面向21世纪课程教材
Textbook Series for 21st Century

信息检索：理论与方法

叶　鹰　主编

高等教育出版社
HIGHER EDUCATION PRESS

附赠学习卡

面向21世纪课程教材
Textbook Series for 21st Century

信息检索：理论与方法

叶鹰 主编
黄敏 刘霞 副主编

高等教育出版社
HIGHER EDUCATION PRESS

内容提要

本书是教育部“面向21世纪课程教材”，是教育部高等学校图书馆学学科教学指导委员会推荐的图书，也是情报学专业核心课程教材。

本书针对新世纪图书情报专业教学的需要编写，兼顾高校信息检索与利用课的要求。全书从理论到方法系统阐述了信息检索的基本知识和核心技术，内容覆盖科技信息检索和文科信息检索，贯通机检和手检，整合古代与现代，并以结合最新的理论探索与技术进展为特色。

本书可作为高等学校的教学用书，也可作为各类图书馆和信息机构的岗位培训教材和业务参考书。

图书在版编目(CIP)数据

信息检索:理论与方法/叶鹰主编. —北京:高等教育出版社,2004.6(2005重印)
ISBN 7-04-015036-0

Ⅰ.信… Ⅱ.叶… Ⅲ.情报检索—高等学校—教材 Ⅳ.G252.7

中国版本图书馆CIP数据核字(2004)第010489号

出版发行	高等教育出版社	购书热线	010-58581118
社　　址	北京市西城区德外大街4号	免费咨询	800-810-0598
邮政编码	100011	网　　址	http://www.hep.edu.cn
总　　机	010-58581000		http://www.hep.com.cn
		网上订购	http://www.landraco.com
			http://www.landraco.com.cn
经　　销	北京蓝色畅想图书发行有限公司		
印　　刷	北京民族印刷厂		
开　　本	787×960　1/16	版　　次	2004年6月第1版
印　　张	25.25	印　　次	2005年7月第3次印刷
字　　数	460 000	定　　价	28.60元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换。

物料号　15036-A1

作者简介

叶　鹰　浙江大学教授，上海交通大学兼职教授，哲学博士，中国图书馆学会理事，浙江省图书馆学会副理事长、学术委员会主任，浙江大学信息资源管理系主任，主要著作有：《Web 信息查询技术》等；主要论文有：《图书馆学基础理论的抽象建构》、《信息科技基础理论的分析建构》、*Triad Philosophy and Triad Science based on Triad Logic* 等。

黄　敏　上海交通大学研究员、情报学专业硕士生导师。主要著作有：《科技文献检索》等；主要论文有：《基于旋律的音乐检索》、《分布式联合虚拟参考咨询系统的建设》、*Music Information Retrieval in the Digital Library* 等。

刘　霞　武汉大学馆员，硕士，在职博士生。湖北省高校图工委用户教育专业委员会副理事长，武汉大学图书馆信息服务中心主任。主要著作有：《因特网基础与网上资源查询》（副主编）；主要论文有：《信息资源网络化对信息服务业的影响分析》、《高校地区性文献传递系统的构建与运行》、《Internet 上图书信息的开发与利用》等。

胡立耘　云南大学教授，硕士，教育部高等学校图书馆学学科教学指导委员会委员。主要著作有：《云南图书馆概览》（主编）、《社会科学文献检索基础教程》（参编）等；主要论文有：《论影响书目》、*Humane Traditions and Technological Innovations – A Study of Academic Library Service to Reader's in the 21st Century* 等。

周建屏　苏州大学副研究馆员，双学士学位，苏州图书馆学会秘书长。主要著作有：《Internet 信息资源检索和利用》（参编）；主要论文有：《知识导航与图书馆软环境建设》、《新形势下情报工作的创新与发展》、《艺术文献检索的系统特征》等。

