计算机情报检索知识丛书③

计算机检索策略和检索效果

孙　凌

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
一九八四年五月

计算机检索知识丛书③

计算机检索策略和检索效果

孙　　凌

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
一九八四年五月

前　言

任何情报检索系统都会遇到检索策略和检索效果的问题，计算机检索策略和检索效果息息相关。检索策略构造的优劣程度直接影响到相关文献的查全率和查准率，关系到能否满足用户的情报需求，关系到情报检索系统的服务效果和系统的生命力。

为了满足广大科技人员和情报工作者学习计算机情报检索的基本知识的社会需要，作者比较通俗地介绍了有关计算机检索策略和检索效果的一般知识。本书特点是结合国情，分析影响检索决策的各种因素及其相互联系，在此基础上论述构造情报检索策略的动态过程与反馈模式，重点分析其中的几个主要环节，并通过实例归纳能提高检索效果的若干调节方法和反馈途径，以促进计算机检索工作更好地满足用户的情报需求，从而更有效地利用浩如烟海的情报资源为四化建设服务。

由于编写者水平所限，加上时间匆促和印刷条件的限制，书中定存许多不足之处，恳请读者不吝指正。

孙　凌

1984年3月于武汉大学

目　　录

第一章　检索决策中的相互关系

§1.1 用户的情报需求与情报提问

尽可能满足用户的情报需求，不仅是情报检索系统运转的动力，也是评价检索系统服务效果的一种准则。我们研究检索策略的主要目的之一也在于尽可能全面而又准确地检出能满足用户情报需求的文献，所以我们先从用户的情报需求与情报提问进行分析。

情报检索系统的用户群常常有科研人员、工程技术人员、教学人员、医务人员以及科技管理人员等等。不同的用户有不同的情报需求，这些不同的检索课题所涉及的学科与专业也各不相同。即使是同一用户，在不同时期也有不同的情报需求。各种类型的检索课题对检出文献的需求是各不相同的。它们有广度上的需求，也有深度上的需求；有现实的需求，也有潜在的需求。从所需求的文献类型看来，它们对会议论文、专利、研究报告及期刊文献等的需要也各有侧重。根据各类用户的各种情报需求以及在情报检索过程中的一些特点，我们可以将用户的情报需求大概分为以下几种类型。

1.攻关类型。

对一些在科研或生产中需要解决某一关键问题的用户，他们往往只要求检出某一主题某一方面的情报资料，这类用户要求检索结果能解决他们的关键问题，要求查准率较高，不一定要求检出的文献量很大。

2.普查类型

对一些编写教材，从事基础理论或应用理论研究的用户，他们往往需要全面系统地收集某一主题范围的文献资料。这类用户要求较高的查全率，他们的情报需求带有横向普查，纵向追溯的特点。

3.探索类型

对一些选择新课题与应用新技术的用户，他们往往需要了解和掌握国内外的最新动态或研究成果，这类用户对查全率和查准率不一定有很高的要求，但他们要求所提供的情报要新而且及时。

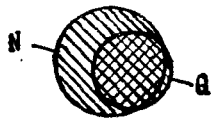
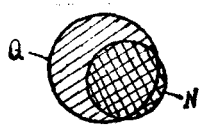
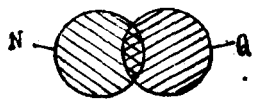
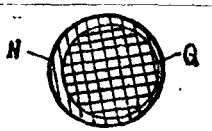
因为情报检索系统不能直接对用户的情报需求给以响应，只能对用户已经表达出来的情报需求，即对用户已经向检索系统提出的情报提问给以响应，如果要想获得成功的检索，情报需求与情报提问必须相当接近，但一般的用户并非总能完全确切地向检索系统表达出他们的情报需求。为了简明起见，本书将情报需求与情报提问之间可能存在的差距状态及其对检索效果的影响列于表一。

