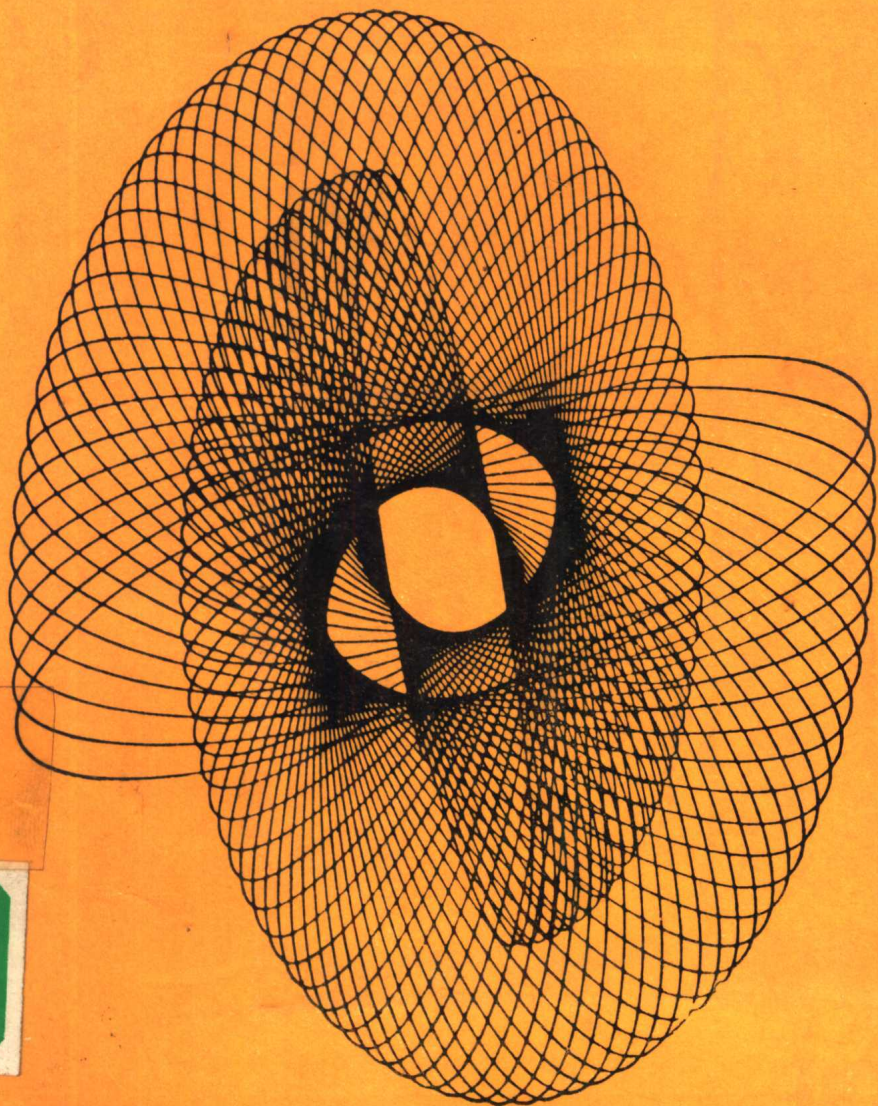


计算机情报检索系统导论

陈光祚 主编

陈光祚 孙 凌 谢新洲 田鹤卿 编著



书目文献出版社

计算机情报检索系统导论

陈光祚 主编

陈光祚 孙 凌 编著
谢新洲 田鹤卿

书目文献出版社

(京)新登字189号:

计算机情报检索系统导论

陈光祚 主编

书目文献出版社

(北京文津街七号)

涿州市新华印刷厂排版印刷

书目文献出版社发行

新华书店经销

850×1168毫米 32开本 13.125印张 332千字

1993年7月北京第1版

1993年7月北京第1次印刷

印数: 1—5000册

ISBN 7-5013-1055-6

G·284 定价: 12.00元

前 言

从1980年以来,我在担任“科技文献检索”课(主要是手工检索)教学的同时,一直在教授“计算机情报检索系统导论”课程。80年代初,我翻译了美国F.W.兰开斯特《情报检索系统——特性、试验与评价》(1980年武汉大学铅印,1984年与王知津、王津生合作在书目文献出版社出版),以此作为这门课程的教材。在教学过程中,深感有必要自己编写一部既反映计算机情报检索学科最新成就、又切合我国国情和需要的教材。1988年初,我主编了《计算机情报检索系统导论》教材,40余万字,在武汉大学作为内部教材出版。一些兄弟院校的图书情报学专业曾采用这个本子作为教材。此后,我们在积累教学经验的基础上,进行了修改,又历经三年多时间,现交书目文献出版社出版。