王宝红　西北大学讲师。主要著作有：《文献检索》（多媒体教材）（参编）、《文献分类学》（参编）；主要论文有：《关于机读目录主题标引一致性的分析》等。

孙滨丽　黑龙江大学（现在北京石油化工学院）副教授，硕士。主要著作有：《科技文献检索教程》等；主要论文有：《化学化工文献标引模式的探讨》、《“机检证明”的问题与对策》等。

胡小君　浙江大学副研究馆员、医学院教师，硕士。主要著作有：《医学

信息检索学》、《医学信息检索》等；主要论文有：《Web形态下引文分析的科学指征研究》、《加菲尔德定律在学科评估中的应用研究》等。

郑江平 浙江大学副研究馆员。主要著作有：《科技信息检索教程》、《农业信息检索》等；主要论文有：《高校文献检索课教材微观质量探讨》、《浙江和台湾农业发展现状及未来趋势比较》等。

王怀诗 兰州大学副教授，兰州大学经济管理学院信息科学研究所副所长。主要论著有：《信息运动规律初探》、《信息素质及其提高途径》、《试论信息概念研究的层次性》等。

王晓鸿 兰州大学副教授，经济学硕士，教育部高等学校图书馆学学科教学指导委员会委员。主要论文有：《我国主题检索语言的发展趋势》、《文献计量学的过去、现在和将来》等。

陈兰杰 河北大学教师，管理学硕士。主要论文有：《图书馆学研究对象学术思想的历史演变与发展》、《基于网络的数学化参考咨询服务系统功能设计》等。

林皓明 上海交通大学副研究馆员。主要论文有（合作）：《关于图书馆利用Chat软件开展实时参考咨询服务的探讨》、《分布式联合虚拟参考咨询系统的建设》等。

徐汝兴 上海交通大学副研究员，情报学专业硕士生导师。主要论文有：《图书馆跨平台信息检索系统初探》、《基于XML的跨平台数据源信息检索》等。

张 辉 山东大学副教授，教育部高等学校图书馆学教学指导委员会委员。主要著作有：《信息管理学》（参编）、《电子商务》（参编）；主要论文有：《知识经济与信息产业》、《网络信息资源建设的现状分析与发展策略研究》等。

张 帆 中国科学技术大学馆员，信息咨询部主任。主要论文有：《图书馆创新建设的重点内容》、《我国信息资源系统建设中的几个问题》、《文献检索课教学改革的理论与实践探讨》等。

目　录

引　言

信息、能源和材料，并称为现代社会的三大支柱。20 世纪以来，人类创生的信息量高速增长，浩如烟海。信息检索，就是从浩如烟海的信息海洋中查找出所需信息的过程。为实现这一过程，就需要适当的理论和方法，本教材因此而撰写。

现代意义的信息检索作为一个独立领域，是 1946 年计算机出现后在国际上逐步得以确立的。信息检索教育的发展，在中国兴起于教育部 1984 年发出的“教高一字 004 号”文件——要求在高等院校开设“文献检索与利用”课程。20 余年来，文献检索教育尤其是手工文献检索教育方面已取得很大成绩。然而，随着计算机、多媒体等信息技术的发展，以缩微品、声像带、磁盘、光盘等形式记录的非纸信息急剧增加，靠“手翻、眼看、大脑判断”的手工检索方式已难以全面适应当今信息社会发展的需要，计算机信息检索应运而生；以 Internet 为代表的全球性计算机网络的迅速普及，则更进一步推动了信息检索的发展，使得网络化信息检索逐渐成为当代信息检索的主流。

因此，在当今这个纸本资源和数字资源并存的信息时代，作为信息资源管理专业以及其他专业的大学生、研究生，既需要了解和掌握传统手工检索的方法，也需要熟悉和掌握计算机检索尤其是网络信息检索的理论与技术，这是信息时代对学者信息素质的必然要求。故本书在系统阐述信息检索基本知识和核心技术基础上，对手检和机检都作了全面介绍，内容涉及各类文献信息源，覆盖科学技术和人文社会科学所有学科领域，并对基于网络的计算机信息检索方法和技巧作了重点介绍。

