JI SUAN JI QING BAO JIAN SUO

计算机情报检索

[illegible]生
王延飞　编著
赵丹群

北京大学出版社

计算机情报检索

赖茂生　王延飞　赵丹群编著

北京大学出版社

新登字(京)159号

计算机情报检索

赖茂生　王延飞　赵丹群编著

责任编辑:段晓青

*

北京大学出版社出版发行
(北京大学校内)
北京印刷三厂印刷
新华书店经售

*

850×1168 毫米　32 开本　11.625 印张　290 千字
1993 年 3 月第一版　1993 年 3 月第一次印刷
印数:00001—5 000 册
ISBN 7-301-02062-7/TP・0170
定价:9.90 元

内 容 简 介

本书系统介绍计算机情报检索的原理和技术方法。全书共分为十三章：概述、情报检索系统、文献数据库、非文献型数据库、联机情报检索系统、检索策略与步骤、检索技术、文档检索技术、光盘数据库及其应用、情报检索系统设计与开发、系统评价、文献处理自动化（自动标引、自动摘录和自动分类）、计算机情报检索的发展趋势。内容新颖、涉及面广、分析阐述深入、注重实用。

本书可作为高等院校情报学、图书馆学和档案学等专业以及与信息管理有关的系科的教材或教学参考书，亦可供各类信息中心、情报研究所（室）、图书馆、档案馆的工作人员及广大信息工作者学习参考。

前　言

计算机情报检索是本世纪50年代出现的新学科。40年来，它在理论和实践方面都迅速发展，在许多国家得到普遍应用和研究，为科学技术和社会经济的发展和图书情报工作现代化作出了重要贡献。我国在这方面虽起步较晚，但近10多年来发展很快。目前，已在全国数十个城市设立了上百个国际联机检索终端，可直接查询欧美地区的联机数据库。我国自建的全国科技情报检索网络也已初具规模，并投入了商业性运营。服务范围也已从原先主要面向科学研究逐渐转向经济建设、经营管理和社会服务领域。同时，许多机构已经开发出或正在开发各种各样的情报检索系统和信息管理系统。从事情报检索和信息服务的人数不断增多。情报检索的基本原理和方法是什么，如何建立和利用信息服务设施，如何开发和管理各种信息资源，已成为我国许多读者关心的问题。

北京大学信息管理学系在80年代初就开始设立“计算机情报检索”课。本书是在总结多年教学实践的基础上，参考了国内外一些有代表性的教材和大量文献资料，经过作者们多年反复锤炼而成的。主要内容包括：计算机情报检索的基本原理，发展概况，情报检索系统构成，文献数据库、数值数据库、指南数据库、术语库和图像数据库，光盘数据库，国内外主要联机检索系统，联机检索的策略、技术和方法，文档检索技术，情报检索系统设计、开发与评价，自动标引、自动摘录和自动分类，以及计算机情报检索的发展趋势。与《科技文献检索》(北京大学出版社，1985年)相比，本书是该书内容的继续和发展。传统的文献检索方法和检索语言方面的知识已在该书中系统介绍，本书一般不再重复。必要时，读者可将这两部教材结合起来使用。

本书共十三章，其中，第一、二、九、十一、十二和十三章由赖茂

生负责编写，第三、四、五章由王延飞和赖茂生合作编写，第六、七、八章由赵丹群负责编写，第十章由王延飞负责编写。全书最后由赖茂生统稿。在编写过程中，得到了系资料室同志和有关专家的热心帮助。北大出版社的段晓青和姚梅生两位老师为本书的出版提供了宝贵的指导和帮助。在此，一并表示由衷的谢意。

由于作者水平有限，书中难免有疏漏之处。敬请各位读者批评指正。

作　者

1992 年春节于燕园

本书系统介绍计算机情报检索的原理和技术方法。概括了计算机情报检索的发展趋势。内容新颖、涉及面广、分析阐述深入、注重实用。

本书系统介绍计算机情报检索的原理和技术方法。概括了计算机情报检索的发展趋势。内容新颖、涉及面广、分析阐述深入、注重实用。

目　次

1 情报检索概述

1.1 情报检索与计算机

电子计算机诞生于20世纪40年代中期。“情报检索”这个术语则出现于40年代末，虽然它作为一门技艺可以追溯到较久远的年代。而作为现代技术，计算机与情报检索几乎是同时问世，且立即建立了非常密切的关系。这也许又是科学史出现的一种很有趣的学科共生现象。

