普通高等教育精品规划教材

高等学校档案学专业系列教材

档案信息检索

肖秋会 编著

ARCHIVAL SCIENCE

WUHAN UNIVERSITY PRESS
武汉大学出版社

高等学校档案学专业系列教材

武汉大学“十一五”规划教材

档案信息检索

肖秋会　编著

WUHAN UNIVERSITY PRESS
武汉大学出版社

图书在版编目(CIP)数据

档案信息检索/肖秋会编著. —武汉：武汉大学出版社,2011.8
普通高等教育精品规划教材
高等学校档案学专业系列教材
ISBN 978-7-307-09078-1

Ⅰ.档… Ⅱ.肖… Ⅲ.档案—情报检索—高等学校—教材
Ⅳ.G273

中国版本图书馆 CIP 数据核字(2011)第 161073 号

责任编辑：詹 蜜 责任校对：黄添生 版式设计：詹锦玲

出版发行：武汉大学出版社 (430072 武昌 珞珈山)
(电子邮件：cbs22@whu.edu.cn 网址：www.wdp.whu.edu.cn)
印刷：武汉市宏达盛印务有限公司
开本：720×1000 1/16 印张：16.5 字数：285 千字 插页：1
版次：2011 年 8 月第 1 版 2011 年 8 月第 1 次印刷
ISBN 978-7-307-09078-1/G·1792 定价：27.00 元

前　言

档案信息检索是对档案信息进行描述、组织、加工，编制档案检索工具，建立档案信息检索系统，并运用各种检索方法和检索技术查找档案信息的一项专门性工作。档案信息检索工作是联结档案基础业务工作与档案利用工作的关键，建立高效的档案信息检索体系是开发和利用档案信息资源，充分发挥档案作用，实现其价值的基本手段。档案信息检索工作对现代信息技术极为敏感，它是现代信息技术和方法在档案工作中运用最为广泛和深入的工作领域之一。在数字网络环境下，档案信息检索的方法和技术不断更新。

《档案信息检索》共分为九章：第一章“信息检索基础”，简述了信息检索的概念和原理，信息检索的发展历史和信息检索模型；第二章“档案信息检索概述”，概述了档案信息检索的内容，档案信息检索系统和检索效率；第三章“检索语言”，简述了检索语言的类型、特点及发展趋势，介绍了国内外具有代表性的文献分类法，以及《中国档案分类法》、《中国档案主题词表》、《汉语主题词表》、《中国分类主题词表》；第四章“档案著录”，简述了档案著录规则，介绍了计算机档案著录和档案机读目录格式的特点和要求；第五章“档案标引”，分别介绍了分类标引、主题标引和自动标引；第六章“档案检索工具”，简述了档案检索工具的种类及编制方法；第七章“计算机档案信息检索系统”，简述了计算机档案信息检索系统的特点和类型、档案数据库组织等；第八章“计算机档案信息检索方法与技术”，简述了布尔逻辑检索、邻近检索、截词检索等常用的计算机档案信息检索方法和相关技术，以及档案检索策略的构造和调整；第九章“网络环境下的档案信息组织和检索”，分析了网络环境下信息组织的主要方式，着重介绍了档案编码描述 EAD（Encoded Archival Description）的产生、发展和应用，并举例说明了部分中外档案网站的信息检索方法。

本书的主要特点是：结构简明，条理清晰，内容丰富，视野开阔，以信息检索的原理、技术和方法为主线，较为全面地阐述了档案信息检索的理论、规则、方法和技术。本书不拘泥于档案信息对象本身，在检索语言、检

索方法等领域适当地介绍了相关的文献信息分类法及常规的计算机信息检索方法。此外，本书对网络环境下的档案信息组织和检索方法较为关注，介绍、分析了EAD等元数据标准，以及档案网站信息检索的方法与途径。

