计算机情报检索导论

计算机情报检索导论

〔加〕H.S.希普斯　著

张承庆　顾慧芳　译

知识出版社

内 容 提 要

本书系统阐述了情报检索的基本概念和一般技术，并反映了七十年代情报检索的新方法和新成果。

本书以教科书形式编写。全书共分14章，详细讲述计算机情报检索的各个重要方面，如：情报检索的基本概念及其所用的文献数据库，提问逻辑及其格式，存储和检索的数据结构，查找程序结构，文献数据库的词汇特性，信息源的研究，数据库的编码和压缩，提问的自动修正，文献自动分类等，各章附有思考题。

本书可供主修计算机科学、情报、图书、档案专业学生以及研究生作教材，同时也是图书馆员、档案、情报工作人员以及从事计算机应用科学人员的一本很好的参考书。

计算机情报检索导论

〔加〕H.S.希普斯 著

张承庆 顾慧芳 译

知识出版社出版

（北京安定门外外馆东街甲1号）

新华书店北京发行所发行 煤炭工业出版社印刷厂印刷

开本787×1092 1/32 印张12.5 字数 262 千字

1984年5月第1版 1984年5月第1次印刷

印数：1—11,100

书号：13214·19 定价：1.30元

译 者 前 言

随着四个现代化的进程，电子计算机在国民经济各部门得到越来越广泛的应用，而计算机情报检索则是其重要的应用方面。鉴于目前国内这方面的教材和参考书还很少，我们翻译了“计算机情报检索导论”这本书，以满足广大读者的需要。

本书原著是一本有影响的书。它是以教科书的形式编写的，内容新颖，由浅入深，自成系统。本书介绍了情报检索的基本概念，又讲述了计算机程序的一般结构以及编制适用的计算机程序所需要的技术，并且采用理论阐述与取自实践的示例相结合的方式进行讲述。若选用不同的章节，则可组配成分别适用于主修计算机科学、情报、图书、档案等专业的学生以及研究生作教材。本书各章还附有思考题，可用来检验读者对各章内容的理解程度，引导读者开阔思路，了解基本原理的进一步应用。

本书讲授了情报检索的很多重要方面，其中计算机检索用的文献数据库，提问逻辑及其格式，存储和检索的数据结构，查找程序的结构，文献数据库的词汇特性，数据库的编码和压缩，提问的自动修正，文献自动分类等各章内容乃是目前设计一个检索系统所必须学习和掌握的。

萨师煊教授和邝桃生、许力群同志在百忙中对本书进行了审校，康金玉同志对本书的翻译工作曾给予帮助，我们在此一并表示深切的谢意。但由于我们的水平有限，译文难免有欠通或错误之处，望读者批评指正。

前　言

本书有双重目的。一方面打算给计算机科学的学生介绍一些情报检索的基本概念，并讲述研制适用的计算机程序所需要的技术。另一方面，讲授有关计算机程序的一般结构，以便图书馆员和情报工作人员等不熟谙计算机细节的人也能理解基本的设计思想。

基于上述目的，本书写成教科书体裁，而不是对当前技术状况作综合的叙述。各章所附的思考题可用来检验读者对内容的理解程度，并引导读者考虑基本原理的进一步发展。

对于已经学过第五章内容的主修计算机科学的大学生，第一至第四章，第六、七、十及十一章的内容适于作为一学期的课程。

对于只具有一般基础知识的学生，可以从第一至第十章中选出部分内容作为两个学期的课程。第七至第十三章的内容构成一门研究生课程的基础。

完全有理由相信，本书的内容足以为非计算机科学专业的学生提供一定程度的基础知识，这些基础知识可能是受过计算机科学教育的读者所熟知的。第三章的开头及第五章的内容便是这种基础知识。同时第三章还为学习计算机科学的学生提供了选自情报检索方面的示例。

要充分地理解第十二章和第十三章，则要求应具备成熟

的数学素养。然而，对于一个不太精通数学的学生来说，在其不理解分析细节的情况下，也能够认识到那些技术的价值及其所得结果的重要性，这一点是毫无疑问的。

第一章用书目数据库将情报检索这门课程给读者作了简要的介绍。第二章中介绍了若干基本的概念。第三章中给出了在情报检索中碰到的数据库类型的示例。第四章着重介绍用户和计算机检索之间称作“接口”的那部分情报检索服务，主要是涉及到提问的可能形式。因此，第一至第四章涉及到情报检索的几个方面与计算的细节及计算机程序结构的组织没有直接关系。

第六章结合第五章中所谈到的数据结构讨论了用于情报检索的计算机检索程序的结构。

第七至第十章是关于文献数据库的词汇特性及通过词汇统计来改进检索系统设计的办法。因此，第七章便介绍了书目数据库的某些一般特点。第八章从理论上对有关情报的存储及传输进行了讨论。第九章中概括地讲述了为了节省存储空间书目数据编码的一些实用方法。第十章则给出了用这些技术设计切实可行的情报检索系统的示例。

