第一章　档案检索系统概述

第一节　档案检索系统的构成

一．档案检索的含义

对“档案检索”一词的解释可以有广义和狭义两种。广义的档案检索包括档案存贮和档案检索两个具体过程。存贮是指将档案中具有检索意义的特征标识出来，然后编制检索工具，建立检索系统的过程。检索是指利用检索工具和检索系统查找所需档案的过程。这两部分内容是密切联系、不可分割的。存贮是检索的前提或基础，没有存贮就无法检索；检索则是存贮目的的实现，没有检索，存贮也就失去了意义。从狭义上讲，检索只限于查找所需档案的具体过程。我们研究检索是从检索的广义概念出发的，也就是说把存贮过程和检索过程作为一个整体来研究。

档案检索的一般过程是，档案管理人员将可供鉴别档案的内容和形式特征记录下来，将档案转换成一种篇幅较短的特殊文献形式，并将这种文献形式按特定的方法加以组织排列，借助它把所需档案排检出来。从这个过程可以看出，档案检索的原理就是采用一种压缩档案信息的方法进行存贮，并在存贮的逆过程中把所需档案查找出来．所谓逆过程，是指查检时的思路与存贮时的思路相一致，只是程序的方向相反，形象地说，即为“怎么放进去，就怎么取出来”。例如根据档案分类法，把有关民族教育方面的档案存贮时归入 B 类(国家政务总类)中第 4 类(民族事务)中第 6 属类(民族教育)，其分类号为 B46，那么检索有关民族教育类的档案也应到 B46 类中查找，这样查检的思路

与存贮的思路一致,即可查出所需档案。如果不一致,存贮时归入国家政务类,查检时到文化教育类中去找,就不能查出所需档案。认识这一原理的目的在于将存贮和查检作为一个整体来看待,在工作中原则一致,互相呼应,以使档案检索过程顺利实现。

档案检索同其他文献检索一样，可以采取三种形式。

(一)数据检索

这种检索是直接回答利用者所需要的有关特定主题的查询。检索的结果是数据。如查询中国共产党第十一届三中全会是哪一年召开的，某城建档案馆中有关立交桥的设计方案共有多少个等等。

(二)事实检索

这种检索检出的结果也是数据。它与数据检索不同的是，对检出的数据进行某种逻辑推理后再输出。如某校教学档案中存贮的数据有‘学生成绩 90 分以上为优，80 分以上为良……”;还有全校学生成绩,如‘张红 90 分,李敏 85 分……”若提问张红、李敏成绩为何档次,输出为‘张红,优 李敏,良”。这种检索在手工检索和机械检索中都存在，不同的是在手工检索中对检出数据进行逻辑推理的是人脑，而在机械检索中则由计算机自动进行。

(三)文献检索

这种检索是要查出记录所需情报的档案文献。事实上这种检索所提供的大部分并不是文献本身,而只是文献目录(二次文献),然后借助目录查到文献的存址,从而最终获得文献本身。

二. 档案检索系统构成

从以上分析可以看出,档案检索的全过程涉及多种因素,这些因素集合而成的整体就是档案检索系统。档案检索系统是任

何一个独立的档案管理系统的重要组成部分，或称为子系统。一般说来，我们把一个档案馆(室)看作是一个档案管理系统，对这个系统的构成因素有多种不同的描述，但无论哪一种都承认档案检索是档案管理中的独立环节，是构成档案管理系统的重要因素。

档案检索的内部也可以分解为各个因素。这些因素不是简单的堆积，而是互相影响、互相制约的。从系统的整体功能出发，去设计规划各个因素的构成方式及其相互关系，可以使整个检索系统的功能得到改善，从而获得较好的检索效果。这也正是把对档案检索过程的研究引向对档案检索系统研究的意义所在。

