



0031632

自动化情报和联机系
统培训班教材之一

情报技术现状及其应用

科学技术文献出版社

情报技术现状及其应用

自动化情报和联机系统培训班教材之一

(限国内发行)

编 著 者：中国科学技术情报研究所

出 版 者：科学技术文献出版社

印 刷 者：中国科学技术情报研究所印刷厂

新华书店北京发行所发行 各地新华书店经售

※

开本：850×1168¹/₃₂ 印张：3.25 字数：85千字

1980年6月北京第一版第一次印刷

印数：1—16,300册

科技新书目：158—19

统一书号：17176·237 定价：0.38元



出版说明

中国科学技术情报研究所、联合国教科文组织和法国科技情报局1979年9月3日至9月28日在北京联合举办“自动化情报和联机系统培训班”。美国佐治亚工学院情报和计算机科学系教授斯拉麦卡应邀来培训班讲授第一周部分课程。斯拉麦卡教授授课的总题目是“情报技术现状及其应用”，历时五天，共18学时。他准备的课程大纲包括“情报技术现状”、“情报技术的应用现状”、“情报技术在科技情报工作中的应用”三大部分。讲课中应中国学员的要求，增加了“国家科技情报系统的设计”部分。

斯拉麦卡教授阐述了情报行业的社会使命和人类加强利用知识的方法，介绍了情报技术的历史、现状和发展趋势，考察了情报技术在各种类型情报系统中的应用，尤其着重从国家的角度探讨了新技术对科技情报工作各个环节已经产生和正在产生的影响，并且介绍了情报科学的教育和人员培训的有关问题。斯拉麦卡教授还针对发展中国家的特点，对利用国际情报资源和建立全国范围的联机情报检索系统的方法步骤以及可以采取的途径，提出了自己的看法和建议。斯拉麦卡教授讲课，内容全面，材料丰富，对我国与科技情报工作有关的各级领导、电子计算机情报检索系统的设计和使用人员、图书馆和情报部门的广大干部以及高等院校图书馆学系或情报专业的师生，均有一定的参考价值。特别是其中的一些观点，值得我们研究和深思。

斯拉麦卡教授的讲课，由王晓初同志担任课堂口译。本教材根据录音进行整理，基本上保持讲课的全貌，个别部分作了调整并且补充了斯拉麦卡教授提供的某些材料。参加本教材整理工作的同志有陈光祚、张晰、陈伟力、黄银成、蔡绍莹、钱华林、庄建南、黄国俊、李占柱、徐如镜和曾民族，由曾民族、陈伟力和徐如镜同志负责编辑和审定。

前 言

这次培训班的主题是联机情报检索系统和自动化情报。当我被邀请到这里来授课时，我建议课程应包括整个情报科学和情报技术的全面介绍。理由是我们属于一个较大的领域，叫做情报部份 (Information Sector)。情报部份中的一个小的部份，叫做科技情报系统；科技情报系统中的一个小的部份，叫做情报检索系统；情报检索系统中的一个小的部份，叫做联机情报检索系统。联机情报检索系统是这个培训班的主题。但是根据我们的经验，许多联机系统的设计者对他们设计的系统所处的领域和环境缺乏全面了解；结果，他们设计的系统不能与别的其它情报系统兼容。因此，设计和使用联机情报检索系统，需要对整个情报部份有所了解。为此，我想在这个培训班今后三周集中探讨联机情报检索系统的设计和使用之前，向你们介绍一下整个领域的某些背景知识。

我想主要介绍三个问题。

第一个问题，介绍情报技术的现状。所谓情报技术，我指的是处理情报的整个技术领域。我可能谈到某些当前在科技情报系统中还没有使用的技术，但是这些技术很可能在不久的将来会应用到科技情报系统中来。我也可能谈到我个人的看法，而且我知道所讨论的技术当前不一定在全球范围内到处都是现实的。我介绍这些技术的目的，不是为了推销这些技术，也不是要别人相信他需要这些技术，我的目的是让你们全面了解当前世界上已经出现了些什么技术。你们设计系统时如何选择这些技术，需要根据你们自己的需要而定。但在做出选择之前必须对当前整个情报技术现状有全面的了解。

