

情报检索理论概述

 Qingbao jiansuo lilun gaishu

4
科学技朮文献出版社

情报检索理论概述

[苏] A. И. 切尔内 著

赵宗仁 许恒泰译

于 桦校

科学技术文献出版社

内 容 简 介

本书以浅显的语言概述了情报检索的一般问题、情报检索效果、各种情报检索语言和实现情报检索系统的重要设备。书后对200多个情报检索中常用术语给出了定义。本书是联合国教科文组织和工业发展组织在莫斯科举办的地区际工业情报和文献工作培训班上讲课的教材。本书可供普通情报检索人员、计算机情报检索人员，广大情报、图书工作者参考，也可作为情报检索训练班教材使用。

А. И. ЧЕРНЫЙ

ВВЕДЕНИЕ В ТЕОРИЮ ИНФОРМАЦИОН- НОГО ПОИСКА

Издательство. Наука. Москва 1975

情报检索理论概述

[苏] А. И. 切尔内著

赵宗仁、许恒泰译

于 桦校

目 录

序 言	(1)
作者的话	(5)
第一章 情报检索基本原理	(7)
第二章 情报检索语言	(22)
§ 2.1 概 述	(22)
§ 2.2 情报检索 语言的种类	(28)
§ 2.3 等 级 分 类 法	(33)
§ 2.4 字 顺 主 题 分 类 法	(40)
§ 2.5 组 面 分 类 法	(49)
§ 2.6 叙 词 语 言	(55)
§ 2.7 语 义 代 码	(81)
§ 2.7.1 佩里——肯特语 义 代 码	(82)
§ 2.7.2 RX代码语言	(85)
§ 2.8 结 构 段 语 言	(87)
§ 2.9 关于利用自然语言作为情报检索 语言问题	(91)
第三章 输出标准	(100)
第四章 情报检索效果	(117)
第五章 检索档的组织方式	(128)
第六章 实现情报检索系统的技术工具	(140)
§ 6.1 书 本 式 索 引	(142)

§ 6.1.1	轮排索引	(142)
§ 6.1.2	双轮排索引	(145)
§ 6.1.3	题外关键词索引	(146)
§ 6.1.4	引文索引	(147)
§ 6.1.5	现组式组配索引	(147)
§ 6.2	目录卡片	(167)
§ 6.3	单元词卡片	(169)
§ 6.4	手工穿孔卡片	(171)
§ 6.4.1	重叠比孔卡片	(172)
§ 6.4.2	边缘穿孔卡片	(173)
§ 6.4.3	开槽穿孔卡片	(176)
§ 6.5	机器穿孔卡片	(178)
§ 6.6	带有文献缩微品的离散式 情报载体	(181)
§ 6.7	带有文献缩微品的连续式 情报载体	(183)
§ 6.8	计算机	(184)
第七章	情报检索系统的工作方式	(189)
第八章	情报检索理论方面使用的主要术语	(204)
附 录	俄汉术语对照索引	(240)
参考文献		(254)

序 言

本书是国内文献中首次企图以简明易懂的形式论述情报检索的各种困难问题。目前在撰写象无线电、电视等书籍时，一般按照通俗易懂的原则来写，可是写别的书籍，例如写《情报检索理论》，一般就按复杂的原则来写了。对此我们需要指出，《情报检索理论概述》一书，是作为联合国工业发展组织（UNIDO）和联合国教育、科学及文化组织（UNESCO）联合举办的地区际工业情报和文献工作培训班的讲座而撰写的。切尔内（А. И. Черный）只是力图使本书的叙述和结构简明，而不是使问题简化。

因此，应当承认，这种态度是唯一正确的。四十年代末至五十年代初提出的情报检索问题，是二十年来的一个中心问题。围绕这一问题正在形成一个新的科学部门——情报学。这一时期解决情报学的道路已经发生了本质的变化。如果最初阶段对这一问题的技术方面比内容方面曾给以更大的注意的话，那末现在，中心显然已转到语义方面。

情报学这一趋向的相对稳定（过去没有重大发现，原则上没有提出新的建议，正在进行许多技术研制），并不意味着情报检索问题已经解决（如某些人所断言的那样），或者已走入死胡同。显然，积累新知识、理解已积累知识的时代已经到来。本书的价值在于，对于情报检索理论的现状、未披露的深度和解决其中心问题途径的极端复杂性作了概括的

论述。读完本书后就会明白，只有综合研究语言和思维的一系列基本问题，才能继续前进。过去和现在，语言学、逻辑学、哲学和心理学还未被成功地应用于研究这些问题。

以上所述只是想引起注意：轻视这一理论的意义或把它复杂化同样都是危险的。许多工程师、科学工作者和图书馆员在未受过情报学专门教育的情况下参加各种科学情报活动，他们只从一种健康的理智立场或传统的图书馆目录概念出发，就想开始解决情报检索问题。这同样也会误入迷途，甚至会在本学科中产生职业偏见。最明显，最普遍的错误之一是他们（特别是图书馆工作人员）确信，组配标引法和叙词情报检索语言与图书馆目录分类法没有本质区别，它们是主题法的一种。

