目 录

张庆国

第四章　检索效果及其改善

4.1　检索效果及其测量指标

计算机情报检索系统的性能可以从两方面衡量。一个是检索效率，一个是检索效果。检索效率通常指的是检索过程中系统的时间、空间及其他资源的耗用相对于检索结果的比值，用户在检索中耗费的精力和其他的代价等等。对效率的考虑是情报检索之所以要用计算机来进行的主要原因之一，也是我们迄今为止讨论的主要问题，典型内容是我们对查找技术的讨论。

检索效果则不同。检索效果指的是检索结果对于用户的满意程度，是指检索行为是否趋向检索目标，是否符合情报提问，以及达到（或背离）检索目标的程度，它相应于英文中的effectiveness，该词又可译为“有效性”，并且这里的“有效”是相对于用户对情报的最终使用而言的。

检索工作是在特定的检索系统中进行的，因此我们不可能离开特定的检索系统来绝对地考察检索结果（检出文献）对用户的满足程度，因为有些问题的解决是系统及其使用者能力范围之外的。通常，对检索效果的评价从以下几个量出发。

a．检出文献中的适用文献数；

b．检出文献中的不适用文献数；（误检数）

c．系统中的未检出文献中的适用文献数（漏检数）；

d．系统中的未检出文献中的不适用文献数；

由上述四个量，可给出评价检索效果的若干指标。

$$\text{查全率 } R = \frac{\text{检出的适用文献量}}{\text{库中存贮的适用文献量}} = \frac{a}{a+c}$$

$$\text{查准率 } P = \frac{\text{检出的适用文献量}}{\text{检出文献量}} = \frac{a}{a+b}$$

$$\text{误检率} = \frac{\text{检出的不适用文献量}}{\text{检出文献量}} = \frac{b}{a+b}$$

$$\text{漏检率} = \frac{\text{未检出的适用文献量}}{\text{库中贮存的适用文献量}} = \frac{c}{a+c}$$

显然，漏检率＝1－查全率，误检率＝1－查准率，因此上述四个指标不是互相独立的，漏检率和误检率可分别从查全率和查准率导出，并且，后者也更直接地说明检索效果，因此，在检索效果的评价中，一般只以查全率和查准率为主要指标。

最理想的情况当然是，对于一次检索作业，检出的文献都是适用的，同时，适用的文献也都被检出，即 $b = c = 0$，这时有 $a = a + c$，$a = a + b$，即查全率 R ＝查准率 $P = 1$，但遗憾的是，不仅这种最理想的情况几乎不会出现，而且连查全率和查准率同时较高的情况也较少出现，实际工作中往往发现，查全率和查准率经常是成反比的。

这是因为在情报存贮和检索的过程中存在着一系列不可避免的误差而导致的。

首先看标引，标引工作把一篇文献转换成代表该文献主题的若干个主题词，在标引人员的精心工作下，这若干个主题词对于原文献来讲不能说是词不达意的，但另一方面，如果认为这若干个语词

全面、准确、客观地反映了原文献的全部内容（或基本内容、主要内容），那也是不对的。一篇文献有其丰富生动的内容，很难（或不可能）把它概括成若干个抽象的概念。也很难（或不可能）找到几个恰如其份地表达这些概念的语词。再者，原文献中的若干个主题往往有着复杂生动的联系和各自不同的重要程度，被转换成若干个标引词后，这些联系和重要程度通常不被反映出来或不能很好地反映出来。还有，对文献的标引质量还受着多种因素的影响，如标引员对该文献主题学科的了解程度，对该课题的熟悉情况，对文献的熟悉情况，对语言（尤其是专业语言和外语）的驾驭能力，词表的质量，对词表的了解，标引员本身的知识水平，标引时的工作条件，甚至标引员在工作时的情绪等等。总之，标引结果（若干个标引词）只是在一定程度上反映了原文献，而不是全部，准确。

其次，情报需求的表达与标引工作类似。情报的需求，对于情报用户来说，是有其丰富的内涵和外延的，但检索工作却只能通过检索词来进行，因此，情报提问式（它由具体的检索词及其相互关系式组成）对情报需求的反映也只是在一定程度上。