档案检索系统是一个动态系统，系统的基本构成，如图 1-1 所示。

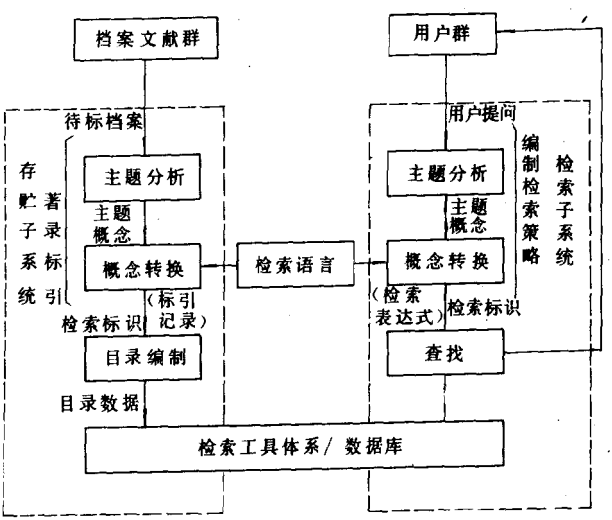

图 1-1 档案检索系统

由图 1-1 可以看出,档案检索系统包括两个子系统:存贮子系统和检索子系统。档案存贮子系统的主要功能是通过著录、标引等手段建立检索工具体系;档案检索子系统的功能是通过编制检索策略在检索工具体系中查找档案文献的线索。具体说来，这两个子系统内部的工作过程是这样的：

在存贮档案时，档案标引人员首先要对档案的内容进行主题分析，使之形成若干能反映档案主题的概念。然后借助于检索语言(分类法、主题词表等)把这些概念转换成规范化的检索词汇，形成检索标识。从对档案的主题分析到检索标识形成的这一段工作就是档案的著录标引工作。然后再把这些检索标识以标引记录的方式加以组织或输入计算机，组成各种检索工具。在手工检索系统中，编制检索工具的结果是形成一套检索工具体系;在电子计算机检索系统中,机读目录组织的结果是数据库的建立。

在检索档案时，档案检索人员首先要根据用户提问确定用户所需档案的实质内容,形成概念,然后同样借助于各种检索语言,把概念转换成规范化的检索词汇,并按实际需求把这些词汇之间的逻辑关系表达出来，形成检索表达式。检索中从用户提问进行主题分析到检索表达式形成的这一过程，就是编制检索策略的过程。如同把存贮阶段形成的标引记录看作是档案著录一样，也可以把检索阶段形成的检索表达式看作是用户提问著录。两种著录的区别在于，前者的结果是对档案内容和形式特征的逐一记录，而后者的结果所表达的各主题概念之间通常含有逻辑。检索表达式形成后,检索人员采用各种检索手段,把检索表达式与检索工具中的文献标识进行相符性比较，将符合检索表达式的结果输出给用户。在手工检索过程中，相符性的比较是由人脑进行的,而在机械检索过程中,则由计算机担负两者

间的匹配工作。至此，一个具体的检索过程结束。至于按照检索结果提取原文传递给用户的过程，在此不把它作为检索的内容加以研究。

从以上两个子系统工作流程的分析中可以看出，二者的主要步骤基本上是一样的。这进一步说明检索是存贮的逆过程的原理。档案检索语言充当了两个子系统的中间桥梁。检索工具是存贮子系统的归宿和检索子系统的查找范围及匹配的依据。由此可见，两个子系统之间的关系是十分密切的，二者的相互作用构成了整个档案检索系统的运动。

存贮子系统的对象是档案文献群，通常是通过把原始文献（一次文献）转化成二次文献的方式进入存贮子系统的。源源而来的档案文献群构成了存贮子系统周而复始的运动过程。检索子系统面对的是档案用户群，用户通过与系统的交互才能顺利地使用系统，满足需求。用户群是使检索子系统处于运动状态的根本条件。

通过以上对档案检索系统的描述和分析，可以看出这一系统的层次态。整个系统包括两个子系统，每个子系统又由若干因素组成。在存贮子系统中主要有著录标引的因素。检索子系统中主要有编制检索策略和检索手段的因素。检索语言、检索工具体系是两个子系统共同涉及的因素。此外，用户与系统的交互方式也有一定的专门内容。据此，本书将分章讲解这些内容。

三．档案检索方式

档案检索方式按照目录或计算机文献档案中数据的排列方式以及检索单元的不同，可区分为文献单元方式和标识单元方式。两种方式在检索工具的组织方法和检索功能上各具特点。

（一）文献单元方式

又称顺检方式、顺排文档。它以一份文献为一个条目，指明该文献的各种特征，即以文献为单元进行检索，其条目的排列按某一文献标识的顺序排列。目前档案部门使用的各种手工检索目录大多采用此种排列方式，如案卷目录、分类目录、主题目录等。下例条目就是采用文献单元方式。

