科技文献检索

教学参考材料

北京大学图书馆学系
情报学教研室

1981年5月

说　　明

本材料系为我系本科学生和函授生学习《科技文献检索》课而编印的。选材时，侧重选收有关文献检索的基本原理、基本方法以及特种文献检索等方面的文章或资料。凡有关介绍某些重要专业文摘索引或国外三大套综合性文摘的资料，因本课程的讲义已系统介绍，故未收入本书。最后收录了一篇系统介绍机读文献库的历史和现状的资料，意在向大家提供有关计算机情报检索的一点基础知识，为以后学习计算机检索做准备。

由于有关文献检索方面的资料非常丰富，选编的时间又颇为仓促，所以未能预先去函征询有关编著者的意见，材料的选择和编排也未必完备、恰当，谨请各位谅解并致歉意。

北京大学图书馆学系情报学教研室

《科技文献检索》教学小组1981.5

目　　次

一门新兴的实验学科——情报检索

沈 迪 飞

第二次世界大战以后，面对数以百万计迅猛增长的文献量，传统的手工检索方式已无力适应，情报的存贮和检索问题越来越引人注目，这就是西方情报界惊呼的“情报危机”的由来。但是，人类是决不会在“情报危机”面前束手无策的。与此同时，一项划时代的新技术——电子计算机诞生了。电子计算机的问世，实现了文献检索的机械化和现代化。从此，一个崭新的概念——情报检索出现了。

一、一门新兴的实验学科

情报检索开创至今，只有将近二十年的历史。尽管时间短暂，但是，情报检索已经建立了许多实验系统和实用系统，情报检索正作为一门新兴的实验学科，以极其迅速的步伐迈入科学之林。

将情报检索看成一门新兴的实验学科，根据何在呢？

第一，情报检索有专门的区别于其它学科的研究领域。

情报检索以计算机在文献检索中的应用为对象，研究适应计算机处理的文献描述、存贮、检索和提供的理论、技术和方法。情报检索这个词是1949年 C.N.Moores 最早使用的。英文原文为“information retrieval”，或者称为“information storage and retrieval”。后来，A.kent 又作了如下解释：“所谓情报检索是指机械化的情报检索(mechanized information retrieval)，它与使用机器的文献检索 (machine literature searching) 几乎是同义的。”英国剑桥大学的情报检索专家 C.J.Van Rigsbergen 也讲“用‘文献’来代替‘情报’就足以论述这类情报检索了”。

国际上，对情报检索有多种定义，从这些定义差异的比较中，更清楚地看到了情报检索特有的研究领域。大致有下面几种定义：

① 所谓情报检索是指情报的存贮和检索。

② 为情报做索引文件以及文献的积累和检索的技术。

③ 从一定的角度出发（为着某一确定的目的），由已经存贮的情报中取得所需要的情报。

④ 根据用户的需求检索情报文件，从文件中取出符合要求的情报交给用户。

⑤ 所谓情报检索，通常是指情报的收集和存贮，并对存贮起来的情报进行检索和分发。

仔细分析一下，这五个定义是有差别的。一类意见认为情报检索就是指情报的“存贮和

检索”，而另一类意见则认为既包括情报的“收集和存贮”，又包括情报的“检索和分发”。这两类意见反映了对情报检索研究领域认识上的差别。前一种看法，单纯指计算机存贮和检索信息的理论和技术，并没有考虑到情报的收集、选择、分析和加工，以及如何在计算机中表示语言情报，也没有顾及到用户的要求、检索途径和检索策略，检索出来的情报如何提供评价等。因此，这种定义仅仅反映了信息处理内容，没有指出情报检索专有的研究领域，当然无法成为情报检索的正确定义。情报检索具有计算机科学同情报科学定义的性质。从情报检索的专门研究领域出发，我们可以得出如下简单结论：情报检索的研究领域就是计算机化的文献检索，第二是指采用计算机的现代化方式，它属于情报工作的范畴。因此，作为一门学科，情报检索也是情报科学的重要内容。

随着计算机应用的发展，相继发展起来的数据检索和事实检索，也都称为情报检索。但这些是情报检索概念的外延，是广义的情报检索，是情报检索研究中旁及的领域。

第二，情报检索不仅有区别于其它学科的特有研究领域，而且有核心研究问题。

C.J.Van Rijsbergen 将情报检索的核心研究问题归结为“内容分析、情报结构和检索评价”三个方面。首先是内容分析，这种内容分析是与手工方式不同的，主要是以适合于计算机处理的形式来描述文献和揭示内容，因而清楚地反映出文献分析加工同计算机技术之结合。目前，有不少情报检索的实验系统在研究文献分析加工的自动化，包括自动抽取关键词、自动标引和自动文献分类等。其次是情报结构，涉及到利用文献间的相关性来改进检索策略的有效性和功能性。这反映了文献检索方法同计算机科学的数据结构和检索方法之结合。顺序检索、倒排文件与聚类文件，各种检索提问逻辑都反映了这个方面的研究进展。检索评价，涉及到检索有效性的测度问题，既从情报学角度测度情报检索的效果，同时又测度了计算机技术的性能和可靠性。