检索理论，是信息检索得以实现的基础，故本书第一章进行系统陈述。

信息资源是信息检索的基础，本书第二章用现代眼光对传统印刷型信息资源和数字化资源进行了整体介绍，算是一个特点。

检索工具，原指具有检索功能的书刊，其古代形态可以上溯到我国汉代刘歆编撰的《七略》。但现代意义的检索工具则是伴随世界信息的增加和科学技术的发展从 19 世纪逐步兴起的，以期刊式文摘和索引为主，也包括部分工具书。本书用第三、第四两章的篇幅对基于书本印刷型检索工具的手工检索给予足够关注，并对部分检索工具的数字化发展或电子版作了相应的介绍。

DIALOG 系统作为国际联机检索的强劲工具，在信息检索中具有极其重要的地位，曾是计算机信息检索的核心技术，也为信息检索理论与技术的发展立下不朽功勋，至今仍在发挥其不可替代的作用，故本书用第五章专门介绍。

20 世纪 90 年代以来，进入以 Internet 的高速发展为标志的网络化信息化时代。对 Internet 上的信息进行检索的需求产生了新型检索工具，其主体是 1995 年以来伴随 Web 技术的发展而迅速兴起的各种搜索引擎，本书也予以足够的重视并专列第六章进行介绍。Web 技术从初出茅庐，到登陆中国，短短几年就深刻地影响到了各行各业的运营模式甚至社会生活的各个方面，正在促成继印刷书刊和广播电视后的第三次媒体和传播革命，因而了解和掌握 Web 信息检索技术意义重大且深远。

图书情报机构作为信息集中地，集成了大量有序信息，基于图书情报平台的检索也就成为一类独特的技术，本书第七章对此进行了专门介绍。

信息检索的目的是为了利用信息资源，故应落实到文献信息综合利用，是为本书第八章。

随着个人信息处理量的激增和信息处理技术的发展，个人信息管理软件开始出现，本书用第九章对此作一个初步引导。

以上就是本书的整体考虑和大体结构，希望能循序渐进地引导读者进入信息检索领域，掌握信息检索的理论与方法。

第一章

信息检索理论基础

本章提要：本章在简要介绍信息检索历史基础上，对信息检索的理论基础进行了系统阐述，重点是：

- 信息检索原理：布尔逻辑检索、向量空间检索、模糊集合检索等理论模型。
- 信息检索技术：布尔检索、截词检索等常用检索技术和当代新兴技术。
- 信息检索系统：信息选择、标引、提问处理等子系统。
- 信息检索评价方法：查全率、查准率、误检率、漏检率等评价参数。

信息检索作为一门学科，其历史可追溯到19世纪下半叶。但在20世纪中期以前，信息存储和传播主要以纸质文献为载体，信息检索活动也围绕着文献的获取和控制展开，因此，信息检索研究关注的是如何检索利用文献中记载的信息，文献检索一度成为信息检索的同义词。20世纪50年代开始了计算机应用时代，信息检索得到迅速发展并使用情报检索一词。由于汉语中“信息”较“情报”的含义更为宽泛，加上英文 Information 可以理解为“信息”或“情报”，随着通信技术与计算机技术的紧密结合，信息载体类型的多样化及传播手段的改进，情报检索研究和文献检索研究逐渐归入信息检索研究这一具兼容性的概念。

随着信息学的快速发展，作为信息储存、获取技术方法的信息检索的研究范围也日趋扩展，主要包括：信息检索原理与技术，信息检索语言，信息检索系统，信息检索评价方法等。本章首先对这些理论基础及一些新的检索理论做一概括介绍。

1.1 信息检索原理

1.1.1 信息检索及其发展

信息检索(Information Retrieval)是在 1949 年国际数学会议上由 Galvin W. Mooers 首次提出的。他在其发表的《把信息检索看作是时间性的通讯》论文中指出："信息检索是一种时间性的通讯形式"，"在时间上从一个时刻通往一个较晚的时刻，而在空间上可能还在同一地点"，并强调"信息接受者是最活跃的一方"。这一看法，揭示了信息存储与获取两个环节是一种延时性的通讯形式。