本书由肖秋会主编，山东大学历史文化学院谭必勇参与编写了第九章第五节的内容，档案学硕士研究生郑健参与了书稿的校对工作。

本书为武汉大学“十一五”规划教材，感谢武汉大学信息管理学院、武汉大学教务部和武汉大学出版社的大力支持和帮助。在本书写作过程中，引用和借鉴了不少专家和同行的论著，在此对他们表示深切的谢意。

由于时间仓促，水平有限，本书难免存在疏漏不当之处，敬请批评、指正。

目　录

第一章　信息检索基础

◎ **本章要点**

① 信息检索的概念和原理

② 信息检索的发展历史

③ 信息检索的模型

◎ **关键词**

文献信息检索　事实信息检索　数值信息检索　布尔模型　向量空间模型　经典概率模型

第一节　信息检索的概念和原理

一、信息检索的概念

所谓信息检索（information retrieval），是指将信息按一定的方式组织和存储，并根据用户的需要从中查找所需信息的过程及所采取的一系列方法和策略。信息检索包括信息存储和检索两个方面的过程。因此，广义的信息检索又称为信息存储与检索（information storage and retrieval），狭义的信息检索仅指查找信息的过程，相当于人们通常所说的信息查询（information search）。

信息存储是指在一定的专业范围内，对信息进行筛选，并对选中的信息进行特征描述、加工、组织和编码，产生信息记录及检索标识，并使之有序化，建立数据库。检索是指借助一定的设备和工具，采用一系列方法和策略从数据库中查找出所需要的信息。信息检索一般包括如下两个环节：①信息内容分析与编码，产生信息记录及检索标识，并将全部记录按文件、数据库等形式组织存储，组成有序的信息集合。②用户提问处理和检索输出。

信息检索最初应用于图书馆和科技信息机构，后来逐渐扩大到其他领

域，并与各种管理信息系统结合在一起。与信息检索有关的理论、技术和服务构成了一个相对独立的知识领域，是信息学的一个重要分支，并与计算机应用技术相互交叉。由一定的设备和信息集合构成的，提供一定存储与检索方法及检索服务功能的服务设施称为信息检索系统，如穿孔卡片系统、联机检索系统、光盘检索系统、多媒体检索系统等。

二、信息检索的原理

信息检索的基本原理是：对大量、无序的各类信息进行搜集、描述、加工、组织、存储，建立各种检索工具或检索系统，并按照一定的方法和技术，从中识别、查找和获取所需的各类信息源。信息存储是检索的基础，其目的是对大量无序的信息进行加工、组织，使其有序化，检索则是从有序的信息集合中找出所需要的信息。存储与检索是相逆的两个过程，类似于“放进去”和“拿出来”的关系（见图 1-1）。

信息检索的关键部分是信息提问与信息集合的匹配和选择，即对给定提问与集合中的记录进行相符性比较，根据一定的匹配标准选出有关信息。

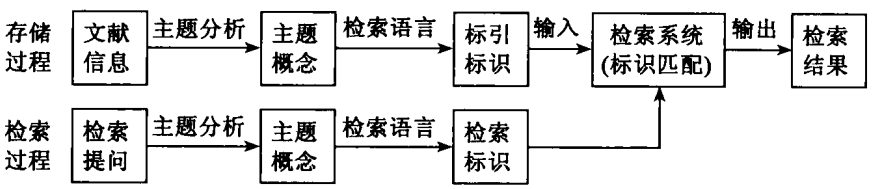

图 1-1　信息检索基本原理示意图

三、信息检索的类型

信息检索可以按不同的标准划分为不同的类型。

1. 按检索对象的形式划分

（1）文献信息检索。通常是指对二次文献信息（题录、索引、文摘）的检索，它们是文献信息的外部特征和内容特征的综合描述，包括文献题名、著者、时间、出版项、文种，等等。信息用户通过检索获取的是原文的“替代物”。与文献信息检索相对应的是书目型数据库。

