吉林省图书馆学会丛书之十一、十二

TUSHUGUAN YEWU ZIXUE DAQUAN

图书馆业务自学大全

（10）情报学导论

周文骏 编著

（11）计算机化情报检索基础

沈迪飞 编著

吉林省图书馆学会

计算机化情报检索基础

沈迪飞 编著

一九八〇年八月

业经吉林省出版局吉印字(80)46号文批准

编辑者：吉林省图书馆学会
出版者：吉林省图书馆
印刷者：长春市印刷厂
出版日期：一九八〇年八月
地　址：长春市吉林省图书馆
　　研究辅导部
电　话：22653
邮政编码：130021

印数：15,000册

前　言

本书以编者今年6月编写的《情报检索基础》讲义为兰本，经过修改加工而成。

这本书有两个目的。一是提供给图书情报工作人员，从手工文献检索角度，了解计算机检索的原理和方法，以及计算机检索和手工检索的关系与异同，从而为了解和从事情报检索工作打下基础；一是提供给将要从事情报检索的计算机专业人员，从计算机技术角度，了解文献检索的一般原理和方法，以及用计算机实现文献检索的基本路子，从而为编制情报检索软件打下基础。由于这样两个目的，书的内容只能涉及最基本的方面，所以取名为《计算机化情报检索基础》。

由于编者的水平所限和时间仓促，本书中资料的收集与归纳，概念的解释和问题的分析，以及内容的系统性都会存在很多的问题，只能供读者参考。错误之处，恳请指正。

沈迪飞

1980.8

目　　录

第一章 序　　言

1.1 情报和情报检索

1.1.1 情报和信息

情报（Information）是一个含义较广，在学术上又没有太弄清楚的一个词，加上外文之Information一词的理解与翻译不同，有时叫情报，有时又称信息，就更令人糊涂了。例如，Information industry一词，有人称为“信息工业”，有人又译为“情报工业”；有人讲，材料、能源与信息是现代社会的三大支柱，也有人将“信息”换为“情报”；有人讲“信息化社会”，又有人讲“情报化社会”。Information这个词涉及到哲学、控制论以及通信（交流）等多个领域，增加了理解的困难。

我们不必多加讨论这个问题，只是因为我们学习情报检索，它直接涉及的两个相关学科是情报学与计算机科学。从某种意义上讲计算机科学同信息科学是同义的，而情报学和信息科学外文是一个词(Information science)，怎么区分？另外，情报与信息两个词在本书中都是常用的最基本的概念，因之有必要讨论一下这两个概念的大体上的异同。

情报是具有保存、传递和加工价值的消息和知识，是人类之间的通信和交流。它是动态的，在传递中的；是有针对性的，是特定的，是有用的，只有当发生源发出的消息被吸收源通过某种形式“理解”时才能成为情报；似乎消息只有被人“接

受”与“有用”时才成为情报，情报与人密切不可分，情报有社会性的特点，只有能为人类服务的消息和知识才能成为情报，否则即使其客观存在，同一个消息也不会成为情报。

信息是指数据和消息中所含有的意义，它不随载荷它的物理设备形式的改变而改变。

可见，情报与信息有同义之处，它们均是指消息中所含有的意义，它们有共同的形式和载荷方式，有不少相同点。所以有人讲“情报就是信息”，有一定道理，在很多地方确实可以通用，可以互相代替。

是不是有不同之处呢？是不是可以讲“信息就是情报”呢？看来不行。“信息论”很显然不能叫“情报论”，计算机科学又称信息科学，能叫情报科学吗？“信息通道”、“信息学”中的“信息”就不能用“情报”来替换。客观存在的消息，只有对特定的人有用时才能成为情报；消息中所含的意义，如果对某人无用，对某人来讲仍然是信息，是对他无关的信息，但却不是情报。

所以，两者的不同点是否可以归结为：

① 学科不同。“信息”用于哲学、计算机科学和通信科学，而“情报”是用于情报学和图书馆学。在这一点上两者含义在学科上是不同的。

② 信息包括情报，广于情报。凡是情报都可以称为信息，但只有对特定的人有用的信息才是情报。情报与人的因素紧密相关，而信息与人的因素无关。

在本书中，一般来讲，这个词属情报方面应用时称为情报，而属计算机范畴时应称为信息。在后面的学习中大家体会一下，看是否可以。

1.1.2. 情报检索（INFORMATION RETRIEVAL）

对于情报检索，也有好多定义，兹列举如下：

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 根据用户提供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。