B33912　　5-18-91-3

印发《关于聘任专业职务时对专业技术人员外语考核的规定》的通知 水电职改字〔1987〕14 号／水利电力部，职称改革工作领导小组。——1987.5.20

这种检索方式的特点是查到某一文献标识即可见到文献的著录事项，了解该文献的概况。但是当查找某一主题文献时需逐件扫描，速度较慢。

（二）标识单元方式

又称逆检方式、倒排文档，即以一个标识为单元，指明含有该标识的全部文献，一般只有标识和文献号（档号）两个项目。各种索引大多采用此种排列方式，如人名索引、地名索引、主题索引等；下例条目就是采用标识单元方式。

政治体制

64-3-50-18
64-17-2-15
67-1-6-3
70-5-2-7
⋮

78-6-190-9
79-2-76-3
89-6-60-1
101-17-6-19
⋮

体制改革	
3–17–6–9	78–6–190–9
17–20–9–1	79–20–76–3
64–17–2–15	88–9–109–17
67–9–19–9	108–17–6–9
⋮	⋮

这种检索方式的特点是，可进行组配检索，如需检索含有 *A*、*B* 两标识的文献，只要将 *A*、*B* 两款目中记载的文献号加以比较，共同含有的文献号即为命中文献，上例中的“64–17–2–15”、“78–6–190–9”就是命中文献。但是这种检索工具查到某一标识后，只能获得含有该标识的文献号(档号)，而不能见到该文献的著录事项，如需了解该文献概况，还需要按照检索后获得的档号再去查找该文献单元款目。

文献单元方式和标识单元方式的结合，可将全部文献数据组成一个文献——语词矩阵(见次页图 1–2)。

这里横的方向 1，2，3…表示文献 纵的方向 A，B，C，… 表示标识。从纵的方向看，第一篇文献包含有 A、D、F 三个标识 第二篇文献包含有 B、D 二个标识……依此次序排列下去，就成为顺排文档，即文献单元方式。从横的方向看，标识 A 包含 1、5、7、8 四篇文献，标识 B 包含 2、6 二篇文献……依此排列下去就是倒排文档，即为标识单元方式。文献——语词矩阵表明，检索任何数据都可以从文献单元方式和标识单元方式两种途径入手。

文献号 / 标识	1	2	3	4	5	6	7	8	9	10	....
A	×				×		×	×			
B		×				×					
C						×		×		×	
D	×	×		×		×			×	×	
E									×	×	
F	×		×			×		×			
G								×			
H			×		×		×				
⋮											

图 1-2　文献——语词矩阵

第二节　检索效率

检索效率是指在检索过程中满足利用者的全面性和准确性程度，它是衡量检索系统性能的一个最基本的指标。就每一个检索过程而言，理想的检索结果是无遗漏无误差地检出利用者所需文献，但由于各方面的因素，实际上很少有可能达到这样的结果。检索效率通常采用检全率和检准率两个指标来衡量和表示。

一、检全率和检准率

检全率和检准率这两个指标是美国情报专家佩里(J. W. Pery) 和肯特(Allen Kent)在 50 年代中期提出的。所

谓检全率,是指根据利用者的需求,检出的有关文献与全部有关文献的百分比。与之相对应的是漏检率，即未检出的有关文献与全部有关文献的百分比。检全率与漏检率是两个相对应的指标,其公式为：

$$\text{检全率} = \frac{\text{检出的有关文献}}{\text{全部有关文献}} \times 100\%$$

$$\text{漏检率} = \frac{\text{未检出的有关文献}}{\text{全部有关文献}} \times 100\%$$

例如,某一利用者要求查找有关‘九·一八’事件的档案,档案馆保存的有关档案是 40件,而检索时检出其中 30 件,有 10 件漏检，那么检全率是 $\frac{30}{40} \times 100\% = 75\%$;漏检率是 $\frac{10}{40} \times 100\% = 25\%$。检全率愈高,说明检出的相关文献愈多,漏检率愈低。