第三，情报检索拥有本身特有的概念和表示这些概念的语言词汇，有自己的规则和规律。

从文献与提问的计算机表示、主题词表与标引、文献数据库、情报逻辑结构及其物理实现、提问逻辑与检索方法，到情报的输出与提供，情报检索拥有反映学科内容的许多专有的概念，以及表达这些概念的独特的语言词汇。在适合计算机的文献描述、关键词的统计、自动原文分析、自动文献分类、文件结构、查找策略与检索评价等方面，已经和正在形成一些规律、规则和方法。

第四，情报检索依据多种理论和相关学科，已经形成了自己的学科体系结构。

情报检索借助相关学科的理论，通过实验方法，创建了一门实验学科；又通过情报检索实践，验证、丰富和发展了实验研究。同许多实验学科一样，实验是情报检索的主要研究方法。情报检索就是在不断地实验和实践中发展起来的。情报检索的历史，就是在实践和实验中创建起来和发展学科体系结构的历史。目前，许多情报检索的实验系统，以美国康奈尔大学的“自动原文检索系统”（SMART）为代表，正在对情报检索的理论和规律进行深入探讨和实验研究。以美国洛克希德公司的 DIALOG 系统为代表的大量的情报检索实用系统，正在全世界运用和服务，情报检索网络遍及全球各大洲。反映这些发展的书籍、报告和刊物大量涌现。据笔者极为有限的统计，情报检索的核心期刊和常用相关期刊，已近40种。情报检索已经在科学、教育和文化生活中发挥出越来越大的作用，已经成为先进国家科研人员不可缺少的助手，并正深入到亿万人民的日常生活中去。

综上所述，情报检索是用实验方法研究计算机在文献检索中应用的一门新兴的实验学科。经过将近20年的发展，这一点已经为国际上越来越多的情报学家所认定。但是这一门新兴的实验学科的发展历史毕竟尚短，它还很年轻和幼稚。目前，情报检索总结出来的理论还不够系统和完善，通过实验和实践所发现的规律也还不够丰富，并且都有一定的局限性。

二、情报检索的特点

1、 情报检索是情报科学和计算机科学的交叉学科

所有交叉学科都具有一个共同的特点，即具有其脱胎而出的那些学科的特征。情报检索也一样，它起源于情报科学和计算机科学，因而具有这二门科学的特征。

情报科学研究情报的性质和情报的各种现象：产生、传输、转换、积累、存贮、检索和提供等，同时也研究情报的增长、老化和废弃，研究各种语言的原文结构对情报量的影响。文献检索作为情报科学的一个分支，它主要研究二次文献的收集、加工、解释和检索、提供服务等等。计算机科学的历史不长，但因它被异常广泛的应用，而产生了许多交叉学科，如计算数学、计算物理学、计算化学和计算语言学等等。情报检索就是文献检索同计算机技术相结合的产物。

情报检索处处表现出交叉学科的特点，它既反映了文献检索原理，又体现了计算机技术的渗入，使文献检索发生了本质的变化。这些变化体现了继承与发展，基础与提高，传统与现代化等方面的关系。以检索途径或称之为提问逻辑而论，文献检索是从传统的作者、分类、先组式的主题等途径检索；而情报检索一方面仍采用这些途径，同时又发展出了与计算机技术相适应的新方式，如后组式的布尔逻辑组配、概念加权、检索词截断、邻接与优先规定、终止模式和通用字符法等等，开拓了手工方式无法实现的途径，使检索方式别开生面。在情报提供方式上，也有了很大的变化，情报检索不仅较容易地通过定题服务实现了对每个用户的对口区别服务，而且通过网络可以实现跨国跨洲的联机检索，达到情报资源共享。计算机应用于文献检索，使情报服务迈入了前所未有的现代化时代。

2． 相关性——情报检索的中心概念

情报检索不是直接解答用户所提问题的本身，而是找出同用户提问相关的文献，所有检索都建设在用户提问同存贮的文献之间比较的基础上。检索应做到把同提问相关的文献与同提问不相关的文献区分开来，区分的越清楚越好。每个具体的检索目标，用集合论的语言讲，都应当“检索出相关的文献 A 而抑制不相关的文献 $\overline{A}$”[2]（A 指文献集合）。这里，核心的问题就是相关性。用户提问同检出的文献相关性，是情报检索的实质，离开了这一点，情报检索就完全失去了意义。因之，相关性是情报检索的中心概念。相关性的含义是指文献的情报内容与提问要求之间的对比数，有时也称之为“关联性”或“相似性”。