我们可以用一句话概括信息检索的基本原理，即对信息集合与需求集合的匹配和选择。如图 1.1 所示。

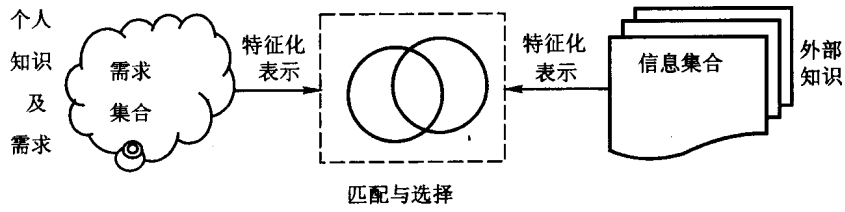

图 1.1 信息检索基本原理示意图

人们为满足某种需要时，往往感到要补充某些知识，因而产生了对信息的需求。信息集合是有关某一领域的文献或数据的集合体，它是一种公共知识结构，可能弥补该用户的知识结构缺陷。而匹配与选择则是一种机制，它负责把需求集合和信息集合进行比较，然后根据一定的标准选出符合需求的信息。

在社会科学化的进程中，信息检索经历了从手工检索到机械检索再到计算机化检索的发展过程。

(1) 手工检索(1876—1954)

信息检索源于图书馆的参考咨询和文摘索引工作。较正式的参考咨询工作是由美国公共图书馆和大专院校图书馆于 19 世纪下半叶发展起来的。20 世纪初，多数图书馆成立了参考咨询部门，利用图书馆的书目工具和索引帮助读者查找图书、期刊及寻找现成答案。随着文献量的激增和读者需求的增长，逐渐发展到从多种文献源中查找、分析、评价和重新组织资料，"索引"突破了以前的狭隘范畴，成为独立的检索工具。到 20 世纪 40 年代，咨询工作的内容又

进了一步，包括事实性咨询，编制书目、文摘，进行专题文献检索，提供文献代译等。“检索”从此成为一项独立的用户服务工作，并逐渐从单纯的经验工作向科学化方向发展。

(2) 脱机批处理检索(1954—1965)

1946 年世界上第一台电子计算机问世后，20 世纪 50 年代初就有人开始研究其在信息检索领域的应用。20 世纪 50 年代中期至 60 年代中后期是信息检索的脱机批处理阶段。当时计算机还没有通过通信网络互相连接，也没有远程终端装置，不能提供实时检索，只能进行现刊文献的定题检索(Selective Dissemination of Information)和回溯性检索(Retrospective Search)，同时利用计算机编辑出版检索性刊物。1954 年，美国海军机械实验中心使用 IBM701 型机，初步建成了计算机情报检索系统，这也预示着以计算机检索系统为代表的信息检索自动化时期的到来。

单纯的手工检索和机械检索都或多或少显露出各自的缺点，因此极有必要发展一种新型的信息检索方式。

(3) 联机检索(1965—1991)

1965 年美国系统发展公司研制成功 ORBIT (On-line Retrieval of Bibliographic Information Time-shared)联机情报检索软件，开始了联机情报检索系统阶段。与此同时，美国洛克希德公司研制成功了 DIALOG 检索系统。至今，DIALOG 系统仍为世界上最著名的信息检索系统之一。20 世纪 70 年代卫星通信技术、微型计算机以及数据库生产的同步发展，使用户得以冲破时间和空间的障碍，实现了国际联机检索。计算机检索技术从脱机阶段进入联机信息检索时期。远程实时检索多种数据库是联机检索主要的优点。联机检索是计算机技术、信息处理技术和现代通信技术三者的有机结合。

(4) 网络化联机检索(1991—今)

