# 情报检索自动化基础

王 永 成 编

知 识 出 版 社

# 情报检索自动化基础

王永成 编

知 识 出 版 社

1984·1·上海

情报检索自动化基础

王永成 编

知识出版社出版
（上海古北路650号）

新华书店上海发行所发行　上海海峰印刷厂印刷

开本 787×1092 毫米　1/32　印张 11.125　插页 2　字数 224,000
1984年1月第1版　1984年1月第1次印刷

书号：13214·1014　　定价：1.25元

## 内 容 提 要

本书依据各国最新自动检索资料并结合我国情报检索的现况编写而成。共分十章；对于排序技术、情报存贮、检索步骤以及联机检索的安排都作了系统的论述和分析。书中并附有不少的国内外有关情报检索的图表。

本书可作为大专院校计算机、图书、档案、情报检索及其应用软件自动化专业的教材，对于科研单位、文字翻译、医疗诊断和管理部门也有参考的价值。

本书特邀编辑：郭明达

# 前　言

电脑① 是20世纪最重大的科技成就之一。如果说，蒸汽机和电动机的发明，大大地解放和发展了人类的体力，那末，电脑的发明和发展则大大地解放与扩展了人类的脑力。近年来，电脑已广泛地应用于科学计算、事务处理、实时控制等方面，它极大地减轻了人们的脑力劳动。可以预测：电脑在人类历史上的作用必将超过蒸汽机和电动机在人类历史上所起的作用。

利用电脑进行检索，是电脑最基本的用途之一。众所周知，科学研究、医疗诊断、文字翻译、科学管理都要用到情报、资料、文献、数据等等的检索。因此，本书系统地介绍利用电脑进行自动检索的基本技术。另外，本书对数据库与排序技术也作了初步介绍。我们指望读者在读完有关章节之后，可以比较方便地研读有关专著。

本书虽以中等文化水平的读者为主要对象，但是，为了给那些有志于自动检索理论研究的读者以一定的帮助，我们在书中介绍了一些较为专深的内容，它涉及到较多的数学知识；对这些数学知识比较生疏或者对自动检索理论不感兴趣的读者，则可略去不读。尽管如此，本书并不失去阅读的连贯性，也不妨碍这样的读者在读完本书以后，能够独立地研制完整的自动检索软件。

---

① 电脑，过去国内大多译为电子计算机。但是，它的作用实际上已经远远超出了计算的范围。作者认为：将“computer”译作电脑似较电子计算机更为合适。

本书是作者在三年前为南京大学数学系情报检索自动化教研室编写的《情报检索自动化基础》讲义的基础上改编而成的，在编写过程中，曾得到许多有关单位和同志们的帮助，在此一并致谢！

王永成

1982年5月

# 目　录

# 第一章　电脑的一般知识

本书将重点介绍现代自动检索技术。所谓自动检索，就是利用电脑进行检索。因此，要掌握自动检索技术，就得对电脑有所了解。为此，我们先介绍电脑的一般知识。本章包括电脑概述、电脑硬件、软件三节。

## 第一节　电脑概述

电脑的出现决非偶然，它是人类生产实践的必然结果，是历史上计算工具发展的最新成果。早在春秋战国（公元前770年至前221年）时期，我们的祖先就用一些小棍摆成不同行列来进行计算，从而创造了“筹算法”。到唐末，我国民间就出现了算盘。公元1274年，南宋时的算盘歌诀一直传到今天。算盘的创造，是我国人民对世界科技的重要贡献之一。到1602年，法国首先制成了能进行加减法的机器，用于税收。1694年莱布尼兹改进了税收机，可做全部四则运算。1938年，贝尔电气公司发明了继电器。到1944年就出现了MARKI——自动逐次控制电算机。1945年，由于冯·诺依曼提议存贮人的指令，终于促成了世界上第一台电脑ENIAC的诞生，它计算速度比人工快30万倍。到1947年左右出现了第一台用程序控制的电脑EDVAC。从第一台电脑诞生至今不过三十多年，但它却经历了多次变革。关于每一阶段的发展特点，可列表概述如下：

| 指标 / 特点 / 代 | 主要元件 | 中　心 | 语　言 | 操作系统 | 应用范围 | 年　代 |
|---|---|---|---|---|---|---|
| 1 | 电子管 | cpu(Central Proccessing Unit) | 机器 | | 科学计算 | 45～58 |
| 2 | 晶体管 | IAS(Immediate Access Store)通道(传送信息的装置) | 程序设计语言 | OS(Operation System)——管理程序 | 科学计算<br>数据处理 | 59～64 |
| 3 | 集成电路 | IAS 总线联结 cpu，通道和I-O (Input—Output) | 会话式<br>终端文件系统 | 大型 OS (包括远程，实时，分时控制) | 普及 | 65～69 |
| 4 | 大规模集成电路，半导体存贮器 | 分布式<br>积木式<br>容错式 | | 通讯网<br>数据库 | | 70 |
| 5 | 固件，激光 | | | | | |