从上面五个定义中看出，彼此是有差别的。大体分二类：一类认为情报检索就是指情报的“存贮与检索”；一类认为既包括情报的“收集和存贮”，又包括情报的“检索和分发”。两种观点都有一定的道理，着眼角度不同。前一种从计算机科学出发，不包括情报的收集和分发，且“情报检索”这个词的起源也有助于这种观点。“情报检索”这个词是1949年 C.N.Moores 首先使用的，原文是 Information Retrieval 或 Information storage and Retrieval，从计算机角度译成汉语应是“信息检索”或“信息的存贮和检索”，重点在于计算机存贮和检索信息的理论与技术方面。第二种是从情报工作角度出发，包括情报的收集与分发。情报界使用这个词，其含义属后一种，它是情报工作的一个部分。

情报检索从广义来讲，包括内容是极广泛的，大体上包括事实检索，数据检索和文献检索三种，用图表示为：

可见，情报检索这个词应用的范围是很广的，但对于我

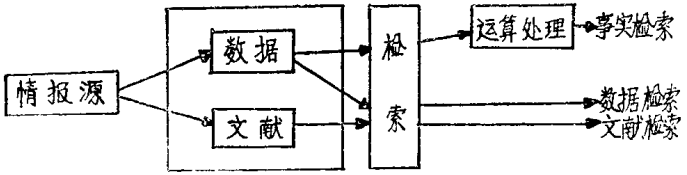


图1.1 情报检索种类

的图书情报部门来讲，情报检索主要指的是文献检索。

情报的收集和存贮，属“情报的管理”，情报的检索和分发是“情报的生产”，情报商品化了。情报的发生源到吸收源的直接传播，已不符合“信息化社会”的要求，作为中间媒介的情报检索系统就越来越发展了。

1.2 计算机化的文献检索——狭义情报检索

传统的文献检索，是在图书馆学目录学的基础上发展起来的，而计算机化的文献检索是传统的文献检索方法与现代化的计算机技术相结合的产物。在文献检索方面最早使用电子计算机的 A.Kent 说：“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。”

广义的情报检索是从文献检索发展起来的，最早的情报检索就是指计算机化的文献检索。随着计算机越来越广泛的应用，情报检索这个概念的含义就随之扩大了，使用同一个词有了广义与狭义之分。本课讲的情报检索是指①文献检索②

使用计算机，是计算机化的文献检索，是狭义的情报检索。

为使大家能先有一个感性认识，我们举一个情报检索提问与回答的实例。

用户提出的问题列成检索式如下：

EITHER { title words (breeze OR wind) AND (sea OR ocean OR waves) BUT author NOT (Jones,A.C) } OR[keywords (marine AND atmosphere)]

由计算机到相应的文献数据库检索，命中了三篇文献：

1. SMITH, P, EFFECT OF WIND ON OCEAN WAVES. J. MARINE PHYSICS, VOL. 6, 1969, PP.11～32.KEYWORDS : METEOROLOGY, OCEANOGRAPHY。

2. JONES, S. AND WILSON, A, RELATION BETWEEN WAVES AND WIND. PROC. SOCIETY FOR APPLIED PHYS, VOL. 14, 1970, PP. 58～60. KEYWORDS: PHYSICS, THEORY, MARINE, HISTORICAL.

3. DAWSON, L. P, STUDIES OF SEA AND AIR. J. ATMOSPHERIC RESEARCH, VOL. 3, 1969, PP. 104～112. KEYWORDS: MARINE, PHYSICS, ATMOSPHERE.