相关性贯穿于情报检索整个过程的始终。在建设文献数据库和编写用户提问式时，就要考虑如何二者相关。表征文献要依据某种规范化了的主题词表和分类法；同样，描述用户提问时也要用同一个主题词表和分类法。这样，表征文献用语同用户提问用语相一致，就为用户提问能同所需文献相关匹配创造了条件，以致当二者相关时，所需文献能被检索出来。

情报检索的多种检索途径，都是从不同角度使用户提问能同文献相关。相关性解决的越好，这种检索途径的生命力就越强。计算机应用于情报检索，发展了布尔检索、加权和截断

等许多检索途径，就在于这些途径同计算机的自动快速处理相结合，能更好地解决提问同文献之间的相关性问题。

不同的情报检索软件系统，可能会采用不同的数据结构和算法，这些结构和算法虽异，但其核心问题都是围绕更好地解决相关性问题的，“在可能受限制的相关决策内构造出一种模式来”。[2]无论批处理方式的顺排文件，随机查找的倒排文件，或是比较新的相似性检索原理和聚类文件结构方法，都是从不同角度，适应机器的特点，利用不同数据结构方法，达到迅速、准确和全面地从存贮文献中找出切合用户提问的相关文献。倒排文件是预先将同每个检索词相关文献的文献号码集中起来，供从检索词迅速查找。“所谓相似性检索原理，就是用数学方法来计算读者提问和库存文献之间的相似程度，并参照读者的意愿来决定文献的取舍。”而聚类方法则是依据“相关文献之间的类似”形成的。

目前正深入研究的使用计算机的文献自动分类（不是用分类法的分类），就是“利用文献之间相似性的类似度同文献属性状态数目成正比的原理”，[2]研究关联度的测度方法。情报检索的一个重要环节反馈，亦是依据检出文献与用户提问的相关性来实现的。

可见，从文献的建设、检索到检出文献鉴定，都贯穿着相关性，它是理解情报检索的关键。

3. 检索效果评价的互逆相关性和模糊性

情报检索的评价不能只从单一角度，而要从多方面来衡量，才能得出比较全面和恰当的结论。归结起来，主要从三方面评价：① 检索效果或有效性的评价，是对检索系统检出相关性文献以满足用户提问要求能力的一种测度，集中以查全率、查准率作为标准；② 检索实用性（Utility）或可用性、适应性（Serviceability）的评价，包括系统对用户是否需要，是否实用，是否“划得来”，有多大“效益”，能发挥多大实际作用等等，这涉及到广泛的社会学问题；③ 检索费用效率（Cost—effective）评价，这相当于计算机的价格性能比（Cost Performance），如同衡量计算机产品优劣好坏、效果高低的指标。这方面的要求是缩短检索时间，少用机时，压缩信息以扩大存贮，减少费用，提高系统性能。系统功能性主要是通过所用计算机资源的时间来测度，这些涉及到建设系统的经济学的问题。限于本文的范围和篇幅，仅讨论第一种测度。

查全率和查准率是情报界非常熟悉的，这二者都是对情报检索系统检索相关文献而同时又抑制不相关文献方面的一种测度。同相关性是情报检索的中心概念一样，这二者也是情报检索评价的中心内容。

情报检索的理想要求是，不检出一条不切合的文献，同时也不漏掉一条相关的文献，检索好似一个过滤器，一个筛子，要找出切合文献，即要求查全率和查准率都等于1，但是，这二者同时都等于1是矛盾的。若要保证查全率到达1，可行的办法是扩大提问主题词或分类的范围，结果，必定会检出许多不切合的文献，造成查准率下降；相反，如果用一个提问范围非常专的检索式来确保查出的文献都是切合的，则必定会导致漏掉一些也是切合的文献，查全率下降。查全率与查准率之间这种互相矛盾的关系，称为“互逆相关性”。国外有人采用决策论研究情报检索，认为在某种条件下，查全率与查准率可以不是互逆关系。这种个别的有条件的论断，还远不能影响“互逆相关性”的特点。

目前情报检索系统广泛使用的布尔检索，是应用经典数学的经典集合论来判断相关性的。检出的每篇文献，或属于切合文献集合，或属于非切合文献集合，二者必居其一。形成

了二值逻辑，非“真”即“假”。但是，人个子之“高”或“低”以及人的“中年人”或“青年人”一样，都是模糊的、不精确的概念。经典集合论的精确性同现实世界中这些不精确性存在着矛盾。这种表现事件概念的内涵和外延的不确定性，就是模糊性，需要用近些年发展起来的模糊集合论来加以解决。用户的情报要求和文献的情报内容，都具有不确定性的特点，二者之间的相关性判断，更强化了不确定性，这就是检索效果评价的模糊性特点。

