

# 第3章

## 文献信息检索的途径、方法与步骤

在浩瀚的文献信息汪洋之中,熟悉文献信息的基本概念和原理,运用科学合理的检索途径和方法,才能高效准确地完成文献信息检索。

### 3.1 文献信息检索的途径

文献信息检索是将用户提交的、包含特定文献信息需求的检索式与文献信息检索系统中的信息集合进行相符性比较的过程,这里的相符性比较是建立在检索语言基础之上的。

文献信息检索的途径就是利用信息的某种特征作为检索标识来查找相关信息的途径。一般要根据已知信息需求、已掌握的文献线索及检索工具的实际情况,有针对性地选择合适的检索途径。根据文献的特征,将检索途径分为内容特征检索途径和形式特征检索途径。

#### 3.1.1 内容特征检索途径

内容特征检索途径又分为分类途径和主题词(关键词)途径。

##### 1. 分类途径

分类途径是按学科分类体系来检索文献。从分类途径检索文献资料,主要是利用分类目录和分类索引。这一途径以知识体系为中心,比较能体现学科系统性,反映学科与事物的隶属、派生与平行的关系,便于我们从学科所属的范围来查找文献资料,并且可以起到触类旁通的作用。

分类途径尤其适用于族性检索,能够保证较高的查全率。如果需要查找的是某一学科领域或某一专题的文献,宜选用分类途径(但不适于查找交叉学科或新学科信息),但要求信息检索人员熟悉学科分类体系,能够正确判断学科所属类目。

##### 2. 主题词(关键词)途径

主题词途径通过反映文献资料内容的主题词来检索文献。几乎所有的检索工具和检索系统都提供主题词(关键词)途径。主题目录、主题索引、关键词索引、叙词索引等是其检索依据。检索时首先要分析主题概念,选择相应的主题词或关键词,再按照字顺查找,进而得

到所需信息。

主题词途径尤其适用于特性检索,能够保证查准率,所以对于一些检索主题新颖、复杂、专深、具体的检索课题宜选用这种检索途径。

### 3.1.2 形式特征检索途径

形式特征检索途径又分为题名途径、责任者途径、代码途径、引文途径和其他检索途径。

#### 1. 题名途径

题名途径是根据文献名称(包括书名、刊名、篇名等)来检索文献的途径。检索时使用各种题名目录或索引,输入题名或题名的一部分,即可获得所有题名中包括该字、词的信息。利用题名途径既可以检索出一篇特定的文献,还可以集中一种著作的全部版本、译本等,因此被广泛地应用于图书、期刊、论文的检索中。通过题名途径检索应注意以下三个问题。

(1) 排检规则。中文字顺排检方法有多种,应留意具体系统采用何种方法,如汉语拼音、偏旁部首、四角号码等。英文字顺排列较为简单,但应对冠词、连词、介词等忽略不计。

(2) 简称缩写。一些检索工具编制时为了节省篇幅,常对刊名、书名进行缩写。若能掌握一般的简写规则,将对理解检索系统中的刊名、书名简写有所帮助。

(3) 与著者相同的字译问题。在根据检索结果去索取原始文献时有可能会碰到。

#### 2. 责任者途径

责任者是指对文献内容负责或做出贡献的个人或团体,包括著者、编者、整理者、译者等。利用责任者途径检索是指按照已知责任者的名称来查找用户所需信息。检索时要以著者目录、著者索引等为依据。一般来讲,每个研究人员的研究方向相对比较稳定,同一责任者名下往往会集中内容相近或相关的文献,可以在一定程度上实现族性检索,并且利用责任者途径,可以及时跟踪研究人员的研究方向,获得最新研究成果。因此,责任者途径也是常用的一种检索途径。通过著者途径检索应注意以下两个问题。

(1) 外国著者姓名的倒置和取舍规则。许多西方国家的个人姓名是名在前姓在后,而在编制著者索引时会倒置为姓在前面而名在后面,因而检索时应照此处理。另外,姓名中的家族称号、宗教称号和其他前缀的处理应根据各检索系统的具体规则而定。