所谓检准率，是指根据利用者的需求，检出的有关文献与检出的全部文献的百分比。与之相对应的是误检率，即检出的不相关文献与检出的全部文献的百分比。检准率和误检率也是一个对应的指标,其公式为：

$$\text{检准率} = \frac{\text{检出的有关文献}}{\text{检出的全部文献}} \times 100\%$$

$$\text{误检率} = \frac{\text{检出的不相关文献}}{\text{检出的全部文献}} \times 100\%$$

例如，利用者欲查找有关知识分子政策的档案。在一次检索过程中共检出 30 份档案，发现在这 30 份档案中有 20 份是相关的，10 份是不相关的，那么检准率是 $\frac{20}{30}\times 100\% = 67\%$；误检率是 $\frac{10}{30}\times 100\% = 33\%$。

任何一次检索结果都可以用图 1–3 表示。

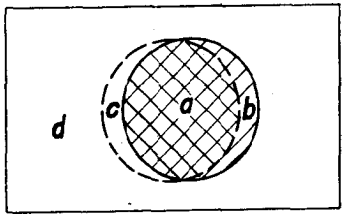

图 1–3　一般性检索结果

图 1–3 中整个大方框内是全部馆藏 $(a+b+c+d)$；虚线圆是关于某一主题的相关文献（$a+c$）虚线圆以外是不相关的文献（$b+d$）；实线圆是在检索这一主题过程中检出的文献（$a+b$）。此图显示的是一次检索过程。按照图 1–3 中的描绘，该检索过程检出大部分的相关文献（a），也遗漏一些相关文献（c），检出一些无关文献（b），正确地排除档案馆藏中大多数的不相关文献（d）。

如果从档案检索系统和利用者两个方面对图 1–3 所示检索结果加以分析，便可用表 1–1 描述出各个因素之间的相互关系。

表 1-1 检索结果 2×2 表

系统相关性测报 ↓	利用相关性判断 → 相关	不相关	总计
已检出	a	b	$a+b$
未检出	c	d	$c+d$
总计	$a+c$	$b+d$	$a+b+c+d$

由表 1-1 可知：

$$\frac{a}{a+c} \times 100\% = \text{检全率};$$

$$\frac{c}{a+c} \times 100\% = \text{漏检率};$$

$$\frac{a}{a+b} \times 100\% = \text{检准率};$$

$$\frac{b}{a+b} \times 100\% = \text{误检率}。$$

表 1-1 从档案馆和利用者两个方面描述了检索情况，通常被称为 2×2 表。从档案馆方面来看，在检索时，档案馆的馆藏总是分为两个部分，已检出文献＋未检出文献＝全部馆藏 $(a+b+c+d)$。从利用者方面来看，在已检出的材料中，根据利用者的需求情况又分为两种情况：相关文献 (a) 和不相关文献 (b)。在未检出的材料中也分为两种情况：利用者需要的、但遗

漏的文献(c)和利用者不需要的、也未检出的文献(d)。从表1-1中可以看出，理想的检索效果应是只检出利用者需要的全部文献，即 $a+c$。在这种情况下，$b=0$，即不相关的文献未被检出，$c=0$，即没有遗漏的相关文献。此时的检全率和检准率都达到100%。这里 a 值（检出的有关材料）对于检全率的高低具有决定性影响。因为相关文献的总数（$a+c$）是固定的，a 值愈大，c 值必然愈小，检全率就愈高。例如，馆内相关文献是80件，检出60件（a），检全率为75%；当 a 值提高，检出70件时，检全率则上升为87.5%。b 值（检出的不相关文献）对于检准率有决定性的影响，b 值愈小，检准率愈高，而 a 值大小不起关键作用。其原因是检准率与馆藏内有关的文献数量没有关系，仅仅测定在每一次检索过程中检出的有关文献在检出的全部文献中所占的百分比。这样如果不控制 b 值，即使 a 值提高，检准率也不会提高。例如，馆藏内有关的文献是80件，在一次检索过程中只检出40件相关的文献，这时的检准率是百分之百（$\frac{40}{40+0}\times 100\% = 100\%$）。之所以达到百分之百，关键是因为 $b=0$。如果检出100件，其中有60件是相关的，这时 a 值明显大了，但 b 值也大了，检准率仅达到60%（$\frac{60}{60+40}\times 100\% = 60\%$）。可见检准率是测定系统阻止不相关文献的能力。