书目数据库情报检索的效率不仅取决于检索程序的效率，同时也取决于用来标引文献的术语适于用作主题标识的程度。这一点在第十一章中进行了讨论。在第十二章和第十三章中，考虑了标引术语的自动选择问题，以及提问的自动修正和文献的自动分类等有关问题。

目　录

第一章 引 言

1.1 知识记录的增长

情报检索的许多现代应用都是以近几十年来形成的理论为基础的，从这种意义上说，情报检索是一门新兴学科。但是，尽管如此，这门学科还是可以追溯到几百年之前。

传统的图书馆通过搜集文献，发展了手工编目，使用卡片索引、书目和发行定购图书、刊物、报告等的标准过程。虽然，传统的图书馆更多地着眼于提供文献，而不是提供情报。倘若图书馆的用户兴趣主要在于少量图书和刊物就能包罗其明确课题的话，那么这个方向还是有效的。但目前却出现了许多新的学科领域，这些领域的研究需要来自若干不同学科的情报，仅仅参考能简单指定的少量文献是不能满足要求的。

正如C.P.Show指出的那样，在本世纪之前的人类历史上，社会变化的速度是如此之慢，以致在人的一生中常常感觉不出来[1]。这种情况再也不会继续下去了。人们普遍地认识到：社会和技术正在迅速地发展着，而这些发展则是少数科学家才能理解的新发现所产生的结果。然而，就是这些

---

[1] C.P.Snow, *The Two Cultures and the Scientific Revolution* (Cambridge, Cambridge University Press, 1959), P.45.

新发现却以前所未有的程度影响着整个人类的生活。因此，越来越多的人对快速得到越来越多的情报有极大的兴趣。

想要具备技术和社会迅速发展的能力，则要求一旦需要便能立即提供大量情报。据估计在1800、1850、1900及1966年时所拥有的科学刊物的数量分别约为100、1000、10,000和100,000种〔2〕。由此可见，要满足上述情报要求所产生的问题的严重性。Holt和Schrank曾指出〔3〕：在1920年到1960年期间，经济方面的期刊论文从每年5000篇增长到40,000篇。同样，心理学方面的期刊论文从每年30,000篇增长到90,000篇。在1868到1966年期间，数学方面每年发表论文的数量从800篇增长到13,000篇〔4〕〔5〕。

Carter曾估计过许多不同的学科在1960年到1970年之间在各种刊物上发表的论文增长的情况，见表1.1。Carter对于心理学方面论文的估计可能与Holt和Schrank不一致，这是由于对什么样的论文才算得上心理学方面的期刊论文的理解不同而造成的。当然，个别的数字不如根据同一标准计算出来的增长率更有意义。

一种新的发现可能会导致出版物数量增加。如1958年有

〔2〕 Proceedings of the Royal Institute of Great Britain, vol. 41, Part I, 1966.

〔3〕 C.C.Holt and W.E.Schrank, "Growth of the professional literature in economics and other fields, and some implications," *American Documentation* 19 (1968), 18—26.

〔4〕 K.O.May, "Quantitative Growth of the mathematical literature," Science 154 (1966) 1672—1673.

〔5〕 K.O.May, "Growth and quality in the mathematical literayure," ISIS 59 (1968), 363—371.

人建议使用激光装置〔6〕。1960年红宝石激光器这一课题大约有20篇论文，1961年关于氮-氖激光器的论文约有100篇，1962年有关固态激光器的论文有325篇，1963年镓砷激光器、脉冲激光器和Q开关的论文有700篇，1964年离子激光器的论文有1000篇，1965年有关N_2-CO_2高效激光器的论文有1200篇。

表 1.1 1960和1970年期间期刊上发表论文的数量

学科	1960	1970	学科	1960	1970
数学	15000	30000	冶金	35000	50000
物理	75000	155000	生物	150000	290000
土木工程	15000	15000	地球科学	91000	158000
机械工程	10000	20000	农业	15000	260000
电与电子工程	80000	150000	医学	220000	390000
航天工程	35000	75000	心理学	15000	30000
工程工业	15000	15000	其他学科	929000	1882000
化学	150000	260000	总计	1985000	3780000

“情报爆炸”一词已为一些科技界人员很乐意地接受，他们对可能不知道的前人所作的工作或其他科技工作者目前同时在作的工作非常敏感。另有一种趋势则是夸大忽略别人工作所带来的恶果〔7〕。

有篇文献讨论了从事研究、管理及教学的物理学家和化

〔6〕 A.Neelameghan, “Theoretical Foundation for UDC, its need and formulation,” *Proceedings of the International Symposium, Herceg Novi, Yugoslavia, june* 28-*july* 1, 1971.

〔7〕 A.G.Oettinger, “An essay in information retrieval or the birth of a myth,” *Information and Control* 8 (1965), 64—79.