(2) 不同语种之间著者姓名的字译系统。英美国家制作发行的检索系统中,对非拉丁语言的文献著者姓名、机关团体名称、刊名、书名等常采用字译方法转换成拉丁字母,再按拉丁字母顺序排列。因此,从著者途径查找英美检索系统时,存在要将非拉丁字母译成拉丁字母的问题。

汉字翻译成拉丁字母,在我国通行使用汉语拼音系统,但海外多用 Wade Giles 系统,它与汉语拼音系统有所不同。俄文字母与英文字母字译都分别有相应的字译系统。

#### 3. 代码途径

有些文献具有独特的代码,如图书有国际标准书号(ISBN),专利有专利号,报告有报告号,标准有标准号等。利用代码途径检索信息就是通过已知文献的这些专用代码来查找信息。检索时要以各种代码索引为依据,如专利号索引、报告号索引等。在已知信息特定代码的前提下,利用代码途径检索信息非常简便、快捷、准确。

#### 4. 引文途径

引文途径就是根据引文即文章末尾所附参考文献来查找所需信息的途径。引文途径较特殊,使用引文途径检索信息时可以通过成套的检索工具(美国的《科学引文索引》、中国的《中国引文索引》等),或者直接利用文献结尾所附的参考文献,查找被引用文献。利用引文途径可以追溯查找相关信息,并依据课题情况实现循环检索,同时也可以作为评价信息价值的参考依据。

#### 5. 其他检索途径

其他检索途径包括出处途径、时间途径和任意词途径。随着检索技术和实践不断发展,必然还会出现其他的检索途径。

综上所述,分类途径和主题途径是信息检索的主要途径,但是任何检索途径都有其优缺点和适用范围,单靠一种检索途径难免会有所疏漏。检索时要根据实际情况灵活运用,并尽量将多种检索途径结合起来使用,以便达到最佳的检索效果。

### 3.2 文献信息检索的方法

所谓检索方法,就是为实现检索计划或方案所提出的检索目的而采取的具体操作方法或手段的总称。在检索文献信息时,可以根据检索课题的要求和对课题有关文献线索的掌握情况选择不同的检索方法,以便达到省时、省力、查全的目的。信息检索的方法一般要根据检索课题的需要和检索系统(工具)的情况灵活选择。一般的检索方法主要有以下三种。

#### 3.2.1 直接法

直接法就是不利用检索系统(工具)直接通过原文或文献指引来获取相关信息的方法。直接法的优点是能明确判断文献所包含的信息是否具有针对性和实用性,缺点是存在着很大的盲目性、分散性和偶然性,查全率无法保证。如果检索课题单一,文献相对集中,又熟悉原始文献,可采用这种检索方法。而对于有多个主题、文献离散较大的课题,则难以获得理想的检索效果。直接法包括浏览法和追溯法。

##### 1. 浏览法

浏览法是指直接通过浏览、查阅文献原文来获取所需信息的方法。该方法的优点是能够直接获取原文,并能够直接判断是否需要文献所包含的信息;缺点是由于受检索人员主观因素的影响,有一定的盲目性和偶然性,难以保证查全率,且费时费力,对检索人员的要求比较高。

##### 2. 追溯法

追溯法又叫扩展法、追踪法,是利用已知文献的某种指引(如文献附的参考文献、注释、辅助索引、附录等)来获取所需信息的方法,这是一种最简捷的扩大信息来源的方法。根据已知文献指引,查找到一批相关文献,再根据相关文献的有关指引扩大并发现新的线索,进一步来查找。在检索工具不全的情况下,可以选用此种方法,但由于这种方法也存在一定的偶然性,因此最好选用质量较高的述评和专著来进行文献追溯。

### 3.2.2 工具法

工具法是一种最常用的方法,即利用各种检索系统(工具)来检索信息。根据具体的检索情况,工具法又可分为以下三种方法。

#### 1. 顺查法

顺查法是根据已确定的检索课题所涉及的起止年代,按照时间顺序由远及近地查找信息的方法。这种方法查全率高,但较费时费力,适用于普查性课题,利于掌握课题的来龙去脉,了解其历史和现状,并有助于预测其发展趋势。

