计算机情报检索原理

姜希强

南京工学院图书情报专业
一九八五年八月

目 录

张庆国

第一章 计算机情报检索概述

1.1 情报与社会的发展

脱离了小生产的现代化社会，不论是生产和社会活动，还是个人生活，都离不开情报。“情报”这个概念的外延是非常广的，人们在科研、生产和生活中所需要的知识或消息都是情报。这些情报有的以独立的形式存在，但大多数以内容的形式存在于文献之中。不管它们的存在形式如何，在需要的时候，必须将其从众多的情报中查找出来，搜集、存贮这许多情报并在必要时找出其中所需部分的技术叫作情报检索（IR Information Retrieval）。

现代科学技术和工农业生产的迅速发展，使积累的情报数量急剧增加。单以科技文献为例，目前已达到“浩如烟海”的地步。而且还继续以更快的速度增长着。据统计，《化学文摘》（CA）发表第一个100万条文摘用了三十年时间，第二个100万条用了十八年，第三个100万条用了八年，第四个100万条只用了四年，而第五个100万条却不到三年，1973年一年就出版了356459条文摘，到1982年，累积已达到1000万条。图1.1是化学文摘的增长曲线。

另一方面，现代科研和生产对适用情报的要求也越来越高了。科学逐渐分化，向纵深发展，科学家和生产者的领域越来越广。任何人都不可能通晓一切，研究一切。研究一个数学课题时，需要的情报决不是笼统的“数学”，而只是“数学”中一个极小极窄的方面。综合性学科的研究同样如此。对于某一情报用户来说，其适用情报在全部情报中的比例已极小极小，并越来越小。

总之，情报（适用情报）的检索越来越困难了，计算机情报检索便是在这种社会背景下应运产生的。

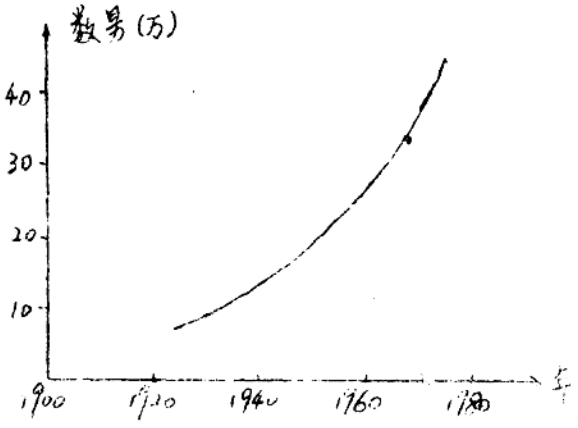

图 1.1　化学文摘的增长曲线

1.2　情报检索与文献检索

按照检索对象的不同，情报检索有以下三种类型。

1. 文献检索

需要查找的对象是科技文献。严格地说，这种类型并不属于“情报检索”，因为情报并不是文献，而只是文献中的部分内容。但由于习惯的原因，我们仍称其为情报检索。目前的大多数情报检索系统实际上也只是文献检索系统，或称书目检索系统。这主要有以下几个原因。

(1) 需要性。科技人员最终需要的是情报，但是，他们直接需要的却往往是文献。从文献中获取最终情报的工作，他们宁愿自己去做。

(2) 易行性。文献检索系统的建立和运行要比真正的情报检索

系统容易。

2. 数据检索

检索的结果是数据，如电话查号，查特性数据，订票系统，查银行帐目等。

数据检索可以通过文献检索系统实现，但更多也更好地是通过数据库。

3. 事实检索

一般要从贮存的情报（数据）中检索出必要的部分后，再加以逻辑推理才能给出答案。事实检索系统已带有智能因素。较强功能的事实检索系统是非常复杂和难以建立的，它们的发展水平目前还很低。