最后，在文献与文献用户之间的关系上也存在问题。文献的作者是为整个社会（或社会的一部分）而撰写文献的，他并未考虑特定用户的特定需求。标引工作同样如此。因此，文献与文献用户之间存在着错综复杂的关系。某一篇特定文献可能为多个用户所需要，某一特定用户可能需要多篇文献。即使是对一个用户需要的一篇文献来说，这种“需要”和该篇文献所提供的材料之间的关系也是非常复杂的。该用户可能需要该篇文献的许多内容（或全部内容），可能只需要该篇文献的很少内容（或极少内容），用户的需要点可能与文献的论述重点相同，也可能不同或有着各种各样的关

该用户对该篇文献的需求程度与用户和文献作者各自的工作特点，知识结构，工作环境，时间，地点，研究问题的角度，深度和广度等都有关系，并且这些关系是没有一个固定模式的。

由此可见，在实际检索过程中，漏检是不可避免的，误检也是不可避免的，这些误差并不是因为计算机系统和检索机制本身的错误（例如，倒排档检索机制本身是很严密的），而是因为上述种种不可避免的不协调。

在漏检的情况下，用户往往要求继续检出原来未检出的适用文献，但由于检索机制的严密性，如果保持原来的检索策略，那么再次检出的结果将毫无疑问地与上次相同。欲检出（相对于上次）更多的适用文献，必须调整检索策略（如选用较宽泛的检索词），但仍然由于误检的不可能，在检出更多适用文献的同时，也检出了更多的不适用文献。因此，查全率的增加与查准率的降低是同时的。

在误检的情况下，用户也必须调整检索策略。同样的道理，在少检出不适用文献的同时，也少检出了适用文献，因此，查准率上升，查全率下降。

上述规律可用下图说明：

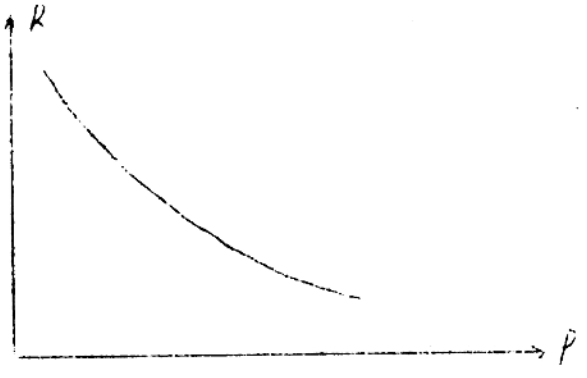


图 4.1　查全率（R）查准率 P 互逆相关关系

上述规律不仅在理论上成立，在实际检索工作中也有验证。Lancaster曾将在50个提问的检索中获得的查全率与查准率互相对比地标在图中，每次检索均用A、B、C、D四个不同的等级进行。（图4．2）

一般检索系统的平均查全率为60—70%，查准率为40—50%，同时达到百分之百的查全率和查准率的情况尚未在大规模的检索中发现，并且，至少在理论上它是很难出现的。

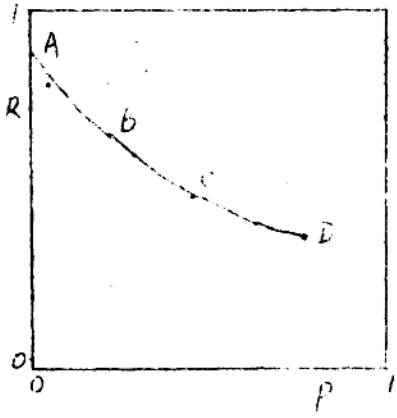


图4．2　五十个检索的平均查全率和查准率

但是，人们对于查全率和查准率的互逆相关关系也不是完全无能为力的。迄今为止，人们已经就为详尽地分析了影响检索效果的各种因素，并制订了各种措施，以尽量改善检索效果，使查全率和查准率得到兼顾。

本章其余部分将介绍一些改善检索效果的方法，这些方法是传统的、定性的，它们在国内外普遍应用，并有较好的效果。第七章介绍改善检索效果的一些定量和自动方法，它们是目前国外正在研究的一些实验性方法。