计算机情报检索是一门发展迅速的学科,各种新的数据库、检索系统、检索方法层出不穷。为了跟上学科的发展,我们力求在本书中反映这门学科的新面貌。另一方面,本书又力求符合循序渐进的教学原则,注意内容安排的基本性、系统性和逻辑性。

计算机情报检索大体可分为两个大的知识领域,一是检索系统的建立与维护,以检索算法和检索软件为中心,需要涉及数学、计算机科学的背景知识;二是检索系统的使用,它以检索语言和检索策略为中心,需要依靠数据库、词表、标引、检索服务、用户需求分析及检索效果的评价等知识。我们认为,作为图书情报专业的教学和研究工作者,有必要将两者有机地结合起来,以体现计算机情报检索学科的全貌。本书正是本着这个认识而编写的。

由于这门学科发展十分迅速,加之我们水平有限,本书的缺

点乃至错误在所难免，敬请读者批评指正。

本书的编写分工如下：

陈光祚：第一章，第二章（除第四节外），第三章，第四章，第五章，第六章的第一至第二节，第七章，第八章，第九章，第十一章，第十二章，第十三章，第十六章。

孙凌：第六章的第三至第五节，第十七章，第十八章。

谢新洲：第十章，第十四章，第十五章。

田鹤卿：第二章的第四节。

在本书稿最后整理定稿时，谢新洲同志给予了大力协助，曾民族、徐如镜两位研究员也对此书的编著给以支持和帮助，在此表示感谢。

陈光祚

1992年9月于武汉大学

内 容 提 要

本书全面论述了计算机情报检索系统的建立、维护与使用,并介绍了相关的数学、计算机科学和情报科学方面的背景知识,概述了国内外计算机情报检索系统的发展历史与状况,以及计算机情报检索工业的现状。全书共分6篇18章,主要内容有:计算机情报检索系统的构成,计算机情报检索工业的发展,各种计算机情报检索系统的构成、使用、分析、设计及有关技术问题,是一部全面介绍情报检索理论和方法的教科书。

本书可作为高校图书情报及有关专业的教材。广大科技人员用以学习情报检索知识和技能也很实用。

目 次

第一篇 计算机情报检索系统及其构成	(1)
第一章 情报检索系统概述	(3)
第一节 情报检索系统的意义.....	(3)
第二节 计算机情报检索发展简史.....	(9)
第三节 计算机情报检索系统的构成及其子系统.....	(15)
第四节 计算机情报检索的基本机制.....	(22)
第五节 情报检索系统与其它相关系统的关系.....	(28)
第二章 机读数据库	(32)
第一节 机读数据库的定义与分类.....	(32)
第二节 机读数据库的发展.....	(36)
第三节 机读数据库的生产过程和环节.....	(43)
第四节 文献磁带的记录格式.....	(45)
第五节 数据库的二次开发与微数据库的建设.....	(61)
第六节 数据库的性能指标和评价准则.....	(64)
第七节 我国书目数据库建库的有关标准.....	(67)
第三章 情报检索软件	(70)
第一节 情报检索软件的重要性及其评价的出发点.....	(70)
第二节 情报存贮与检索软件的评价及选择准则.....	(72)
第三节 CDS/ISIS软件系统.....	(84)
第二篇 计算机情报检索工业和检索服务	(99)
第四章 计算机情报检索工业	(101)
第一节 概述.....	(101)