第二个问题，我想介绍情报技术在各种类型情报系统中的应用情况。我要谈到的不完全是局限于科技情报系统，而要涉及社会领域

各个方面正在设计或使用的情报系统。从发展趋势看，科技情报系统正在逐渐地与别的情报系统融合。我要提醒你们注意别的情报系统的存在，当融合的那一天到来的时候，你们才不致于感到突然。

第三个问题，我想从国家的角度来考察新技术在科技情报系统中的应用。

总之，你们应该全面了解整个情报技术领域及其应用情况。我希望你们结业之后，如果有人同你讨论科技情报检索时，你们能站在更广阔的视野上发表你们的意见。

目 录

前 言	(I)
-----	-------

第一章 情报技术现状

一、概论	(1)
二、存贮技术	(4)
三、输入输出技术	(8)
四、处理技术	(16)
五、通信技术	(22)
六、选择情报技术应注意的问题	(25)

第二章 情报技术的应用

一、情报行业的社会使命	(30)
二、提高人类产生新知识和解决问题的效率	(33)
三、扩大人类利用现有知识的能力	(39)

第三章 情报技术在科技情报工作中的应用

一、概说	(45)
二、情报资源的管理	(47)
三、面向用户的情报服务	(54)
四、加强情报的传播	(64)

第四章 建立国家情报系统的若干问题

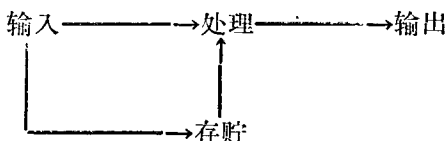
一、发展中国家和工业化国家的国家情报系统的区别	(67)
二、设计国家情报系统的一般步骤	(68)
三、发展中国家建立国家范围的联机情报系统的一些 途径	(72)
四、对国家范围联机情报系统设计的几点意见	(74)
(附) SDC联机情报检索系统的使用情况调查	(75)
五、关于情报科学的教育	(92)

第一章 情报技术现状

一、概 论

1. 情报系统的基本模型

情报系统最简单的模型由以下四个部份组成。



情报首先输入存贮区，经过加工处理，然后按一定的形式输出。情报系统的这一概念是非常广义的，不仅计算机是情报处理系统，人脑也可以看成是情报处理系统；不仅科学技术情报系统，而且政府、企业、医院的情报系统都包括在内。以下，我将按照情报系统的这四个功能来考察情报技术。

2. 名词定义

作为讨论问题的共同基础，首先我要对几个术语给出定义：

- 符号 (Sign) 通信是通过符号进行的，符号代表某种概念。
- 信息 (Message) 符号的集构成信息。
- 通信 (Communication) 通信是有意义的信息的一个或多个参数的交换。通信的三种形式：情报、教学和诱导，都是由信息而产生的交换。
- 数据 (Data) 表示事物性质的符号，叫做数据。
- 情报 (Information) 有用的数据或被认为有用的数据，叫做情报。对情报的衡量一是数量，一是价值。人们通常根据以下各点情报的属性来衡量情报的价值：特定性、识别性、获得性、可靠性、有用性、时间性、全面性、相关性。

3. 情报的表现

可用于通信的符号或符号系统有很多种。最原始的，例如动物摇动尾巴，也是一种通信符号。但是复杂的通信需要能够表现复杂概念的符号系统。这种符号系统必须具备两个性质：第一，能够表达复杂概念；第二，能够被接收的一方所理解，换句话说，通信的双方必须采用同一符号表达同一概念。几千年来，对复杂通信的需要，导致了语言的发展。有些语言至今仍采用象形文字，但绝大多数语言采用字母。字母是一种代码，代码是规定了意义的符号。我们有了符号系统，又有了使用这个符号系统的规则，就有了语言。在自然语言中，规则便是语法。有两种语言，自然语言和人工语言。人工语言是为了通过机器进行通信而设计的语言。

