词表等）的基础上的。一般而言，数据库中所收录的信息，需要依靠系统词表中的标识来加以表征和组织，用户的检索提问也需要借助系统词表中的标识来加以表达，这样才能使存储与检索得到有效的沟通和控制，保证检索系统达到令人满意的查全率和查准率，减少漏检和误检。但随着计算机技术、通讯技术等现代信息技术的迅猛发展及其在信息检索领域应用的不断深入，关键词、单元词、自由词等一些无词表系统控制的检索语言已得到广泛应用，图形、图像、音频、视频等多媒体检索技术也取得了长足进展②③④。但多媒体检索技术离实用化水平特别是在广大用户中的普及尚有距离，目前的信息检索大多数是针对文本信息的。为此，作者也试图借助流程图来表示现代文本信息检索的一般原理，如图 2-1 所示。

首先，需要建立文本数据库，并使其成为可检索的信息集合。这是信息检索的基础，将

① 陈光祚：《计算机情报检索系统导论》，书目文献出版社，1993 年，第 1 页。

② 李恒峰，李过辉：《音频信息检索》，载《计算机工程》1999 年第 8 期。

③ 毛力，张晓林：《基于内容的图像检索技术与系统》，载《现代图书情报技术》1999 年第 5 期。

④ 王玉波：《多媒体信息检索技术概论》，载《情报科学》1999 年第 2 期。

从底层决定信息检索系统的检索利用方式。建立文本数据库前，数据库管理者需要明确文本的搜集范围、对文本的操作方式，以及文本模型（如文本结构和可检单元等）。文本操作是要实现原始文本的转换，以产生文本的逻辑视图。逻辑视图一旦形成，数据库管理者将利用数据库管理模块建立文本索引。对于大规模的信息集合而言，索引是必不可少的。索引结构可能各不相同，但常用的还是倒排档。所谓倒排档就是建立在文本数据库基础之上的倒排索引的文件形式。数据库管理者所确定的文本模型（如图 2-1 所示）中有许多能反映其外表特征和内容特征的属性值，如文本记录的标识号、篇名、文摘、关键词、叙词、自由词、作者、期刊名、出版年、语种、产品代码、机构名称等。从检索的角度讲，这些属性值可能有检索意义，并有可能成为检索入口。倒排档一旦建立，我们可以认为，文本数据库已成为可检索的信息集合。

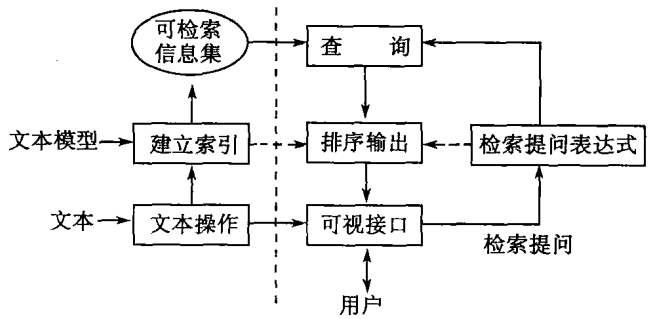

图 2-1　信息存储与检索原理①

在此基础上，现代意义上的信息检索才能开始②。用户首先要明确自己的信息需求，然后运用用于文本信息的文本操作方法对信息需求做语法分析和转换，以产生检索提问，再运用系统所要求的用户信息需求的表达方式，构造检索提问表达式，然后提交检索提问表达式，利用已建立的索引（主要是倒排档）以实现快速查询，并产生检索结果。

在将检索结果提交用户之前，检索系统将运用系统所确定的对相关性评判的算法对检索结果进行排序，并试图将与用户信息需求密切相关的排在前面。这样，用户就可以浏览“排序”输出的检索结果，以找出自己感兴趣或需要的信息。如果用户对检索结果不满意，可以修改检索提问表达式，以产生新的“排序”输出的检索结果。此过程可以循环反复（具体过程如图 2-2 所示），直到检索提问表达式更好地表达用户真实的信息需求，产生令用户满意的检索结果。

