情报检索系统

——特性、试验与评价

（第二版）

（美）F·W·兰卡斯特 著

陈光祚 王知津 王津生 译

书目文献出版社

1984·北京

F. Wilfrid Lancaster

Information Retrieval Systems: Characteristics, Testing and Evaluation

2nd Edition

情报检索系统——特性、试验与评价

（第二版）

（美）F·W·兰卡斯特 著

陈光祚 王知津 王津生 译

科学文献出版社 出版

（北京文津街七号）

秦皇岛市第二印刷厂排版

涿县辛庄印刷厂印刷

新华书店北京发行所发行 各地新华书店经售

850×1168毫米 1/32开本 12$\frac{7}{8}$印张 300千字

1984年5月北京第1版 1984年5月北京第1次印刷

印数 1—9,300册 定价：1.60元

图书分类号：G354 统一书号：7201·45

目　次

译本前言
序　言 …… 1
第一章　情报检索系统的功能 …… 1
§1　情报服务的作用 …… 4
§2　情报需求的类型 …… 6
§3　情报检索系统 …… 7
§4　情报检索系统的组成 …… 13
第二章　匹配子系统 …… 15
§1　先组式系统的局限性 …… 15
§2　后组式系统 …… 22
§3　文献／词语矩阵 …… 32
第三章　计算机在情报检索中的应用：脱机批处理系统 …… 37
§1　MEDLARS标引 …… 39
§2　书目引文的加工处理 …… 50
§3　查问的回溯检索 …… 54
§4　定题情报提供 …… 69
§5　计算机系统的优点 …… 73
第四章　联机情报检索 …… 75
§1　联机检索系统的若干特性 …… 75
§2　检索步骤 …… 77
§3　文档结构 …… 79
§4　联机检索系统的若干其它特点 …… 81
§5　简史 …… 82
第五章　机读数据库的增长 …… 85
§1　机读文档的可用性 …… 85

第六章　缩微资料与缩微资料检索系统……98
§1　文献提供系统……99
§2　缩微资料检索系统……101
第七章　情报中心与情报服务……104
§1　咨询工作……104
§2　关于正在进行中研究项目的情报……105
§3　文献服务中心……108
§4　情报分析中心……111
§5　网络……113
§6　国际情报服务……114
第八章　情报服务的评价标准……118
§1　查全率与查准率……121
§2　查全率与查准率的替换测度……128
§3　其他性能标准……130
第九章　情报服务效果的评价……133
§1　评价的主要步骤……134
§2　性能指标的推导……138
§3　评价结果的分析……146
§4　影响情报检索系统性能的因素……152
第十章　潜在的需求与表出的需求……153
§1　用户／系统交互……160
第十一章　数据库的选择与检索……167
§1　检索策略……169
§2　联机检索实施指导……178
§3　加权词检索……187
§4　输出的筛选……189
§5　影响检索成败的因素……190
第十二章　词表控制……193

§1 叙词表……195
§2 词表对检索系统性能的影响……201
第十三章 标引子系统……210
§1 标引的穷举度……211
§2 标引的质量与准确度……213
§3 标引的一致性……214
第十四章 根据评价结果改进情报服务的性能……217
第十五章 机读数据库及其情报服务的评价……221
§1 服务中心的评价……228
第十六章 费用／效果评价与费用／效益评价……236
§1 数据库的费用／效果问题……240
§2 标引的费用／效果问题……246
§3 索引语言的费用／效果问题……251
§4 检索步骤的费用／效果问题……253
§5 情报系统的利弊平衡问题……257
§6 费用分析诸因素……262
§7 效益与费用／效益研究……266
第十七章 国家情报系统的评价……272
第十八章 适用性与相关性……278
§1 相关性……283
§2 适用性……284
§3 论述相关性的文献……289
第十九章 情报系统评价概述……297
第二十章 情报检索中的自然语言……304
§1 自然语言系统与控制词表系统的比较……310
§2 自然语言的检索……315
第二十一章 自动系统……321
第二十二章 非正式交流的作用……329

第二十三章　用户与用户需求 ……342
第二十四章　情报服务的设计 ……350
§1　设计上的一些想法 ……351
第二十五章　未来：无纸情报系统 ……356
§1　当前科学交流中的问题 ……357
§2　自动化的成就 ……360
§3　未来的展望 ……362
§4　结论 ……366
主题索引（英／汉）……367
主题索引（汉／英）……379
人名机构名索引 ……389
缩略语词表 ……395