关于这一问题，一个情报检索理论的先驱和创造者穆尔斯（C. N. Mooers）不久前指出：《叙词表明——他写道——局限于分类法是不必要的。叙词的出现是分类法进一步发展的结果。并且令人信服地证明，叙词可以迅速有效地实现‘工业生产’，而且类似的叙词系统无论对于小型文档还是大型文档的检索，同样都是有效的。但是，作者奇怪的是，完全出于他所预料，大部分图书馆专业的代表仍然不能理解叙词的意义……在作者深信这一点没有其他解释之前，十年过去了，当他敢于提出这一深刻的假设到现在又是十年。诚然，我们人类的大多数，不可能离开表达思想的词语来交流思想（除初级水平以外）》¹。

注：1. C. N. Mooers, Descriptors, ——载：Encyclopedia of library and information science, vol. 7. New York, Dekker, 1972, 40—41页。

可以不赞成穆尔斯 (C. N. Mooers) «耽溺于词语», 或用色盲来类比, 说成«思想色盲», 比作某些人失去了识别颜色的能力。但不能不承认, 用一切可能的方法阐明组配标引情报检索法的新内容是完全必要的。在这一方面, 本书比起以前的一些名著——范尔特尔 (С. Р. Фэртон)、肯特 (A. Kent) 和维克里 (B. C. Vickery) 关于情报检索理论的专著大大前进了一步。切尔内 (А. И. Черный) 以严格公正的形式 (考虑到专业上的自负这是很重要的) 客观地评价了传统的和新的情报检索语言的相对成就和缺点。

我们认为, 书中可以找到, 对不理解叙词分析真正本质的事实作出的十分令人信服的说明。这一说明是, 到目前为止对于选词的过程和原则还未给以足够的注意。«某些作者认为——切尔内 (А. И. Черный) 写道——各种分类法 (其中包括叙词情报检索语言) 之间的相互区别仅仅在于: 其中排列的次序不同, 而利用相同的词和词组。产生这种错误认识的原因在于, 原则上没有考虑下述重要情况: 某种情报检索语言基本词表中所有词与词组的选择, 在很大程度上是由这类情报检索语言对它们的使用方法所决定的»。特别是书中正确地指出, 叙词是这样一些词、词组和短语, 它们不能由该语言的基本词库中已有的成分通过逻辑乘单义地加以创造。这一简要公式中包含有可能获得好结果的许多新研究和研制的来源。

应当特别强调本书中的另一个重要问题是: 骤然看来, 在对情报检索语言的习惯分类中, 除按先组语言和后组语言划分外, 也没有多大变化。有人认为, 题录引文语言与这些情报语言也没有区别。认可题录引文语言 (或称引文语言)

是合格的情报检索语言，不仅是由于利用这种语言分析文献内容和相互关系的方法获得了广泛的应用。很明显，这种语言是一类新情报检索语言的代表。这些语言建立在利用作者与被引用某篇文献有这种或那种相互关系的基础之上。最新的研究表明，利用合作者的关系和作者在撰写序言和跋，编辑和评论等过程中的相互关系，可以解决一系列情报检索问题。

此外，计算机在情报工作中的广泛应用使情报检索中有可能利用文献的许多题录特征，这些特征与文献中所包含的情报有某种关系。首先，这些特征是指文献标题、子标题和题头事项。毫无疑问，这一情报检索趋向是有发展前途的。可以继续列举本书具有革新意义的各章目次。切尔内（А. И. Черный）明确提出关于利用自然语言作为情报检索语言的问题，并以新的观点分别讨论一般只顺便提到的问题，如输出标准、情报检索系统的工作方式。

本书是根据现代科学成就撰写的，书中包括作者多年钻研和他所领导的全苏科学技术情报研究所科学工作者小组研究的成果。同时，本书可供不具备这一专业知识的读者阅读，也可推荐给广大读者阅读。

米哈依洛夫（А. М. Михайлов）

吉里雅列夫斯基（Р. С. Тилиревский）

作 者 的 话

《情报检索理论概述》一书是一部未经重大修改的讲义。这是作者1972年9月为地区际工业情报和文献工作培训班的学员撰写的。本培训班是全苏科学技术情报研究所受联合国工业发展组织（UNIDO）和联合国教育、科学及文化组织（UNESCO）的委托于1972年8月28日至11月12日在莫斯科举办的。作者编写本讲义的目标是扼要叙述情报检索理论的原理——自然包括对这一理论的理解和说明——，并力求使本书简明易懂。论述的顺序是从简单到复杂。在对专门术语未作解释或下定义以前，作者没有使用任何专门术语。作者究竟在多大程度上达到了上述目标，将由本书的读者加以评定。