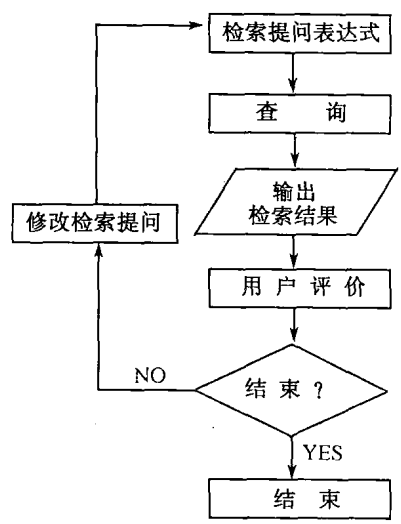

图 2-2　检索提问表达式调整、修改流程图

① 参见：http://www.sims.berkeley.edu/~hearst/irbook/1/node5.html[Sept 14, 2000]。

② 如同“手翻眼看”的原始方式一样，数据库的“浏览”也是一种从数据库中获取信息的方式，但从严格的意义讲，它不是现代意义上的“检索”。

2.2 信息检索技术的历史发展

考察信息检索技术的历史发展时，人们习惯于按年代进行划分。大体说来，20 世纪 40 年代之前，人们主要利用书本式或卡片式检索工具，完全通过手工式检索。50 年代出现了穿孔卡片检索及缩微胶卷检索，其中，穿孔卡片检索包括重叠比孔卡检索方式和边缘穿孔卡检索方式，这一时期的检索技术存在的时间很短，对信息检索理论和方法没有产生多少影响。60 年代出现了通过磁带，以脱机批处理方式为主，以计算机为基础的信息检索。这一时期可以称为“计算机信息检索的萌芽期”，出现了一些专门学科领域的情报中心。它们从数据库生产者那里购买机读磁带，装入自己的计算机，并生产必要的软件，以进行最新资料报道、定题服务(SDI)和追溯检索的批式处理。70 年代，信息检索技术步入联机检索新阶段，这一时期，分时计算机、带终端的远程处理系统、廉价的大容量随即存储器、分组交换网，以及数据库生产的迅速发展为联机检索的发展提供了良好的技术条件，Dialog、Medline、Orbit、Brs 等联机检索系统先后投入商业化运作。80 年代，信息检索发展为多机联网检索及光盘检索。到了 90 年代，信息检索发展到基于因特网检索的新阶段。

为了便于深入探讨计算机、通讯等现代信息技术在信息检索领域的应用对检索策略的影响，同时，也为了论述的方便，我们根据信息存储手段和载体形态的变化，粗略地将信息检索技术划分为手工检索、联机检索、光盘检索以及因特网信息检索等阶段。本节在回顾信息检索技术的发展历史的同时，将对每一阶段的特点进行归纳。

2.2.1 手工检索

主要借助书本式、卡片式等各种形式的先组式索引，通过人们的手翻，眼看，脑子作出判断而进行，检索结果主要是文献线索，然后依据文献线索获取原始文献。这种检索方式的最大优点是，人们边查找、边浏览、边思考，可以随时得到新的启发或者激发潜在的信息需求，甚至可能有意外的发现。但其缺点也相当明显，主要表现在：效率低、速度慢、耗人耗时多，当检索者注意力不集中时，随时可能出现对命中的文献款目“视而不见”的漏检现象①。另外，虽然人们试图对文献信息的内容特征和外部特征予以充分的揭示与处理，提供尽可能多的检索入口，但实际上能提供的检索途径非常有限。

2.2.2 联机检索

联机检索是计算机技术、数据库技术和卫星通讯技术共同发展的产物，是检索终端上的用户通过通讯网络用“人机对话”的方式从主机上获取所需信息的过程。传统的联机信息检索系统有如下特点：

(1) 大多数是集中式管理，有专人负责维护整个系统，定期更新数据库中的信息。联机检索系统一般是由联机检索中心、通讯网、检索终端三部分组成。联机检索中心是联机检索系统的“神经中枢”，主要包括中央处理机、联机数据库、检索软件等部分。中央处理机又称“主机”，是联机检索系统硬件的核心部分，在系统软件和检索软件的支持下完成信息的存储、处理和检索等任务。系统中大多数联机数据库是从国际上权威的数据库生产商那里购买或租借的，信息来源可靠，信息质量较高。也正是由于联机数据库实现了专业化生产，因此更新速度快，用

① 陈光祚：《计算机情报检索系统导论》，书目文献出版社，1993 年，第 6 页。