1.1.1 情报检索的涵义

今天，计算机的概念几乎是家喻户晓。情报检索也已经突破图书馆这个狭窄的圈子，走上了整个社会大舞台。然而，何谓情报检索？对这个基本概念问题，要获得一个准确、统一的解释，也并不是很容易的。本书的姐妹篇《科技文献检索》(北京大学出版社1985年出版)曾将它定义为“将情报按一定的方式组织和存贮起来，并根据用户的需要找出有关情报的过程。”即把它解释为人类信息活动的一种过程，其中包括存与取两个环节，但又不是简单机械的存取。在这里，存是指一种面向来自各种渠道的大量信息而进行的高度组织化的存贮。而所谓取，就是面向随机出现的信息需求而进行的高度选择性的检索，且尤其强调快速便利地检出与需求有关的信息。

从本质上讲，情报检索也是一种通讯。穆尔斯(Calvin W. Mooers)1949年首次提出此术语时，就把它定义为一种“延时性通讯形式”，“在时间上从一个时刻通往一个较晚的时刻，而空间上可

能还在同一地点”。它的某些方面可以和基于统计学的通讯理论相比较，即可以把从范围广泛的知识海洋中找出适用信息的问题看成类似于有噪声干扰的情况下探测信息脉冲是否存在的问题。

情报检索还是一个发展的概念。随着有关技术的进步，应用领域的扩大，它的内涵也会更加丰富。在今天的信息社会中，人类获取信息的行为常常都离不开情报检索。后者包括或涉及了一切有目的和组织化的信息存取活动。“存取”(access)已成为一个更为流行的、既含有情报检索过程又能体现社会信息化特征的术语。“情报检索服务”也有可能被“存取服务”(access services)这一术语所取代。在我国，“信息”这个词比“情报”使用得更为普遍，所以，“信息检索”这个术语的流行恐怕只是个时间问题，但不管怎样，它与情报检索是同义的。

最后，还值得一提的是，可供信息存取的媒体多样化(如超文本和多媒体数据库的出现)，也使得情报检索已不能简单地区分为文献检索、数据检索和事实检索了。人工智能的研究成果已开始应用于情报检索，出现了“智能情报检索系统”或“智能数据库”。知识的采集、表示、存贮、处理和检索利用问题，将成为情报检索的重要内容之一。

1.1.2 情报检索的基本原理

人类的情报检索行为总是随特定的情报需求产生而开始，并在特定的环境和情报检索系统中完成。这里所说的环境包括产生需求的环境，情报检索系统的运行环境和其他制约因素。特定的检索系统则包括完成检索过程所需的一定设施和工具，它可以是图书馆、信息中心或情报经纪人，也可以是某种工具书(如文摘索引、目录、资料集、手册、词典等)或机读情报源(如各种机读数据库)。

人类的情报需求千差万别，获取信息的方法也各种各样，但情报检索的基本原理却是相同的。我们可以把它最本质的部分概括

为一句话：对情报集合与需求集合的匹配与选择（见图 1-1）。

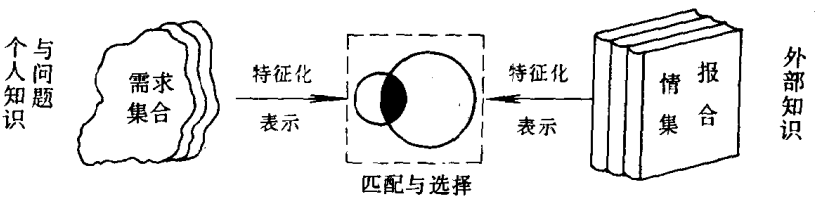

图 1-1　情报检索基本原理示意图

人们在完成某一任务或满足某种需要时，往往会觉得缺少某些知识，因而产生了需求，要访问情报检索系统。情报集合就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，有可能弥补某个特定用户的知识结构缺陷，即可以向用户提供所需要的知识或事实，或获取知识的线索，或者提供某种信息去激活人脑中存贮的知识。而匹配（match）和选择则是一种机制，它负责把需求集合与情报集合进行相似性比较，然后根据一定的标准选出符合需要的情报。这种机制至少包括两个要素：执行匹配的动因和选择的标准（或称匹配标准）。前者可以是人或机器，或二者同时作用；后者则要依需求性质和系统的智能水平来确定。