Internet 的雏形初显于 20 世纪 60、70 年代，到 80 年代末才开始在世界范围内流行。在通信和网络技术支持下，出现了各种信息利用和检索平台。如：WAIS，允许用户检索整个因特网上文本信息资源；Gopher，像地鼠一样，能使用户十分容易地获取因特网上的信息资源；还有针对 FTP 资源的 Archie；BBS 等等。特别是 20 世纪 90 年代崛起的 WWW (World Wide Web)——优秀的万维网技术有取代其他信息工具的趋势，使传统的联机检索向因特网上迁移。90 年代联机检索的发展进入了一个重要的转折时期。随着互联网的迅速发展及超文本技术的出现，基于客户/服务器的检索软件的开发，实现了将原来的主机系统转移到服务器上，使客户/服务器联机检索模式开始取代以往的终端/主机结构，联机检索进入了又一个崭新的时期。

计算机技术的不断进步和信息量成倍地增加，使人们对信息检索技术的要

求也越来越高，尤其是网络技术和多媒体技术的出现，促使信息检索技术也不断地发展。目前，信息检索技术正向两个方向发展：一是传统信息检索向全文本、多媒体、多载体、多原理等新型信息检索的发展，在深度上提高管理和组织信息的能力，如探索自动抽词、自动索引、自动检索、自动文摘、自动分类、自动翻译等；二是信息资源的网络化和分布化，面向 Internet 中浩瀚无垠的资源，在广度上提高管理和组织信息的能力。在信息检索技术研究领域中，基于概念、超文本信息和多媒体信息检索技术的研究最为活跃，并已取得了突破性发展。网络的发展给信息的获取提供了广阔的空间，而检索技术的发展为人们利用信息提供了更方便快捷的手段。网络信息环境的出现，使信息检索研究对象和范围不断扩大，研究队伍也突破了原有的以图书情报领域的专家学者为主的框架，众多的计算机界专家、信息技术专家也加入到研究开发信息检索系统的行列。可以说，网络使计算机信息检索技术进入一个崭新发展阶段，而网络信息检索又使网上信息资源的利用率进一步提高，使信息组织更为有序和高效。基于互联网的检索系统成为网络信息检索系统的代表。

1.1.2 信息检索模型

信息检索系统的实际运行性能，在很大程度上依赖于设计过程中所采用的信息检索模型的优劣。因此，信息检索模型是信息检索理论中最重要的研究内容之一。

信息检索的模型，就是运用数学的语言和工具，对信息检索系统中的信息及其处理过程加以翻译和抽象，表述为某种数学公式，再经过演绎、推断、解释和实际检验，反过来指导信息检索实践。

最简单的信息检索模型是单项检索模型。它将文献集合中的每一篇文献用 1 个或多个主题词标引，提问式由单个主题词构成。系统对提问的响应是：检中或不检中。匹配标准是：若提问式中的主题词属于某文献标引词集合中的成员，则该文献为检中；反之，为不检中。此模型由于检索过程简单，较为人们熟知且广泛使用。但此种模型的检索效果往往不好，尤其当文献集合很大时，检中的文献很大部分是无用的文献。

1957 年，Y.Bar-Hillel 最先探讨了布尔逻辑应用于计算机检索的可能性，10 年后，布尔逻辑模型正式被大型文献检索系统所采用，并逐渐成为各种大型联机检索系统甚至是网络搜索引擎的典型、标准检索模式。为弥补布尔逻辑模型的不足，相继出现了向量空间模型、概率检索模型、模糊集合模型、扩展布尔逻辑模型等。在介绍这些模型前，我们先以 $S=(D,T,Q,\rho)$这个四元组的方式来描述一个信息检索系统。其中：

$$D = (D_1, D_2, D_3, \cdots, D_n)$$

$$T = (T_1, T_2, T_3, \cdots, T_m)$$

$$Q = (Q_1, Q_2, Q_3, \cdots, Q_l)$$

$$\rho: Q \times D \rightarrow R$$

D 为系统中经过标引的文献集合，T 为所有可能存在的标引词，Q 为提问集合，ρ 为匹配函数，R 为函数值集合。

1.1.2.1 布尔逻辑检索模型(Boolean Model)