（2）数值信息检索。它是以数值或数据为检索对象的检索，如各种统

计数据、自然现象观测数据、市场行情数据、企业财政数据、公式，等等。检索系统不仅直接提供有关的数据或数值，还能提供对数据的运算推导功能，以及制表和绘图功能，信息用户可用检索到的数值信息作进一步的定量分析。与数值信息检索相对应的是各种数值数据库和统计数据库。

（3）事实信息检索。这是以某一客观事实为检索对象的检索，查找某一事件（事实）发生的时间、地点和过程（情况）等方面的信息，其检索结果主要是客观事实或为说明事实而提供的相关资料。例如：通过公司黄页检索××公司的销售业绩、人员组成、工资情况、市场规模等信息。与事实信息检索相对应的是各种指南数据库和全文数据库。

另外，近年来，出现了一种以信息检索对象的形式为划分标准的新的三分方法，即文本检索、数值检索、音频与视频检索。其中，文本检索以各种自然语言符号系统所表示的信息作为主要检索对象，是传统文献检索方式的延续，目前在信息检索领域仍占据主要地位并得到新的发展，检索对象既包括早期的结构化书目信息，也包括越来越多的非结构化或半结构化的自由文本信息，检索方式包括关键词检索、概念检索及语义检索。音频与视频检索主要是针对各种数字化音频与视频信息而迅速发展的一种新兴的信息检索类型，随着媒体数字化技术和网络技术的发展，人们对数字音频与视频信息的分析和查找的需求越来越突出，基于内容的音频与视频信息检索成为信息检索研究领域的热点之一。

2. 按系统中信息组织的方法划分

（1）全文检索。检索系统中存储的是整篇文章或整本书，用户检索时可根据自己的需要，从中查找、获取任意的字、句、段、节、章等信息，还可以进行各种频率的统计和内容分析。随着计算机存储容量的增大和运算速度的提高，全文检索已经由最初的法学、文学领域迅速向其他学科和专业扩展。

（2）超文本检索。超文本是由节点（Node）和节点之间的逻辑链路（Link）所构成的一种信息组织方式。节点与节点之间通过链路相互连接，形成了错综复杂的信息网络。超文本是一种非线性的信息组织方式，超文本检索能够提供浏览式的查询，通过链路的指引，信息用户可在浏览节点内容的过程中选择进一步阅读或查询的方向。

（3）超媒体检索。超媒体系统的存储对象突破了文本，集成了图像、动画、声音等多种媒体的信息，信息的存储结构从单维发展到多维，存储空

间范围在不断扩大。超媒体是对超文本的补充和发展。

3. 按检索工具和检索方式划分

（1）手工检索。手工检索对应于印刷型文献和检索工具。通过人工方式对文献进行著录和标引，建立著录卡片，并按一定方式编排，建立卡片式或书本式检索工具。在检索时，用手翻找著录卡片或书本式目录，眼睛查看其内容，并动用大脑思考，从而作出判断来完成检索过程。其特点是检索者可以边查边看边思考，并随时修改检索策略，但检索速度很慢，一次检索只能采用一种检索途径，检索效率低下，检索工具的更新慢。

（2）机械检索。即机械穿孔卡片检索，是在手工穿孔卡片基础上发展起来的，依靠探针及其辅助设备，对代表检索标识（分类号或主题词）的穿孔卡片进行选取的一种检索方式。与纯手工检索方式相比，机械检索在一定程度上提高了检索效率。但由于设备笨重，操作复杂，适用范围较窄。

（3）缩微品检索。它是以缩微胶片和缩微平片为存储载体，利用相应的光学或电子技术设备处理信息的一种检索方式。这种检索需要借助于缩微显示设备。

（4）光盘检索。光盘是继缩微品、磁盘存储器之后的一种新型信息存储载体。它的原理是利用激光束改变存储介质对激光束的不同效应来识别和读出信息。按照数据存取方式，光盘检索可分为只读光盘、交互式光盘、一次写入式光盘、可擦写式光盘 4 种类型。按存储信息的类型可分为音频光盘、视频光盘、数字光盘和多媒体光盘等。光盘与其他载体相比，其显著特点是存储容量大、易保存、便携带、可套录、有限花费、无限检索。光盘检索特别适于开展专题检索和定题服务。