众所周知，现实世界中的情报发行量和累积量都非常庞大。仅以文献为例，每年发表的科技文献达数百万篇。即使是在一个较小的学科领域，其文献量也是数以万计，且来源广泛。所以，要想进行有效的匹配和选择，首先必须对大量的原始情报进行收集和加工处理，使之从无序到有序，使每件情报都获得某种特征化表示，即让原来隐含的、不易识别的特征显性化。这种加工处理作业通常称为内容分析与标引，其结果是使每件情报都得到某种标识（分类号、主题词等）。这样，人们就可以按照这种标识来组织和检索情报了。其次，由于原始情报往往篇幅很长，不便于管理和检索操作，故人们又发明了各种情报压缩方法（如编目、做文摘或提要），把原始

情报压缩为一条简短的书目记录，或从中摘取各种有用的数据或事实，使情报更便于组织、存贮和匹配选择。

另一方面，对用户提出的情报需求（问题或检索课题）也需要做类似的加工处理，即分析需求的内容，提取出主题概念或其他属性，并利用与情报集合相同的标识系统（检索语言）来表示需求中所包含的概念和属性。经过这样加工处理的情报需求称为提问（query）。

这样，原先的情报需求与情报集合的匹配就简化为提问与有序的、特征化表示的情报集合之间的匹配，即两组有限的语词符号化特征之间的匹配比较。这种简化显然可以提高匹配和选择的效率。不过，它也会带来一些问题，如漏检和误检问题。如何减少乃至避免这类问题，就成了情报检索领域中一个经常性的课题和奋斗目标。

1.1.3 情报检索与计算机联姻

人类早就盼望有一种机器能代替人去做情报匹配选择工作，或者进而代替人去做内容分析工作。“情报检索”一开始就含有机械化检索的涵义。20 世纪初开始使用的穿孔卡片检索装置，就是人们在这方面的一种努力成果。但它的功能和效率都很有限，不是理想的检索机械。

电子计算机的出现使人类的宿愿成为现实。计算机能存贮大量的信息和数据，处理速度快、运算准确、可靠性高。人们可以通过编制程序来指挥计算机完成规定的操作，存取各种数据。这些特性非常适合情报检索的需要。情报检索过程中要存贮大量的数据，要对这些数据进行各种组合，有大量的排序和比较操作，这些都很适合计算机去做。于是，当第一台电子计算机诞生后不久，有些情报学家就开始考虑它能否应用于情报检索，并开始着手这方面的应用研究。随着计算机运算速度的提高，尤其是存贮能力的迅速扩

大，它作为情报检索的主要工具的地位就确立了。计算机与情报检索的正式联姻，产生了计算机情报检索这一新的知识领域。

所谓计算机情报检索，就是在人和计算机的共同作用下去完成情报存取操作，从机器存贮的大量数据中自动分拣出用户所需要的部分。在这里，情报检索的本质没有变，但情报的表示方式、存贮结构和匹配方法变化了（见图 1-2）。要用计算机可以识别的代码来表示情报，用便于计算机快速存取的方式存贮情报。匹配方法亦由人工比较变为机械匹配。匹配标准由隐式变为显式。在这种机械匹配过程中，原先表达概念的语词符号变为没有内涵的字符串。检索过程就是字符串匹配和逻辑运算的过程，表示用户需求的字符串与计算机内存贮的大量字符串（情报集合）的比较和运算的过程。若二者一致或部分一致，并符合给定的逻辑运算条件，即为命中，然后将命中的情报输出给用户。