布尔检索模型采用布尔代数和集合论的方法，用布尔表达式表示用户提问，通过对文献标识与提问式的逻辑运算来检索文献。在传统的布尔模型中，每一文献用一组标引词表示。如，用表达式 $D_i = (T_1, T_2, T_3, \cdots, T_m)$ 表示文献 i，式中 T_1，T_2，T_3，$\cdots$，T_m 表示文献 i 中的所有标引词集合。

每个提问式 Q 除表示用户需求中的标引词组合外，还有各标引词的布尔组配。系统在对提问进行处理时，输出一个包含有该提问式的组配元(标引词)且符合组配条件(逻辑运算符)的文献集合。

常用的布尔逻辑组配运算符有：逻辑“与”(AND,常用符号“*”表示)、逻辑“或”(OR,常用符号“+”表示)、逻辑“非”(NOT,常用符号“-”表示)。图 1.2 为这些运算符的图解，阴影部分即为两个集合的运算结果。

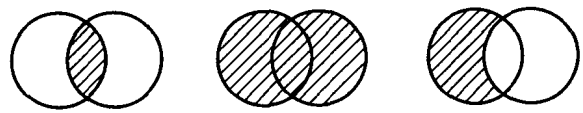

图 1.2 布尔逻辑运算符文氏图

如，对于一个表示为 $Q_i = (T_1 \text{ AND } T_2) \text{OR } (T_3 \text{ AND } (\text{NOT } T_4))$ 的提问式，系统的响应必须是这样一组文献集合：这些文献中都含有标引词 T_1 和 T_2，或者含有标引词 T_3 但不含有标引词 T_4。

布尔检索模型因其简单、易理解、易实现、能处理结构化提问等优点，在信息检索系统中得到了广泛的实际应用。然而，由于它所采用的准确匹配策略太僵硬，将一些有可能满足提问需要的文献排除在命中文献集合之外，所以，检索结果常常不能十分令人满意。传统布尔检索模型的具体缺陷主要表现在以下五方面：

(1) 布尔检索式的非友善性，即构造一个好的检索式是不容易的。尤其是对复杂的检索课题，不易套用布尔检索模式。

(2) 易造成零输出或输出过量。检索输出完全依赖于布尔提问式与系统倒

排档中文献的匹配情况，输出量较难控制。

(3) 无差别的组配元，不能区分各组配元的重要程度。

(4) 匹配标准存在某些不合理的地方。由于匹配标准是有或无，因此，对于文献中标引词的数量没有评判，都一视同仁。

(5) 检索结果不能按照重要性排序输出。

为了克服上述缺陷，人们对传统的布尔模型进行改进和扩展，建立了一些新的模型。

1.1.2.2 向量空间检索模型(Vector Space Model)

向量检索是以向量的方式确定检索内容的方法，系统中的每一篇文献和每个提问均用等长的向量表示。如，文献集合中的第 i 篇文献用 $D_i=(T_1,T_2,T_3,\cdots,T_m)$ 表示，其中，T_1，T_2，T_3，…，T_m 为系统中所有标引词集合；提问集合中的第 j 个提问用 $Q_j=(T_1,T_2,T_3,\cdots,T_m)$ 表示；T_k 表示文献向量或提问向量中的第 k 个分量，即文献表示或提问式中所含的第 k 个标引词或检索词。传统的向量空间模型将 T_k 取值为“0”或“1”，现在大多在[0,1]区间取值。这样，就可以构成一个向量空间，把信息检索中文献与提问的匹配处理过程转化为向量空间中文献向量与提问向量的相似度计算问题。某一文献与某一提问的相关程度通过计算该向量对之间的相似度来测定。这种方法自然引入了检索的柔性和模糊性，从理论上使检索更为合理，一出现即备受关注。

计算相似度的函数式有几十种，其中有一些来自数值分类领域，有些是用于文献自动聚类或关键词聚类的，而不是用于检索排序输出的，最简单的计算方法就是用点积函数。较常用的方法是用余弦函数，这种方法的实质就是计算 m 维空间中文献向量与提问向量之间的夹角余弦。当两个向量完全相同时，它们在该空间中相互重叠，即夹角为“0”时，函数(相似度)达到最大值。