二. 检全率与检准率的关系

英国情报学家 C.克勒维当(C.Cleverdon)根据1963年美情报专家对7万篇文献的研究结果做出检全率和检准率这两个指标之间存在着互逆关系的结论。也就是说，如果放宽检索以达到较好的检全率，那么检准率总是下降反之，若是限制检索

范围以改善检准率,则检全率总是变坏。

图 1–4 所表示的全部检索是在四种不同的“等级”上进行的。从图可以看出，如果进行很广泛的检索时（点 A),检全率很高,可以达到 90%左右，而这时检准率则很低；相反，当检索范围小，很专指时（点 D), 检准率较高，检全率又较低。点 B和点 C 的检全率和检准率都比较平均。这条曲线是美国情报学家 F.W.兰卡斯特（F.W.Lancaster ）根据 50 次检索的调查结果绘制的，所以称它为经验曲线。

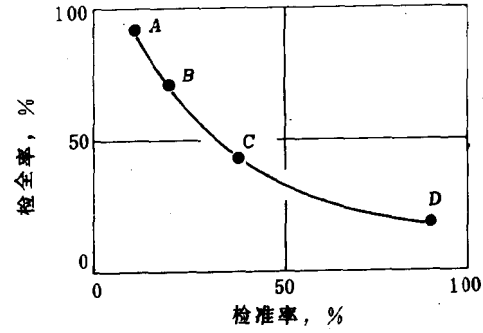

图 1–4 检索结果经验曲线

这条经验曲线实际上是一条平均曲线，也就是说它是根据若干次检索结果的平均情况绘制而成的。因此不能以此理解为每一个检索过程均为如此。在实际工作中，常常会遇到这种情况，有时检全率和检准率都可能达到百分之百。例如，某档案馆只保存了一件关于某主题的档案，检索时就被找出了。而有时，检全率和检准率都可能是 0,检出一大堆材料，均属无关文献。如果把每次检索的结果具体标出来，就形成了如图 1–5 的散点图。

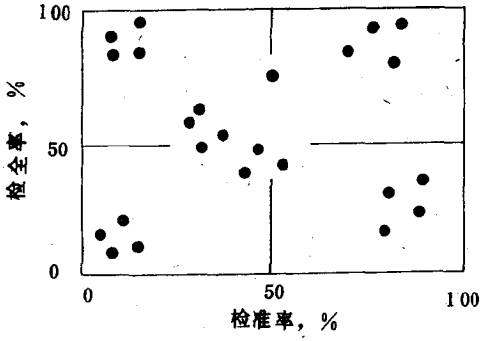

图 1-5　检索结果散点图

在图 1-5 中，每个“ · ”代表一次检索结果。实际上每次检索结果不一定都是互逆的。有时检索效率很好，检全率和检准率均很高(右上角)有时检全率和检准率又都很低(左下角)某些结果是检全率高，检准率低 某些结果是检全率低，而检准率高。这些结果平均起来，就获得检索效率的经验曲线，显示出检全率和检淮率之间的互逆关系。

三．影响检索效率的因素

究竟有哪些因素影响检索效率，这是一个复杂的问题，有待于深入研究．目前可以归纳出以下几个因素。

(一)检索人员的素质

不论是手工检索系统还是机械检索系统，都要由检索人员来参与和控制检索过程，因此检索人员的素质对检索效率有直接的影响．检索人员的基本素质主要是应具有一定的科学文化知识水平(包括准确地表达情报需求、正确地进行主题分析)和检索技能(包括对检索工具的熟悉和使用，对检索策略的灵活运

用等）这两个方面。据国外一次试验表明(MEDLARS系统试验),在检全率失败的原因中,查找过程的失误占 35%,在检准率失败的原因中，查找过程的失误占 32.4%。检索过程是检索人员控制的,可见,检索人员的素质对于检索效率的影响是很重要的。

（二)检索语言的性能

人们在存贮和检索档案时都要借助于检索语言。某一种检索语言的词汇、语法对于据此而构成的目录的功能有直接的影响,如分类语言中的交替类目、参照类目,主题语言中的语义参照系统有助于提高检全率；检索语言的专指性能则对于检准率影响极大。采用性能好的检索语言可以使检索系统具有较理想的检索效率。