本课程主要只讨论文献情报检索。

1.3 文献情报检索系统的基本功能

图 1.2 是计算机文献情报检索系统基本功能的简略示意图。

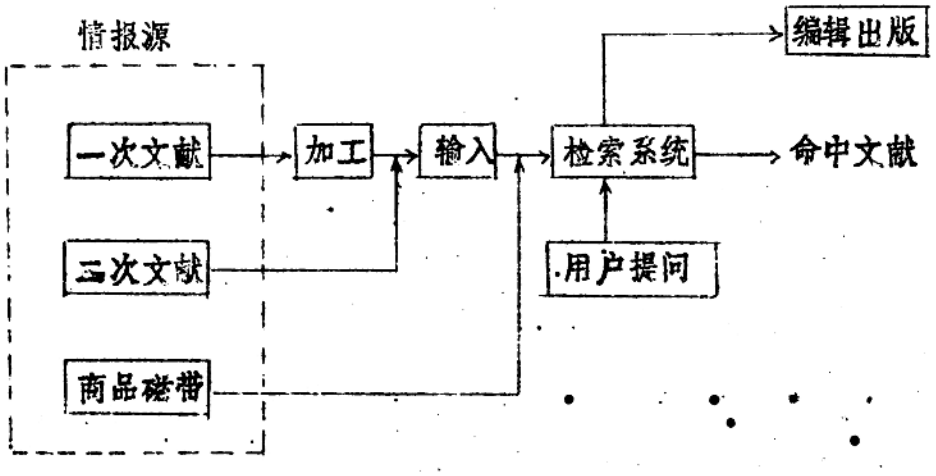

图 1.2 计算机文献情报检索系统

图中，情报源部分一般包括三种情报来源：

一次文献。一次文献要进入系统，要经过以标引为主的加工工作，这一工作目前主要还是人工进行的。标引完之后，可穿孔输入。

二次文献。有系统之外的机构承担了一次文献向二次文献的转换工作。如二次文献中已有标引词，直接输入即可。

商品磁带。是其他计算机情报检索系统的产品，本身已是机器可读形式，只要通过计算机合并进自己的文献数据库即可。

编辑出版是某些计算机情报检索系统的例行工作，其产品为整个社会服务（而不仅仅是为本系统的用户服务），二次文献和商品磁带可以是其他系统的情报源。

检索服务是在用户提问的启动下进行的，通过对文献数据库进行检索，提供给用户以适用的情报。

1.4 文献情报检索系统的基本原理

1.4.1 文献与文献标识

无论是机械的还是电子的文献检索系统，都是建立在文献标识及其代码化的基础上的，因为只有对文献进行标识并将这种标识代码化，才能由机械的或电子的方式加以识别判断。显而易见，描述文献主题最自然的方式是以主题词作为文献标识。用主题词标识文献的工作过程称为“标引”。

例如，本讲义的主题内容可以用“计算机”和“情报检索”这两个词来进行标识，这两个词便成为本讲义的标引词。

用来标识一篇特定文献的标引词可以来自文献篇名，也可以来自文献的正文或文摘，或同时来自它们。当文献内部没有合适的词能标识该篇文献的主题时，标引词还可以来自文献外部，如主题词

典。

文献的标识用词在标引时称为“标引词”；在检索时，同样是用这些词汇检索。只不过这时称它们为“检索词”。标识用词的其他名称还有“主题词”，“标识词”，“关键词”等等。在本书中，我们将不加区别地使用它们。

用主题词标识文献的直接目的是为了使机器能够识别文献。然而，对文献进行标识和存贮的最终目的是为了检索。因此，对这些主题词，这些主题词之间的关系以及用这些主题词标识文献时所遵循的规则有研究的必要。上述主题词，词间关系和标识规则在研究和限定之后成为一种语言——检索语言。有效的检索语言能保证标引者和检索者的一致，保证存贮进系统的文献记录能够被检索出来。

在某一个（若干个）学科内，或者对某一特定的检索系统，可以将其所用检索语言的外延明确化，即制订一检索语言实体。这样的检索语言实体称之为“主题词典”。主题词典仍然包括：

① 经过选择规范化了的主题词汇。主题词汇一般包括正式主题词和非正式主题词两种，其中非正式主题词在使用时要转换成正式主题词。

② 词间关系，一般指词与词之间的用、代、属、分、参等关系。

③ 使用规则，如组配规则等。

在计算机情报检索系统中，主题词典可以是书本式的，也可以存贮在计算机内。前者一般作为标引和编写提问单时的共同依据。后者则可随机查询，随时解决标引或检索中与词汇有关的问题。