（5）计算机检索。这是把信息及其检索标识转换成计算机可阅读的二进制编码，存储在磁性载体上。由计算机根据程序进行查找并输出结果。根据检索者与计算机之间进行的不同的通信方式和计算机信息检索的发展阶段，分为脱机检索和联机检索。

（6）网络信息检索。网络信息检索是指互联网用户在网络终端通过特定的网络搜索工具（搜索引擎）或是通过浏览的方式，查找并获取信息的一种检索方式。网络信息检索以互联网基础设施、卫星通信技术、网络检索软件、网络标准通信方式等为基础。基于 WWW 的网络信息检索比传统的联机信息检索方便、快捷、低廉，但其稳定性、安全性、数据的准确性和权威性得不到保障，检索冗余度较大。

第二节　信息检索发展的历史及模式的演变

信息检索经历了手工检索、机械检索、脱机批处理检索、联机检索和网络信息检索五个阶段。早期的信息检索主要依靠信息分类，但是随着文献量的增加和文献类型的多样化，单纯依靠分类难以解决快速查找文献信息的需要，因此，出现了能基于某一线索深度揭示某一具体内容的工具——索引，以及能在一定程度上浓缩和揭示文献内容的工具——文摘。目录、索引和文摘等书目型检索工具成为手工检索时期查找印刷型文献最基本的工具。

机械检索开始于20世纪50年代，它采用各种机械工具和设备进行检索，是手工检索向计算机信息检索的过渡。

20世纪中期以来，计算机技术在信息检索领域的应用不断取得突破，计算机信息检索经历了脱机批处理检索、联机检索、网络信息检索等主要阶段。计算机具有强大的信息存储与信息处理能力，与手工检索相比，计算机信息检索具有检索速度快、检索途径多、检索范围广、检全率高等优点。

一、手工检索（19世纪70年代—20世纪40年代）

信息检索活动起源于图书馆参考咨询工作和文摘索引工作。从19世纪下半叶开始发展至20世纪初，信息检索逐渐成为图书馆的一项独立的用户服务工作。1876年，美国图书馆协会（American Library Association，ALA）成立并召开了第一届大会，马萨诸塞州伍斯特（Worcester）公共图书馆员塞缪尔·格林（Samuel S. Green）在会上首次提出了开展参考咨询服务的建议。1883年，美国波士顿公共图书馆设立了第一个专职参考咨询职位——参考馆员，并设立了参考阅览室，从此，作为信息检索起源的参考咨询工作成为图书馆的一项正式业务而得到发展。20世纪初，绝大多数图书馆都成立了参考咨询部门，其服务内容包括：利用图书馆的书目工具帮助读者查找图书、期刊，进行文献分析与翻译等。

在手工检索时期，独立性的文摘性刊物出现并得到发展，索引也成为独立的检索工具，而索引与文摘的结合使用，使各种手工检索工具的查询功能得到提高。在这一时期，出现了一批高质量的文摘和独立索引等大型检索工具，包括：美国工程信息公司编辑出版的《工程索引》（*Engineering Index*，EI）（1884年），英国电气工程师协会编辑出版的《科学文摘》（*Science Abstract*，SA）（1898年），美国化学文摘社编辑出版的《化学文摘》（*Chemi-*

cal Abstract，CA）（1907 年），美国生物科学信息社编辑出版的《生物学文摘》（*Biological Abstract*，BA）（1926 年），美国科学信息服务社编辑出版的《科学引文索引》（*Science Citation Index*，SCI）（1961 年），等等。上述学术性的文摘和索引工具为各国科研人员提供了重要的文献信息源和检索服务。

手工检索的特点是：以印刷型文献为主要检索对象，以各类题录、索引、文摘等书目型工具为主要检索工具，以图书馆的参考咨询部门为检索服务的主要机构。手工检索操作简单，费用低廉，但检索效率很低。随着文献信息量和信息类型的增加，计算机信息检索技术的迅速发展，传统的利用印刷型文献进行手工检索的方式逐渐退出了检索的主流。

