21世纪高等院校教材
信·息·管·理·系·列

信息检索进展

THE DEVELOPMENT OF INFORMATION RETRIEVAL

焦玉英 主编

科学出版社
www.sciencep.com

21世纪高等院校教材——信息管理系列

信息检索进展

焦玉英　主编

科学出版社

北京

内 容 简 介

本书内容包括信息检索基本理论、现代信息检索技术、信息检索服务三个部分。本书简要概述了从文献到网络环境下信息检索的发展进程中，国内外在信息检索基本理论，特别是数据信息描述机制和网络信息检索原理、工具与策略方面的最新进展、主要学术见解和研究成果；介绍了全文检索、多媒体检索、超文本及超媒体检索、联机及光盘检索、网络信息检索等当代信息检索技术的发展趋势及主要研究成果；总结了传统信息检索服务方法与策略，对计算机、光盘和因特网的检索服务系统及其相关技术研究与应用进行了评价。

本书可作为高等院校信息管理、情报学、文献学等专业的研究生教材，也适合于程序员、互联网从业人员、企事业单位情报人员、图书馆管理人员以及相关人员参考。

图书在版编目(CIP)数据

信息检索进展/焦玉英主编. —北京：科学出版社，2003

(21世纪高等院校教材：信息管理系列)

ISBN 7-03-011646-1

Ⅰ. 信… Ⅱ. 焦… Ⅲ. 情报检索 Ⅳ. G252.7

中国版本图书馆CIP数据核字(2003)第055476号

责任编辑：陈 亮 / 责任校对：包志虹

责任印制：安春生 / 封面设计：耕者设计工作室

科学出版社 出版

北京东黄城根北街16号

邮政编码：100717

http://www.sciencep.com

双青印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2003年8月第 一 版 开本：B5(720×1000)

2003年8月第一次印刷 印张：20

印数：1—3 000 字数：379 000

定价：32.00元

(如有印装质量问题，我社负责调换〈环伟〉)

参与本书编写人员

主　编　焦玉英

编　委　索传军　李法运

李进华　刘伟成

柯　青　郑　琳

刘　颖

前　言

以因特网为核心连接起来的全球计算机网络对传统的相对集中和以规范为基准的文献信息数据库及其服务系统构成了严峻的挑战。网络信息浩瀚几乎没有什么范围限制；内容覆盖面之广，涉及到科技、经济、军事、政治、贸易、商业、金融、教育、娱乐、人文社会科学等各个领域；多媒体技术的应用使静态和动态图形(像)、音频及超频特征的信息得以经过数字化处理后以位串形式存放，采用点线、区域、关联、形状等多种形式再现出来；网络化与数字化技术使分布在世界各地主机上的信息资源联为一体，构成了跨时空、跨行业的高效快速的国际化知识信息平台，赋予了每一个用户“我为人人，人人为我”的得天独厚的信息共享环境；网上客户通讯、聊天、个人简言、即时沟通、个人邮件传递等已成时尚；多功能搜索工具及导引方式，在线帮助等使用户甚至足不出户，不必经过专门培训即可从网上获取自己所需要的信息。

在这种环境下，传统的“提问-检索”模式已被网上“浏览-查询”模式所取代。人们传递与获取信息的主动性得到空前提高。但网上“浏览-查询”检索模式运行至今仍有诸多不尽人意之处。用户采用“点击”式浏览远远未达到预期的查询效果。这就涉及到一系列检索理论、技术、方法问题的研究。诸如多种网络信息检索语言的兼容、综合问题；各类型网络搜索工具搜索功能的提高、策略的优化及其效果评价；网络环境下检索服务理论及其系统建设；网络检索理论模型探索等，都是目前网络信息技术提供商、信息内容提供商、信息管理理论领域关注的焦点。

从检索理论发展角度看，网络环境下“浏览-查询”检索过程中遇到的上述问题均与传统“提问-检索”方式中的理论课题密切相关。它们往往在彼此交叉、互补、借鉴中才能得到真正意义上的优化。本书旨在反映从文献信息检索到网络信息查询的发展过程中涉及到的国内外主要研究成果，以及今后的发展趋势。