机器中的符号通常采用电的或电子脉冲的信号来表示。根据信号特点的不同，我们将情报机器区分为模拟式的或数字式的两类。模拟信号是连续的电信号；数字信号是离散的、单个的电信号。所以，模拟式机器又称为连续型机器，数字式机器又称为离散型机器。

数字式机器只能识别两种状态，即有电和无电。换句话说，数字式机器只能使用只有两个符号的语言，这两个符号通常用“0”和“1”表示。我们为这种机器设计了一种只有两个符号的语言，叫做二进制代码。二进制代码用的符号是0和1。要用机器处理自然语言（例如英语），就要设法把自然语言翻译成只有两个符号的语言。也就是说，自然语言的每一个字母必须使用这两个符号来表达。我们的作法是：首先对字母进行编号，例如， $A = 1, B = 2, C = 3 \dots Z = 26, a = 27, b = 28, c = 29, \dots z = 52$ 。然后，只要把这十进制的数变成二进制的数，就成了机器处理自然语言的代码系统。

4. 数字器件的发展

数字计算机对二进制码的运算的功能，基本上可分为存贮、代数运算和逻辑运算三种。代数运算实际上只是一种加法运算，因为减法、乘法和除法都是通过加法运算来实现的。对情报处理来说，更为

重要的是逻辑运算，而逻辑运算基本上就是布尔代数运算。

数字计算机是如何处理情报的呢？

数字计算机是以逻辑门来表示对情报的描述。逻辑门代表最基本的逻辑功能。逻辑门由电子元件物理地制成的数字器件来实现。这种电子元件也叫开关元件，基本作用相当于一个开关，能表示“开”、“关”两种状态。把几个这样的电子元件组合在一起，就相当于一个晶体管。为了执行某种功能，把若干晶体管按要求相互连接在一起，叫做一个电子电路。

十五年前，数字器件的技术还很年轻，晶体管都是一个一个分立的。六十年代初期，人们把晶体管及其相互之间的连线同时制作在一块半导体硅片上，构成一个整体电路，称为集成电路。七十年代以来，人们又学会把许多电路组装在一块基片上制造出了大规模集成电路。今天，大规模集成电路已经发展到实际上可把一台计算机所有的电子电路制作在一块4平方厘米的基片上。根据采用的技术，我们把数字器件划分为几代。五十年代采用真空管，称为第一代。六十年代采用晶体管，称为第二代。六十年代末七十年代初采用集成电路，称为第三代。最引人注目的技术发展是数字器件制造费用的下降。制造集成电路的费用，过去几年内每年下降25%，十年内降价100倍。存贮器的价格在十年内每年下降38%。打个比方，如果50年代初，买一辆高级小汽车需要大约人民币16万元，如它的价格也像存贮器那样降低的话，那么今天只要4角钱就可以买到了。

这些是过去10年内发生的事情。现在的问题是今后价格还会不会继续下降。种种迹象表明，今后10年内的价格将会以同样的速度继续下降。完全有理由期待，今后10年内数字器件的功能将提高大约50倍，而成本也将下降大约50倍。但是，这类技术在今后10年将达到极限。我们相信在这10年内，基础研究将会有新的突破，以使技术得以继续发展下去。现在，基础研究已经取得若干重要的进展，人们相信，再过十年之后，技术将可以照样继续发展下去。

二、存贮技术

存贮可分生物存贮和非生物存贮两类。非生物存贮又可分电子存贮和非电子存贮两类。非电子存贮的代表是纸和胶片。下面介绍存贮技术某些新的发展。

1. 缩微胶片

从技术角度看，缩微胶片遇到了竞争。最大的对手是电子存贮。缩微胶片有很多缺点，其中最主要的是，记录是不能改变的，不能抹除和更新。目前这一缺点暂时还找不到改进的办法。另外的缺点还有：它除了邮寄外不能传输；胶片记录的信息不能机器处理。这两个缺点，今天已经有了补救办法。缩微胶片的信息已可以进行传输，这个系统叫做电话传真（Telefacsimile）又称为缩微传真。过去传真只能传输印刷品，而今天传真已经能够传输缩微胶片。缩微胶片上的信息不能处理的缺点也在逐渐得到改进。这种技术叫做数字扫描和光学字符识别。这种技术能够把缩微胶片上的信息经过电传送往远地，在另一端变成印刷品或仍然输出缩微胶片。迄今为止，缩微胶片和缩微胶卷没有普遍采用的一个重要原因是通信上的困难。今天已经能够高速传输，这是一个突破。