**电脑工业发展迅速，产品种类繁多。兹将电脑的分类概述如下：**

1. 按类型分
   - 数字电脑：以离散数表示被运算量。它解题精度高，便于存贮，使用灵活。
   - 模拟电脑：用连续变化的物理量（如电压、长度等）来表示运算量。它速度高，且能模拟物理量。
   - 混合型电脑：它结合了数字与模拟电脑二者的优点，但造价高。

2. 按功能分
   - 专用机：它比较精巧，但功能狭窄。
   - 通用机：现在大多采用通用机。

3．从采用的数制上分

- 定点机：数限在（−1,1）范围内，小数点位置固定，精度高，多用在小型机或自控专用机上，借助软件也可进行小数点位置可移动的浮点运算，但功效大大降低。
- 浮点机：即每个数皆用二部分表示：一个作为阶，一个作为尾数。

4．从结构上分

- 串行机：即每一位是逐个串行进行的，因此速度慢，已渐渐淘汰。
- 并行机：数的运算是在数的各位同时并行处理的。

5．从大小上分

- 巨型：
- 大型：
- 中型：

  划分的标准不一。巨型机已成为各国追求的对象。
- 小型：一般称存取周期2毫微秒、字长8～18字位、内存不大、结构简单、体积小、重量轻、操作容易的电脑为小型电脑。
- 微型：以微型电路或微处理器为基础的、字长4～16字位、结构简单的电脑通称为微型电脑。
- 电算器：指不能修改程序的小型台式或袖珍式电脑。

综合世界电脑工业的发展，微型与巨型机是很受重视的机型。

不管是哪一种类型的电脑，现代电脑在机内表示数大多都不是用十进制，而是采用所谓的二进制，就是逢二进一。采用二进制的理由主要有三：

1．可以证明：为表示一定范围内的数，用三进制可使机器内部相应的元件的各种不同状态的总数最少，二进制仅次之。

2．但是，从机器结构上看，使用二进制只需要构造极简单的机器元件。譬如，可用电路的接通表示“1”，断开表示“0”，

因此就可以表示二进制数。

3．二进制的运算规则非常简单，它只要：

乘法：$0\times0=0$，　$0\times1=1\times0=0$，　$1\times1=1$；

加法：$0+0=0$，　$0+1=1+0=1$，　$1+1=10$

所以在二进制中做乘法，实际上只要做移位与加法就可以了。

二进制虽然有上述优点，但是因为它往往要用很多位才能表示一个数，因之对人的书写非常不便。譬如，为了表示十进制的 47 这个数，它就要六位的二进制数 101111 来表示（犹如十进制中对 365 可写成 $365=3\times10^2+6\times10^1+5\times10^0$）。同样，对二进制数 101111 有：$10111=1\times2^5+0\times2^4+1\times2^3+1\times2^2+1\times2^1+1\times2^0=47$。以后为了避免混淆，我们常常用足码来注明该数是在什么进制下的表示。如十进制数 47，我们写成 $(47)_{10}$，它的二进制表示为 $(101111)_2$，等等。

由于三位二进制可用来表示一位八进制数：

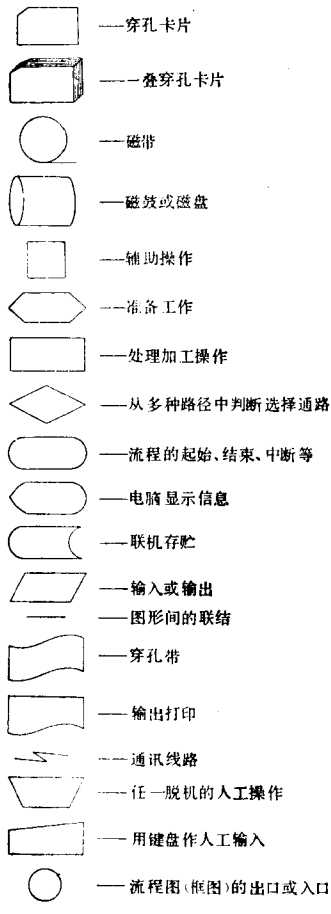
| 二进制数 | 000 | 001 | 010 | 011 | 100 | 101 | 110 | 111 |
|---|---|---|---|---|---|---|---|---|
| | ↕ | ↕ | ↕ | ↕ | ↕ | ↕ | ↕ | ↕ |
| 八进制数 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