本书由焦玉英组织编写，负责全书的策划、统稿、定稿工作并撰写前言、第1章总论。索传军、李法运、李进华，刘伟成、刘颖、柯青、郑琳参与编写。按他们编写的章节顺序分工如下：刘伟成负责编写第2章；李法运负责编写第3章；柯青负责编写第4章；郑琳负责编写第5章；李进华负责编写第6章；刘颖负责编写第7章。索传军参加由焦玉英主持的国家自然科学基金项目(70073022)的研究。他负责编写的第8章是该项目成果报告的核心内容。

本书的出版得到武汉大学信息管理学院院长马费成教授、胡昌平教授的举荐。科学出版社的陈亮同志为本书的出版付出了辛勤的劳动。在此向他们表示诚挚的

谢意。

本书引用和参考了国内外专家学者的论著，在此一并致谢。限于水平，疏漏之处，在所难免，欢迎批评、指正。

焦玉英

2003 年 7 月于武汉大学

目　录

前言
第1章　总论…… 1
1.1　从文献检索到网络信息检索的理论研究进展 …… 1
1.2　网络信息检索展望…… 13
第2章　情报检索理论模型应用与发展 …… 17
2.1　概述…… 17
2.2　布尔检索模型…… 19
2.3　向量空间模型…… 24
2.4　概率检索模型…… 29
2.5　模糊检索模型…… 33
2.6　逻辑检索模型…… 40
2.7　概念检索模型…… 44
2.8　案例检索模型…… 47
2.9　检索模型的比较及其评价…… 50
2.10　情报检索模型比较指标体系 …… 51
2.11　情报检索模型研究发展的趋势 …… 51
第3章　信息检索语言 …… 58
3.1　信息检索语言基础理论研究…… 58
3.2　受控检索语言的分析、比较 …… 61
3.3　自然语言处理与自然语言检索…… 72
3.4　检索语言的网络版及其应用…… 91
第4章　检索系统功能、策略、评价体系研究…… 109
4.1　信息检索系统历史沿革 …… 109
4.2　联机检索研究进展及其评价 …… 117
4.3　光盘检索研究进展及其评价 …… 123
4.4　网络检索系统研究进展 …… 129
4.5　网络检索系统性能评价指标体系 …… 142
第5章　搜索引擎研究…… 152
5.1　搜索引擎的发展概论 …… 152
5.2　搜索引擎的分类研究 …… 154
5.3　搜索引擎的应用与比较研究 …… 157
5.4　搜索引擎的质量评价 …… 164

5.5 搜索引擎与检索策略 …… 168
5.6 搜索引擎发展中的论题 …… 172
第6章 信息检索技术进展 …… 181
6.1 全文检索技术 …… 181
6.2 多媒体信息检索技术 …… 188
6.3 超文本及超媒体信息检索技术 …… 196
6.4 联机信息检索技术 …… 202
6.5 光盘检索技术 …… 207
6.6 自动分类与自动标引技术 …… 209
6.7 信息检索技术的发展趋势 …… 213
第7章 网络环境下以用户为中心的信息服务体系研究 …… 225
7.1 以文献收藏为中心的传统信息服务体系 …… 225
7.2 网络环境下信息服务面临的挑战 …… 226
7.3 以用户为中心的信息服务体系的建立 …… 229
7.4 从用户出发的信息服务质量评价研究 …… 250
第8章 基于Push技术的网络主动信息服务 …… 259
8.1 Push技术的国内外研究概述 …… 259
8.2 Push技术的工作原理与实现方式 …… 269
8.3 基于Push技术的主动信息服务系统建设 …… 278
8.4 基于“推-拉”技术的网上主动信息服务系统的设计方案 …… 289
8.5 Push技术现存问题与发展前景 …… 305