2. 录像盘（Video disc）

最近，市场上出现一种电子存贮技术，叫做录像盘，形状像唱片，但能以数字形式记录照片、图像等模拟信息。录像盘是用激光记录信息，记录密度很高。一个录像盘可记录 56×10^9 位的信息，而且价格很便宜。录制第一张录像盘需用1,000美元，但在此后复制第二张时就只需要2—3美元。存贮密度高和价格便宜加在一起，使它成为存贮科技情报和其他情报的极为有希望的介质。 56×10^9 位的信息相当于15,000篇科技文献的全文，再加上所有有关的索引、主题词典和计算机程序的总容量。而15,000篇文献相当于世界上一个学科一年所发表的文献量。一台小计算机加上显示器和键盘带一个录像盘，

价格大约是5,000美元,而这就是一个完整的情报检索系统。这样的系统为什么目前还没有实用?原因在于要把15,000篇文献打字输入计算机,这笔费用很大。但是事情已经开始有了希望。光学符号识别设备等可以很快地把全文输入计算机,而且目前市场上已有了产品,只是还没有同录像盘结合起来。一旦结合起来,至少理论上可以设想,只需花5,000美元,就可以在自己的办公室建立自己的情报检索系统。这种技术将来很可能成为科技情报系统的主体,而且还可能完全改变目前情报传播的方法。

3. 全息照像

全息照像是另一种很有希望的存贮技术。全息照像采用胶片介质,利用激光相干光源记录三维信息,在胶片上形成全息图。一张和缩微平片同样大小的全息片,可记录20万个全息图,也就是可存贮20万页全文文献。市场上已出现全息照相的记录和阅读设备,价格是300,000美元。全息照相阅读器则在2,500—25,000美元之间。2,500美元的阅读器是一般的阅读器,25,000美元的则是可以连接多个阅读器的整套设备。全息照相技术是一种非常有希望的存贮技术,将来完全有可能代替其他某些存贮介质。

4. 各种存贮介质的比较

- 纸 纸是几百年来存贮信息的最普通的介质,从用户角度看是最好的,但从效率上看则是最坏的。

- 穿孔卡片 一张穿孔卡片记录80字符的信息。通常一个穿孔卡片盒存放2,000张卡片。据估计,美国国会图书馆馆藏的总信息量为 10^{12} 位,如果采用穿孔卡片存放,则穿孔卡片盒直线排列将达198公里长。

- 磁带 一盘磁带的存贮量相当于667盒穿孔卡片。美国国会图书馆的馆藏如采用磁带存放,每年费用估计为100万美元。总的来说,采用数字形式存贮信息是经济的。磁带由于存贮密度高,价格便宜,仍将继续使用下去,但将主要用于存放档案性的保存资料。

- 缩微平片 一张 4×6 吋²的缩微平片可存贮60页以上的

文献，超微平片已可存贮2,000页以上。缩微平片的优点是可以存贮图像、照片。一盒平片的存贮量，相当于1560盒穿孔卡片，上述美国国会图书馆 198公里长的成盒穿孔卡片如用缩微平片代替，则缩微平片盒排列的长度仅为237米。

• 计算机存贮 从图1.1可以看出，计算机存贮的每位存贮费用在过去十年内大大下降。随机存贮费用（图1.2），过去十年内每年下降28%。计算机存贮费用下降是非常有意义的，因此将越来越多地采用新型的电子存贮。