在情报检索实践中，可以利用互逆相关性和模糊性特点，对不同职业不同要求的用户，采取有针对性的区别服务。

4. 语言信息

在情报检索中，文献的表征用语和用户提问的描述用语，大多是语言信息。一部分是人工语言，大量的是自然语言，而且涉及到计算机字符集所能表示的多个语种。作为情报检索的表示，所用语言必然要影响到文献库、用户提问和检索。

当前的情报检索系统，一般均是人工语言同自然语言兼用，检索时以人工语言为主。使用主题词表和作者规范文件等，就是将某些自然语言加以标准化和规范化的人工语言集合；另一方面，也大量使用标题、文摘、出处等原文的自然语言。采用主题词等人工语言方式，需要事先编制主题词表，也需要大量人力对文献进行分析和加工，进行复杂的标引工作。更由于标引者的知识面、水平和倾向性，而使文献加工结果带有很大的主观随意性，当然会影响检索质量。自然语言由于没有规范化和形式的复杂多变，带来了新的问题。自然语言的同一词干具有多种前缀和后缀形式，同一词的单复数区别，不规则的拼法(如英文颜色之Color和Colour二种)，多义词（Plasma和Plant等），同义词和缩略语等，使检索词间要采用复杂化的形式。词与词之间形成的不同含义，如“don't care”，词的顺序所产生的相关关系，词在原文中（句子、段落、字段）的位置，都要影响到机器处理。现在已有的截断、优先、邻接和终止模式、通用字符等检索方式，大都是用于处理自然语言的。无疑，用自然语言在计算机中描述文献，是理想的方式。联机检索的命令语言，也在向自然语言和多语种方向发展，方便用户自由询问，由系统进行规范。这一切，都要求尽快发展自然语言处理的研究，有的国家已在设计能理解自然语言的计算机。

情报检索的语言信息，形成了文献库的某些特点。由于受计算机字符集限制，有些语言无法用计算机表示，尽管采用拼写等方法，也还是影响到文献库的收录范围。

在语言信息这一点上，文献库有别于70年代迅速发展的数据库。数据库是从文献中或从实验室的实验中选取的经专家鉴定过的一次数据情报，采用结构化的方式，存贮在计算机中，检索后可以直接使用。数据库的信息，大多数是数值数据或数值化的图谱数据，也有语言信息，但比例较少；因而，其整个数据量和每年增长的数据量，都大大低于语言信息的文献库。二次情报与一次数据情报，大量而迅速增长的语言信息和经过专家评价的数量有限的数值信息，这是文献库同数据库的主要区别之一。

5. 半结构数据库 (Semi—Structured Database)

文献库和数据库的另一个主要区别是数据库结构方面的差异。气象、地震、人口和经济管理等方面的数据库，大多数为数值数据，采用方便计算机处理的严谨结构。有层次模式、网状模式和关系模式等结构方式，数据项的格式固定，称为结构数据库。文献库都是原文(full—text)数据库，或称为书目数据库或图书馆数据库，是由非标准的结构性不强的原文数据和语言信息组成的，因而使文献库形成了自己的结构特点，称之为半结构数据库。

半结构数据库和结构数据库，二者在计算机处理上，差别极大。原文数据中包括标题、作者、出处、主题、分类等多种项目，不同记录所采用项目数和每个项目长度数都是不相同的。因此，建立标准字段是困难的，无法组织固定格式、固定长字段的结构数据库。文献库大都须用可变格式、可变长字段的记录结构数据库。文献库大都采用可变格式，可变长字段的记录结构，与此相适应的是，为了方便处理，加入了目录区字段，通过目录字段来存取相应的数据字段，这又同直接存取数据字段的结构数据库不同。国际标准《书目信息交换用磁带格式》，就是采用目录方式的可变长字段的半结构数据库，反映出文献数据库的上述特点。

半结构数据库加上语言信息的特点，在计算机处理中，不适于采用数据库结构的各种模式，大多采用文件结构，组织多种文件，适应多途径检索询问的需要。组织文件的主要依据不是从结构入手，而是考虑如何使提问同文献相关联。这里不是不要结构性，而是要使结构能适应于情报的逻辑关系，“十分要紧的是要研究出一种文件结构，它能有效地解决更复杂的文献同询问的描述”。以属性为基础组织的倒排文件，它同顺排文件相结合，是目前情报检索软件典型的文件组织方式，属性就是检索询问的入口。但是，庞大的倒排文件的担子压在开创至今的传统计算机结构上，大大影响了检索效率。国外许多实验情报检索系统正在研究一种新的文件结构——聚类文件。它从“紧密关联的文献往往与相同的检索有关”设想出发，研究文献之间的相关性，将类似的文献组织到一个类里。检索时不是对个别文献检索，而是对聚类文件中的类进行检索。这不仅能提高检索速度，而且也会使检索更为有效。