4．2　影响检索效果的主要因素

4.2.1 情报提问对情报需求的表达程度

在情报提问与情报需求之间可能存在着种种差距。如下表所示，表中N是情报需求，Q是情报提问

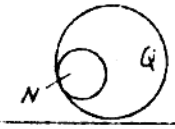
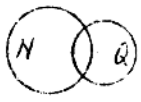
表达的差距	失误原因	失误状况
	提问比需求更专	查全失误
	提问比需求泛	查准失误
	提问偏离需求	查全与查准都失误
	提问与需求无关	检索效果不好
	提问接近需求	检索效果较好

表 4.1 情报提问与情报需求的表达差距

情报需求向情报提问的转化是在用户与系统的交互接口上实现的。影响这一接口功能的主要因素是：

1. 用户明确表达出情报需求即形成情报提问的能力。这与用户的专业知识与情报素养及其以往进行检索所获得的经验有关。

2. 系统与用户的交互方式。即用户将自己的需求向系统提出“协商”地处理的质量。其中包含检索者与用户交谈协商的方式及其工作态度等；

3. 用户对检索系统能力的估计。即用户不问化真正想要的东西，而提出要求检索他认为系统所能给予他的东西；

4. 系统所能提供的辅助检索的措施。如让用户参阅检索词表、用户手册、SDI定题须知、联机中的测览、词表显示等等；

5. 检索服务的方式。即采用委托式检索还是人机对话式检索。

根据用户的情报需求和情报提问的特点以及系统所用的检索方法，我们可以将情报提问分为下列几种类型：

1. 已知检索入口点的情报提问

对于这类检索课题，用户常常能提出他已经知道的某一位著者、某一篇相关文献、某一种分子式等等，以用户提出的已知线索为起点，不断扩充检索，判断检索的深度和广度，最后达到用户的检索目标为止。

2. 主题范围明确的情报提问

这类检索课题，用户常常没有提出任何已知线索，但能够明确表达出他们的情报需求范围。这可以从主题途径检索，用检索词进行逻辑组配，使检索式表达的内容与情报提问的内容相符，或者用分类号辅助检索，最后再筛去无关文献。

3. 检索范围不确定的情报提问

这种检索课题或者是因为用户的情报需求尚不明确，或者是因为缺乏检索知识，需要系统提供一些辅助检索手段，给检索者以启发帮助。这样在检索过程中，尤其是联机时的词表显示、文献测览、判别检出文献等步骤，将使用户模糊的情报需求逐渐明确，检索范

围也就会逐步确定。

4.2.2 数据库的选择和比较

文献数据库是检索工作赖以进行的物质基础。在检索系统能够提供的众多文献数据库中，如果选择错误，则无论检索工作怎样精心，也不能取得满意的效果。

选择数据库的主要依据是用户的检索课题的主题内容及其检索要求，这与用户和检索者把他们有关数据库的知识同他们理解的情报提问进行匹配的能力有关。选择数据库的方法常用经验判别法和试验反馈法。具体选择时，主要考虑下列准则。

1. 数据库的可检索性。

2. 数据库的收录范围及特点

这是判别数据库检索价值的因素，要了解的问题包括数据库摘录的文献种数，来源和类型，数据库中资料的数量，数据库跨越的时间范围，与其他数据库相比较而言的独特性和交叉性。

3. 数据库的标引与词表的因素

包括数据库所用词表的控制程度，即该数据库的检索是用规范词还是自由词，是严格控制还是稍加控制，词表中的词的专指度，标引的网罗度，确切性与一致性，所提供的检索辅助手段。

4. 数据库的大小与增长率及其更新周期

各个数据库的规模相差十分悬殊，如化学文摘等数据库的记录已达数百万条，它被输入DIALOG数据库时就被分成5个文档。而有的高度专业化的数据库可能只有几千到几万条记录。因此要注意比较数据库输入记录的总量，起止年代，每条记录的平均长度，数据库中记录的增长率，数据库的出版周期和更新周期。通过上述项

