计算机检索知识丛书①

计算机检索概论

陈光祚

武汉市科学技术情报研究所
武汉计算机检索及其用户协会 编印
一九八四年一月

计算机检索知识丛书①

计算机情报检索概论

陈 光 祚

武汉市科学技术情报研究所
武汉计算机检索及其用户协会
一九八四年一月

前　言

当前，文献资料数量浩如烟海。依靠手工查找的方式越来越不适应于科学工作者及时、系统、准确地获取有关科学情报的需要。在这种情况下，采用电子计算机这种现代化的手段进行文献资料的存贮与检索，是完全必要的。电子计算机情报检索的实现，使得过去需要耗费大量时间和精力的资料查找工作，变成了高速、准确和自动化的过程。

国外从五十年代末期开始，电子计算机检索登上了图书情报工作的舞台。六十年代赢得了社会的承认，正式向公众提供服务。七十年代更与电信技术相结合，联机情报检索蓬勃发展，成为社会上的“信息产业”的一部分。八十年代，电子计算机正向网络化发展，使得情报的传递方式在速度与范围方面产生了一场革命。

我国也从１９７４年开始研制计算机情报检索系统。进入八十年代，国内有十几家科技情报单位和图书馆向社会提供计算机情报检索服务。

对于科学工作者来说，了解计算机情报检索的原理与使用方法，掌握获取科学情报的新手段，是当前一项应当具备的基本功。

本书是为广大科学工作者、图书馆与情报资料工作者而写的。目的在于简单介绍电子计算机情报检索的初步知识，重点是站在检索用户的立场，说明计算机情报检索的产生、发展、功能及使用方法的梗概，并介绍了国内开展计算机情报检索服

务的情况。至于检索系统的设计、硬软件的配置及文档结构等方面，考虑到与用户关系不很直接，因而从略。总的来说，本书的重点不是论述如何建立情报检索系统的问题，而是叙述如何使用情报检索系统的问题。

由于本人水平所限，加之时间匆促，本书中可能有不少缺点乃至错误，请读者予以指正。

本书承研究生王兵、黄祥喜、覃光、张进等同志校对，在此表示感谢！

陈光祚

1983年12月于武汉大学

目　　次

一、学习计算机情报检索的现实意义……1
二、计算机情报检索出现之前检索技术的演变……3
三、计算机情报检索系统的构成与种类……8
四、机读数据库……12
五、检索系统的主要检索功能……19
六、机检的两种方式——脱机检索和联机检索……27
七、计算机情报检索服务的种类……36
八、检索策略……39
九、国外主要的联机情报检索系统……52
十、我国机检服务系统概况……53
十一、小结……56

一、学习计算机情报检索的现实意义

计算机，电信和情报三者的结合，使得情报的搜集、存贮、加工、检索和利用的流程，产生了革命性的变革。情报的传递出现了崭新的模式。同传统的情报工作方式相比，计算机检索不仅速度快、效率高，而且打破了人们获取情报方面原先所存在的地理上的障碍和时间上的延迟，极大地提高了文献情报的可获得性。今天、计算机情报检索已经在科技生活中站稳了脚跟、赢得了信誉，对科技研究与发展工作产生了越来越大的实际影响，并必将日益加速科学进步的过程。

对于图书馆和情报资料单位来说，计算机情报检索使它们获得了为情报用户服务的新手段。情报检索服务不再仅仅依靠本单位的收藏，而可以获得远比本馆藏书丰富的情报源，从而具备了在更大的深度与广度上满足情报用户需求的能力。计算机情报检索在这个方面导致了两个结果：其一、是使情报检索服务由过去的各馆单独进行的小规模作业方式变成了社会化的服务系统，产生了象DIALOG,ORBIT等商业性的、集中化的计算机情报检索服务中心；其二、是使各个图书情报单位的检索人员日益变成情报用户同计算机情报检索服务中心之间的中间人。他们受情报用户的委托，同计算机检索系统打交道，为用户拟定检索策略，协助用户评价检索结果，并把评价成果作