• 联机存贮 联机存贮就是任何时间都与主机连接。目前磁盘的存贮量为10亿字节，而且一个主机可连接多个磁盘。科技文献每页按600个词（约等于3000个字符）每本书按333页算，则磁盘的联机存贮费用为 5 美元/百万字节/月，等于60美元/一本书/年。磁盘

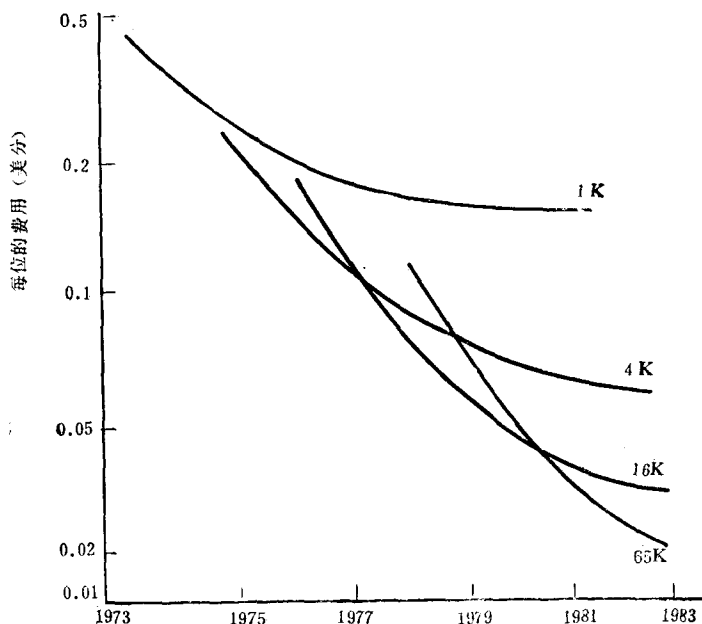


图1.1 计算机存贮器每位的存贮费用

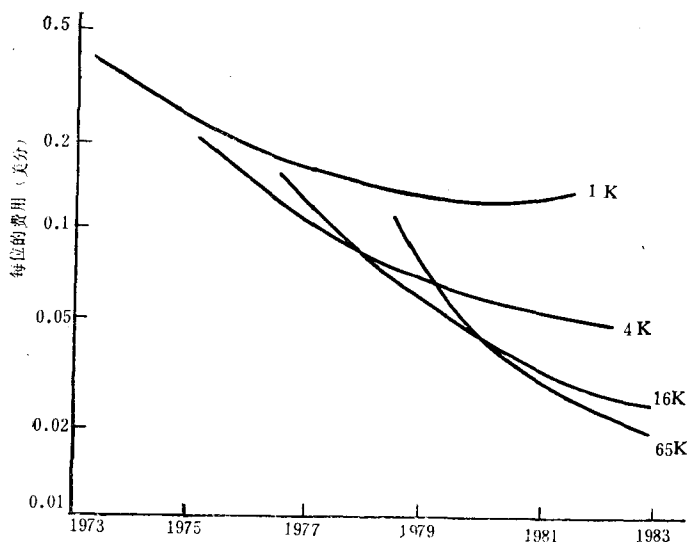


图1.2 随机存取器的存贮费用

存取速度很快，容量也大，但费用很高。因此，我们只能把随时需要利用的情报存放在磁盘上。

• 特大容量档案存贮器 (Very Large Archival Storage)

不经常使用的情报可存放在特大容量档案存贮器上。这种存贮器采用蜂窝状的磁带筒，需要时通过与计算机联机的机械方式把磁带调出，将其信息调入磁盘。一个完整的特大容量档案存贮器可存放5,000亿字节；存贮1百万字节每月只需0.10美元，也就是说，一本书的存放费用每年为1.2美元。

表1.1

存 贮 介 质	单位存贮容量 (字节)	百万字节的月存 贮费用 (美元)	一本书 (333页)的 年存贮费用(美元)
磁 盘	10^9	5.00	60.00
特大容量档案存贮器	500×10^9	0.10	1.20
录 像 盘	7×10^9	0.008	0.01