二、机械检索（20 世纪 40—50 年代）

20 世纪 50 年代，机械检索开始得到使用。1954 年，现代情报学创始人美国的万尼瓦尔·布什（Vannevar Bush）博士在其论文 As We may Think 中首次提出了设计自动的、在大规模存储数据中进行查找的机器的设想，他与美国农业部图书馆馆员拉尔夫·肖共同制造了一台快速检索机——布什·肖检索机。它利用光电原理，对缩微复制在胶卷上的文献信息进行检索。

机械检索的原理是，通过设计和制作特定的机械装置，改进信息的存储和检索方式，通过控制机械动作，借助机械信息处理机的数据识别功能部分地替代人脑，在一定程度上实现了信息检索的自动化。但它只是采用单一的方法对固定的存储形式进行检索，成本高，检索复杂，检索效率不甚理想。机械检索系统很快被计算机情报检索系统所取代。

三、脱机批处理检索（20 世纪 50—60 年代）

1946 年世界上第一台计算机问世，20 世纪 50 年代，计算机技术开始在书目情报检索领域得到应用，1954 年，美国海军军械实验中心利用 IBM701 机将 4 000 篇技术报告进行了计算机存储与检索的实验，建立了世界上第一个计算机文献情报检索系统。20 世纪 40—60 年代，在计算机应用领域“穿孔卡片”和“穿孔纸带”数据录入技术及设备相继得到应用，以它们作为存储文摘、检索词和查询提问式的媒介，使得计算机开始在文献检索领域中得到了应用，20 世纪 70 年代之后，“穿孔卡片”和“穿孔纸带”被磁性媒介（磁带等）所代替。

计算机应用于信息检索的早期阶段主要以脱机检索方式为主。脱机检索利用单机的输入和输出装置，用磁带作为媒介进行检索。以脱机方式检索，

计算机只能顺序检索磁带上记录的信息，每检索一次都必须从头到尾读一遍磁带，因此，一般采用批处理方式实施检索。脱机批处理检索的具体表现为：输入计算机的待检文献信息（文献题录、文摘等）存储在磁带上，检索提问则存储在穿孔纸带或穿孔卡片上，其特点是不对一个检索提问立即作出回答，而是集中大批提问后再进行处理。即由检索人员集中一批用户的检索提问，预先编制检索策略，存储在计算机检索系统中，定期地检索数据库中新增加的内容，然后把命中的文献信息分发给各个用户。脱机检索过程中，人机不能直接交互、对话，处理的周期较长，因此，检索效率往往不够理想。脱机检索方式适用于面向科技人员的定题服务。定题服务是登记用户提问并存入计算机中形成一个提问档，每当新的数据进入数据库时，就对这批数据进行处理，将符合用户提问的最新文献提交给用户，可使用户随时了解课题的进展情况。

四、联机检索（20 世纪 60—90 年代）

联机检索是在远程终端设备上借助通信线路与远距离数据库系统的一种问答式检索。联机检索产生于 20 世纪 60 年代中期到 70 年代初，由于计算机分时技术的发展，通信技术的改进，以及计算机网络的初步形成和检索软件包的建立，用户可以通过检索终端设备与检索系统中心计算机直接进行人机对话，从而对远距离之外的数据库进行检索。

1965 年，美国系统发展公司（SDC）研制成功了联机检索软件——书目情报分时联机检索（Online Retrieval of Bibliographic Information Time Shared，ORBIT），标志着联机检索的诞生。1966 年，美国洛克希德导弹与宇航公司研制了世界上第一个人机对话的信息检索系统——DIALOG 系统，开始了联机文献情报检索。此后，其他大型联机检索系统如 BRS 系统（存储和信息检索系统）、欧洲的 ESA-IRS 系统（欧洲航天局信息检索系统）等都开始研制并逐步发展起来。