第一章　情报检索系统的功能

本书是讨论“情报服务”、“情报中心”，特别是“情报检索”的，本章试图对这些贯穿整个教程的术语的涵义，给出定义或说明。

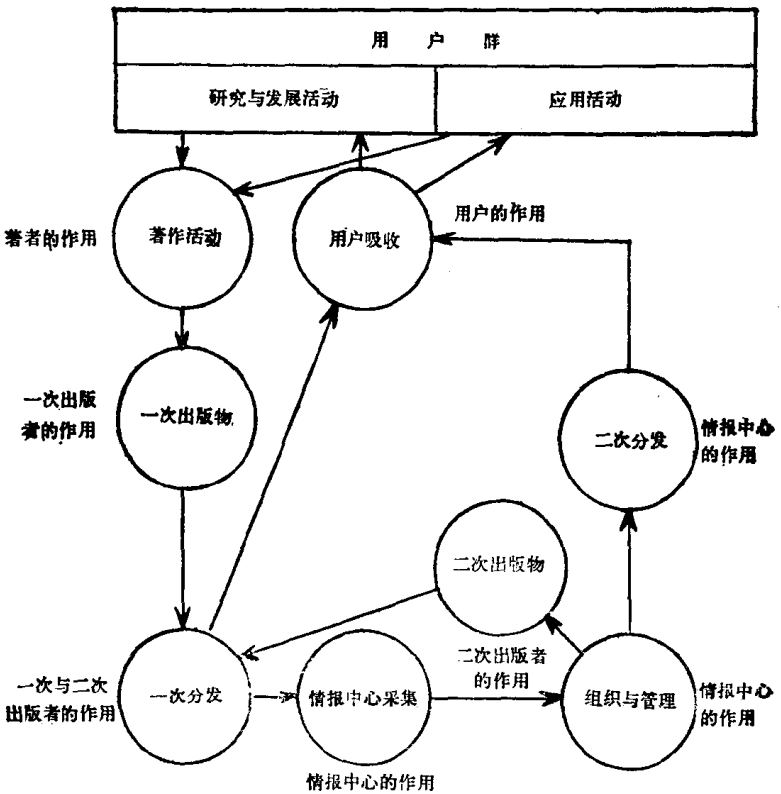
当从情报传递的正式渠道的完整循环圈这个更广泛的范围来看问题时，对情报中心所起的重要的功能也是最容易理解的。图一表示该循环圈的主要部分。简而言之，“用户群”就是在特定学科领域里工作的人群。他们之中有些人从事“研究与发展活动”，有些人则从事图一中不确切地叫作“应用活动”的其他各种活动。从某种意义上说，他们都是情报用户，并且其中有些人也是情报产品的创造者。这意味着，假如用户群中的某些人的活动使另一些人感兴趣，那么，这些人则以报告的形式，阐述他的经验、研究或看法。这便是交流循环圈中的“著者的作用”。但是，著作活动本身并非是一种交流形式。在把著者的著作大量复制并正式分发（即出版）之前，它对专业团体影响很小或没有影响。复制与分发就是交流循环圈中的“一次出版者的作用”。一次出版物可以是图书、杂志、技术报告、学位论文、专利等等。

在图一中，用两条线表示一次出版物的分发：

1．通过个人的预订和购买直接到达用户群。

2．通过情报中心的预订和购买间接到达用户群。

在情报传递循环圈中，情报中心——在图一中，通常用这个词表示图书馆、其他类型的情报中心和二次服务出版社（商）——具有非常重要的作用。通过采集与存贮方针，图书馆提供专业成就的永久性档案和存取这种记录的有保障的资料。此外，通



图一 情报传递循环圈

过编目、分类、标引及其他步骤，图书馆与其他情报中心一起组织和管理文献。文摘索引服务社和国家书目出版社所起到另一种组织与管理的重要作用。这些机构负责出版与分发“二次出版物”*。某些二次出版物可以直接到达用户群，但是，对于绝大多数二次出版物来说，不是到达个人手里，而是到达公共机构订户，即情报中心。