用主题词典明确化了的语言称之为规范化检索语言，与之相对的是自由检索语言。自由检索语言中使用的词汇称之为自由词，并

且各种规定也少得多。自由检索语言的优点是，方便标引者和检索者；主要缺点是，标引与标引，标引与检索之间的一致性较差，系统工作难度较大。目前世界上运行的检索系统，有使用规范词的，有使用自由词的，也有同时使用两种词汇的。

1.4.2 文献——语词矩阵

在文献用其标识化的形式——标引词表示之后，我们可以认为，对于一个有N篇文献，M个标引词的文献检索系统，存在着如下一个N×M矩阵

$$D=\begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ D_N \end{array}\begin{array}{c} \begin{array}{cccc} T_1 & T_2 & \cdots\cdots & T_M \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \cdots\cdots & a_{1M} \\ a_{21} & a_{22} & \cdots\cdots & a_{2M} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ a_{N1} & a_{N2} & \cdots\cdots & a_{NM} \end{pmatrix} \end{array} \qquad (1.1)$$

其中a_{ij}表示第j个标引词T_j对第i篇文献D_i的适用程度，称为T_j对D_i的权值。于是，矩阵D的每一行表示一篇文献与各标引词的关系，每一列表示一个标引词与各篇文献的关系。

权值a_{ij}的取值可以是离散的，如指定为0，0.25，0.5，0.75，1中的一个，也可以是连续的任意实数，但取连续数值时，一般都限制在一个范围内，如规定在开区间（0，1）内取值。

a_{ij}有一种非常有用的取值方法，即取离散值0和1，具体取法是，当文献D_i用标引词T_j标引时，$a_{ij}=1$，当D_i未用T_j标引时，$a_{ij}=0$，这种方法把文献与标引词的关系限制为“完全相关”或“完全无关”两种情况，因此它在很大程度上简化了文献与标引词之间的关系，不过a_{ij}的这种取值方法给文献——

语词矩阵的运算以及计算机文献情报检索中的许多过程带来了便利，因此是一种重要方法。

例如，有三篇文献 D_1、D_2、D_3 和四个标引词 T_1、T_2、T_3、T_4，其中 D_1 用标引词 T_1 和 T_3 标引过，D_2 用 T_2 标引过，D_3 用 T_2、T_3、T_4 标引过，权值取离散值 0 和 1，于是相应的文献——语词矩阵如下：

$$D=\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{pmatrix} \qquad (1.2)$$

a_{ij} 取离散值 0 和 1 时，$\sum_{i=1}^{N} a_{ij}$ 表示用标引词 T_j 标引过的文献篇数，$\sum_{j=1}^{M} a_{ij}$ 表示标引过文献 D_i 的标引词个数。

对于一个较大规模的文献情报检索系统，D 的体积一般是很大的，通常不可能存贮这样一个矩阵的原形，同时也无必要，因为 D 是一个稀疏矩阵，其中有大量的零元素。在计算机技术中，稀疏矩阵有很多很经济的存贮方法，我们根据文献情报检索的需要，可以选用一些。

文献——语词矩阵是理解计算机文献情报检索的基础，很多重要的检索机制都可以认为是从这个矩阵导出的。我们以后会经常用到这个矩阵。

1.4.3 三种基本的文献检索方式

1. 顺序检索

在文献——语词矩阵中从上到下地检查每一行（每篇文献），逐篇判断文献是否满足用户的提问要求。

顺序检索是一种比较简单但也比较原始的检索方式，在历史上曾经起过重要作用，但现在已逐渐不被使用，其原因一个是效率太低，文献数据库中有大量文献记录，而其中符合特定用户要求的部分极小；另一个原因是，顺序检索的盛行主要是因为磁带在计算机存贮设备发展历史上一段时期内的重要地位。随着近年来磁盘等随机存取设备的发展，人们已摆脱了顺序存取方式的限制。

不过顺序检索在现在也还有一定地位，如用于SDI。在SDI中，每批文献（新到文献）的数量不是太多的，而每次检索又可以针对多个提问同时进行，即每扫描到一篇文献，检查它与多个提问的满足与否，这样效率还是比较高的。