第二节	情报数据库生产业·····	(103)
第三节	联机情报检索系统经营业·····	(105)
第四节	几点基本认识·····	(113)
第五章	联机情报检索系统·····	(117)
第一节	联机情报检索的技术设备·····	(117)
第二节	联机情报检索的文档结构·····	(124)
第三节	联机情报检索步骤·····	(137)
第四节	联机情报系统的检索手段·····	(159)
第五节	检索系统的选择·····	(174)
第六章	用户检索需求和检索策略·····	(183)
第一节	情报用户的需求类型与需求状态·····	(183)
第二节	检索提问的表达·····	(881)
第三节	检索策略的构造·····	(192)
第四节	SDI和联机检索的反馈·····	(202)
第五节	检索策略的调节方法·····	(207)
第三篇	非书目数据库检索系统·····	(215)
第七章	数值数据库及其服务·····	(217)
第一节	数值数据库的迅速增长·····	(217)
第二节	数值数据库的结构与编制特点·····	(218)
第三节	数值数据库情报服务的发展道路·····	(221)
第四节	数值数据库检索系统的功能·····	(223)
第五节	国内外主要的数值数据库及其服务·····	(225)

第八章 全文数据库及电子出版物.....	(227)
第一节 全文数据库的范围.....	(227)
第二节 全文数据库的检索性能.....	(235)
第三节 全文数据库检索的发展趋向.....	(237)
第四节 电子版图书及其设计要求.....	(238)
第五节 电子版图书的制作技术.....	(243)
第四篇 基于新型载体的检索系统.....	(261)
第九章 光盘情报检索系统.....	(263)
第一节 CD—ROM检索系统.....	(263)
第二节 光盘数据库的软件设计和检索.....	(268)
第三节 CD—ROM光盘的网络化.....	(282)
第十章 超文本与多介质情报检索系统.....	(284)
第一节 超文本系统.....	(284)
第二节 多介质情报检索系统.....	(290)
第五篇 计算机情报检索若干技术研究.....	(301)
第十一章 用户友好的情报检索系统界面.....	(303)
第一节 用户友好的意义.....	(303)
第二节 用户友好的设计要求.....	(305)
第十二章 专家系统在情报检索中的应用.....	(310)
第一节 专家系统的意义.....	(310)
第二节 专家系统在情报检索中的体现.....	(313)
第十三章 单汉字检索模式.....	(320)
第一节 单汉字检索模式的历史发展.....	(320)
第二节 单汉字检索模式的可行性.....	(323)
第十四章 文本自动处理.....	(333)
第一节 自动标引.....	(333)

第二节	自动文摘·····	(347)
第三节	汉语文本自动处理·····	(349)
第十五章	数据通信网络·····	(356)
第一节	国际联机检索系统的数据通信·····	(357)
第二节	我国联机数据通信网络现状与动向·····	(359)
第十六章	联机情报检索教学模拟系统·····	(367)
第一节	模拟系统的发展·····	(367)
第二节	模拟系统的特色·····	(373)
 第六篇	 情报检索系统的分析与设计·····	 (377)
第十七章	情报检索系统的分析·····	(379)
第一节	开发计算机情报检索系统的阶段划分·····	(379)
第二节	系统分析概要·····	(381)
第三节	结构化系统分析的步骤·····	(386)
第四节	逻辑模型及其图表工具·····	(390)
第十八章	情报检索系统的设计·····	(393)
第一节	检索系统的总体设计·····	(393)
第二节	输入设计与输出设计·····	(397)
第三节	文档设计·····	(400)
第四节	系统设计的优化和模拟·····	(403)
 参考文献	 ·····	 (409)

第一篇 计算机情报检索 系统及其构成

第一章 情报检索系统概述

第一节 情报检索系统的意义

情报检索是指从信息集合中识别和获得所需信息的过程，同时又是作为情报科学技术一个分支的学科名称。虽然信息的查找早就萌芽于图书馆的参考工作，但“情报检索”一词开始出现于50年代。由于现代记录下来的知识急剧增长，需要给以合理的存贮，