关于如何减少情报需求转化为情报提问时的差距问题，这一方面牵涉到用户能否将自己的情报需求充分准确地表达出来的能力，另一方面也关系到对于用户已表达出来的情报需求，系统的检索者能否全面正确地理解并且用系统的检索语言再完整地表达出来的能力。具体说来，在形成提问阶段，影响情报需求转化为明确的情报提问，影响用户与系统的交互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.系统与用户交互的方式。即用户将自己的需求向系统提

表一　情报提问与情报需求的表达差距

表达的差距	失误的原因	失误状况
情报需求  情报提问	提问比需求更专	主要表现为查全失误
 	提问比需求更泛	主要表现为查准失误
 	提问偏离要求	查全失误与查准失误都较大
 	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

出“协商”地处理程序的质量，其中包括检索者与用户交谈协商的方式及其工作态度等等。

3.用户对检索系统能力的估计。即用户不问他真正想要的东西，而提出要求检索的是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，联机中的浏览，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一机对话式检索。

在将情报需求转化为情报提问的过程中，不论是委托式检索还是人一机对话式检索，其中都有一个情报传递并再现的对话过程。在这个过程中可能遇到某些智能困难，因此，检索系统应该有一些辅助检索的措施。例如提供用户手册，详细介绍检索系统的特点与功能，包括检索操作指令，基本的检索技术，所能提供检索的数据库或文档的收录范围，所用词表的标引特点，词表结构的说明以及有助于用户将情报需求转化为情报提问的启发式的情报提问登记表等等。

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，本书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功能，需要检索者给以启发帮助。这样在检索过程中，尤其是联机时的词表显示，文献浏览，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。

§ 1.2 检索目标与检索效果

构造检索策略前，必须先对情报提问进行详尽的主题分析并确定检索目标。是否充分理解用户的情报提问实质，检索目标是否确定得恰当，这些是决定系统的检索服务能否满足用户的情报需求的前提条件。

如果没有明确的检索目标，则所采取的检索行为将是盲目的并且也是不利于反馈的检索行为。因为在检索过程中存在着一系列产生检索失误的潜在根源，欲图获得较好的检索效果，就应象医生诊断疾病那样来分析情报提问并确定检索目标；像医生开处方那样来对症下药地构造检索策略。

由于影响情报检索效果的系统因素比较多，而我们的重点是探讨检索策略及其反馈调节问题，所以只分析检索策略对检索效果的影响。需要指出，本文所谈的检索效果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，它主要通过查全率 R 和查准率 P 来体现。

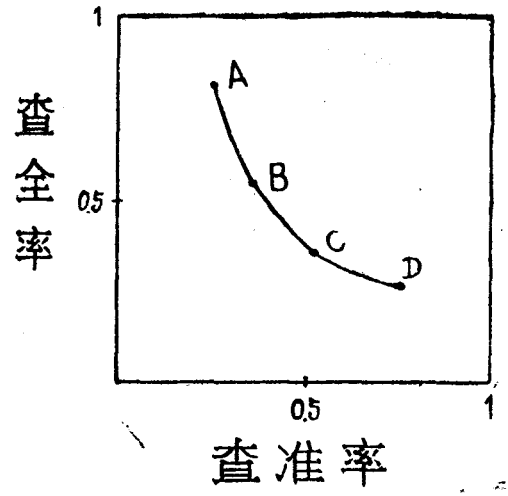
所谓查全率是衡量符合提问要求的文献中实际检出文献的数量，查准率是衡量检出的文献中有多少符合情报提问的文献数量，即

$$\text{查全率 } R = \frac{\text{检出的相关文献量}}{\text{库中存贮的相关文献量}}$$

$$查准率 P = \frac{检出的相关文献量}{从库中检出的文献量}$$

查全率R与查准率P结合起来，可以表示检索系统的筛选能力，即让需要的文献通过，而阻止不相关的文献通过的能力。

一个理想的检索过程将最大限度地满足用户的情报需求，一个好的检索策略应该能为用户提供全面而准确的相关文献。然而，由于查全率与查准率之间倾向于互逆关系，为了提高查准率而限制一个检索范围时，查全率就会相应降低，反之，要提高查全率而放宽检索条件，查准率也会相应降低。美国的情报学家F.W.Lancaster曾将在50个提问的检索中所获得的查全率与查准率互相对比地标在下面的图中，每次检索均用A、B、C、D四个不同的等级进行。