由于情报检索的文件系统和数据库的结构系统的种种不同特点，看来用数据库管理系统（DBMS）代替情报检索软件是有许多困难的，要了解具体的文献库结构特点和DBMS功能才能确定。这一点是值得注意的。

6. 特殊的数据处理

数据处理是计算机应用的一个方面。最初只指由计算机加工商业、企业的信息，现在则泛指非科技工程方面对任何型式数据资料的计算、管理和控制。如企业管理、库存管理、报表统计、帐目计算以及情报检索等。数据处理问题的特点是存贮数据所需的存贮空间远远大于控制数据的程序所需要的空间。

同一般数据处理比较起来，情报检索具有一些特殊的地方。除前面提到的语言信息处理，特别是自然语言处理和半结构数据库的特点之外，还有：① 数据量极大并以惊人的速度增长，这就需要大容量的内存和外存，特别是外存。美国DIALOG系统到80年底，磁盘量已达60,000MB，即600亿字节，1978年磁盘量仅为150亿字节，两年时间增长了3倍。这样大的存贮量与增长速度，对一般数据处理来讲，是难以想象的。② 大量频繁的内外存数据交换。③ 大量输入和输出数据，要求多种多样输出介质。为满足上述两点要求，在计算机的配置上要考虑多通道大流量和多种外部设备，甚至要配备一些专用外部设备，要有大小写字符的快速打印机以及计算机控制的排版设备，要有大小写字符的快速打印机以及计算机输出缩微胶卷（片）设备。④ 联机要同时接得大量用户，要保证一定的响应时间。为此，机器要有较强功能的分时系统，有些国际情报系统都拥有3000个以上的专用终端。

目前计算机的体系结构基本上还是四十年代计算机问世时的冯·诺依曼的结构思想，它适用于科学计算和一般的数据处理，但却难于支持情报检索这类特殊的数据处理。情报检索的大量的语言信息和原文数据，需要非常强的字符串处理（特别是比较）功能。目前计算机是

用两个字符串逐字符相减为零则相等的办法以增加循环控制指令的方式来实现字符串处理，大量的字符串比较，必然会极大地减低机器效率。比中型机 IBM370/150 快十倍的大型机，处理300亿字节的检索仍要 8 个小时，处理费需要 1000 美元。此外，外存到内存数据交换，由慢速到快，形成瓶颈的现象。软件方面，处理机主要都花费在倒排文件、索引文件等目录结构文件的组织、管理和传送上，都大大影响程序系统的功能。

上述情况表明，现在计算机的体系结构，对情报检索确实有不少不适合之处，有许多难于解决的困难。近些年来，国外的计算机的体系结构，如大容量快速内外存数据交换的相联存贮器，其特点是以内容定地址，还有检索词比较器，询问解算器，原文扫描检索系统，以及倒排文件(或索引文件、目录)处理机等。用迅速发展的便宜的硬件取代软件的一些功能，使情报检索专用机达到高速、低费用、大容量要求的目标。估计在80年代会投放市场。

7. 学习和自适应系统

计算机学习功能，是指计算机依据以往经历改进其自身程序的过程，机器根据以往参数进行判断，即借助它以前运行历史的记录分析，随时改善它以后的处理方法。自适应是指一个系统改变自身的性能以适应环境的能力。

如前所述，情报检索效果的测度，依据用户对检出的文献切合性的估价。检索过程中要依赖于一些参数，这些参数在系统内是依不同情况而变化的。如果回答某些提问的参数形成了最佳的检索效果，那么，无论何时再碰到类似提问时，计算机就可以依据以往经历的记录分析，自动赋予控制参数以保证最佳检索效果。这种能力称为情报检索的学习和自适应系统。

现有情报检索的“反馈”、实时修改检索式和主题词扩检等，是建设学习和自适应系统的基础。“反馈”概念是在生物学系统和自动控制系统中建设起来的。它是将输出的结果，同输入进行比较，用比较的结果控制系统，即根据过去的执行情况来改进系统的性能。通过反馈可以进行检索式的自动修改，从而建设起情报检索的学习和自适应系统。

人们进行推导是根据不同概念间的联系，以发现新概念，这一过程某种程度也可以计算机来完成。如果一个概念常用第二个概念联系，同时也常同第三个概念相关联，则第二与第三个概念之间有极高的概率是具有某种联系的，这可以由计算机检索词连接矩阵来实现。建立在人的许多判断和经验基础上的相当数量的情报，亦可以编入到一个纯粹的机械系统，检索词的自动关联和文献自动标引，就以此为原理。这些都在研究与发展当中，学习与自适应系统将为情报检索开创更为美好的前景。