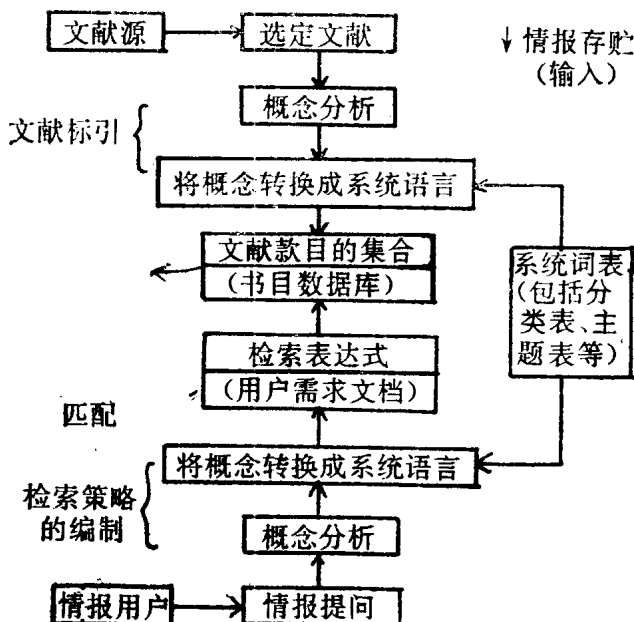


图1—1 情报检索系统流程图

人们查找的着眼点更多地指向包含在文献中的情报而不是文献本身,以及电子计算机在情报工作中的应用使得人们拥有强有力的信息处理手段,从而促使了情报检索走上计算机化的道路。

广义的情报检索包括信息的描述、加工与有序化——建立数据库即信息的存贮和从数据中查寻所需信息即信息的检索。因此广义的情报检索又称为信息的存贮与检索。而狭义的情报检索仅指后者。这里所谓的“情报”,主要是文献信息——书记录和文资本料。

为了实现情报的合理存贮与检索,必须建立情报检索系统。所谓系统,是目标、资源、设备、方法和功能的有机结合体。情报检索系统,是指为满足情报用户的需求而建立的,存贮经过加工了的文献情报信息、拥有一定的存贮、检索与传送的技术装备、提供一定存贮与检索方法及检索服务功能的一种相对独立的服务实体。

情报的存贮与检索流程,如图1—1所示:

从上图可以看到,文献情报的存贮过程,主要包括从文献源中选择一定范围和水平的文献(由此产生检索工具的学科覆盖面、文献摘贮率等指标);对被选定的文献进行内容主题的概念组面分析,并将每个概念组面用系统词表中的词(包括分类号)加以标引(由此产生检索工具的标引深度、标引专指度及一致性等指标)。对被标引的文献款目进行系列化,从而形成有序的、可供检索的书目数据库。而检索则是存贮的逆过程,就是情报用户将自己的需求加以明确化,形成检索提问,并且必须对提问进行概念组面分析,将每个组面用系统词表中的词加以表达,从而构成包含有检索词以及各检索词之间逻辑关系的检索策略,有次序地在书目数据库中查寻同检索词相一致的标引词(标识)。这里,查寻的过程实际上是匹配的过程,检索词同标引词两者一致,就算找到了符合要求的文献,否则就不能算“命中”。匹配

的过程包含了对书目数据库的扫描。

也就是说,任何一个情报检索系统,①都有明确的目标,即服务对象、专业范围及用途;②其次必须搜集、加工、存贮一定数量的文献信息,这是情报检索系统不可缺少的情报资源;第三,③必须拥有实现存贮与检索的技术装备与技术手段,即存贮记录文献信息的载体和用以进行文献标识同检索提问相对匹配的机具以及信息的传送设备;第四,④必须提供一定的存贮与检索方法和措施,如保证达到一定的查全率和查准率,从而实现情报服务的功能。总之,目标、资源、技术装备、方法及功能是构成一个情报检索系统的要素。

在这里,技术装备(即信息载体与检索机具)和匹配手段具有相当重要的地位。技术装备的每一次革新,都有力地推动着检索方式的改变,赋予文献情报检索系统以新的面貌。在某种意义上来说,文献情报检索的发展史,是以其所采用的技术设备与手段来划分阶段的。