---

* 本文中的若干图表，取材于美国兰卡斯托的有关著作，例如《情报检索系统：性能、试验与评价》（陈光祚译）。

为对进一步修改检索策略的“反馈”。承担起用户与检索系统之间的桥梁作用的图书情报工作人员，需要掌握比过去手工检索更多的知识，他们不仅要熟悉有关词表、标引、检索工具（数据库）的品种、检索途径的选择等知识，而且要掌握同计算机检索系统打交道的有关询问语言与布尔逻辑，了解计算机检索系统的性能，充分利用系统的潜力拟定恰当的检索策略。计算机情报检索不仅没有取代图书情报单位工作人员的作用，相反，要求他们必须具备新的素质，掌握新的基本功。

计算机情报检索促使了情报学这项新生学科的出现与发展。计算机情报检索的理论与方法，是在传统的图书馆学、目录学基础上，同计算机科学相结合而产生和发展起来的。计算机情报检索是情报学赖以建立的核心与基础，并且是整个情报学中发展最为迅速、最为活跃的一个分支。情报学这一概念，本身就内含有计算机的应用和计算机化的信息处理。如果没有后者，前者（情报学）的存在是不可理解的。

就当前国内的情况来说，计算机情报检索已经不是什么“空谈”，而是已经从研究、试验阶段转入了实用的阶段。从1975年我国计算机检索事业起步以来，至1982年下半年为止，先后有近70个单位进行了这方面的试验与服务工作，所参与的人员约为500多人。专门用于情报检索的大、中、小型计算机有8台。目前已引进32种国外发行的文献磁带。邮电部、化工部、地质部、机械工业部、石油部等部门的情报所，北京文献服务处，上海科技情报研究所，以及南京大学等单位已正式开展计算机情报检索服务，拥有数千个情报用户。国家正在制订全国性的机检规划。此外，1980年开始，八部一局租用了一个香港终端，向美国DIALOG及ORBIT两个检索系统进行远距离的联机检索，已为国内各部门检回了数百个课题资料。近年来，还

在北京北方科技文献服务中心设立了上述美国两大检索系统的电传终端，以及北京电报大楼的电传终端，可在北京直接对其进行联机检索。最近，在北京又设立了 ESA（欧洲航天局）系统的终端，已可开展检索服务。汉字文献的计算机检索，也取得了可喜的进展，中国科技情报研究所同医药总局合作，建立了“中国药学文摘”的汉字数据库，并由计算机编排了“中国药学文摘”印刷版。不少图书馆已经成为情报用户同计算机检索系统之间的“中间人”。计算机情报检索的知识技能，正在逐步地向广大图书情报工作人员和科技人员进行普及。有关计算机情报检索的理论探讨和经验分析，也正在逐步发展。创办了《计算机与图书馆》杂志。中国科技情报学会计算机情报检索组每年举行一次学术讨论会。研究的课题日趋广泛，从检索软件，逐步扩大到检索策略、汉字检索和检索效果评价，以及文献的自动标引、分类等等。以上事实表明，尽管从总体来说，我国计算机情报检索尚处于初具规模的阶段，但是其发展的势头是确定无疑的。