三、情报检索的相关学科

情报检索是一门交叉学科，是借助于相关学科的理论和方法而逐步发展起来的。因此，学习与研究情报检索，除掌握它自身的规律和特点外，应下功夫学习与研究其相关学科。

情报检索的相关学科涉及到情报科学、计算机科学、数学、语言学和系统科学等多种领域。

在情报科学领域，同情报检索相关的有情报学、文献检索、目录学以及主题法和分类法等等。因为情报检索是用计算机技术来实现文献检索功能的，因此，这些学科和专题，是情报检索的出发点和基础。

在计算机科学领域，情报检索的相关学科涉及到许多方面。在硬件方面，涉及到计算机系统的构成、专用设备和网络，这是使用一个计算机系统和建立一个情报检索系统首先要接触和了解到的。只有了解计算机系统和网络构成、性能和工作原理，结合情报检索的特点，才能明了情报检索用计算机的特征，这对使用机器和购置系统都是非常必要的。在软件方面，要涉及程序设计、汇编语言、COBOL 和 PL/1 等语言、数据结构编译技术、操作系统和数据库管理系统许多方面的知识。数据结构是数据处理的基础，编译技术是情报检索主要借鉴的软件方法，操作系统是情报检索软件同计算机系统软件的接口，数据库管理系统是情报检索非常相近的软件系统。情报检索软件就是在这些理论和技术基础上发展起来的。此外，在自动控制方面，情报检索在许多地方同人工智能是一致的或类似的。例如，情报检索确定切合文献以响应系统用户询问很类似于模式识别，学习与自适应系统要应用人工智能的许多技术。可见，情报检索几乎应用了当前计算机科学发展的主要成就。尽管计算机技术不是情报科学的组成部分，但它是情报科学的一个极为重要的手段，是情报处理的主要工具。因此，欲理解今天的情报科学，就必需了解今天的计算机特征和能力。

数学同样是情报检索赖以发展的理论基础。数学领域的离散数学，包括布尔代数、集合论、图论、组合分析等，以及概率论，模糊数学中的模糊集合论，效用论 (utility) 和线性代数，都是情报检索的科学依据。以这些理论为基础，情报检索才能实现各种相应的数据结构、文件结构、检索算法和检索提问逻辑，也为评价检索效果提供了数学方法。很清楚，随着情报检索的发展和研究的深入，会更多地应用数学科学成就，从而也会加速情报检索的理论化和成熟化。

电信科学方面的信息论，语言方面的数理语言学 (mathematical liguistics) 或称为计算语言学 (Computational liguistics)，都是情报检索应用很多的学科。信息论对研究情报的信息量之测定与变化，提供了理论和方法。通过机器翻译而发展起来的数理语言学，是语言学与计算机科学的交叉学科，而情报检索要处理大量的语言信息，数理语言学为此提供了极大的方便。语言的句法，语义和符号语言学的研究，以及相应的自然语言处理的研究，对情报检索的软件设计和改进处理方法有着决定性的意义。

系统科学方面的系统分析与系统设计，是建立一个情报检索系统必须掌握与遵循的方法。只有对原手工系统进行周密的调查研究和细致的系统分析，进行自动化系统可行性研究，才能在此基础上进行计算机系统的基本设计和详细设计。诸如详细地考虑系统的组成、需要的条件、未来的扩充以及同其它系统的关系等等。系统的全局是决定情报检索成败与好坏的关键。

英国剑桥大学情报检索专家 Rijsbergen，1975年所著《情报检索》一书的序言中写道，他决心“去研究作为一门实验学科的情报检索”。[2] 情报检索是一门学科，这对许多人来讲，是难于理解的，即使对于初从事这一工作的人，头脑中也是一个问号。我带着这个问题，查阅文献，收集资料，综合体会，总结实践，写成此文。我越来越相信，Heaps 所讲的“情报检索学科”[1] 由于其强大的生命力和迫切的客观需要，必将以更快的步伐发展成熟，并指导实践，为人类的科学事业做出更大的贡献。

参 考 文 献

[1] Information Retrieval, H.S.Heaps, 1978, P.344.

[2] 情报检索（英），C.J.Van Rijsbergen·郭瑞枫等译，南京大学数学系。
[3] Annual Review of Information Science and Technology (ARIST)。Vol.14, 1979.
[4] 相似性情报检索系统，上海交大31室，1979.12。

（选自《情报科学》1981年第2卷第2期）

今后二十年的情报检索

C.N.Mooers

情报检索的含义应当有发展的观点。现在，情报检索已不仅限于寻找和提供文献，已经涉及到发现和提供非文献形式的情报。将来情报检索的概念必须随着工作的发展而加以扩大。

情报检索，实际上已经在应用机械。应用机械的目的，是要用来代替人工所进行的繁重的和难于胜任的工作。必须注意，任何时候人们利用情报机械检索系统是为了提供服务而不是为了机器。