#### 2. 倒查法

倒查法是按照时间顺序,由近及远地逐年查找,直到找到所需信息。利用该方法能够获取较新的信息,把握最新发展动态,因此较适用于检索新课题或有新内容的课题。

#### 3. 抽查法

一般来说,任何一个学科的发展都具有波浪式的特点,在学科处于兴旺、发展期时,成果和文献较多。抽查法就是根据检索需求的特点和学科发展的实际情况,抽取这一段时间的文献进行检索。抽查法能够获得较多的信息,但要求检索人员必须熟悉该学科的发展情况。

### 3.2.3 综合法

综合法是指综合利用上述各种检索方法来查找信息的方法。利用各种检索方法,使其互相配合、取长补短,进而得到较为理想的检索效果。

## 3.3 文献信息检索的步骤

文献信息检索是一项实践性很强的活动,它要求检索人员善于思考,并通过经常性的实践,逐步掌握文献检索的规律,从而迅速、准确地获得所需文献。一般说来,文献信息检索的基本步骤包括以下几步:分析信息需求、选择检索系统(工具)、确定检索途径与方法、编制检索表达式、获取信息线索、获取所需信息。如图 3.1 所示为文献信息检索的步骤。

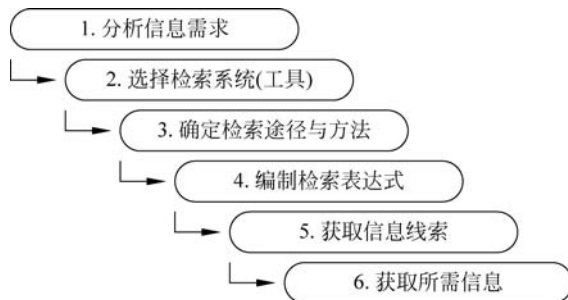


图 3.1 文献信息检索的步骤

### 3.3.1 分析信息需求

在检索前,要周密地分析信息需求,主要是根据用户的表达,明确本次信息检索的主要内容、所涉及的学科范围及所需信息的文献类型、语种、地区、时间等方面的要求。分析信息需求的目的在于理清检索的基本思路,明确检索的目的、要求与检索范围,并从检索需求中发掘检索的已知条件。一般来说,分析信息需求时可以参照表 3.1 中的要素进行逐一明确。表 3.1 为信息需求的要素与内容。

表 3.1 信息需求的要素与内容

| 信息需求要素 | 内 容                                   |
|--------|---------------------------------------|
| 主题概念分析 | 提炼课题中的核心概念,这一步是课题检索的重点和难点             |
| 检索目的   | 本次检索所服务的目标,如申报课题、开题报告、学术论文、成果查新、商业需求等 |
| 文献类型   | 期刊论文、会议论文、科技报告、图书、专利、标准、数据等           |
| 结果形式   | 全文、文摘、题录、数值、事实                        |
| 检索年限   | 如近五年、近十年的文献                           |
| 语种     | 中文、英文、日文                              |
| 检索结果数量 | 根据需要的结果数量,对上述检索条件进行调整                 |

分析信息需求是信息检索成功与否的关键,信息需求分析得越深入细致,越准确,后面的检索效果越好。

### 3.3.2 选择检索系统(工具)

由于特定的检索工具与信息检索系统往往有着明确的文献收录范围,因此明确检索的学科与主题属性有助于选择适用的数据库。在选择检索系统(工具)时,应着重考虑以下几个方面:在内容和时间上,该检索系统(工具)对检索需求的覆盖程度和一致性;在手段和技术上,应选择便捷性和效率较高的计算机检索;在成本和可获取性上,通常选择成本低、获取性高的检索系统(工具)。例如,查找比较专、深的信息最好选用专业性强的检索系统,查找专业性较低、时效性较高的信息则可以通过搜索引擎进行检索。另外,在有多种检索系统可选择的情况下,要选择最权威、最全面、最方便的检索系统。