第1章 总 论

1.1 从文献检索到网络信息检索的理论研究进展

在信息处理与传递的电子化与网络化进程中，浩瀚的知识信息海洋，快速、便捷、高功能、简便易用的多元化网络信息查寻工具，形成了世界范围内真正意义上的资源共享格局。据称，世界上最大的因特网能为1亿多用户提供50多万个科研机构、学校、图书馆和信息系统的网络资源：WWW覆盖了2000万个大小不等的信息源；预计不久的将来将有100万个网络，1亿台计算机和10亿用户入网。CNNIC2002年发布的第十一次《中国互联网发展统计报告》指出，中国以各种方式上网者有5910万网民，总数在世界排名第二。全球网络化的大趋势使信息系统及服务方式面临严峻的挑战：网上"浏览-查询式检索"模式对传统的"提问-检索"模式的替代使专职检索代理机构的业务逐步缩小；终端用户对系统的直接介入带来用户成分的巨大变化；系统为用户提供的在线即时获取方式，自组织信息功能使一种以满足个性化需求为特征的信息服务模式成为主流。在这种环境下，以文献单元描述结构为基础，手工检索方式为主导的传统文献检索向以信息单元组织结构为基础、网上浏览式信息检索发展成为必然的选择。

1.1.1 以"提问—检索"模式为核心的文献检索相关理论与方法基础

传统文献检索系统由存储与检索两个核心部分组成。存储是选择文献、按规范化语言文本描述文献内外特征并使其有序化；检索是系统根据用户提问按规范化语言文本进行概念转换，经逻辑匹配输出与提问相关的文献。围绕系统存储与检索开展的研究主要涉及到下列方面：

(1) 理论要素

对此颇有建树的国内外专家(如陈光祚、吴慰慈)对文献信息流动态特征及其内在规律[1]颇有研究。他们的观点促进了以知识为核心的文献生产域到知识的替代、重组、综合、检索、利用过程中文献特征的揭示、文本检索工具的生产以及文献信息传播理论体系的发展。张琪玉、侯清、俞君立、陈树年、曾蕾、曹树金等对规范化检索语言及其与自然信息的兼容研究与试验[2,3,4,5]，不仅奠定了传统文献的特征揭示及其有序化组织理论，而且对目前网络环境下信息资源的组织与有效利用具有现实的指导意义。

Lancast、Gilman以及Smithson的检索系统评价研究[6,7,8]、曾民族教授提出的系统评价指标体系[9]、Grenfield确立的标引语言及评价方案、Salton研制的矢量空

间模型及文献聚类技术[10]、Ottaviani 的用于文献聚类的基于分数几何的模型[11]、Roberson[12]和国内康跃红博士等对检索概率理论模型的研究[13]等，为检索系统构建及检索效率的提高提供了一系列理论根据与检索试验系统数据与算法。

20 世纪 90 年代，以数据库结构驱动检索逐步过渡到以用户为中心的研究，成为学术界关注的热点。Watters 指出了信息检索系统中的数据库应以满足用户感知的信息需求来驱动信息检索，并主张构建以用户为中心的检索试验系统[14]。此外，关于传统文献检索涉及到的最佳匹配策略，以及查全率、查准率及其相关性测度的研究一直十分活跃。

周智佑教授在论述检索领域诸多问题时指出了其“理论研究薄弱、理论与实践两张皮现象”，呼吁我们的研究应“力图使检索理论摆脱单纯思辨的局面，使之与现代科学技术研究方法结合起来”[15]。他的观点应对目前网络信息检索的理论与实践研究有启示作用。

(2) 结构要素

结构要素包括检索工具与检索系统部分。其研究主要围绕数据库与文档结构的分析与评价，以及与此密切相关的标引技术与索引系统款目描述标准，联机与光盘检索系统的应用分析等来展开。

在文献检索的结构要素中，20 世纪 80 年代关于科技文献检索工具的结构体例分析研究十分活跃，其成果见著于全国各大专院校的 600 多种《科技文献检索与利用》教材之中。它们对高校信息管理专业乃至自然科学与工程技术的大专学生、研究生、教师、科研人员、工程技术人员打开世界科技文献宝库，为充分利用国内外科研成果发挥了重要作用。