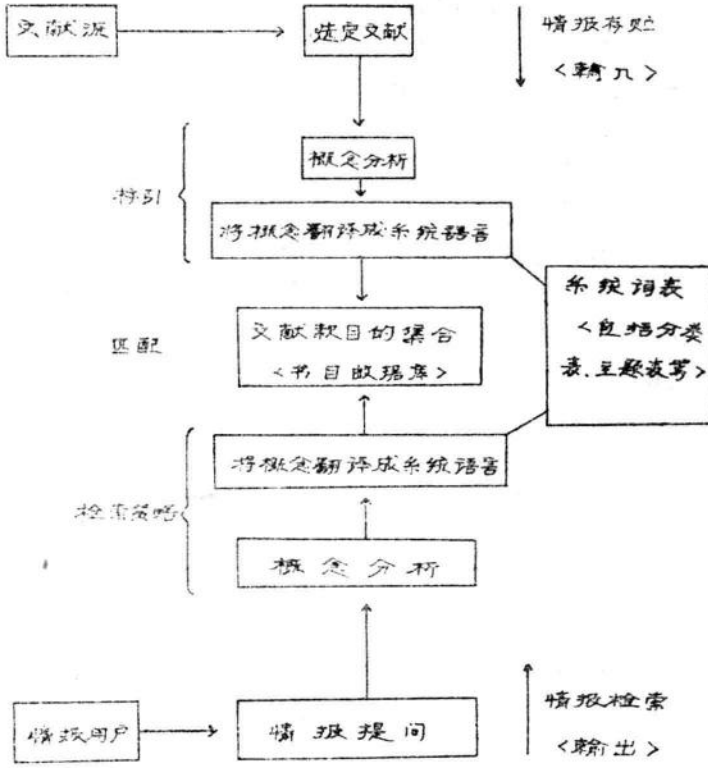
因此，今天我们来学习计算机情报检索的理论和技能，不是什么“脱离实际”的“清谈”，而是有其现实意义的。

二、计算机情报检索出现之前检索技术的演变

计算机情报检索的出现不是偶然和突然的。从检索技术来说，有它自己的逐步进化的过程。这种进化的过程，主要的推动力量，是文献情报的不断增长和人们检索要求的不断提高。此外是可供应用的技术的不断发展。而检索技术的进步，集中

地表现在情报提问同文献标识之间匹配方式的变革上。

情报的存贮与检索流程，如下图所示：



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复盖面、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引得深度、标引的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。这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的分析，将每个概念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。

为了实现这种匹配和扫描，就需要把文献款目记载在一定的物质载体上，并且要用一定的符号（包括文字或代码）来表示文献标识。另一方面，检索策略也需要记录在一定的物质载体上，以便进行同文献标识的匹配。

用来进行扫描和匹配的最原始方式，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。

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之点，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，使扫描与匹配的效率有所提高。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变成相应的一系列轧孔置位，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。人们不需要也不参予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对分离，成为一种由机具（当然这是非常简单的）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。

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说，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。这种代码化，是出于用机具扫描以代替人的浏览的需要。

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，逐渐地同人的参予相分离。随着扫描匹配速度的加快，人也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、明确化、形式化。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。

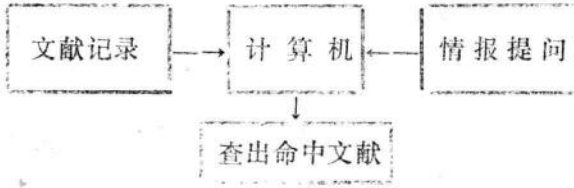
以上三个发展趋向，在计算机情报检索中进一步地得到加强。文献标识的代码化进一步变成整个文献款目的机读化，产生了计算机可以阅读的文献记录和机读数据库；扫描匹配过程的相对独立化，导致由计算机根据程序进行的高速准确的运算作业，人不可能参予，也不必参予，检索者可以把计算机内部进行的扫描匹配过程甚至整个计算机情报检索系统看作是一个

“黑箱”；检索策略的相对固定化和形式化，在计算机情报检索中则是要求把检索提问写成布尔逻辑检索表达式，交计算机去执行。计算机只是起一个巨大而高速的匹配器的作用。它严格按提问表达式中所指定的条件进行扫描匹配。检索策略的相对固定性更加明显。

由此可见，检索的扫描匹配技术是有其自身的演变历史的。计算机情报检索的出现，不是偶然的。可以这样说，计算机情报检索是发展到现阶段的检索方式。

三、计算机情报检索系统的构成与种类

计算机情报检索并不神秘。它是模拟人的手工检索的。计算机情报检索的最简单的概念如下图所示：



从图2可以看出，计算机一方面接受文献记录（即文献款目及其标识），另一方面接受情报提问（即检索提问表达式），然后进行两者之间的匹配，以找出符合检索要求的文献。计算机在这里的作用，只是代替人的手翻、眼看、脑子判断的工作，即“匹配”的功能。作为一个巨大而高速、准确的“匹配器”，计算机使情报检索的过程实现了“电子化”，其检索效率是手工检索所不能比拟的。