因此，要把二进制数表示成八进制数也很方便，只要三位一转换。如 $(101111)_2=(57)_8$，反之，要把八进制化为二进制同样是方便的，这只要把一位八进制化为三位的二进制即可。例如：$(36)_8=(011110)_2$。

在实际工作中，有时人们还嫌八进制书写较长一点，因此，就利用下述的二进制与十六进制对照表将二进制数写成十六进制数（下表中A代表 10，B代表 11，C代表 12，D代表 13，E代表 14，F代表 15）：

| 二进制数 | 0000 | 0001 | 0010 | 0011 | 0100 | 0101 | 0110 | 0111 |
|---|---|---|---|---|---|---|---|---|
| | ↕ | ↕ | ↕ | ↕ | ↕ | ↕ | ↕ | ↕ |
| 十六进制数 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

| 二进制数 | 1000 | 1001 | 1010 | 1011 | 1100 | 1101 | 1110 | 1111 |
|---|---|---|---|---|---|---|---|---|
| 十六进制数 | 8 | 9 | A | B | C | D | E | F |




因此，要将二进制数化成十六进制数，也非常容易，只要四位一转换即可。如$(47)_{10}=(101111)_2=(2F)_{16}$。

上述的二进制数与八进制数及十六进制数之间的转换，在电脑中经常遇到，读者应熟练掌握。

电脑的发展，尽管已经历了若干阶段，但其工作的基本过程和原理还是大致相同的。下面，我们以一台具体的机器为背景，来解剖一下它执行一条指令的过程，顺便对一些名词作出说明。