(3) 经验要素

传统文献检索的实质是用户提问与系统中的标识之间的匹配。影响检索效果的因素，除了取决于系统本身的文献覆盖面、文献摘贮率、索引系统的完备程度之外，更与来自检索提问及其操作策略、代检人员的才学与经验密切相关。其研究主要围绕查全率与查准率的目标与设置，以及对检索相关性的认识而展开，如检索课题主题内容相关性分析与判断；检索提问策略的编制与应用技巧；检索语言语义确切性分析；对各种不同检索工具内容范围、检索手段的熟悉程度；检索策略及效果评价指标体系与方法(含联机检索系统与光盘检索系统)等。

总之，传统的赋值标引，规范化检索语言，线性的书目数据结构，预定的检索策略，面向专家的文献信息服务系统，以及回答检索提问的被动式服务方式与策略等构成了以“提问—检索”模式为主的传统文献检索理论与方法学基础。

1.1.2 以“浏览—查询”检索模式为主导的网络信息检索相关理论体系

信息环境的巨变极大地提高了人们获取信息的主动性。网上“浏览—查询”模

式逐步取代传统的“提问—检索”模式是不争的事实。但这并不意味着信息检索的问题都解决了。恰恰相反,网络信息检索理论属于探索性研究,它应具有超前性,以对网络检索实践起指导与推动作用。

本章将网络环境下与“浏览—查询”检索模式相关的理论研究焦点归纳如下:

1.1.2.1 网络信息资源组织与揭示的研究[16,17,18]

目前,关于网络信息资源组织与揭示的研究主要包括以下几个方面:

(1) 自由文本组织方式

自由文本以全文数据库存储为基础。它将一个信息的全部内容(而不是信息的线索),转化为计算机可以识别、处理的信息单元而形成数据集合,适应了对Web网页中非结构化信息处理的需要。它必须对全文数据库进行词(字)、句、段落等深层编辑、加工,允许用户用自然语言表述、检索,直至直接查看一次信息。由于自由文本组织方式占用的空间大,系统响应速度慢,因而关于全文数据库压缩技术的研究,关于超高密度磁盘、光盘及芯片技术的研究,以及关于自然语言后控机制的研究和给标引语句加权的研究等,将是自由文本组织方式中需要解决的问题。

(2) 超维组织方式

这是一种基于知识单元的新型信息组织与揭示方式,它借助超文本技术来实现。超文本技术将文本信息存储在无数节点(node)上,一个节点就是一个相对独立的“信息块”,节点之间用“链”(link)联接,由此组成信息网络;它也可以链接声音、图像(形)、影视等多媒体信息,构成超维检索点。在这种超维系统中,用自然语言分析、抽取知识单元,不仅减轻了专业标引人员的负担,而且打破了传统系统线性序列的局限性,允许用户按个人兴趣以熟悉的语言浏览、查询信息。目前,国外研制的基本理论参考模型有:

① Dexter——超维交换格式的标准模型,它由运行层、表现层、存储层、锚定及内成员层等构成。该模型中用于标准化的主要层次是运行层、存储层与内成员层之间的接口分别采用的规范、锚定机制。运行层显示多媒体信息以及用户与系统的交互;存储层由成员和链组成超网;内成员层描述成员的内容和结构。它在保持系统各层次之间的充分独立性、面向全局、有效的接口机制,以及层与层之间的通信交流方面有突出的优势。

② 适用于多种类型的超维系统模型(HAM)。该模型包括数据库层、超媒体抽象层(HAM)和表现层。数据库层主要处理信息存储中的传统问题,保证信息的存储对高层的透明性。抽象机层决定节点和链的基本特征,记录节点和链的关系,并保存节点的结构信息。表现层则处理抽象层中信息的表现,作为人机交互的窗口。

(3) FTP 组织方式

FTP 是以文件系统保存和组织网上信息资源的最简便方式。它传送的文件包括文本、图像、声音、多媒体、数据库以及可执行二进制的代码文件,其操作类似于在网络上两个主机间拷贝文件。

(4) 主题树组织与揭示方式

主题树方式主要是通过人工发现信息,选择并对其进行归类(包括网址主题类及子类、子子类等),从而构建一个层次分明的等级结构体系。主题树方式由于采用人工编制,具有科学性、专题性特征,能较好地满足人们按类浏览专题信息的需求。