大体来说,本世纪40年代之前:先组式索引,完全手工式的检索,检索工具是书本式、卡片式的纸上印字的载体。

50年代:穿孔卡片系统;缩微胶卷检索系统。

60年代:通过磁带的、以脱机批处理方式操作的、以计算机为基础的系统。

70年代:计算机联机检索系统。

80年代:开始向多机联网检索系统发展。

以上的划分是粗线条的,不是很精确的,而且这种划分阶段的依据,并不是指各种技术的萌芽时间,而是它们在社会生活中实际存在并产生服务效果的时间。

①用来进行扫描和匹配最原始的方式,就是手工检索。文献款目记录在纸这种载体上,文献的标识是供人阅读的文字或数码,检索策略由脑子记忆,扫描和匹配是通过人们的手翻、眼

看，脑子作出判断而进行的。这种匹配和扫描的过程，效率低、速度慢，当检索者注意力不集中时，有时对应命中的文献款目“视而不见”地漏检。但是，手工检索有一个可贵的优点，就是人们边查找、边浏览、边思考，可以随时得到新的启发，随时调整检索策略。检索策略的执行——扫描匹配——检索初步结果的评价与反馈——调整检索策略，这几个环节是有机地联系在一起的。在整个扫描匹配过程中，人的意识始终存在，并且占主导地位。

④重叠比孔卡的检索方式，虽然也是把文献标识以供人们阅读的文字或数码的形式，记录在卡片（当然也是纸上），检索策略也可以是记忆在脑子里，但是它有一个重要的不同点，就是用孔位来表示文献的号码（地址）。进行组配检索时，由重叠的孔位（透光的孔位）来表示命中文献。这种检索方式，使扫描与匹配的 efficiency 有所提高。然而，检索过程中的反馈有所减少。用孔位表示文献号码，已初步包含了代码化的意义。

边缘穿孔卡检索方式，在扫描与匹配的技术上又向前跨进一步。虽然文献款目也是以供人们阅读的文字或数码的形式记录在卡片上，但是文献标识却变成了一系列不同排列、组合的轧孔。另一方面，人们的检索策略也必须转换成相应的一系列轧孔位置，并且用穿针来识别命中的文献款目。这种检索方式的扫描与匹配，穿针这个“机械”的动作代替了人们的手翻、眼看、脑子作出判断的过程。人们不需要也不参与对每篇文献款目及其标识的审视。检索策略一旦拟订下来，扫描与匹配的过程可以是同人的意识相分离，成为一种由机具来操作的过程。与此相适应的是，检索结果的反馈与检索策略的修改变得更为迟钝。

⑤缩微胶卷（片）的检索方式，则是用不同排列组合的粗细线条或不同排列组合的黑白点来表示文献标识。另一方面，也用同样的方式表示检索策略，将其存贮在扫描镜头中。扫描与匹配的

过程，是由扫描镜头对准胶片上表示代码的一定部位、让胶卷迅速在镜头前通过而进行的。当代表检索策略的代码图案同代表文献标识的代码图案相一致的时候，就产生光电效应，从而驱动有关机具将命中胶片上的画面显示出来，或复印下来。这种方式使扫描与匹配的速度大大提高，但是在扫描匹配过程中，人不可能参予，因而是同人的意识更加分离的。

从上面几种检索方式的发展来看，已经显示出如下三个倾向：

1. 文献标识的代码化，即由供人们阅读的文字或数码，逐步变成穿针能够“识别”的孔位，或能够产生光电效应即扫描镜头能够“阅读”的光学图案。

2. 扫描匹配的过程逐步变成由检索机具独立完成的过程，也就是说，扫描匹配日益成为一种由机具执行的、事先加以规定的作业，逐渐同人的参予相分离。随着扫描匹配速度的加快，人不可能对每条文献款目进行浏览审视。

3. 检索策略相对固定化、形式化及可运算化。也就是说，在手工检索中，检索策略是记忆在检索者的脑子里的，可以边检索边得到启发，策略的修改可以随机应变、不断调整。而在比孔、穿孔、光电等检索方式中，检索策略必须事先确定下来，并用孔位、光学图案的形式加以表现。也就是说趋于形式化。检索策略的修改不很灵活，逐步趋于固定化。当然，当扫描匹配工作告一段落、对命中文献进行相关性评价之后，也可修改检索策略。但是这不是在扫描匹配过程中的随时调整。

以上三个发展趋向，在计算机情报检索中进一步得到加强。文献标识的代码化进一步变成整个文献款目的机读化，产生了计算机可以阅读的文献记录和机读数据库；扫描匹配过程的相对独立化，导致由计算机根据程序进行的高速准确的运算作业，人不可能参予，也不必参予，检索者可以把计算机内部进行的扫描匹配过程甚至整个计算机情报检索系统看作是一个“黑箱”，在计