顺序检索的匹配算法很多，其中较有影响的是日本人菊池敏典在一九六八年提出的"表变换法"，近年来还出现了一些其他算法，但我们在本课程中将不对顺序检索方式作更多的讨论。

2 倒排检索

用户的提问往往由多个因素组成，具体地说，用户的检索提问式往往由多个检索词及一些词间逻辑关系组成，这些检索词和词间逻辑关系共同构成了用户提问的复合主题概念，在检索时，我们可以先对检索提问式进行分解，析出每个因素（检索词），然后根据文献——语词矩阵中相应各列的情况，找出分别满足这些检索词的文献集合，最后根据原检索提问式中的词间关系，经过文献集合之间的逻辑运算，确定满足用户提问的文献集合

例如，用户要求检索同时具有标引词T_2和T_3的文献，在文献——语词矩阵（1—2）中，相应的两列分别是

$$\begin{matrix} T_2 \\ \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \end{matrix} \qquad \begin{matrix} T_3 \\ \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \end{matrix}$$

这表示，满足检索词 T_2 的文献有 D_2 和 D_3。我们能够很容易地判断出来。同时满足这两个检索词的文献只有 D_3 一篇。我们再检查文献——语词矩阵（1—2）中的 D_3 行，便可以知道 D_3 这篇文献的各标引词情况，从而知道这篇文献的主题内容。

倒排检索的过程是很经济的，首先，它只访问文献——语词矩阵中与检索词相对应的列，这样的列的个数是有限的；其次，它只访问文献——语词矩阵中与命中文献相对应的行，这样的行的个数也是有限的。因此，倒排检索是一种效率较高的检索方式。

倒排检索也是目前技术最成熟，应用最广泛的检索方式。现在世界上运行的商业性检索系统几乎都是基于倒排检索原理的。在本课程中，我们将对倒排检索方式作重点介绍。

3. 聚类检索

在文献——语词矩阵中，每一行表示了一篇文献与各标引词之间的关系，这种关系是通过各标引词对该篇文献的权值给出的。比较不同的行，我们可以考察行与行（文献与文献）之间的相似程度，并可以通过数学方法把这种相似程度表现为匹配度值。匹配度值说明了文献与文献之间的相似情况，因之我们可以把文献聚成主题相近，内容相关的类。把用户提问也表示成文献——语词矩阵中行的形式，我们便可以计算提问与文献之间的匹配度值，并把匹配度值较高的文献作为命中文献。

聚类检索技术目前还不十分成熟，但它至少在以下两点上优于倒

倒排检索。

1. 倒排检索是一种“完全匹配”方式，一篇文献只有“命中”和“不命中”两种情况。而聚类检索中，文献与用户的关系却是通过“适用程度”来刻划的，后者似乎要更符合文献情报检索的本质。

2. 对于改善检索效果过程的自动化程度，聚类检索有较深的潜力。

本课程将以一定篇幅对聚类检索及一些有关问题作些介绍。

1.5 联机情报检索

情报检索系统的发展主要取决于情报存贮手段的发展，可分为以磁带为基础的脱机处理方式和以磁盘为基础的联机处理方式。

脱机情报检索在六十年代占主导地位，那时，磁盘之类的随机存取存贮设备还未得到有效的发展，文献资料记录主要存贮在磁带上，因为磁带的检索效率较低，所以，一方面不能在用户容忍的时间内对检索作出回答，另一方面也不可能（得不偿失）为每一个提问而将整个文献数据库扫描一遍，因此提问多是委托式的，系统在积累了一批用户的提问之后，一次性地在文献数据库中检索，然后将检索结果分别送给各用户。

七十年代及其之后，由于计算机技术的发展，尤其是磁盘等随机存取设备的迅速发展，联机检索方式在情报检索中盛行起来。联机检索方式是指，检索者可以同他想访问的文献数据库及负载该数据库的计算机进行直接的通信（联机），实时提出检索要求，实时得到回答，并且在检索过程中，检索者与系统可随时交换信息。因此，联机系统也叫作交互式或会话式系统。

脱机检索系统有几个缺点，①检索者必须预先制订好检索策略，