(5) 计算机索引数据库组织与揭示方式

这种方式与主题树方式的主要区别是非人工构建,即主要依据于 Spider 或 Robots 的计算机软件程序的运作,是目前网上二次信息组织的主要方式之一。其组织方式有:① Spider 根据数据网络协议在网上漫游,发现新的网址、网页信息,抽取、排序、归并建立网络索引数据库;② 数据库按一定方式、结构存储,提供特定处理系统需要的相关信息(包括网址及相关描述性信息、计算机可识别的字段标识符)。这种方式的自动化程度高,更新速度快,并可提供位置检索、概念检索、截词检索、嵌套检索等。

(6) 指引库组织方式

指引库常用于组织网上专题性强的二次信息。它是一个由语义信息、文献信息、链接信息组成的语义——文献双层数据结构。其中语义信息与语义链集合构成双层结构的第一层;文献信息和结构链构成另一层;链接信息由不同的语义节点的语义链、不同文献节点的结构链以及链接于语义节点和文献节点之间的链组成,穿行于第一、二层之间。指引库不是有关网址的堆积与拼凑,而是对网上专业信息资源的重组和开发,在内容上必须符合专业人员研究的需求,在功能上具备动态性,能及时更新数据,反映学科前沿的情况。

1.1.2.2 检索语言兼容与整合的研究[19,20]

检索语言是建立和利用检索系统必要的语言。传统的文献检索系统是采用对自然语言事先规范而形成的受控语言(如分类表、主题词表),来描述文献信息特征,生成概念及其概念标识系统,人们通过分类表中的分类符号或主题词表中的主题词(或叙词)作为控制检索的入口格式进行检索。在网络环境中,文献尤其是非文献信息数量急剧增多,受控语言的专业性太强、应用范围有限及更新维护困难等不足日益突出,而自然语言恰恰可以解决这些困难。所谓自然语言,是指作者的书面用语。采用自然语言,可以减少概念间转换产生的误差,检索入口词多,操作简单、方便、灵活,也适合专业人员之外的广大用户群。但从网上自然语言使用的情

况看,问题并不那么简单。如选词不加严格控制,会导致词语量过大,过多占用磁盘空间,从而影响主题的集中,降低查准率。同时,由于自然语言对多义词也基本不加控制,往往使相关主题内容的文献分散,从而造成漏检。受控语言与自然语言存在的这种互逆相关性,恰好说明它们在网络环境中兼容、整合的必要性。近年来,国内外学术界提出的检索语言兼容、整合措施与方法,主要有以下几种:

(1) 对各种数据库采用的不同检索语言进行综合、集成的方法

网络检索实践表明,试图用一种检索语言统一各种数据库的不同分类体系及叙词法是不现实的。G. Riesthuis 提出了词语、句子和主题三个层次的兼容方法[21];艾奇逊的《分面叙词表》、《中国分类主题词表》、张琪玉先生的"学科-事物概念-组配"模式、《联合国教科文组织叙词表》、英国《科学文摘》中的主题指南、美国《生物学文摘》中的主题分类表等都较好地解决了检索语言的兼容问题。

美国的 UMLS(一体化医学语言系统)是世界医学领域的一个高度专业化、分类主题一体化、检索语言综合化的典范。它由超级叙词表、语义网络、情报源图谱、专家词典等组成。2001 年它的超级叙词表集成了 70 种叙词表、标题表、另类表、词典和专家系统;包括 80 万个概念、190 万个不同词汇;语义网络包括 132 种类型、53 种语义集。

(2) 采用中介语言来实现多种检索语言之间的兼容

国内外不少专家提出建立一种转换系统,即中介语言来实施网络检索系统中多种数据库查询语言的兼容。如 Dahlberg 所进行的以《情报编码分类法》(ICC)作为转换系统与国际著名分类法 UDC、DDC、LCC、LBC 等兼容的可行性研究,A. P. Chamis 研制的词表转换系统(vocabulary switching system)[22],兼容了物理、商业、社会科学、生命科学等专业领域的 12 部叙词表。国内侯汉清研制的基于集成的叙词表词汇转换系统,分类-叙词表转换系统,是不同检索语言之间的转换、兼容较成功的实验系统[23,24]。