· 录像盘 录像盘存贮一本书的存贮费用为每年0.01美元。
几种存贮介质的性能和费用的比较见表1.1;

三、输入输出技术

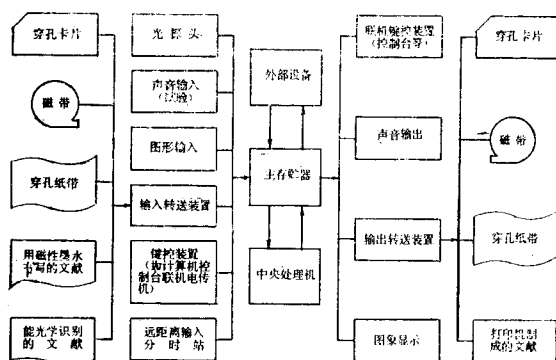


图1.3 输入输出介质和硬件

1. 输入技术

输入技术大致分为图象 (Visual) 的输入和声音 (Audio) 的输入两个方面。图象的输入又分为字母数字输入和图形的输入两类，而图形的输入包括线条图输入和活动图象输入 (如电视、电影等)。

过去，对字母数字形式的输入技术作了很多研究工作，取得了很大成绩，技术已比较完善。现在主要的研究方向是声音与图象相结合的输入技术。今天，我不打算讲打字机、穿孔机这样的传统输入设备和技术，而将着重介绍光学字符识别 (OCR)、数字扫描 (Digital Scanning) 和声音输入 (Voice input) 的技术。(参见图1.3)

1.1 图象输入

1.1.1 光学字符识别

光学字符识别具有识别符号即把印刷字符转换成数字代码的能

力。其工作原理是将一个输入的字符信息同存贮在机器内的所有字符的信息作比较,如相同则算识别。但字母有大写、小写;字号也有区别。如果要把这些各种各样的字符信息统统存贮在机器里,这样机器就必然非常昂贵。因此,人们只好选择某种打字机上的字符,作为存贮字符,只有这种型号的字符才能被机器所识别。目前,光学字符识别装置要求某种特定的打字机或特定的字盘,原因就在这里。但是,大量已存在的情报——报纸杂志上的文章以及各种打印件,其字符型号是各种各样的,因此,应使各种原始文献中的字符,不需经过特定打字机的再加工,而直接变成数字化信息输入机器。这是向光学字符识别技术提出的研究课题。近来,这方面的研究有所突破。Kurzweil发明了一种读出机 (reading machine),它能够识别几乎所有各种印刷的标准字符。将书面字符输入后,然后输出声音。这种机器可以为盲人读文献。在这种机器里,扫描是通过一种五百单元的电荷耦合器件 (CCD) 在两毫秒的取样时间里完成的。一种“分解分析”模块程序把字符实体分离出来,然后识别这个字符并通过一种小型不变性的取样方法将其转换成适当的数字代码。再通过一种“单义”模块程序完成最后的选择:按规则和用同一规则的例外词典作为辅助,来决定字符顺序所形成的语言;并用句法分析器加上全句的韵律,以便发出的声音不会出现单音。控制信号驱动一组电子合成器来产生讲话的声音。这种机器可为盲人阅读各种印刷材料,产生人们能够接受的讲话,机器的价格约 85,000 美元。这种 Kurzweil 机器的发明人所发明的一种能使光学字符识别器识别各种字符、并使之变成声音的算法,是非常简单的,可存贮在一片微型存贮器上。我们可以把这种存贮器用到别的机器上去,例如可供自动标引中识别字符之用。

1.1.2 数字扫描

与光学字符识别相反,数字扫描不是识别整个的字符,它以一束光对整页文稿进行高速扫描。扫描器只能识别黑的颜色和白的颜色,使之变成 1 和 0 的信息。一长串的数字不代表别的什么意思,

而仅仅表明载体什么地方是白的，什么地方是黑的。数字扫描能将模拟信息转换成数字信息，并且密度相当高，它可以不考虑白的部分，从而对信息进行压缩。这种技术用于传递字符和图象信息比较合适。