在循环圈中，情报中心也起到“提供与传播”的重要作用。这些活动构成出版物和出版物消息的二次分发的一种形式，其中包括资料流通以及现期通报、参考咨询和文献检索服务。七十年代，在提供各种情报服务的过程中，二次出版社（商）的机读数据库发挥着日益重要的作用。

如图一所示，循环圈的最后阶段是“吸收”。这就是用户群吸收情报的阶段，但是，这个阶段是最不确切的。这里，要对“文献传递”和“情报传递”加以区别。后者，只有在用户研究了文献，吸收其内容，以达到这文献能改变他对该文献主题内容认识的程度时，才会发生。专业团体吸收情报可以通过一次或二次发行完成。不同的文献具有不同的与其相联系的吸收级别与速度，某些内容由于从来也不使用，而根本不曾为人们所吸收。

之所以把正式交流过程作为一个循环圈提出来，是因为它们是连续的和反馈的。通过吸收过程，读者可以获得能用于他的研究与发展活动的情报，反过来，这些研究与发展活动又产生出新的著作和出版物，从而，循环圈又继续下去。

在一个重要方面，图一是过于简单化了。它表示出正式渠道的情报传播，但没有明确说明非正式过程。然而，非正式渠道一般也传播正式渠道所传播的情报。两者传播相同的经验和研究成果。非正式渠道与正式渠道的区别在于：非正式渠道传播情报的方式不同，或者方式相同但时间早得多。例如，草稿和预印本的发行。非正式渠道是重要的，因为它传播情报比正式渠道快，至少在专业团体内没有隔阂。另一个原因是，它们把情报传播给某些人，但由于种种原因，这些人不选用正式渠道。

简短地讨论情报传递循环圈，是为了弄清情报中心与服务在全循环圈中所起的作用。本教程主要讨论图一中标有“情报中心

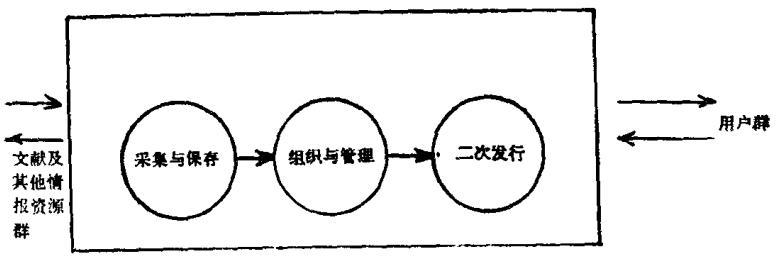
• “二次出版物”或译为“二次文献”——译者注。

的采集”、“组织与管理”、“二次分发”等环节所说明的作用。由于某些情报中心也建立了用于提供各种情报服务的二次数据库，所以，本教程至少部分地包括了“二次出版物”和二次出版物的“一次分发”等活动。图一所描述的其他活动，即有关著作活动、一次出版物及其分发和用户吸收情报等，不在情报中心与服务直接控制的范围之内，尽管本书有些地方也触及到这些活动，但是，本教程对它们不予详细讨论。

情报服务的作用

任何情报服务的主要作用都是在某一用户群和印刷型或其他形式的情报资源群之间充当接口（interface）（图二）。用户群通常以地域、机构会员、学科兴趣以及它们的某种结合来确定。就国家情报系统而言，用户群就是在该国工作的科学家和其他专业人员群。另一个群就是情报资源群。对于大多数情报中心来说，最重要的情报资源以文献的形式出现，这里，在最广泛意义使用“文献”一词。情报服务的作用就是尽可能有效地经济地把这两个群汇集起来（接口）。这就是说，在多少有点被动作用的意义上，情报服务的作用就是要保证作到：当用户群中的某一成员需求文献或情报时，应当尽可能地让他得到，在其较为积极作用的意义上，情报服务的作用就是根据用户解决问题或决策的需要，按要求进行文献检索，并且连续地向用户通报他们感兴趣的领域中新发表的文献（“现期通报”——Current awareness），从而引起用户群对文献或数据的注意。

一个有效的现代化的情报服务应当能够保证：使可以利用的文献群中的任何文献或文献中的任何数据，实际上都可以让被服务的用户群中的任何成员得到。这意味着，应当让用户群在可得性（accessibility）的不同级别上得到“文献资源群”。因为没有一个图书馆或情报中心能够收藏一切，所以，重要的是，补充藏