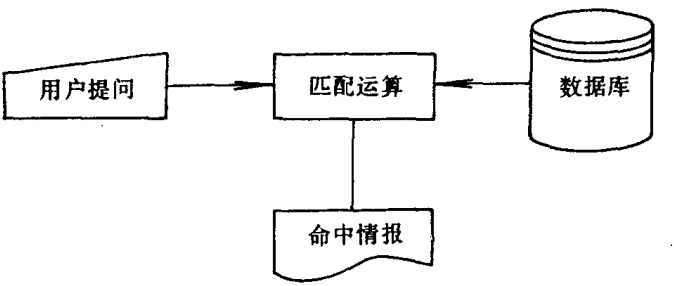

图 1-2　计算机情报检索原理示意图

计算机情报检索的实现，开辟了情报管理现代化的新纪元。它使人类得以从过去那种大海捞针式检索的苦恼中解脱出来，显著地提高了情报选择的效率，节省了大量时间，使人们有能力去应付情报的爆炸性增长。它不仅扩展了图书情报机构的资源和服务能力，使传统的图书馆和情报中心的概念发生了变化，产生了一种新的职业——情报经纪人（即联机服务商、检索专家等）。而且，它还

使情报检索服务走出了象牙之塔，进入了各种办公室和千家万户，直至面向整个社会，成为一种社会公用事业。情报检索的社会化促进了各行业的信息管理现代化，而后者又反过来推动情报检索的发展，给它提出了许多新的课题。最后，还值得特别指出的是，情报检索的计算机化使情报信息成为一种现实的战略资源。信息资源管理（包括信息的生产、分配、利用与保护等）已成为各国普遍关心的重要课题。

1.2 情报检索发展简史

如前所述，情报检索作为一种专门技艺，它的历史可以追溯到图书目录和文摘索引产生的年代。即使是机械化的情报检索，其出现时间也比一般人的想像要早得多。据报道，在20世纪初，穿孔卡片检索装置就已经应用于人口普查了。然而，情报检索真正作为一个科学概念和成为一个学科领域，那还是近40多年的事。

1.2.1 50年代：探索与试验时期

1951年，人们首次利用计算机进行文摘检索试验，并初步证明了它的技术可行性。之后，IBM公司的研究中心和美国海军兵器中心图书馆分别在IBM 701机上开发出计算机情报检索系统。

1958年，具有批处理能力的第二代计算机（晶体管计算机，用磁带存贮数据）问世。科技情报界以召开“国际科学情报会议”（1958年，华盛顿）为契机，掀起了研制开发文献处理自动化系统的小高潮，并在会上展出了许多机编关键词索引以及自动标引、自动摘录和机器翻译等方面的研究成果。它们为后来的机读数据库和机检系统的研制奠定了技术基础。1959年，基于KWIC索引的计算机化定题检索服务（SDI）诞生。

1.2.2 60年代:实用化时期

60年代是计算机情报检索进入生产性开发和实际应用的年代,又是联机检索的试验时期。

1960年,美国国家医学图书馆(NLM)开始建造"医学文献分析与检索系统"(MEDLARS),1964年建成并投入使用。美国化学文摘社(CAS)也于1961年开始发行《化学题录》(CT)机读磁带版。它们的成功使检索刊物的生产和查阅实现了计算机化。其他一些二次情报服务机构也从中受到鼓舞,先后实施了这种变革性的计划,发行各种机读版的文摘索引和目录,即后来被称为书目数据库的机读磁带。

据统计,到60年代末,市场上流通的数据库已有50—100种。批式检索(batch searching)是这一时期计算机情报检索的主要方式。许多机构开始利用市售机读磁带为用户提供SDI服务和回溯检索服务。

批式检索虽比手工检索便利了很多,但用户还是不能与系统进行实时对话,及时修正检索策略,系统的响应速度也慢。所以,人们在60年代初就开始研制更便利的联机检索系统。

1960年,美国麻省理工学院(MIT)开始实施有关联机检索系统设计的"技术情报计划"(TIP),系统发展公司(SDC)也在它开发的全文检索系统"Protosynthex"上进行了首次联机检索表演。SDC的检索程序采用了倒排档技术,可使用位置检索和截词检索方法,但不能用布尔检索法。所用的检索终端用金属丝与计算机连接。1963年起,SDC分别为美国国防部和空军系统司令部研制联机检索系统,于60年代中期开发成功,分别命名为BOLD(Bibliographic On-Line Display)和CIRCOL(Central Information Reference and Control On-Line Experimentation)。SDC后来推出的著名联机情报检索软件ORBIT(On-Line Retrieval of Bibliographic