当然，上面的概念图是过于简单化的。例如，文献记录事实上是以机读代码的形式存贮在磁带上的，或存放在计算机磁

盘上的，这样，计算机才能“阅读”和理解。磁带上或磁盘上机读记录的集合，叫做文档，或称为“数据库”。对文献检索系统而言，更确切地说，应当是叫书目数据库。如果用常规的、供手工检索用的文献检索工具来相比的话，一条文献记录，就是一条文献款目。即一张目录卡片，或书本式检索工具中一条文献的完整著录。但是一条记录除了文献款目的内容外，还加上了一些供计算机处理所必需的符号，如指示符、分隔符、字段或记录结束符等，此外还有记录长度、各字段（如书名、著者等项目）标号，长度与起始地址等的说明。书目数据库是计算机情报检索系统的情报资源。也就是计算机进行查找与处理的对象。它是检索系统的原料。当然，对于情报检索系统来说，除了书目数据库外，还有事实型或数值型数据库（即各种数字或事实），或者是文献的全文，即文本数据库。因此情报检索系统按照其数据库种类分，可以有书目检索系统、事实或数值检索系统以及文本检索系统。总之“库”（数据库）是建立计算机情报检索系统的必要条件之一。

其次，计算机是一种机器，它只是各种电子元件与钢铁等材料的结合，即“硬件”。光凭“硬件”，它是不能动作的。计算机的一切行动，都要听从“指令”的指挥。许多指令编在一起，这叫“程序”，即规定计算机做各种动作，并规定其先后次序，以及遇到某些情况时应该如何判断，下一步执行什么指令等等。情报检索程序，就是为了实现人的检索目标，事先周密地规定了计算机的各种动作，让计算机一步步地按照一定的次序（它大体上是模拟手工检索的）来达到查出文献，并编辑打印出来的目的。情报检索程序的好坏，决定机器处理速度的快慢和检索功能的高低。当然这些也同硬件的潜力有关。总之，硬、软件（各种程序）是计算机情报检索系统的另一个必