(3) 通过标准化手段实现各种检索语言的兼容

标准化是检索语言兼容的最高层次。建立一种开放式、多功能、多种语言的分布式概念和术语知识库,来集中和规范概念间各种关系及其使用规则是解决这一问题之关键;C. Moore 和 J. Chain 等学者对多文字/多元化环境下的标题表数字化问题的研究对网络环境下检索语言兼容有重要的参考价值[25]。

随着自然语言处理技术的发展,跨国语言的检索将越来越普遍。在对检索语言兼容、整合进行研究时必须看到,受控语言仍有自然语言无法取代的优势,多种检索语言(受控语言)和自然语言的多种使用方式的结合,将共存于信息检索系统之中。

(4) 构造自然语言控制机制

寇钧锋提出的控制方法有[26]:

① 事先控制法。当检索要求输入时，即加以控制，而对输出不加限制。这样，用户可以先选择自己所需要的词汇，再通过一种入口词表将提问词转换成受控词汇，以提高查准率。事先控制法能减轻标引人员和检索者的负担。

② 事后控制法。此法与上述方法相反，它既不对检索提问词进行任何控制，也不对输出结果进行太严格的控制，是一种在很大程度上接近自然语言，又保留受控语言许多特点的控制方式。

张琪玉教授指出，自然语言的后控机制即是从系统内部看检索语言是高度专门化的，从系统外部用户接口而言又是高度自由化的[2]。吴广印、胡亚莉提出的自然语言后控制机制[27]："标引不控制 + 检索控制"模式，即输入标引阶段使用自然语言，不对标引加以严格控制，在检索输出阶段进行后控制。后控措施包括：截词、位置逻辑、标引词加权和编制后控词表库手段。后控词表能较好地解决大量存在的词的等同关系、等级关系、相关关系。实践中，系统通过学习功能，将出现的新概念、新术语及时组成动态词表，使用户借助该表浏览词汇。国外具有后控词表功能的系统可参考：

美国国防技术信息中心(DTIC)的科技报告文库(scientific and technical reports database)。该库的后控词表在功能上类似于传统的叙词表，体系清晰，结构合理，易用性较好。其不足之处表现在：① 为用户只提供浏览与选择功能，未提供词的添加和修改功能；② 没有自动构造检索提问式功能。

美国教育资源信息中心(ERIC)数据库系统。它是美国国家教育部、教育研究和发展者的一个联邦基金项目(http://www.ericae.net/scripts/ewiz/amainz.asp)。该系统中的后控词表具有自动添加检索词的功能，是目前网上运行的比较成功的后控词表系统。其不足之处与 DTIC 一样，未向用户提供词表更新、维护和自学习功能。

美国生物科学情报社(BIOSIS)的词表系统中的 ZR(zoological record)提供了具有后控功能的词表：① 主题词表(subject thesaurus)；② 生物体系词表(systematic thesaurus)。它们是专为用户提供帮助的词表。由于它在结构上使词表与全文检索相分离，应该说并不是严格意义上的供全文检索的后控词表。

国内方面：张琪玉教授、曾蕾教授从事多年后控词表系统的研究，并发表了许多有影响的研究论文；武汉大学信息管理学院臧国全的硕士论文《后控词表系统研究》；空军政治学院周全明的硕士论文《全文检索后控制技术研究》；北京大学信息管理系韩东梅的硕士论文《后控词表的开发设计与利用》等都丰富和推进了网上自然语言的后控机制的理论与应用研究。从目前情况看，大部分实验都是针对单机系统的，适用于网上全文检索系统的自然语言后控词表的研究正在成为目前和今后的研究热点。