（三)检索途径

任一档案在存入检索系统之后，该系统向利用者提供的检索途径愈多,它被查到的概率也就愈高,如某一份档案在检索系统中只向人们提供一条检索途径，那么人们只有找到这唯一途径,才有可能获得这一份档案。如果有六条检索途径可供查检，那么只要找到其中任一途径便可获得,这样检全率、检准率自然都会相对提高。检索途径的多少，就使用单一的检索工具而言，取决于标引的深度，就使用整个检索系统而言，除标引深度外，还取决于目录的种类及数据库内部的数据结构。当然，如果检索途径过多,则会加重系统的负担。

（四)著录、标引质量

检索标识是组织检索工具、进行检索的依据,因此,检索标识的准确性对于检全率、检准率也是一个重要的因素。通俗地说,没有存进检索系统或存的不准确,就难以取出来。如果在著录标引时主题分析不全面,有漏标现象,就会造成漏检。漏标就

是指标识出来的主题概念少于档案中论述的主题概念，也就是该提炼的主题概念没有提炼出来。例如，《××同志在三级干部会议上关于镇反、土地改革和抗美援朝运动的讲话》应标出镇反、土地改革、抗美援朝三个主题词。如果漏标‘土地改革’这一主题词，那么意味着‘土地改革’的类目下未能显示出这份档案，按照这一途径检索，就会产生漏检。如果主题分析或概念转换有误差，形成误标，就会导致误检的发生。误标就是指标识出来的主题与原档案主题不符。例如，有关对某单位违反经济纪律行为进行立案审查的文件，其主题应标为‘违纪审计’，如果标“财政制度”即为误标，这一份文件就会在‘财政制度’类中被错误检出。

第三节　档案检索系统评价

如前所述，档案检索系统由诸多因素构成，每一个因素的质量水平以及诸因素之间的协调方式构成了检索系统的总体性能。检索系统总体性能的优劣，可通过系统评价得出结论，因而系统评价对于检索系统的建立和完善是十分重要的。系统评价指标可作为建立检索系统的标准，其基本和主要的指标表现为质量、时间和费用三种因素。

一、质量因素

档案检索系统的质量因素就是指系统的检索效率，即利用者获取所需档案的完整、准确程度，具体表现为检全率和检准率的高低。

一般说来，检索效率是系统评价中最重要最根本的因素。因为检索效率实质上反映了系统运行结果与利用者检索要求的

吻合程度。利用者使用档案检索系统的根本目的是要获得所需档案,系统愈是齐全、准确地检出有关文献,就越接近利用者的预期目的，利用者的满意程度就愈高。只有在能够达到目的满足需求的前提下，利用者才需要考虑和选择达到目的的难易程度,即系统的时间、费用等方面的因素,反之,如果系统不能基本满足利用者的需求,提供所需文件,那么即使响应速度很快,费用低廉，对于利用者来说也是不足取的。

通常情报检索系统的一般水平是，检全率指标达到 60 ～ 70%左右,检准率指标达到 40 ～ 50%左右,但是具体到各个检索系统则不尽然。因为选择检全率和检淮率指标的对应关系应由所建检索系统拟解决的任务来决定。对于某一利用者和某一具体的检索过程来说，对检索效率的要求也常常各有侧重。例如,历史学家和编史修志人员,为了尽可能穷尽某一时期、某一领域、某一地区的文献,他们对检全率有很高的要求 而某些从事具体工作的人员，为了查证某个事件和问题，要求高检准率，而无需获得全部相关文献。检索系统尽管选择了检全率和检准率的指标,但是针对上述这些不同的需求,还应具有一定的调节能力，如检索人员可通过选择不同专指度的检索标识和不同的逻辑表达式等方法控制检索范围和检索深度，以接近或达到利用者的不同需求。

采用检全率和检准率的指标，虽然可以比较理想地描述出某一系统的检索效率，但也有一些局限性。也就是说这两个指标只能认为是对检索效率的近似反映。其主要原因是：

1. 在计算检全率时，是以检出的相关文献为基本数据的，即检出的相关文献在全部有关文献中的比例。在这里相关文献被同等看待,不作任何区分。实际上,在一组相关文献中,每一份文献的相关程度是有差异的，即文献中相关内容的多寡、深