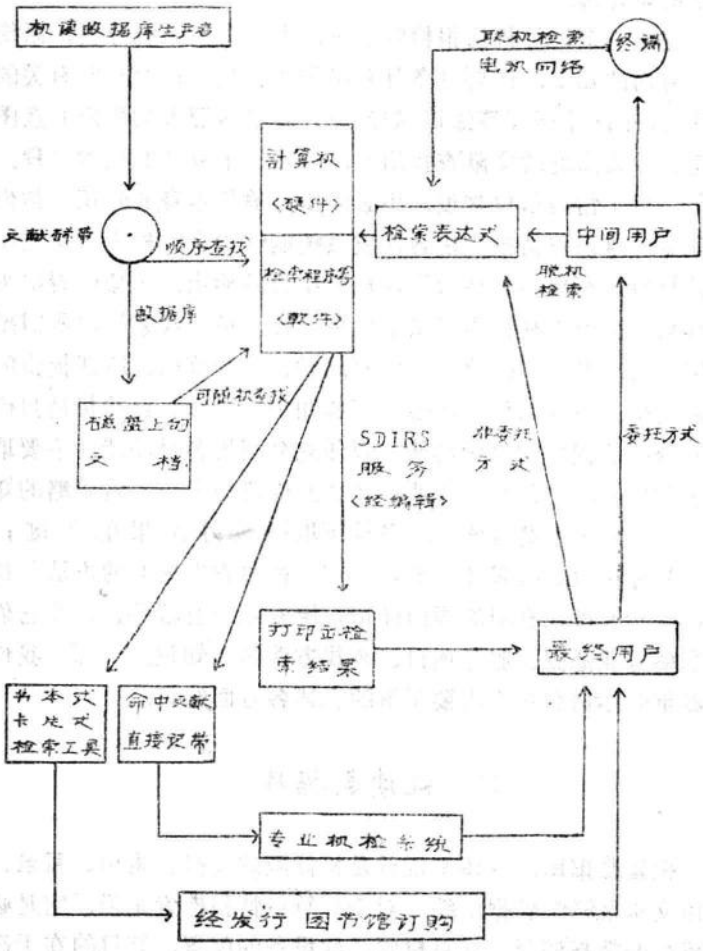
要条件。

再次，情报提问这一方面，为了适应计算机检索，必须将提问的内容变成计算机能够处理和运算的检索表达式（目前还不能使用自然语言来表达），并且使其代码化。检索表达式目前主要有布尔逻辑的表达式，加权检索的表达式等。检索表达式也可叫做“检索策略”，它相当于人们检索的意图。如果检索的意图不清楚，检索表达式拟订得不好，再好的情报检索系统也难以取得好的检索结果。检索表达式（检索策略的表达形式）的构造、调整及其输入也是计算机情报检索系统的必要条件之一。

再其次，人们的检索是在计算机旁进行呢？或是在远离计算机的地方借助于终端设备与通讯线路向计算机进行提问呢？前者称为脱机检索，后者称为联机检索。对于联机情报检索来说，终端设备及通讯线路也是计算机情报检索系统的必要条件之一。

最后，情报的需求者（即最终用户）是自己来从事检索呢？还是把自己的情报需求告诉中间人，委托中间人来进行检索呢？后者称为委托检索，前者称为非委托检索。这种中间人一般是图书馆员或情报工作人员。对于检索系统来说，他们也是用户，当然是中间用户，而不是最终用户。从理论上来说，最终用户自己面对检索系统进行检索，是比较好的。因为最终用户最了解自己的情报需求，也最能判断检出文献是否对口，便于及时修改自己的检索策略。但是，最终用户往往不熟悉检索系统的询问语句、不大了解词表的结构与标引规则，同时在联机检索的条件下，自己不能熟练地操作键盘，因而会延长检索时间、增加检索费用。因此，通常的情况是把自己的需求委托中间人来进行检索。委托式检索固然能发挥中间人熟悉检索作

业的优点，但是也会带来不少问题，即中间人有时不明白最终用户真正的检索要求、学科专业知识也受限制。这种委托式检索和非委托式检索，是用户—系统交互子系统的两种方式。一



般来说，脱机检索是委托式检索，联机检索可以是非委托式检索，也可以是委托式检索。

现在，我们可以把上述情报检索系统的构成，画出下列较详细的概念图：

建立一个计算机情报检索系统，不仅需要系统分析、系统设计方面的知识，还要具备计算机硬件、软件以及与此有关的算法、程序设计语言等知识。如何使计算机按照人们规定的意图将符合检索命题的文献查找出来，这是一个复杂的技术过程。但是，对于情报用户来说，不必懂得计算机本身的知识。他们的任务，就是将情报需求转换成系统能够接受的检索表达式，向计算机进行提问，然后等待检索结果的输出。正象收看电视的观众不懂得电视机内复杂的电路一样。情报检索用户所需要了解的是：（1）机读数据库的品种、收录范围、所能提供的各种索引、以及如何选择数据库的知识；（2）计算机情报检索系统所能提供的检索功能，以便充分利用各种功能（主要取决于软件）来进行检索作业；（3）构造与调整检索策略的知识。这是在一定的检索系统条件下取得较好结果的关键；（4）评价检索结果（主要是查全率与查准率）的方法与技能；（5）可以利用的国内外情报检索系统有哪些、以及它们的学科专业范围、服务项目、查找方法等等知识。下面，我们将着重介绍情报用户需要了解的上述各方面知识。

四、机读数据库

机读数据库，具体来说就是机读版的文摘、索引、目录、自由文本及科学数据汇编。对于计算机情报检索来说，它是必不可少的情报资源。计算机硬、软设备的配置，其目的在于高