1.1.2.3 构建用户模式的研究

基于网络的信息查询模式主要有:① 超文本浏览;② 搜索工具查找。前者对信息的组织以信息单元为基本单位。各单元之间的概念按非线性方式组织与存放。用户既可按顺序也可跳跃式浏览查询。此方式方便、灵活,但易入"迷途";后者由 Robots、搜索引擎、索引数据库、查询服务等模块组成。用户只要转入搜索引擎地址,组织好关键词及其相互逻辑关系、位置关系等,将检索式输入查询框内,发出搜索命令,系统即会把搜索到的结果,如网址、文摘乃至全文显示出来。这两种查询的完成都离不开人的参与。为了提高人的参与系统的效果。人们一直致力于以用户为中心的检索系统的研究,构建用户模型就是其中之一。其重点正从数据库为中心转向以用户为中心同时,强调数据库要面向由用户感知的信息需求驱动式检索。即允许用户动态地表明其观点,允许用户通过判断选定检索中心数据流模式、集合模式、关系模式、等级模式等,从而形成个人需求。

用户模式用于捕捉网上用户的需求兴趣点,管理用户兴趣,优化用户查询行为与结果。用户特征模板是解决问题的核心。

一般来讲,可将用户兴趣信息分为两类:① 用户提供的主动信息;② 被动的用户信息。前者指的是用户背景信息。例如用户在申请注册个性化信息服务时登记的诸如专业、职业、兴趣爱好、上网方式等方面信息。系统的处理办法是对其属性进行描述;给出个性化检索词列表;用户提供常用检索关键词等。这些词被存入用户兴趣库,备系统进行分析。后者是指在 Windows 操作中,系统对用户访问的网站点保留,作为历史记录。根据用户历次查询的场景及用户行为,通过用户兴趣学习算法,可以被动地推导出用户兴趣特征。提出一个基于关键词为主体的用户对象兴趣模型。其模型应包括:① 关键词的父对象信息;② 子对象信息,即某用户兴趣所在的分类中的主题信息或关于该主题的所属分类信息;③ 扩展联系信息;④ 相关领域信息等。即根据主题对象的彼此联系度建立用户兴趣网,通过与数据用户兴趣信息相互作用关系程度来判断对象的权值,决定被选顺序。该模型还能通过学习机制获取用户反馈信息,进行评价、修复和优化,提高查准度。

用户需求兴趣特征描述应注意稳定性、可获取性、可归纳性、可描述性、可推理性等原则。并在此基础上提出用户需求兴趣的获取办法,友好接口,建立良性反馈机制,组织用户查询历史记录,通过客户端观察等手段与策略来发现用户兴趣构建用户模型。如,目前的网络信息资源指引库就是追踪用户兴趣特征的措施之一。

西南财经大学图书馆网络上运行的一个基于智能检索导引服务系统中设置了一个用户帮助机制,系统可以通过客户方脚本来设计,以服务器端脚本来创建动态页面,追踪用户状态,进行输入,纠错分析。我国交通部已建立的"交通专业网络信息资源指引库"已输入运行,并初见成效。

目前受到广泛关注的个性化智能 Agent 技术；个性化网上信息过滤技术；网上信息主动推送技术；自动发现技术等都是与用户模型构建密切相关的研究课题。

1.1.2.4 完善网络信息检索机制及其应用研究

储荷婷教授针对 WWW 站点资源的组织过程与方式，提出了检索机制的三个组成部分，即采集标引机制、数据组织机制和用户检索机制[1]。其中，以 Robots 为核心的网络信息资源自动采集，旨在以 HURL、HTTP 为基础，集中不同类型的信息产品（纸质型、缩微型、计算机可读型、录像带、光盘等），开发使全球范围内的各种信息资源能实时及时地进入信息系统。其自动采集机制提供的网页样本，为网络检索工具的量化标引、量化评价提供理论根据。数据组织机制以数据采集为依据，直接对网上索引数据库系统的动态维护与管理产生影响。用户检索机制涉及用户界面友好、检索策略的合理程度、检索执行以及检索结果的相关性处理等。因此，完善检索机制可以说是网络信息检索领域的核心课题，国内外许多专家学者主要围绕下列有关检索机制问题开展研究。

（1）网络信息检索工具分类研究

网络检索工具处在发展之中，从不同角度对其进行类型划分的研究很难统一。其方法归纳起来大致有：

① 按信息内容组织方式划分为分类范畴搜索引擎和词语搜索引擎两大类。前者主要包括 Yahoo、Infoseek、Galaxy、GNN，以及 WWW Virtual Library；后者主要有 Webcrawler、Lycos、Alta Vista、Excite、Open Text 等（刘静的划分方法与此相一致）。