学家对于情报的需要[8]。Urquart[9]和Bernal[10]分别论述了物理学文摘服务以及化学论文摘要的作用。Bottle对化学家和物理学家用的情报源的问题进行过全面的论述[11]。Guttsman则论述了在社会科学中迅速利用文献的必要性[12]。

1.2 情报检索学科

尽管用文字记录下来的情报通常是通过表示文献的存储数据来进行检索的，但其强调的是与请求有关的情报，而不是文献的直接说明，这就是情报检索这门现代学科的特点。情报检索这门学科除了与可供使用的计算机化的检索系统的实际设计有关以外，同时还包括测量理论与定义方面的问题，以及信息量与适用性方面的问题。

情报检索的某些方面可以与三十年来电子工程师和数学家们所应用的统计的通信理论相比拟。从范围广泛的知识中找出适用情报的问题类似于在有噪声干扰的情况下探测是否

---

〔8〕 Survey of information needs of physicists and chemists.The report of a survey undertaken in 1963-4, in association with Professor B.H.Flowers, on behalf of the Advisory Council on Scientific Policy.Journal of Documentation, vol. 21, pp.83—112, 1965.

〔9〕 D.J.Urquart, "Physics abstracting use and users," *Journal of Documentation* 21 (1965)：113—121.

〔10〕 J.D.Bernal, "Summary papers and summary journals in chemistry," *Journal of Documentation* 21 (1965)：122-127.

〔11〕 R.T.Bottle, "A user's assessment of current awareness services," *Journal of Documentation* 21 (1965)：177—189.

〔12〕 W.L.Guttsman, "The literature of the social sciences and provision for research in them," *Journal of Documentation* 22 (1966)：186—194.

存在着信号脉冲的问题。诸如Wiener均方根判别准则、匹配滤波器、反馈以及相关探测器等方面的概念在情报检索理论方面均有其相应的东西。说起来也许十分令人费解，Norbert Wiener这位曾在把线性预测技术应用到控制理论和控制论中显示出伟大洞察力的人物，却对情报检索的研究价值产生了怀疑，他断言他自己课题研究所需要的任何情报只要他写信给这方面的6位世界专家中的任何一位便可以极容易地得到。

Wiener的观点是很有趣的，可用它来说明现代学者与前辈学者所处的环境差别。正如目前活着的人的数量占自古以来投生到地球上的人数的相当大的一部分一样，以文字记录形式出现的情报量也要比以往任何时候都大许多倍。更有甚者，不仅自然科学技术和社会科学方面的专家需要情报，就是那些连主要权威名字都未必知道的非专家们也需要情报。

有效情报增长的速度具有许多哲学上的意义。很久以前，人被看作是与他人相互作用的孤独的旅行者，后来又被看成是被机械的宇宙所包围，现在，人被视为情报接收者。由于情报无须涉及到物理量或可精密测量的单位，因此可以这样推测：为情报检索和情报评价而发展的技术最终应朝着进一步揭示人类组织思维和理解科学概念及人文概念的过程这个方向发展。

人类的口头通信是通过声波进行的，其速度约1000英尺/秒。计算机之间或计算机与外围设备之间的通信则是由速度为1000英尺/微秒的电子脉冲来完成的。至于计算机之间或通过通信网络而进行数据传输的原理的简单介绍，读者可以

参考Kallenbach的论文[13]。

现代计算机中的继电器和开关设备响应时间为几个毫微秒（1毫微秒 = 10^{-9}秒）。因此，计算机可以在很短的时间间隔内处理和传送大量的情报。对于使用如此高速传送可能产生的影响，以及计算机在这样速度下吸收情报的能力进行科学的、精辟的、富于想象力的推测，导致出现了许多科学小说，例如天文学家Hoyle的那些作品[10]-[16]。McLuhan则从通俗的角度出发特别强调了即时通信对当时社会影响的程度[17]。

计算机化的情报检索系统必须是既经济又实惠的。同样，经济上的考虑已导致了工程结构和化学工程处理上更加严格、因而数学上也就更复杂的设计。正是这些经济上的考虑，要求情报检索原理有更严谨的数学公式，以保证在经济意义上可以使用计算机和计算机可存取的存储装置。同时随着新的理论的发展，研究可供使用的检索系统的性能，以便深刻了解从系统用户角度提出的问题也是重要的。

在某种程度上，计算机化情报检索系统的效率取决于计算机的硬件。由于计算机硬件继续不断地改进，而且随着效

---

[13] P.A.Kallenbach, "Introduction to data transmission for information retrieval," *Information Proceesing and Management* 11 (1975) :137—145.

[14] F.Hoyle and G.Hoyle, *A for Andromeda* (Greenwich, Connecticut: Fawcett Publications Inc.).

[15] F. Hoyle, *The Black Cloud* (London: Heinemann, 1957).

[16] F.Hoyle, *October the First Is Too Late* (London: Heinemann, 1966).

[17] M.McLuhan, *Understanding Media: The Extensions of Man* (New York, McGraw-Hill, 1964).