众所周知，情报检索理论还处于形成阶段。这一理论的基本概念和术语体系还未最后形成。因此，作者深知本书中某些观点和论述带有讨论的性质。但是，不管读者是否同意本书中对某一问题叙述的观点，作者仍然期望，本书能促进读者再一次批判地考虑他所感兴趣的情报检索问题的立场；或者坚定自己的立场，或者改变自己的立场。因此，本书将有助于明确情报检索理论的基本概念和术语，也有助于促进这一理论的研究。

在编写本书时，作者考虑到了我国近年来就某些专题（如国际十进分类法和实现情报检索系统的技术设备）出版的

不少专门论著和文集。因此，本书对这些问题只作了扼要的论述。

当然，作者非常感谢读者对本书的内容与形式的一切批评指正。

第一章 情报检索基本原理

目前，全世界每年出版约三百万篇自然科学与技术期刊论文、七万至七万五千册书籍、约四十五万件专利和作者证说明书以及几万种其他科学技术文献。因此，要在浩如烟海的科学技术文献中查找我们感兴趣的主题或问题的知识这类出版物，正在成为愈来愈复杂的问题。瓦维洛夫 (С. И. Вавилов) 院士认为，「…现代读者就像沙里淘金一样站在图书馆的喜马拉雅山面前」¹。

如何在如此大量的文献中查找我们感兴趣的主题或问题的出版物？最简单可靠的方法是阅读或仅浏览一切出版物，从中选出符合我们情报需要的文献。但是，由于文献量太多，即使是任何一个科学或技术领域，采用这样的方法检索需要的文献，实际上也是完全不能接受的。因为这种方法要求耗费很多时间、花费大量的脑力劳动。不难计算，假设每周用40个小时，每小时浏览论文30篇，那末仅以1972年发表的化学论文为例，要看完它就需要五年时间！

克服这些困难的唯一方法是按类给每篇文献作一个「标签」，在这个标签上写上这篇文献讨论的主要问题或主题的简要特征。在检索理论中，这种「标签」称为文献的检索标志。这样，为查找我们需要的文献必须阅读或浏览的就不是某

1. С. И. Вавилов. Несколько замечаний о книгах. «Советская книга», 1947, №1, стр. 15.

一图书馆全部文献的全文，而仅仅是检索标志的全文。因此，我们可以大大加速检索过程。用文献检索标志表示的文献内容愈简要，检索时愈能节省更多的时间。但是我们必须注意，表示的文献主要问题或主题愈简要，检索就愈不全、不准。这是因为，文献的检索标志中只给出了文献的主要问题、主要主题，而没有给出文献的次要问题或主题。因此，检索时将找不到那些非主要问题或主题而是次要问题或主题的文献。

例如，发表了一篇主要论述“卡米尔”飓风的论文（1969年）。当时，在这篇论文的检索标志中没有给出这篇论文中报导这一飓风袭击毗连墨西哥湾的美国许多地区所遭到的破坏消息。因此，查找1969年飓风对世界各国居民带来的损失文献时，这篇关于“卡米尔”飓风的论文就不可能找到。

因此，为了大大加速文献检索，我们不得不降低查准率和查全率，因为放弃了阅读或浏览文献全文，只阅读这些文献的检索标志。我们标引的检索标志愈少，检索速度就愈高，但同时查准率和查全率就愈低。

但是，文献检索问题不只是用阅读文献的检索标志来代替浏览文献全文就算完事。为了解决这一问题，必须使对于检索的问题或主题只有一般概念的情报机构或图书馆工作人员能够进行检索（即阅读检索标志，并将检索标志与情报提问比较），而不是科学家或工程师自己进行检索。如果用机器来完成检索标志同情报提问的这种比较，那就更好。

目前，情报检索机械化问题由于一系列原因愈来愈尖锐了。其中主要原因是：

1. 必须进行情报检索的文献库迅速增长。这使检索程

序愈来愈繁难了。

2. 由于学者与专家的数量及其所研究的科学技术问题的数量和复杂性不断增长,使情报提问的数量、复杂性和紧迫性迅速增长。

因此,及时、准确和全面地提供科学文献来答复大量的和各种各样的提问是一个具有广泛服务性质的问题,这一问题只有采用相应的机械化和自动化设备才能有效地解决。

不研究检索标志的意义,怎样达到使文献检索标志同情报提问的比较过程实现机械化?为此,文献检索标志和情报提问必须用这样一种语言表达,即这种语言的词和短语是单义的,不允许有多种解释。这种语言叫做**情报检索语言**。那时,用情报检索语言表示的情报提问可以同也用同一种语言表示的文献检索标志进行机械比较。如果这两种表示完全重合(或第一种表示完全包含在第二种表示之中,即是第二种表示的一部分),那末就认为该篇文献能够答复情报提问并提供给用户。据以确定用同一种情报检索语言表示的文献检索标志和情报提问之间必须充分一致程度的那些标识总和称为**输出标准**。