在这之前，我们再把常见的国际通用的图形及其所代表的含义介绍如左。

现代电脑的内部构成图一般仍然与图1.1所示类似。

我们来看看电脑计算10＋8＝18的工作过程（请对照图1.1）。

假设命令电脑计算10＋8的指令已通过输入装置存放在

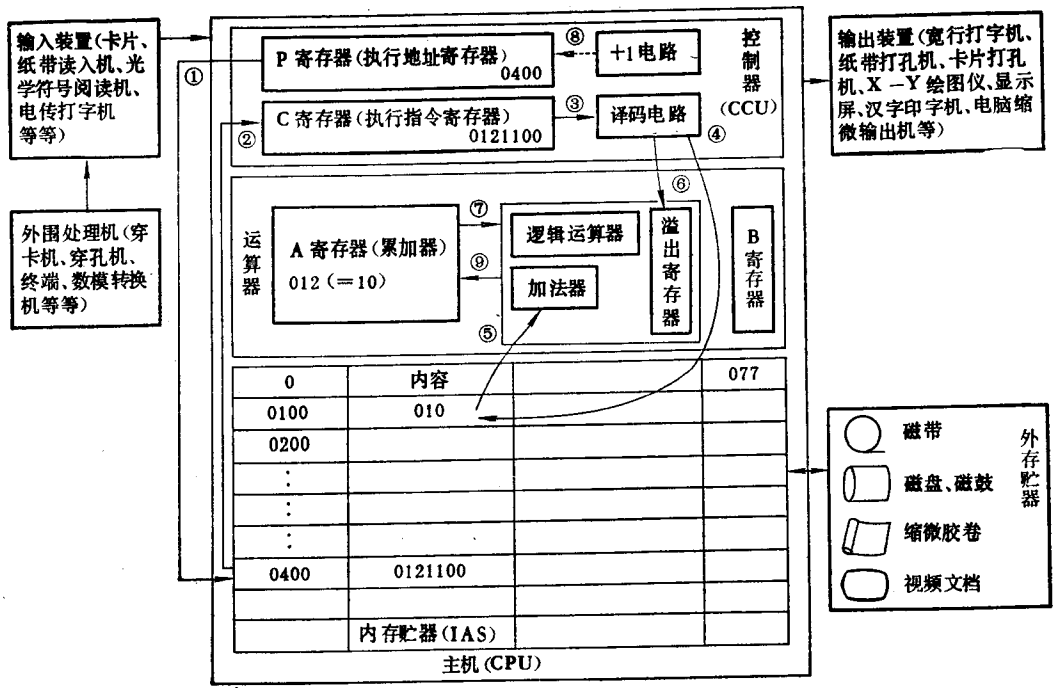

输入装置(卡片、纸带读入机、光学符号阅读机、电传打字机等等)
外围处理机(穿卡机、穿孔机、终端、数模转换机等等)
①
②
P 寄存器(执行地址寄存器)
0400
C 寄存器(执行指令寄存器)
0121100
⑧
+1电路
③
译码电路
④
控制器(CCU)
⑥
运算器
A 寄存器(累加器)
012 (=10)
⑦
⑨
⑤
逻辑运算器
加法器
溢出寄存器
B 寄存器
0
内容
077
0100
010
0200
0400
0121100
内存贮器(IAS)
主机(CPU)
输出装置(宽行打字机、纸带打孔机、卡片打孔机、X-Y绘图仪、显示屏、汉字印字机、电脑缩微输出机等)
磁带
磁盘、磁鼓
缩微胶卷
视频文档
外存贮器

图 1.1

0400 号内存单元中了，并且数字 8（其八进制表示为 010）已输入到 0100 号单元，在A寄存器中因执行上一条指令而存放着结果 10，P 寄存器中根据操作员的操作已指明电脑从 0400 号指令开始执行。

第一步——控制器根据 P 寄存器中的内容(0400)到内存相应单元去取信息。

第二步——将取出的信息（指令）送入C寄存器（指令假设写成 0121100 形式）。

第三步——将指令内容交译码电路分析。

第四步——根据分析，要到 0100 号内存单元中去取数 8[$=(010)_8$]。

第五步——将取出的数送入运算电路。

第六步——译码电路根据对指令的分析，命令运算电路准备做加法。

第七步——从A寄存器中取出上条指令的执行结果送运算电路，与 8 相加。

第八步——控制器中的“加 1 电路”使 P 寄存器内容自动加 1，即命令机器准备下一指令。

第九步——将运算结果 10 + 8 = 18( = 022) 送入A寄存器。

第一步到第二步执行完的时间称为“取指令周期”，第三到第九步所需时间为“执行指令周期”，而“操作周期”为“取指令周期”与“执行指令周期”之和。

## 第二节 硬　　件

目前生产的电脑是一种能自动高速进行大量数据处理的电子设备。

所谓“自动”，就是能存贮人赋予的动作指令，并忠实地按人的指示工作。由于它记得牢、不怕繁复、动作迅速，它在有些方面的工作能力已大大地超过了人手人脑。它主要能做五方面的基本工作：读（输入）、写（输出）、算、记、存（外存⟶内存）。“读”、“写”目前仍通过机械方式实现，速度以每字千分之一秒——毫秒为单位；“算”则以百万分之一秒——微秒为单位。现代电脑除了速度快和准确性高之外，它与人相比，仍为“双目失明”的聋哑者。即使将来电脑有了极大的发展，我们也可断定：它永远不会完全代替人脑和人手的劳动，而只不过是“人脑和人手的延伸”而已。

至于“高速”，现代电脑运算速度，有的已高达亿次/秒。电脑的运算速度，是反映电脑性能的重要指标。目前有四种算法：

1．最常见的是指电脑做加法的速度；

2．是指电脑做四则运算的平均速度；

3．比较严格一点的是所谓的吉布松（GIBSON）混合频度法计算出的平均速度，即对四则运算进行加“权”平均计算，“权”数由统计获得；

4．用存取周期表示。

可用下面两个例子来说明现代电脑的运算速度之快：

现代火箭可谓快了，但电脑对其运行轨道的计算速度却比飞行本身快得多。

如果人以百岁计算，一生只有30多亿秒。因此，一个人一生的计算劳动，不抵现代电脑几秒计算。历史上有位数学家W. Shank，他用了毕生的精力计算$\pi$的值，达到707位数。临死前，他要求家人把他一生的劳动成果铭刻在他墓碑上，引以为骄傲。但在电脑广泛使用之后，短短几秒钟内的计