情报检索的全过程包括下面五个方面的内容：

① 情报的收集和选择；

② 情报的分析和加工；

③ 情报的存贮；

④ 情报的检索；

⑤ 情报的提供和评价。

下面画出这五个方面构成的示意图，以便有助于大家理解：

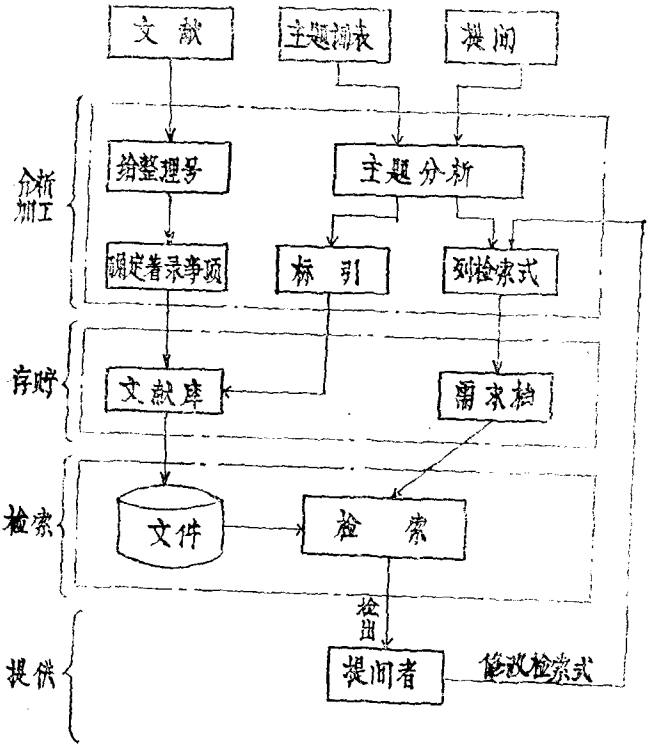


图1.2 情报检索流程图

作为一个新兴的学科，情报检索有其本身的特点：

① 情报检索是图书馆学、目录学、情报学同计算机科学交叉的学科。至今在检索系统的实践方面有了一定发展，但尚未建立起一套独有的系统的理论。它是借助相关学科的

理论和方法而逐步发展起来的，所以学习情报检索，除掌握目前已总结出的其自身的规律之外，要下功夫学习相关学科。其相关科学主要是：

图书情报方面的目录学、情报学以及主题法和分类法等；

计算机科学方面的数据结构、数据库管理系统、程序设计、模式识别、编译技术、操作系统以及计算机系统构成与计算机网络等；

数学方面的离散数学（布尔代数、集合论、图论、组合分析）、概率论以及线性代数等；

系统科学方面的系统分析和系统设计；

通信科学方面的信息论；

语言学方面的计算语言学等等。

② 情报检索侧重在情报是否切合需求者的要求，而不是对文献的直接阐述，因此情报检索的内容特征，除涉及到检索系统的实践，还包括信息量和切合性的测定与定义的理论方面。情报需求的切合性，由于对文献的情报内容本身和需求者的要求不能精确测定，所以这种切合性具有不确定性或模糊性的特点。这反映在检索效果的查全率与查准率互相矛盾的关系上，只能从模糊集合的角度求得最佳化的检索效果，这给检索的评价造成困难。

③ 由于文献信息的特点，使情报检索的数据处理形成了自身的特征：数据量大，大量频繁的数据输入输出，处理上多为字母数据的识别、存贮、重新排列和分类合并，大量的计算机内的数据传送和逻辑运算，在数据的使用上需长期保存和大量的不定时的查询等。在这个特点方面，情报检索相类似于统计通信理论，在大量存贮的数据中检索情报，有如在嘈杂纷乱的噪声环境中检测需要的信号脉冲一样，在许

多概念上都能互相对应。

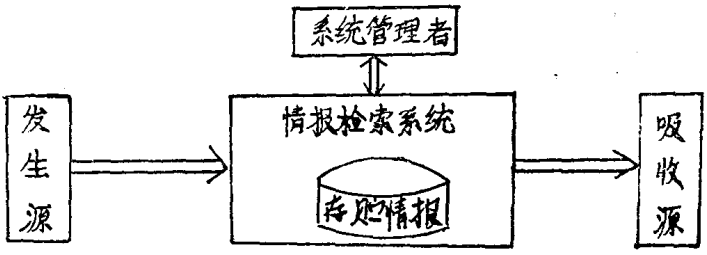
情报检索的这些特点就给硬件和软件提出了相应的要求。

④ 情报检索可以发展成为计算机学习和自适应系统。检索效果的测定是依据于用户对检出文献的切合性和未检出文献的不切合性的估价。如果检索过程依赖于一些参数，这些参数在系统内是变化的，某些参数可能是形成了最佳检索效果。这样，无论什么时候迁到类似需求时，计算机可以根据以往的经历记录，自动改变控制参数以保证最佳的效果，这就形成了计算机学习和自适应系统。现有情报检索系统中的“反馈”，实时修改检索式和主题的扩检等是一些初级的形式。检索式的自动修改和文献的自动分类等是名符其实计算机学习和自适应系统，这给情报检索的发展开创了更为美好的前景。