20 世纪 70 年代开始，联机检索由实验转向商业化运营，对社会公众提供服务。20 世纪 80 年代，随着空间技术的发展，信息检索进入了国际联机检索的新时期。国际联机信息检索是指商业性的联机数据库检索服务机构通过国际（卫星）通信网络，为世界各地的用户终端提供人机对话式的检索服务方式；用户利用计算机终端设备，通过国际（卫星）通信网络，与世界上的大型计算机检索系统的主机联结，从而能检索世界范围内各个计算机检索系统的信息资源。国际联机信息检索使信息检索超出了一个国家和地区

的范围，促进了全球信息资源的共享。

在此期间，出现了光盘检索。与国际联机检索相比，光盘检索不需要支付国际通信费，具有方便、易操作、费用低、寿命长，以及其海量存储可实现原文检索等优点，因此在我国十分普及。但光盘数据库更新速度慢，不能完全取代国际联机检索。

五、网络信息检索（20 世纪 90 年代—至今）

网络信息检索是在国际联机检索和光盘检索基础上发展起来的，通过 Internet 对远程计算机上的信息进行的检索。20 世纪 90 年代，随着卫星通信、光纤通信等现代通信技术以及信息高速公路等网络基础设施的迅速发展，基于 Web 的网络信息检索开始出现并得到迅猛发展。在这一时期，因特网资源爆炸式增长，网络搜索引擎技术的发展应用令人瞩目，同时，传统的联机检索系统如 Dialog 及各类数据库检索系统的信息服务也逐渐建立了 Web 服务平台，面向互联网终端的用户提供服务。

Internet 技术发端于 20 世纪 60 年代，在经历了早期的军事、科技与教育等专门领域的试验和应用之后，于 20 世纪 80 年代末开始在全球范围内飞速发展。文件传输（FTP）、远程登录（Telnet）、电子邮件（E-mail）成为当时 Internet 上广泛使用的三大基本服务。进入 90 年代，WWW（“World Wide Web”，又称为“Web”、“3W”）在 Internet 上获得迅猛发展，它所支持的超文本技术和浏览器技术能够使用户获取联入 Internet 的世界上任何一台计算机上的文本、图形、声音以及视频等各类信息，WWW 信息检索迅速取代了 FTP 和 Telnet，成为主流的技术应用平台。

随着因特网资源的增长，搜索引擎技术产生并不断创新。1994 年 4 月，美国斯坦福大学的两名博士生 David Filo 和美籍华人杨致远创办了网络资源目录 Yahoo!，同年 7 月，具有现代意义的机器搜索引擎 Lycos 诞生。搜索引擎技术从此进入了高速发展时期。搜索引擎分为两类：①以 Yahoo! 为代表的目录式搜索引擎，主要采用人工方式搜集网络信息，进行分类整理，并以分类目录浏览的方式提供服务；②以 Lycos、AltaVista、Excite 等为代表的机器人搜索引擎则通过特定的搜索软件对 Web 信息进行搜索，建立索引文档，为用户提供关键词查询服务，具有自动化程度高、收录范围广、功能强大的优势。由于目录式搜索引擎成本高、更新不及时、数据量有限，逐渐被具有现代意义的机器人搜索引擎所超越。

在信息检索由手工检索、机械检索、单机检索、联机检索向网络信息检

索的发展过程中，信息检索模式也随之发生着演变，即由传统文献信息检索的“提问—检索”模式，转向以 Internet 为基础的“浏览—查询”模式。

第三节　信息检索的基本模型

信息检索需要解决的核心问题之一，就是明确用户的信息需求形成机制以及如何最大限度地满足用户的信息需求。随着信息量和信息类型的快速增加，人们在信息检索过程中开始采用科学的方法。信息检索模型是应用数学知识和工具，对信息检索系统中的信息及信息处理过程进行概括、翻译、解释和抽象，并用特定的数学公式描述其基本原理，从而指导信息检索实践活动。