#### 1. 确定检索途径与方法

要根据信息需求分析的结果和已选定的检索系统(工具)的情况,确定适当的检索途径。检索途径的确定在很大程度上受到检索系统(工具)的制约,但如果有多种检索途径可选择,一般来讲,如果信息需求的范围较广,最好使用分类途径;如果要求的信息较专、深,最好使用主题途径;如果事先已经掌握了信息的责任者、题名等信息,可选用相应的途径。为了提高信息检索效果,还要根据以上分析结果确定适当的检索方法。

#### 2. 编制检索表达式

在计算机检索系统中,有时需要编制检索表达式,即用布尔逻辑算符、位置算符等计算符将两个或两个以上的检索词进行组配,以表达式的形式来确定检索词之间的关系,准确地将信息需求提交给计算机。

### 3. 获取信息线索

一些检索工具(目录、文摘等)在完成上述步骤后并不能直接提供所需信息,而只能提供信息线索。

### 4. 获取所需信息

检索的执行一般都由计算机自动完成,也可以由手工完成。所需信息的获取有时需要按照信息线索的指引才能获得,有时可直接从检索系统(工具),如全文数据库、网络搜索引擎等中获得。按照预先制定的检索策略进行实际检索,但仍要根据检索的阶段性成果或碰到的实际问题适当调整策略和进程。灵活运用检索工具、检索途径和检索方法是检索成功的保证。如果检索结果与检索需求存在差距,则要对检索进行再分析,使用多种检索方法对检索策略(包括检索途径与方法、检索表达式等)进行优化处理。

最后获取的检索结果,可采取复印、复制、打印、下载、E-mail等多种方式收集。对于收集到的文献资料进行认真整理,说明检索结果,按要求给予答复,或提供原始文献。至此,一个完整的文献信息检索过程就完成了。

## 3.4 文献信息检索效果评价

信息检索过程是一个复杂的过程,从表象上看,信息检索过程是信息检索提问式与信息集合标识之间的匹配运算,但实际的机理问题却要复杂得多,它不仅涉及以用户认知结构为基础的信息需求唤醒、提问表达与转换、检索标识的形成和检索结构的相关性与适用性判断,而且还涉及对检出信息的理解与吸收利用。信息检索过程是系列过程组成的综合体系,其各个阶段和环节都可能产生不确定性。信息检索的不确定性是指由于忽略次要因素、相关性不确切或不完全、知识不成熟、证据本身可能错误,或是仅注重对物的研究而对信息传递主体与信息接收客体的关注不够等原因而产生的检索过程的模糊认识。信息检索过程中不确定性的产生机制已经成为信息检索研究中的重要课题。

信息检索的不确定性是由于人们对信息和信息检索过程认识的类属不清、状态不明造成的,用户与文献作者之间知识结构的差距是导致信息检索失败的主要原因。因而作为情报系统来说,必须能够响应用户带有一定缺陷的知识结构,反映和支持用户在信息需求表达中所利用的领域知识和语言知识。

有关信息检索的不确定性研究是将检索过程建立在一种理想化的假设之上,即从需求唤醒到提出问题再到情报吸收、利用的一系列检索过程能在用户与系统的交互作用中顺利进行。

### 3.4.1 文献信息检索效果判断环节

计算机检索时,首先要由用户向计算机信息检索系统提交查询表达式,系统经过查询匹配后把检索结果输出给用户,再由用户进行判断是否满足自己的信息需求。可见计算机检索的相关性判断有两个环节:一是系统相关性判断,即系统自动对相关度进行计算,并输出检索结果;二是用户相关性判断,即用户在选择系统、拟定检索表达式及在系统命中的结果中进行取舍时所做出的主观判断。

### 1. 系统相关性

系统相关性指的是文档标识与用户提问之间的相符程度,其量化指标为相关度。检索系统的输出结果一般按照相关度从大到小排列。相关度的算法因系统而异,是决定系统检索性能优劣的主要因素。各检索系统评判结果是否相关及相关程度的方法虽有不同,但归纳起来主要有词频方法、位置方法、引用率方法、大众单击率方法、分类或聚类方法等。