目的比较。可判断数据库提供追溯检索的能力。判断数据库提供最新情报的速度。

5. 检索费用。包括检索数据库中每条记录或每个可检项目的费用。注意比较使用各个相关数据库价格的差别。

在国外的一些大型情报检索系统，如DIALOG和ORBIT系统，为了帮助检索者对系统所拥有的文档进行选择，在系统中编制有一种“介绍性数据库”，应注意利用。

我国现在已经从国外引进了几十种文献磁带，并且已建成相应的机读数据库来提供给广大用户进行计算机检索。在使用这些系统时，也要参考有关的指导性工具书。

4.2.3 检索途径的选择

主要根据检索课题的要求和待检数据库中的可检项目。一般可分为主题途径和非主题途径。主题途径是最常使用的，它通过检索词、分类号等反映文献主题内容特征的可检索项目来查找。非主题途径则是对反映文献外表特性的一些可检项，如著者、著者所属机构、期刊代码、文种、报告号、专利号等可检项目进行查找。

对于不同的数据库，其提供的可检项目是不完全一样的，具体选择时要根据所检数据库的特点来决定。

在实际检索中，常常将主题途径与非主题途径结合起来使用。

4.2.4 检索词的选择与调节

主要指主题词的选择和调节。主题检索是最重要的检索途径，而在主题检索的检索式中，主题词又是最基本和最重要的成份，因此，对主题词的选择具有重要意义。

主题词是用来表达所要检索的主题概念的，在主题词和主题概念之间，同样存在着表 4 —1 所示的各种差距关系。此外，由于主题词与词表直接有关，故选词也受词表的制约。选词当中经常出现的一些失误如下：

1. 所选检索词与词表不符合。这是最常见的选词失误现象。这类情况经过检索者对照词表或印刷本文摘即可修正。

2. 所选的检索词太专指，致使检不出相应的文献。这常需要将其扩充到上位词或增加相关词。

3. 选词太泛指，不能准确地表达用户情报提问的实质，因而检出的无关文献较多，查准率低。这时需要选用专指度较高的词，或对原检索词或检索式加以各种条件限制。

4. 选词不全面，不准确。很多检索词之间具有语义相关关系，忽略这种关系，就会忽略一些应该有的检索词，从而造成检索结果的不全面。要避免这种情况，除了要求检索者具有较多的专业词汇知识外，还需要系统提供一些辅助选词的措施。

除了上述常见的几类选词失误外，还有关于检索词的单、复数处理不当及截词处理不当产生的失误等等。

一般性地说来，选择检索词的前提是对情报提问概念分析的正确性和全面性。这要求不仅从字面上析词，更重要的是从词的含义上析义，注意所选检索词的全面性、专指性、一致性，并要考虑下列选择词的基本准则：

1. 要从词表规定的专业范围出发，选用各学科内具有检索价值的基本词汇；

2. 选词要适应待检数据库的检索用词规则；

3. 尽多选常用的基本词汇进行组配。

4.2.5 检索式的结构

根据检索课题的要求及其词间关系，选定了检索词之后，就可以用布尔逻辑算符和一些检索指令将提问中各有关概念之间的关系表达为布尔检索式。布尔检索式是用户（及检索者）向系统提交的请求信息的直接形式和最终形式。

检索式中的逻辑运算符主要有A N D、O R和N O T。容易理解，两个用A N D连接的概念是各自对对方施以外延限制，而两个用O R连接的概念则是各自对对方以外延扩展。因此，从理论上说，某些概念复杂的课题，用A N D越多，限制条件越多，查准率也越高。但是，由于检索前并不知所检文献的具体特征，加上考虑有标引失误、选词失误等因素影响，所以查全率一般在这种情况下是不会太高的。用户宁愿采取的方法是，多用一些O R，先取得较多的适用文献，然后再用各种辅助方法筛去其中的无关文献。

为了合理安排检索式中A N D和O R的比例，先明确用户对查全率和查准率的要求是有好处的。我们可要求用户根据下列几种情况对他的检索目标给予必要的说明。

1. 要求高查全率，希望获得所有的相关文献。这一类用户属于普查类型，主要为编写教材、综述、从事基础理论研究或应用理论研究等目的而收集资料。

2. 要求高查准率，希望有一定范围的文献量，不限定篇数，但不希望有误检。这一类用户属于攻关类型，他们要在科研或生产中解决某一关键问题，只要求检出某一主题某一方面的资料，对文献和相关性要求高，而不一定要求检出的文献量很大。

3. 一般性的要求，希望有一定比例的相关文献，不具体限定