20 世纪中期，数学工具被引入到信息检索领域，研究人员先后提出了不同类型的信息检索数学模型，这些检索模型在信息检索实践中得到了不断的发展和完善。信息检索模型可以分为基于内容的检索模型和结构化数学模型两大类。其中，信息检索的三个基本模型是：布尔模型（Boolean Model）、向量空间模型（Vector Space Model，VSM）和概率模型（Probabilistic Model）。在布尔模型中，文献和查询用标引词集合来表示；在向量空间模型中，文献和查询用空间向量图来表示；概率模型则应用概率论原理来构建文献和查询机制。此外，使用比较广泛的检索模型还有模糊集合模型、扩展布尔模型，以及近年来广受关注的基于本体的检索模型、跨语言信息检索模型，等等。

一、布尔模型

布尔模型是产生最早、应用最广泛的经典信息检索模型。1957 年，Y. Bar-Hillel 首次探讨了布尔逻辑模型应用于计算机检索的可能性，至 20 世纪 60 年代中期，布尔模型正式被大型文学检索系统所采用，此后经久不衰，被各大联机检索系统以及网络搜索引擎所采用，成为各类信息检索系统都普遍采用的经典模型。

布尔检索模型采用了布尔代数和集合论的方法，用布尔表达式表述用户提问，通过对文献标识与提问式的逻辑运算来检索文献。检索提问往往涉及多个概念，同一个概念又可以表达为多个同义词或近义词，为了全面、准确地表达检索提问，检索系统采用布尔逻辑运算将不同的检索词组配起来，形成一个个具有简单概念的检索单元，将这些检索单元进一步组配，形成一个

具有复杂概念的布尔表达式，从而表达用户的信息检索需求。

布尔检索的主要优点是形式简洁、结构简单、易学易用；主要不足之处是：检索词没有权重区别，不能体现检索项的主要程度；采用非是即否的精确匹配方式，无法描述与查询条件部分匹配的情况，导致检索结果不够精确，查全率受到影响。

二、向量空间模型

向量空间模型是一种利用统计学方法而建立的数学模型。20 世纪 60—70 年代，G·萨尔斯顿基于部分匹配（“partial matching”）的检索思想，在其开发的实验性检索系统 SMART 中首次提出了向量空间模型，其工作原理是将检索文档和检索提问式（关键词）都看做是一组数值向量，形成向量空间图，将检索文档向量与检索提问式向量进行相似度测定，对检出的文献按文档与检索提问之间的相似度降序排列，实现文献与查询的部分匹配。

向量空间模型的主要优点是：标引词加权处理，可以灵活地定义标引词与文献的关系深度，从而改进检索效果；部分匹配策略能检出与查询条件接近的文献，避免了布尔逻辑模型非是即否的僵化的缺点；余弦公式可对检索结果按照与提问的相关度排序输出，便于用户修正检索提问。其主要缺点是：检索过程转化为向量的计算方法，不能完全反应文献之间的复杂关系；标引词加权和检索词加权是分离的，随意性大，质量难以保证。

三、经典概率模型

概率论模型利用概率论原理来理解和解决信息检索问题，是基于文档与提问式是否相关的概率来进行信息检索。经典概率模型（Classic Probabilistic）是最早出现的概率模型，此后还出现了推理网络模型（Inference Network Model）和信念网络模型（Belief Network Model）。

经典概率模型由 S. E. Robertston 和 K. Sparck Jones 于 1976 年提出，它的基本指导思想是：给定一个检索提问，则检索系统中存在着一个与该提问相关的理想命中结果集合，如果已知该集合的主要特征及其描述，则用户的检索要求不难实现。但在现实中，用户并不知道这个理想结果集的特征，因此，需要在初始检索时对其进行猜测，并获得一个初步的命中结果集合。用户或者系统对这个初始检索的命中结果集合的文档进行相关性判断，并根据反馈的信息，不断优化和改进后续的检索策略，从而逐步使检索结果接近该提问的理想命中结果集合。