② 按专业范畴划分为通用性和专业性查询引擎两类。

③ 按检索功能划分为常规（或单一）查询引擎与多元查询引擎。后者是多个单一搜索引擎的集合，又称集成搜索引擎。元搜索引擎是网络检索工具的后起之秀，它没有独立的数据库，主要依靠系统提供统一界面，构成一个一对多的分布式且具独立功能的虚拟逻辑机制。主要的元搜索引擎有 Web Search Engines、Savvy Search、All-in-one、Best Search、Metacrawler 等。

此外，也有将网络搜索工具从功能角度划分为目录式、索引式、指南式三大类的；分类主题目录式、搜索引擎式、主题式及多元式的；按网上信息资源组织方式，将其划分为 Web 式搜索引擎与非 Web 搜索引擎两种类型的等。目前，对网络信息检索工具的研究已从上述的类型划分进入深开发阶段，即在更高程度上优化检索工具。例如，关于构建网上专业指引库的研究、网上资源自动跟踪与发现技术的研究等。

(2) 网络信息检索工具与检索系统比较评价研究

检索工具是检索系统的核心组成部分,其比较研究主要集中在系统功能设置、用户界面、数据库内容结构与更新,以及对国内外搜索引擎的准确性、易用性、可选择性、检索效果的分析、比较等方面。研究的目的,一是帮助用户选择优质检索工具,二是为网上信息检索工具的量化评价提供理论依据。除了元搜索引擎外,大部分搜索引擎都有自己独立的数据库。因此,其比较研究主要以数据库资源和搜索引擎性能评价为依据。对于数据库资源的定性评价因素,应该说用于印刷型资料的标准基本上适用于网络信息资源(包括数据库资源)。目前,定性评价研究主要侧重在:① 热门或精选站点推荐;② 设立网络编辑或网络评价员。定量评价主要包括:① 学科站点分布;② 用户访问数量;③ 站点被引用情况。在评价数据库资源的指标体系方面,内容范围、适用对象、编排方式、权威性及价格;可存取性、交互性与愉悦性、信息的广度与深度、主页链接的可靠性、版面设计质量、信息时效性、主页的可操作性及读者对象等,都是可参考的指标。在搜索引擎性能的评价方面,曾民族认为 Lancast 提出的针对传统系统的涵盖范围、查全率、查准率、响应时间、用户方便性、输出格式等标准,基本上适用于网络信息检索工具性能的评价。在此基础上,他提出了数据库规模与内容、索引方法、检索功能、检索结果(相关排序、内容显示、输出数量选择、显示格式选择)、用户界面、检准率和响应时间等评价指标[9]。

(3) 网络信息检索策略研究

网络信息检索策略对实现检索工具的功能至关重要。"功能"强调其静态性,"策略"则强调其动态性。网上各类型搜索引擎所采用的检索策略除个别特定符号规定外,大部分都认同布尔逻辑、截词手段、相邻度检索、位置逻辑检索、字段检索、加权检索以及自然语言检索、相关信息反馈等的使用。全面了解这些策略有利于提高对检索的准确性。但灵活运用策略主要取决于用户的直觉、经验而不是逻辑思维。用户除了采用网上各种搜索引擎规定的特定检索策略标记符号及组配原则外,还需要熟悉和掌握其他策略。如黄晓斌提出的策略是:一般性查询选用 Yahoo,自然语言查询用 Infoseek,不确切知道关键词时用 Excite 进行概念检索,全文检索用 Open-Text 和 Excite,反向检索用 Webcrawler,短语检索用 Open-Text,查期刊论文选用 the Electric Library 等。雷燕则从检索方向、检索细节、最可能查到相关信息、搜索站点评论、搜索标题和 URL、检索用户小组 6 个方面,归纳了可选用的搜索引擎及其策略。

智能化是网络信息检索工具研究的方向性课题,即由"智能代理"充当用户检索工具的中介,包括用户的检索工具选择、策略的灵活运用、搜索并整理检索结果等。