### 2. 用户相关性

用户相关性是一个灵活、相对的概念,它表示的不是检索出的文档与用户检索表达式之间的一致性,它衡量的是文档与用户需求的一致性。当用户不知道某些相关信息的存在或对检索课题不甚了解从而不能形成完整的信息需求表达时,某些与用户的信息需求相符的文档却可能与检索提问不符。反之,检出的文档与检索提问相符却不一定能满足用户的需求。系统相关不一定意味着用户相关。用户相关性由用户本人来判断,它具有强烈的即时性和明显的个性化特征:用户对于文献相关与否的判断会因条件、时间的不同而有所变化,还会因用户知识背景、知识结构、兴趣爱好不同而有所不同。

## 3.4.2 信息检索的效果评价方法

检索效果是指检索结果的有效程度,反映了检索系统的检索性能和检索能力。评价检索效果的目的是为了准确地掌握检索系统的各种性能和水平,找出影响检索效果的各种因素,为检索策略的改进调整提供依据,才能够快速、全面、准确地查找出所需要的文献信息。对计算机检索系统来说,通过对检索效果的评价,以便改进系统的检索性能,提高系统的服务质量。

任何检索系统都有存储和检索两个功能。就存储而言,保证某一学科或专业领域信息收集全面并不十分困难;而对于检索来说,从系统中输出全部相关信息,排除所有无关信息则比较难以实现。通常情况下,在查找信息时,不可避免地会带来一些无关信息,而漏掉一部分相关信息。在网络检索系统中,这种情况尤为突出。其主要原因是系统相关性匹配算法的机械性、用户提问的模糊性及其与信息需求的偏差等。

根据 F. W. Lancaster 的阐述,判定一个检索系统的优劣,主要从质量、费用和时间三个方面来衡量。因此,对计算机信息检索的效果评价也应该从这三个方面进行考量。质量标准主要通过查全率与查准率进行评价;费用标准即检索费用,是指用户为检索课题所投入的费用;时间标准是指花费的时间,包括检索准备时间、检索过程时间、获取文献时间等。在这三个方面中,查全率和查准率是评价检索效果的主要指标,也是用户在实际检索时最关心的问题。下面对目前采用最为普遍的检索效果量化评价指标——查全率、查准率、漏检率、误检率进行简单介绍。通常使用  $2 \times 2$  表格对这 4 个指标进行描述,如表 3.2 所示。

表 3.2 检索结果评价指标

| 用 户   | 系 统      |            |                 |
|-------|----------|------------|-----------------|
|       | 相 关 文 献  | 非相关文献      | 总 计             |
| 被检出文献 | $a$ (命中) | $b$ (噪声)   | $a + b$         |
| 未检出文献 | $c$ (漏检) | $d$ (合理拒绝) | $c + d$         |
| 合计    | $a + c$  | $b + d$    | $a + b + c + d$ |

### 1. 检索结果评价指标

查全率是对所需信息被检出程度的量度,用来表示信息系统能满足用户需求的完备程度;查准率是衡量信息系统拒绝非相关信息的能力的量度;查全率的误差即是漏检率;查准率的误差即是误检率。其数学表达式分别如下。

查全率( $R$ )=被检出相关文献数/系统中的相关文献 $\times 100\%=(a/(a+c))\times 100\%$

漏检率( $O$ )=未检出相关文献数/系统中的相关文献 $\times 100\%=(c/(a+c))\times 100\%$

查准率( $P$ )=被检出相关文献数/被检出文献总数 $\times 100\%=(a/(a+b))\times 100\%$

误检率( $N$ )=被检出不相关文献数/被检出文献总数 $\times 100\%=(b/(a+b))\times 100\%$

查全率和查准率是评价检索效果的两个重要指标,查全率、查准率越高,说明检索效果越好,但在实际检索中,查全率和查准率是不可能同时达到100%的。克莱夫登通过克兰菲尔德实验证明,在同一个信息检索中,当查全率和查准率达到一定的阈值,即查全率为60%~70%、查准率为40%~50%后,二者呈互逆关系,即查全率与查准率在某种程度上成反比例关系,一方的提高往往导致另一方的降低,偏重哪一方都是不妥当的。在检索实践中,需要根据课题的具体要求,合理调节查全率和查准率,找到最优平衡点,保证适度的查准率和查全率。因此,在检索过程中,可以从以下几点考虑。