当全部文献向量与某个提问向量的相似度都计算完毕后，系统就把相似度超过某一规定阈值的文献(或者根据预定要检出的文献数量)按相似度大小降序排列输出。因此，排在最前面的文献从理论上讲是和提问最相关的文献。

采用这种向量检索模型的典型系统就是 G. Salton 等人在 20 世纪 60 年代中期开始研制的实验性系统 SMART (System for the Mechanical Analysis and Retrieval of Texts)。与采用布尔模型的普通检索系统相比，该系统有以下几个特色：

(1) 采用自动标引技术为文献提供标引词；

(2) 改变了布尔检索非“1”即“0”的简单判断，标引词和文献的相关程度可在[0,1]闭区间中取值，使标引者和检索者都可比较灵活地定义组配元(标引词)与文献的关系深度，改变了布尔检索模型僵化的缺点；

(3) 由于以其相似的程度作为检索的标准，可从量的角度判断文献命中与否，从而使检索更趋于合理；

(4) 检索结果可按与提问的相关度排序输出，便于用户通过相关反馈技术修正提问，控制检索量；

(5) 布尔模型的逻辑关系依然可以使用，保留了直观性和方便性。

向量空间模型为揭示信息检索的基本原理做出了重要贡献。但是，向量模型也存在着某些明显的缺陷：如检索过程转化为向量的计算方法，不能反映出文献之间的复杂关系；由于对任何一个提问都需要计算全部文献库中的每一篇文献，因此，计算量大、算法复杂性较高；由于标引加权和检索加权是分离的，因此，随意性较大，难以保证质量。萨尔顿也承认文献中的标引词实际上并不是相互独立的，它们之间存在一定的语义联系。为此，有人又致力于研究基于词相依性的向量模型。例如，有人提出了广义向量空间模型，用一组经过挑选的正交基向量来表示词向量，词间关系可直接由基向量表示给出较为精确的计算，而且没有在假定标引词相互独立的前提下给出文献矩阵和提问向量。

虽然向量空间模型也有缺陷，但其检索方法仍具有一定的科学性，尤其是它引入了模糊相关的概念，将匹配工作定量化，有利于拓展检索自动化的思路。因此，向量空间模型也常用在目前的网络资源检索系统中，如搜索引擎。

1.1.2.3 概率检索模型(Probabilistic Model)

概率检索模型基于概率排序原理，即根据文献与提问的相关概率来排序输出。有证据表明，在一定条件下，它可以产生优良的排序结果。

事实上，对于某个特定的检索提问，文献集合中的某一文献是否符合用户的信息需求(即是否是相关文献)可以看成是一个随机事件，每篇文献是相关文献的概率各不相同，综合信息需求的概率和文献与标引的相关概率，才能更为合理地划分检索结果。概率检索模型正是基于这一思想建立起来的。目前提出和建立的概率检索模型大多数建立在 Bayes 概率与统计决策理论基础上，基本上是一种决策理论的自适应模型。与前两种模型不同的是，它的提问式不是由用户直接给出，而是由系统通过归纳学习(相关反馈)来构造决策函数，表示信息提问。

概率模型主要关心的是对应一个提问 Q，一篇文献 D 出现时它为相关(或不相关)的概率。概率模型正确处理了文献相关的随机性，因而体现了更为先进的检索思想，并且向用户提供文献的分等级输出，因此，从客观上讲，概率模型使检索更为合理。其主要优点是：

(1) 采用了理论上更为严密的方式进行决策；

(2) 容易与加权方法结合起来使用，为人们提供了一种理论基础；

(3) 不涉及布尔逻辑运算符，回避了构造布尔提问式的困难；

(4) 文献可按用户的期望值输出排序；

(5) 吸收了相关反馈原理，可开发出理论上更为合理的方法。

但是，它也有明显的不足：如增加了存储和计算资源的开销；参数估计问