图一：查全率与查准率的对比标示

当检索很泛指时，可达到高查全率，如图中A点，但查准率相当低。当检索很专指时，可达到高查准率，低查全率的D点，而B、C两点表示介于两个极端之间的折衷的检索策略，其查全率与查准率各维持在一定水平上。

世界上任何一个检索系统均不可能得到百分之百的查全率和百分之百的查准率。一般的检索系统的平均查全率为60～70%，查准率为40～50%。不同的用户对查全率和查准率有不同的要求。例如，从事理论研究的科研人员一般都具有较高的文献查阅能力，他们密切注意着国内外有关研究的新动向，新进展和新成果，他们要求高查全率，不怕文献多，就怕漏检，而工程技术人员关心的是有关新产品，新技术的十分具体的情况，一般要求高查准率。尤其那些从事应用研究和工艺流程的用户，他们感兴趣的是有关专利文献和技术报告之类 他们不要求高查全率，但不希望有误检现象。还有一些管理人员和专业情报人员，其检索要求则十分广泛，他们的检索目标有很强的动态性和机遇性，他们根据所从事的工作性质的不同以及研究阶段的不同，而对查全率和查准率有着不同的要求。

总之，对于高查全与高查准这两者不可得兼。对于以查全为重点要求的用户来说，可以将他的检索目标表示为：允许检出文献中有某些不相关的文献 但数量要限制在某一容许的查准范围内。即在给定了比较具体的查准率的条件下，要求尽可能提高查全率，所检出的相关文献越多越好。

对于以查准为重点要求的用户，其检索目标可以描述为：希望所提供的文献尽可能符合情报提问，对检出篇数没有具体的指标，即希望在某一查全范围的基础上尽可能提高查准率，使检出文献符合课题要求。

关于检索目标的制定，首先需要搞清楚不同类型的用户具

有不同的情报需求的特征，就是同一用户在不同阶段也会有不同的情报需求。在为这些千变万化的情报需求和各种各样的情报提问制定检索目标时，就需要根据具体情况进行分析、比较，考虑选择什么样的检索方案，怎样构造检索策略才能实现检索目标。

为了进一步明确检索目标，还需要提出一个很直观的检索指标，即期望文献量，也就是用户希望能提供给他们的相关文献数量。这通常取决于用户的阅读消化能力，他所接受任务的紧迫程度及其求知欲望等客观条件以及心理因素等等。

当然，也应该指出，由于用户和检索者在未检索前一般是不知道所检数据库中关于某个检索课题的实际的相关文献的含量的，因而难以计算出具体的查全率。此外还应考虑到查找相关文献时所遇到的一些困难，要用户直接提出具体的期望文献量往往很难做到。但如果在不考虑数据库中关于某个检索课题的实际相关文献含量的情况下，我们只是把它作为由用户所给出的一个主观的检索指标来看待，并且在情报提问登记表上用一种方便用户填写的形式给予必要的说明，这是在一般情况下，用户愿意也能够填出这个主观的检索指标的。例如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。