(1) 作为检索人员,要确定自己是对查全率更关心,还是对查准率更感兴趣。据此选择不同的检索策略。

(2) 了解检索系统和数据库的特点和规模。对专业性强、规模小的数据库,要注意提高查全率;对数据量较大的系统,如网络搜索引擎,由于其结果输出量比较大,保证查准率则显得更为重要。

### 2. 调整查全率和查准率的方法

影响查全率的因素从文献信息存储来看主要有:数据库收录文献不全;索引词汇缺乏控制和专指性;词表结构不完整;词间关系模糊或不准确;标引不详;标引前后不一致;标引人员遗漏了原文的重要概念或用词不恰当等。从信息检索来看主要有:检索策略过于简单;选词和进行逻辑组配不当;检索途径和方法太少;检索人员业务不熟悉和缺乏耐心;检索系统不具备截词功能和反馈功能;检索时不能全面地描述检索要求等。

提高查全率,即进行扩检,可以按照如下方法调整检索提问式。

(1) 选全同义词并以“or”的方式与原词连接后加入到检索式中。

(2) 降低检索词的专指度,从词表或检出的文献中选择一些上位词或相关词。

(3) 采用分类号进行检索。

(4) 删除某个不甚重要的概念组面,减少“and”运算。

(5) 取消某些过严的限制符,如字段限制符等。

(6) 调整位置算符。

影响查准率的因素从文献信息存储来看主要有:索引词不能准确描述信息主题和检索要求;组配规则不严密;标引过于详尽;检索系统不具备逻辑“非”功能和反馈功能;检索式中允许容纳的词数量有限。从信息检索来看主要有:选词及词间关系不正确;组配错误;检索时所用检索词(或检索式)专指度不够,检索面宽于检索要求;截词部位不当;检索式中使用逻辑“或”不当等。

若要提高查准率,即进行缩检,可按如下方法调整检索提问式。

- (1) 提高检索词的专指度,增加或换用下位词和专指性较强的自由词。
- (2) 增加概念组面,用“and”连接一些进一步限定主题概念的相关检索项。
- (3) 限制检索词出现的可检字段,如限定在篇名字段和主题字段中进行检索等。
- (4) 利用文献的外表特征限制,如文献类型、出版年代、语种、作者等。
- (5) 用逻辑非“not”来排除一些无关的检索项。
- (6) 调整位置算符。

### 3. 其他评价指标

- (1) 用户负担:即检索工具的用户友好性及用户在使用该工具时的方便和易用程度。
- (2) 新颖率:从检索系统中检索出来的对用户而言含有新颖信息的文献数量与文档中总相关文献数之比。
- (3) 覆盖率:在某一特定时间里,从某一检索系统中检索到的涉及特定主题领域的所有文献数与该主题领域相关的实有文献总数之比。
- (4) 检索结果的满意度:包括检索结果相关命中数、重复链接数、死链接等。
- (5) 响应时间:即完成一个检索要求所用的时间。
- (6) 相关性排序:即将输出结果根据与检索词的相关度进行排序。
- (7) 输出数量选择:即限定或改变输出量。
- (8) 输出方式:标题的有无、类目位置、网页文本大小等。
- (9) 检索界面:用户界面的易用性情况,包括是否含有检索说明文档、是否有帮助文件、是否有查询举例等。

## 学 与 练

1. 文献信息可以分为哪些类型?在日常生活和学习中利用最多的类型有哪些?
2. 一次、二次、三次文献有什么样的区别与联系?
3. 文献信息检索的途径和方法有哪些?分别适用于什么样的情况?
4. 文献信息检索的步骤是怎样的?尝试提出一个检索需求,设计一个完整的文献信息流程。
5. 布尔逻辑检索的“与”“或”“非”各有什么意义?尝试提出一个检索需求,运用布尔逻辑检索式进行检索提问。
6. 熟练掌握截词、位置、限制、加权和聚类检索技能,并尝试设计出检索需求,运用相应的技能进行检索。