1.3 情报检索系统

情报检索系统是从情报发生源收集情报，并向情报的吸收源即利用者提供情报。其基本作用是对情报的发生源起吸收源作用，对情报的吸收源（利用者）起发生源的作用，使二者互相联系，成为情报发生源和情报吸收源的媒介。也可以讲，情报检索系统的作用，就是作为吸收源的总代表向各色各样的情报发生源收集情报，又作为发生源的总代表向各色各样的情报吸收源提供情报，使情报交流畅通。情报检索系统不仅要准确地检索和提供情报，并且要处理加工成利用者所要求的形式。为此，要使情报检索系统开展工作，还必须有系统管理者。

我们是作为系统管理者来学习情报检索的。



情报检索系统的基本功能是①情报的收集与存贮②情报的检索与提供③情报检索系统的管理。

一个情报检索系统由三个要素构成：硬件、软件和数据库，简称硬、软、库。硬件指计算机及其外围和通信、终端设备，由于情报检索的特点和各系统的规模，对硬件是有一定要求的。软件包括系统软件与应用软件，情报检索也有特定的要求。这些内容我们在第七章情报检索系统的建立中较详细讨论。数据库问题将在第三章中详细讨论。

随着情报的商品化，情报管理和情报生产大大社会化了。信息包括情报作为一种重要的资源，产生了“信息工业”。近些年来美国成立了庞大的“信息工业协会”（IIA—Information industry association）。其1978—79年会员指南上登记有117个会员单位。在今天这个“信息时代”，信息工业负起管理“情报爆炸”的任务，由生产者、分配者、传播者、管理者组成的信息工业，已经具有了其他产品工业的经济结构。现在，“信息同财经、人员、空间、材料一样，是一种具有价值的资源”。IIA从1971年起每年颁发信息产品奖。美国洛克希德（Lockheed）和系统发展公司（SDC）等情报检索系统都是IIA的成员。

1.4 情报检索发展简史

情报量的迅速增加与计算机技术的突飞猛进是情报检索发展的条件。

1．50年代到1964年脱机检索阶段：

54年美国海军兵器中心 NOTS，使用 IBM701 机，进行单元词组配检索输出文献号码。有的认为，情报检索最早使用计算机的是美国Western reserve大学的 J．W．Perry 和A.kent，他们开头用继电器为元件的WRU 选择器（selector）进行试验。

50年代大都是机械检索，使用穿孔卡、比孔卡、磁卡片等，出现了各种检索机器。1958年第二代电子计算机出现，60年起发展了计算机的租用方式，应用的范围大大发展了。

H．P．luhn 开创并发展了叙词法、自动文摘法和KWIC 索引法，1961年他首次用计算机为 CA 编出了KWIC(Keyword in context) 索引。

G.Salton 的全自动化情报检索系统（SMART）的研究，从理论和方法上推动了情报检索的发展。一方面计算机编制索引的系统发展起来了，并开始出售文献数据库，如：CAS (Chemical absrtacts service of the american chemical society)、CADRE 等。美国医学图书馆用计算机编排《Index medicus》，并在世界上首次使用光电照排机，由计算机控制进行自动排版。另一方面开展了批处理SDI 服务，美国医学图书馆的 MEDLARS (Medical literatureanalysis and retrieval system)、美国科学情报所 ISI(Institute for scientific information)、美国国防

部的 DDC(Defense documentation center for scientific and technical information从前叫 ASTIA)、美国国家航空及航天局NASA (National aeronautics and space administration)、以及英国 Derwent 出版公司，日本外务省等都开展了情报检索工作。

这时期在技术上是单机批处理，内存容量较小，外存用磁带，磁盘最大容量为29MB。

2．1965年到71年的联机检索阶段：

随着计算机分时系统 TSS (Time Sharing system)由实验阶段进入商业使用，开始有了一个主机带多个终端的联机系统。系列机 IBM360投入市场，大大加速了情报检索的发展，最早进入商用的是 Keydata 公司的 Tss 系统。

麻省理工学院的 TIP (Technical information project)是联机检索较早期的代表，1966年建成。当前很有名的大的情报检索系统在这时期纷纷建立，1970年建立了美国洛克希德(Lockheed)的 DIALOG 系统和美国系统发展公司 SDC (System Development Corporation)的 ORBIT系统，1971年 MEDLARS 发展了联机系统 MEDLINE (MEDLARS on—line)。在日本则有日本情报处理研究中心的联机检索系统JOLDOR (JIPDEC on—line document retrieval system)，电气通讯研究所的文献会话情报检索系统CIRES(Conversational information retrieval system on document) 等实验系统。这时期内存容量扩大了，磁盘为较大容量的 100MB，建立集中型的文献数据库。

3．1972年到现在的网络化检索阶段：

1968年美国 ARPA 网建成。70年代初期计算机网络进