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。

3.一般性的要求，希望有一定比例的相关文献，不具体限定范围。

4.提出某一固定的期望文献量，如想要 i 篇相关文献。

5.提出某一浮动的期望文献量，希望在某个范围，如检出1～n篇相关文献。

6.不提出任何要求，随便系统怎样处理均可。

§1.3 检索决策的动态过程及反馈模式

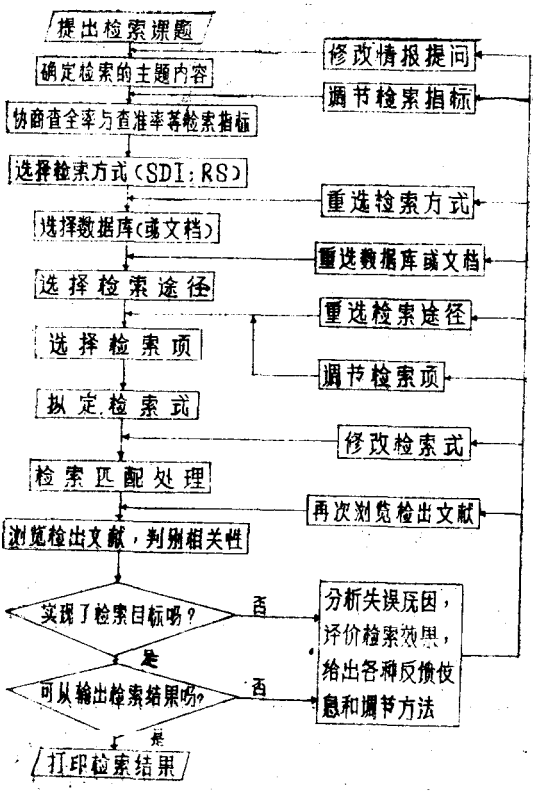
为了更有效地构造检索策略，应该研究构造检索策略的模式。所谓模式是关于所研究对象和过程的本质属性及其相互关系的描述。欲图解决某个问题，一般都要借助与其相应的模式，以了解并分析各种有关因素及其相互影响的方式，然后利用模拟、实验或计算等途径作出预测或决策，并估计可能产生的效果，以选择能达到预期目标的合理策略和方法，由于用户的情报需求具有随机性和动态性，也由于检索系统本身功能的不够完善，影响检索系统性能的各种因素的变化又比较复杂，这样就使得在构造检索策略的各个环节中必然会存在一些产生检索失误的潜在根源。为了达到检索目标，当然就应该采取各种反馈途径和各种调节检索策略的方法。

本书所用的反馈概念的定义是指在情报检索过程中，系统（包括检索者和用户）针对检索课题和检索目标而制定的检索方案和检索策略所采取的检索行为的信息返回。系统以这些反馈信息为参照，随时调整修改下步的检索方案及其检索策略，以求符合用户的情报需求并达到既定的检索目标，或者是根据这些反馈信息调节总检索目标或制定新的检索策略。

在手工检索中，这种反馈和调节工作是通过检索者眼看、脑想、手查的过程来完成的。手检某一课题时，也要弄清楚真正所要求的检索角度、深度和广度，选择恰当的检索工具；要确定从何种检索途径着手，使用什么索引，该查什么类目，该用什么主题词或关键词，并且也需要拟定查找的步骤及在可能检不出相关文献时可以选用的后备方案等等。但是在手检条件

下，检索策略往往只存在于检索者的脑子里，其构造模式是漫画式的，不定的。检索者在具体的查找过程中边查边看边考虑，可以灵活地改变策略。通过这种不断浏览，不断进行思维活动，从而不断给出反馈信息，不断进行调节并作出选择，直至找出一定数量的相关文献为止。

在计算机检索的条件下，情报提问与文献标识之间的对比



图二：拟定检索策略的动态过程及反馈模式

匹配工作是由计算机进行。机器不具备人脑那样的思维活动，而是严格地执行人们事先安排好的程序。尤其是考虑到SDI服务的反馈条件较差，联机检索的费用比较昂贵时，欲提高检索效果则不论是脱机检索还是联机检索，都应该事先研究检索策略。为了简明起见，下面用框图形式描述出构造检索策略的动态过程及反馈模式，见图二。

由于在构造检索策略的各个步骤中均可能产生误差，因此有必要分别采取相应的反馈途径和调节方法来改善检索决策的方向、强度及内容，使整个检索过程变成一个不断进行动态平衡的流通回路。