现在最新的发展是把光学字符识别和数字扫描结合在一个机器里。用光学字符识别技术识别字符，再用数字扫描中的压缩信息的算法将信息进行压缩。这样，使电子建档 (Electronic filing) 变成可能。在任何一个办公室里，过去是采用穿孔机和电传打字机等把信息存入文档的。现在可以使用光学字符识别和数字扫描技术进行输入，后一种技术更为经济有效。光学字符识别和数字扫描相结合的机器，在一年内将由Burroughs公司制造出来。别的公司也宣布了制造电子建档设备的计划。这种设备是很昂贵的，原因是输入量非常大。要预见这种技术对科技情报事业的影响，目前还很困难。因为今后新的情报在其产生的时候，就已经是电子化的信息了。如现在出版的杂志，已经出现电子形式的版本。

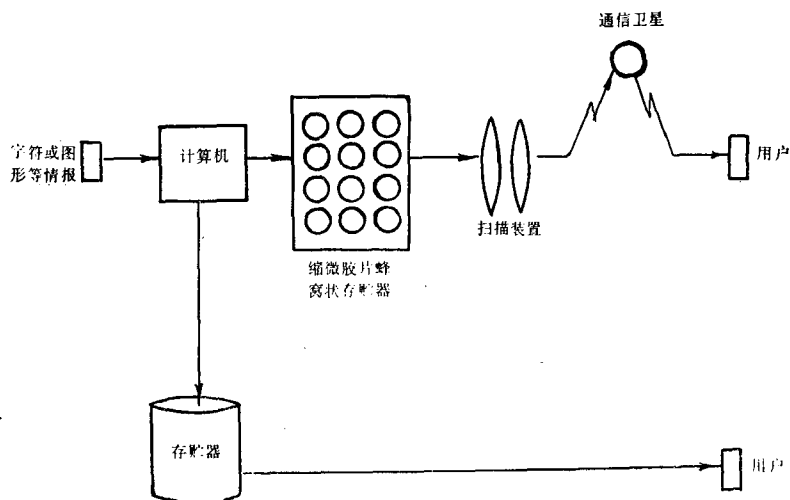


图1.4 没有纸的系统

对缩微胶片的数字扫描技术，使得缩微品可以用电子形式进行传输，而不必象以前那样必须通过邮寄的办法来传递。

1.1.3 没有纸的系统

现在正在向没有纸的时代发展。

没有纸的系统可以有两种方式。一种是把书面形式的字符图象先变成缩微平片或缩微胶卷，存入蜂窝状存储器，然后用扫描装置变成电信号，通过卫星等传输，使用户得到情报。另一种方式是将情报通过打字输入直接送入存储器，通过直接传送而到达用户手里（见图1.4）。

两种方式，主要根据被传送的情报的使用频率来选择。使用频率低的情报，主要使用第一种方式。经常使用的情报，采用后一种方式，即把信息直接存入存储器进行直接传送。由于索引、文摘是需要经常使用的，作为联机情报系统来说，希望使用这后一种方式；但全文信息不希望采取这种方式存储，因为费用较高。一般来说，全文信息存储在缩微胶片上，采用脱机系统传送。还有一种办法是把全文信息存储在录象盘上。录象盘复制容易，价格便宜，并且存储量很大；通过发行录象盘，许多全文信息就不再需要远距离传送。

现在还不清楚这几种技术在今后几年里到底哪种技术会占上风，这是国家情报系统的设计者面临的一个困难。

1.2 声音输入

另一种输入技术是声音的输入。

当我们谈到语音（speech）处理时，涉及到两个方面。一是声音识别，即机器能识别声音；另一个是声音输出，即机器能说出声音，讲给我们听。目前有许多机器只具备一个方面的功能。我们希望有一种机器能把二者结合起来，即我们既可以对它说话，它也能对我们说话。

现在，较易做到的是声音的输出，而声音识别则要困难得多。这是因为人们说话时发出的声波图形（pattern）是非常复杂的。