图二　情报服务的接口作用

书的文献应尽最大可能对其用户有价值。然而，同样重要的是，情报中心应能通过采集、照相复制或馆际互借，尽快得到用户群合理需要的任何其他文献。此外，在情报中心自己的藏书中，必须按所要求的级别组织文献，使最常用的文献最易获得。因此，可以说，情报中心的用户按可得性级别得到文献资源群，习惯上，可得性级别如下：

1．中心入藏的开架文献。

2．中心入藏的闭架文献。

3．离开收藏地点的文献。

4．中心未入藏的文献。

一般地说，这是可得性递减的顺序。但是，在某些情况下，这过于简单化了，获得中心未入藏的文献复制品，可能要比放错位置、送去装订或已借出的文献容易些。再者，中心未入藏的文献其可得性并不都一样，这是因为到该中心的距离可能对文献的可得性产生某种影响。还有另外一些因素也可能对可得性产生影响，其中包括出版物是否还在印刷中，是否收录在某种联合目录里，邮寄的可靠性与速度等等。

为了发挥情报服务的接口作用，情报服务需做三项工作（图

一和图二）：文献的采集与保存；文献的组织与管理；用各种服务手段（流通、文献检索、照相复制等），向用户提供文献或文献情报。

情报需求的类型

实际上，情报服务用户的主要需求分两大类：

1．对查找或获得某一已知著者或题目的文献的需求。

2．对查找某一主题或能够回答某一问题的文献的需求。

第一类需求可以称之为“已知文献的需求”，第二类需求显然是“主题的需求”。情报中心提供已知文献的能力，就是它的“文献提供能力”。情报中心检索某一主题的文献或回答某一问题的能力，就是它的“情报检索能力”。文献提供和情报检索这两个功能就是情报服务的主要活动。这两个功能密切相关，许多对已知文献的需求可能直接来源于早期的情报检索活动。

主题需求也可明确地分为两大类：

1．帮助解决某一问题或便于作出某一决定的情报需求。

2．有关某一专业领域新进展的情报。

后者通常称之为现期通报，但是，对第一类情报需求的描述，还没有一个被普遍接受的词。只能称之为解决问题的情报需求。实际上，通常由情报服务检索以往的文献，回答用户的某一问题，从而满足这类需求。因此，通常把这类检索叫作“回溯检索”（retrospective search）或“查问检索”（demand search）。

解决问题的检索在许多方面不同于现期通报检索。前者目的性较强——用户必须采取主动，而在现期通报的情况下，情报服务可以采取主动——；前者比较专门，可能包括相当多的文献——即回溯许多年，用户很可能觉得，这种检索结果比现期通报检索结果更令人信服。

解决问题的情报需求的本身也可分为几类：

1．需要单项的事实数据。这是典型的由图书馆处理的“快速参考”（quick reference)询问。虽然通常利用文献来满足这类需求，但提问者不一定接受任何文献——可以用电话回答这个问题。

2．需要一篇或多篇论及某一主题的文献，但要少于已出版的或可从某中心得到的全部文献。这是典型的图书馆工作，例如，某一提问要求超声波焊接方面的最新论文。有一类需求很特殊，即找到第一篇某类文献，就会完全满足需求。举例来说，专利审查员可能只需要文献中以前申请过的一个事例，以便拒绝某一专利权限。

3．需要全面检索（comprehensive search），即尽可能多地检出某一时期内已出版的某一主题的文献。正在编书或写评论文章的人，或者正在开始一项新的研究计划的科学家，正需要这种全面检索。还有一种特殊类型的需求，检索的目的是为了证实文献中不存在某一论题，即提问者认为这方面的文献从未发表过，并着手证明之。发明者想要证实某发明的首创性，就是这种需求的明显一例。