历史过程

在介绍机械检索方法的成就以前，需要回顾一下历史。1915年泰勒（Taylor）发明了近年所谓的主题卡（Peek-a-boo）的方法。1920年英国人苏潘（Soper）在改进泰勒方法的基础上发明了一种新的装置，用于情报检索时已较先进。后来，由于电影和电影机的发展，1931年哥尔盘（Goldberg）得到了胶卷阅读器和照相复制机的专利。1935年，美国人戴维斯和屈雷权（Davis and Draeger）对缩微胶卷文献以及应用十进位分类法的问题进行了研究。1938—39年，麻省理工学院的布许（V.Bush）及其学生制造了第一台反射式缩微胶卷阅读器的样机。后来又出现了伊斯特曼缩微卡片机，这种机器是快速选择机和荷累利斯（Hollerith）穿孔卡片机的结合。但是，这种机器以及潘金（Perkin）的边缘穿孔卡片检索机等的检索效率不高，于是近几年来又以电子计算机用于情报检索。由于现在设计的计算机不符合情报检索工作的要求，所以效果仍然不能令人满意。如果需要贮存和检索的情报量过小，那就可以用简便的方法而不必用电子计算机来进行检索。相反，如果情报量较大，就可能把计算机的记忆容量全部占满。因此，为了完成有效的大量的情报检索，需要有特殊的装置。

到目前为止，实际上还没有一台能够胜任地检索和选择具有五万到十万项或更多情报的检索机。如果需要贮存和检索的情报数量达到百万项，那么选择情报单元时就要在多到10笔的记录上进行选定。这种装置将要建立而且已经在建立，但是采用什么分类方法（如主题、叙述词还是其它类似方法）的研究，却没有成功。现在的新发展是采用“叙述词”。虽然“叙述词”本身还有缺点，而机械检索的分类问题已接近解决。

发展现状

如果有一台情报检索机能够贮存和检索上百万甚至超过一亿的项目，但是如何编制叙述

词，却是一项困难和费时的工作。作者认为，阅读和标出一项专利符号，仅仅确定发明者的意图平均需要十五分钟。而美国专利局大约有三百万份专利。因此，这项繁重的工作应当由机器来做。虽然，这是一项不容易机械化的工作，但是鲁恩（Luhn）创制的自动文摘机，说明这项工作是能够成功的。鲁恩的方法是使用计算机“消化”论文的正文统计检出不平常的字，然后选出含有这些字的句子编成（自动的）文摘。如果这一过程仅做到选出所需要的字为止，那就是单元词。假如在这种形式所选出的字能代以具有大体上相同意义的标准字，如果能略去同义语，那就产生叙述词。这种在检索工作中处理同义语的方法实际上已应用多年，最近称这一方法为“叙词表法”。

这种扩大了的鲁恩的方法并不能最终解决检索问题，因为鲁恩所设计的机器虽然可以检出那些在本文中高度重要的字以后，但是还不能转换成标准的叙述词以便进一步检索。因此，需要建立这样一种机器，它不仅能将有关的词归类，而且能够“识别”叙述词的性能。现在这种具有初步“智慧”的机器是可能出现的，例如苏仑诺夫（Solomonoff）所说的“学习机”（Inductive inference machines）。这种机器，当它被“教会”正确解决一系列问题的例子后，机器就能解决同类的其它问题。

编制文献中的叙述词，也就是对文献中的情报用检索符号来描述，换言之这就是最简单形式的文字翻译。所以说是最简单，因为机器所翻译的文字不涉及文法，叙述词的文法是不存在的，至多也是初步的。如果“学习机”能够建造出来，既可能贮存文献的主题索引和分类号，也可能贮存图书的章节和段落的标题，目的是要使机器至少具有一些初步的图书馆的标题目录。可以期望，几年后“学习机”必将获得有意义的发展。

近　期　目　标

为了情报检索，还有一系列的其它种类的机器要应用。比如情报的贮存、输送，以及图书馆中有关分类、排架、登记等，都需要利用机器，本文均不讨论，只着重讨论情报检索机的问题。

当人们来到情报检索系统，最重要的而且是不了解的问题，是如何获得情报。首先，在检索系统中所用的名词、术语，和他所习惯应用的可能稍有不同。其次，在检索时还需要一种工具作为机器的助手。

某些简单的分类卡检索系统中提供了解决上述问题所需的系统和方法。有些分类卡系统在这方面做得很好；但也有一些不是很有效的。在柴道符号系统（Zatocoding System）中，使用了数百个简单的叙述词。在叙述词表中罗列了某一词的有关的或接近的词，以便交叉参考。这样，即使人们开始时用他自己习惯了的词，也能找到进入检索系统的方法。当情报检索者查找到一个叙述词，他会发现有一个说明，叙述了和这一叙述词有关的意义范围。在叙述词字典的另一处，叙述词本身被分成十五或二十组。在每组的头上列明关于所包括的叙述词的有关问题，以便查找。