正是由于在检索过程中各个环节都是紧密相关彼此制约的，前面的决策失误将给后面的决策带来很大的损失，因为这些检索失误是累积性的。这里可举出交通部情报所利用香港终端检索的一个课题。这个检索课题是关于“高速公路设计的最新进展”的，其检索结果和失误原因已经经过该部专业人员的详细分析。该课题原来想从国外有关文档进行普查，考察各文档对交通专业文献含量的多寡及所含文献的具体著录，作为以后引进数据库时的选择依据。由于提问单上没有把这个目的与要求写清楚，致使在香港终端的检索人员视作解决某一特定研究课题的提问者来对待，只选一个文档查找，该课题普查的意图从检索开始时就落空了。该课题具体在检索时，选词也不恰当。该题检索第一步所选的检索词Motorway不符合美国用词习惯。美国的高速公路在技术上与管理上各用不同的称呼，常用的有 Expressway，Freeway，Interstate highway systems 等等。据79年的HRIS文档共约4000条文献为准，其中同高速公路直接有关者有23个词，共计 100 条，而同高速公路间接有关者约 500 条。据此推算，16年可达一万条，但其中

竟无一条用Motorway这个词。由此可见用词不当把99%以上的条目漏检了。

关于该课题的第二步检索是以Highway，Design和Development这三个词用AND组配起来，这样由于检索式拟定不当又把大量有用文献都排除出去了。结果虽检出了59篇文献，真正对口的只有3篇。即查准率为5%。正如该部专业人员所概括的："选库、用词与定逻辑式三者中有一步落空，就会造成重大失误；三者都有偏差，就必然满盘皆输了。"

上面这个例子说明了在为一个检索课题构造检索策略时要通盘考虑，应该为检索课题构造出一些可资比较的检索方案和备用策略。有时选库、选词稍有差错，或在拟定检索式时逻辑关系组配不恰当，就会产生大量漏检与误检，并且浪费机时。所以在机检前要作好充分的准备并要将检索过程中的动态变化尽可能纳入系统的控制之下。在作出检索决策时要尽量疏通各种反馈渠道，使整个检索过程变成一个不断进行动态反馈的流通回路，从而使检索目标与检索效果之间的距离逐步缩小，使检出结果与用户的情报需求趋向一致。

第二章　构造检索策略的几个主要环节

§2.1 构造检索策略的要素

由检索决策过程中一些相互关系的分析及检索实例分析得知，检索策略构造的优劣程度是与许多因素有关的。充分表达

了用户的情报需求并确立了具体的检索目标这是构造一个良好的检索策略的基础。而选择恰当的数据库及检索方式，这是制定合理的检索策略的前提条件。具体在构造检索策略时，则必须考虑下列要素：

1.所选择的检索途径是否恰当；

2.情报提问是否被准确地进行了概念分析；

3.所选择的检索项是否恰当，即情报提问的概念组面是否被转换成了一组合适的检索词或分类号等检索项；

4.检索表达式的专指度与网罗度是否恰当，是否考虑了比较条件和一致原则；

5.所选择的关于检索结果的输出格式，它表示的内容是否能满足用户的相关性判断；

6.是否充分利用了各种反馈途径和调节方法；

7.是否有目的，有针对性地进行了检索效果的调查分析。

尽管影响检索策略的因素较多，而决策的关键在于选择与比较，没有正确的选择就难以构造出合理的检索策略，没有充分的比较分析就难以给出恰当的反馈信息，从而难以形成新的经过改进的检索策略，也难以获得好的检索效果。所谓“选择”即按照某个原则或某种标准从m个物体中选择n个物体（$m \geqslant n$）。

选择是检索行为中最普遍的活动，如选择数据库，选择检索途径，选择检索词等等，但是选择的普遍性并不意味着选择的等同性。在不同的检索步骤中，选择具有不同的作用，各有其必须遵循的特殊原则及其必须考虑的相关要素。

由于构造一个良好的检索策略往往牵涉到各方面的知识和技能，诸如，是否了解检索系统的特性和功能，是否熟悉所检数据库的标引规则及词表结构，是否掌握了必要的检索方法与