情报检索系统

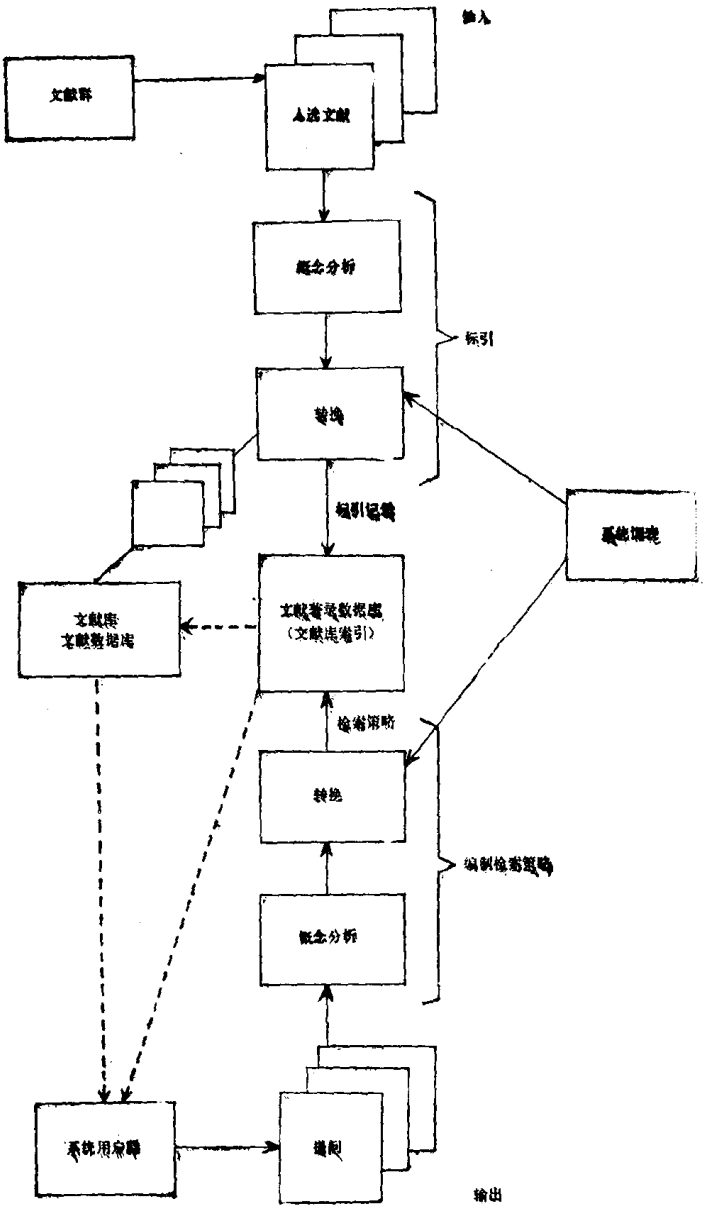
图三以较简单的形式，表示出许多情报系统的主要工作。该系统输入的是文献，即由情报中心采集到的文献。这意味着，存在着某种选择标准和方针，选择标准和方针的存在又意味着，详细而准确地了解被服务的用户群的情报需求。采集到文献之后，必须对它们进行“组织与管理”，以便能够识别与查找，回答用户的各种提问。组织与管理包括：分类、编目、主题标引和作文摘。如图三所示，主题标引过程包括两个很不相同的智力步骤：文献的“概念分析”——也可称“内容分析”——以及把概念分析

“转换”成词汇。一般不把这两个步骤区分开来。这实在令人遗憾，因为各个步骤都提出不同的约束，并且都产生影响该系统性能的不同因素。为了有效地进行概念分析，标引员必须既要明白文献是关于什么的，即理解文献的主题内容，又要透彻了解该系统用户的需求。弄清文献是关于什么的和用户为什么对它感兴趣，即文献的什么方面最引人注目，是构成了概念分析的主要之点。文献的概念分析可以记录在纸上，但是，最大的可能是只存在于标引员的头脑中。

标引工作的第二个步骤是把概念分析转换成某些词汇或“标引语言”。在大多数系统中，这一步要使用“控制词表”（controlled vocabulary）即必须表达文献主题内容的限定的一套术语。这种词表可以是主题标题表、分类表、叙词表或者是“经认可”的关键词表。十分明显，“非控词汇”（noncontrolled vocabulary）不能限定标引员采用的术语。非控词汇一般使用被标引文献中出现的词或短语。无论取自控制词表，还是取自非控词汇，在本书中，一般把标引员用来表达文献主题的词语都叫作“标引词”（index terms）。

标引过程一旦完成，文献就进入某种形式的文献库（数据库），而标引记录则进入二次数据库，在二次数据库中，按便于检索并回答各种主题的和其他的提问的方式，对标引记录加以组织。标引记录或“文献记录”（document representations）可以象卡片文档或印刷型索引那样简单。不过，在现代化系统中，它更有可能是磁带或磁盘上的机读文档。这种数据库可以称之为文献库的“索引”。