这种简单的卡片系统，还能在另一方面有助于情报检索者，因为这个系统能在检索开始后一分钟或更少的时间内就能将卡片选出。如果检索的方向错了，情报检索者在看到卡片或文献后，即可迅速发觉。但是，这种系统现时所用的机器动作太慢，检索效率不高。

将来，情报检索者在使用大的机器检索系统所需要的工具问题可能是严重问题，但必须

面对这一问题。如何使检索者的要求转换成适合机器工作的形式，将需要一种工具。荷尔脱和特仑斯基（Holt and Turanski）认为，现在已经发展起来的数学自动编码系统，就是这样一种工具。将书面语言变为机器语言，也即成为机器能进行操作的抽象符号和机器的详细命令。因此，在使用更大和更复杂的机器检索系统，必须在检索过程中，在人-机器之间有多次的信讯往还。

鲁恩对处理情报检索者的输入问题有不同的看法。他建议情报检索者将他需要的情报写成短文，再将短文中选出的字和从文献中选出的字进行比较。当发现有足够的相似性时，就开始进行检索。这对于情报检索者从机器所贮存的情报中获得最大检索效果，是极有帮助的。

另一可能解决情报检索者输入问题的是利用“学习机”，这种机器可能“学习”做很多种工作。可以预言，将来一定会有计划的建立对情报检索者有足够讯息能力的情报检索系统，使得情报检索者能更好地指导机器检索。

检索作为“教学”的过程

假如机器能有助于情报检索者并指导他使用检索系统，机器必须“教育”情报检索者。这一观点提出了一系列新的要考虑的问题。

可以预言，将来某些情报检索系统不仅会提供线索或文献，并将远远超过这一范围。某些装置将帮助情报检索者挑选或阅读那些由机器所提供的文献。这一预言不是幻想，将来是可能实现的。

哈佛大学斯金纳（Skinner）的试验说明，利用机器来“教学”是可能的。他使用了机器方法来教写现代语言和大学数学。斯金纳机是使用人们预先写好的材料，机器的动作是根据固定的讯息而运转的。与此相同我们也可以使机器“运用”从文献上指定叙述词的技术。因此，发展这一“教学”机器，使机器本身能从检索出来的文献记录中检出情报单元，并且在情报检索者面前展出。

这就是说，对情报服务中心而言，贮存情报的机器就是检索装置，对情报检索者而言，则是“教学”装置。检索者所需情报的输入方法以及给出方法，就是情报检索者和机器之间的讯息往还。当然，情报检索者一旦发现对他特别有兴趣的情报，他就极愿意看到原始文献。这样就能做到使情报检索者选择他所愿意直接阅读的文献。

这种机器和翻译装置所固有的可能性结合起来，那么将来应用的范围还会扩大。当然人们可以期望将来的情报中心能提供翻译，将检索系统中提供的文献，从某种文字译成另一种文字。更进一步，还要求翻译机按需要的形式译出流利的口语式的译文。例如从德文译成英文，必须将德文句中置于句末的动词放到英文句中正确的位置上。可以预期，由“教学”机产生的情报单元输入到这样的机器中，它对这些情报单元进行处理，其一半的任务是文字翻译。这样得到的结果，在某种程度上将是可以接受的写成的文章。

如此，一个情报中心的来访者，当他叙述了对情报的要求后，例如规定要有一篇800字的关于日光对高分子变质的论文，所要求的理解能力不高于大学化学系毕业的水平。过一些时间，机器就能产生这样一篇论文。

大家知道，在技术文献以及许多期刊中都有重复的情报。为此，可以考虑在机器本身仅仅贮存新的情报。一个机器能仅仅贮存新的情报而且不是全文，就可避免重复同样的情报，

也就有可能增加机器的贮存能力。

机械检索还可进一步发展。例如科学家在实验室里将他新得到的成果或是接近新的成果直接输入机器中进行计算，校核其可接受程度，对照早期的成果进行改正，最后贮存起来，这是有可能的。那就代替了几乎仅以文献形式存在的科学资料，将来一部分资料可能是机器的形式。这种机器资料能被采用的唯一的条件将是要求机器能作出提要或个别重要章节。

将来这类的情报机，不要想象是一台巨大的机器中心，而是由大量的装置所构成的，类似于现在已经使用的大量的电子计算机装置。有大的也有小的情报机。某些机器相互间有内部的联系，而某些机器是独立工作的。在不同的时候，从一个或几个机器的记忆系统中，将这些记忆转录到磁带上来，含有大量情报的磁带记录能被转送合并到许多其它情报中心的记忆系统中去。

如果机器能和实验室的人员与成果相联系并直接贮存情报，那就可以期望，机器能够指出实验室人员在某一领域需要进行更进一步的试验。当然，要多久才能实现这类"主动工作"的机器，现在仍然难于估计。

（顾坚节译自《American Documentation》,1960, Vol.11, p.229—236）

（选自《综合科技动态》1964年第1期）