实际上，系统输出端上的这一步骤和输入端非常相似。接受服务的用户群向情报中心送交各种提问，中心的工作人员则为提问编制检索策略。检索策略的编制也包括概念分析和转换这两个步骤，这种看法是合适的。第一步是提问分析，确定用户实际上



图三　多数类型情报中心所起的主要作用

要找的是什么，第二步是把概念分析转换成该系统的词汇。转换成了该系统语言的提问的概念分析，就是“检索策略”，象把标引记录看作文献著录一样，也可以把检索策略看作提问著录。两者唯一真正的区别在于：后者通常含有“逻辑”，即指定各标引词间的逻辑关系，而前者通常不含逻辑，没有把各标引词间的逻辑关系明确地表示出来。

检索策略编制出来后，就以某种方式将其同文献著录数据库进行“匹配”。这可以包括检索卡片文档、印刷型索引、缩微胶片或磁带、磁盘，从数据库中，把同检索策略匹配了的（即满足检索的逻辑要求）文献著录检出来，并送交提问者。这个过程可以反复进行，当提问者对检索结果表示满意时，该过程才算完成。在某些情况下，这个过程可能意味着，对于数据库中没有同提问者的需求确切相关的文献，而提问者感到满意。

图三描述的各个步骤说明了“委托检索”(delegated search)的情况，即需求情报的人把检索数据库的责任委托给情报专家。在“非委托检索”（nondelegated search）的情况下，由于用户直接接近数据库，检索过程多少有点简单。然而，既使在这种情况下，用户也必须对他自己的情报需求进行概念分析，并将他的分析转换为系统语言。当然，在检索多种系统时，构造检索策略不能脱离数据库和检索工作本身。在检索卡片式目录、印刷型索引或联机系统时，检索策略很可能是交互式地或启发式地发展，也就是说，概念分析和转换工作或多或少地同文档的查找工作同时发生。不过，既使在这种情况下，也需要某种形式的概念分析／转换工作。回溯检索服务和现期通报——如定题情报提供（SDI）——之间的唯一真正的区别在于，后者的检索策略或“用户提问档”（user interest profiles）代表着系统用户当前的研究兴趣，将这些检索策略定期地同新到的文献著录进行匹配，也就是说，数据库在时时更新，匹配的结果也定期地送交用户。

当然，对情报服务的某些提问不是针对某一主题的情报或文献，而是针对已知著者或题目的某一特定文献。在已知这些提问项目的情况下，通过藏书的索引或目录中的著者或题目中的存取点（access points），或者通过其他一些途径，例如，报告号或专利号，向文献库提问（见图三）。中心提供所需文献的能力叫作它的文献提供能力。

虽然以上讨论已对情报检索系统的范围作了一些暗示，但到此为止，仍然没有给出这个词的精确定义。在最普遍地使用情报检索一词时，它确实是文献检索的同义语。当在最广泛的意义上使用文献一词时，情报检索是查找某一文献库的过程，以便找出那些某一主题的文献。为便于文献查找工作而设计的任何系统，可以合理地称之为情报检索系统。图书馆的主题目录是一种情报检索系统，印刷型索引也是一种情报检索系统。

情报检索系统的输出，通常由一条或多条文献书目构成，书目也许还附带一些补充的情报，例如，文摘或标引文献的词语表。一般把这些文献著录送交要求检索的人。然后，这个“提问者”可以要求情报中心或其他某一中心，向他提供文献检索输出中的某些或全部文献。在某些情况下，情报中心删去中间步骤，直接向提问者提供文献本身或者提供工作人员认为最可能相关的那些文献。情报检索功能与文献提供功能偶尔也结合在同一个系统里。例如，缩微胶卷检索系统或计算机系统可以存贮短文献的全文（如电报或报纸文章），检索输出是文献本身的打印件，而不是文献著录。但是，大多数情报检索系统停留在提供文献著录阶段上。提供文献则是完全独立的一项工作，文献检索部门是否提供文献本身都可以。

从以上的讨论中，应当明确，情报检索并不是一个描述通常用于这一工作的特别令人满意的词。情报检索系统并不检索情报。的确，情报是个无形的东西，不能看到它、听到它、摸到它。