最简单的输出标准是情报提问同文献的检索标志完全一致或检索标志中完全包含有情报提问。

这样,文献检索过程就是将文献检索标志同情报提问进行比较。在技术上如何实现这种比较?看来解决这一问题的最简单方法,就是把文献检索标志连同其题录或地址存贮在专门的载体上,以便将检索标志同情报提问进行快速比较。作为这种载体使用的有普通目录卡片、各种穿孔卡片、缩微胶卷、磁带、磁鼓和磁盘以及磁卡和视频磁卡。

现在可以引入更严格的定义。**情报检索**是在某一文献集合中查找某些文献的过程，这些文献是在某一问题(或主题)的情报提问中指出的或包含有用户需要的事实、知识。在这里，我们把记录在某种物质载体上具有一定逻辑完整性并包含有来源或作者信息的任何能理解的文字称为**文献**。按照这一定义，文献不只是指图书、论文、著者证或专利说明书等，而且指这些文字的独立片段——章、节、段等等。

按照查找对象划分有两类情报检索：**文献检索**和**事实检索**。**题录检索**是文献检索的一种。文献检索时，查找对象或者是文献的题录（题录检索），或者是这些文献本身或它们的复制品。事实检索时，查找对象是具体事实或现象（例如，某些机器的技术特点，机器图，物质与材料的性能等等）的说明，即事实记录。

情报检索通过**情报检索系统**实现，有手工的情报检索系统或利用相应的机械化和自动化设备的情报检索系统。具体的文献情报检索系统的例子可以举书本式索引、图书馆中的图书目录或书库等等，而最简单的事实情报检索系统可以举电话号码簿、通讯录、产品目录等等。这两类情报检索系统的原则区别仅仅在于，两种系统的检索对象不同：文献情报系统的检索对象所需情报的文献（或这些文献的地址），而事实情报检索系统的检索对象是以某种形式记录的事实、知识。

有时，某些作者把事实情报检索系统划为一种特殊的情报检索系统，仿佛从复杂性来看，这些系统比文献情报检索系统具有更高的水平。我们认为，这种观点是完全没有根据的。

任何事实要成为检索对象必须取某种文献形式。虽然事实本身可以脱离文献而存在，但是关于事实的知识必须用文献记录，因为检索的概念必须是及时传递情报。

事实用文献记录意味着，这一事实必须用某种自然语言或人工语言表示或记载，而且在这种记载中必须指出这一事实何时何地产生。因此，与文献检索一样，进行事实检索时我们也要检索文献。如果事实情报检索系统和文献情报检索系统或多或少有所区别的话，那末这些区别只应从检索对象和检索任务的特点中寻找，而不应从建立这类情报检索系统的基本原则和方法中寻找。

某些事实情报检索系统由于检索对象和检索任务的特殊性，要求研制专门的情报检索语言。例如，为了检索具有所需性能的化合物，要求按这些化合物的结构式特殊片段建立专门的情报检索语言。1964年全苏科学技术情报研究所建立的事实情报检索系统«ФТОР»中采用的就是这类情报检索语言。利用这种情报语言能迅速查找一定化学形式的反应、获得某类化合物的方法、某类化合物参加的反应、进行化学反应的条件反应、某类化合物和关于化合物性能的知识等²。

但是，事实检索的许多对象都有类似的特点。例如，电工与自动装置的仪器和设备零件检索系统（1967年英国建立）使用一般情报检索语言，与各种文献检索语言并无原则

2. Г. Л. Мищенко, Е. М. Рохлин, Л. В. Тарасова. Информационно-поисковая система для химии органических соединений фтора (Описание и правила пользования). М., ВИНТИ, 1969.

区别³。

实践表明,如果把事实检索理解为查找过去已输入情报检索系统的事实说明过程,而不理解为从其他事实中逻辑输出所需事实的过程,那么在建立文献和事实情报检索系统的理论和方法中不存在多少本质区别。因此,为了使事实检索同文献检索对立起来,特别对事实检索赋予特殊的理论意义是没有任何根据的。

事实检索必须与逻辑信息处理相区别,没有逻辑信息处理,要直接向人回答所提出的问题是不可可能的。事实检索时,所查找的(可能找到的)仅仅是已输入情报检索系统的事实或知识,不可能找到没有以明确形式输入本系统的情报(虽然这些事实或情报可能合乎逻辑地从已输入系统的文献所包含的内容中得到)。

为了在用一种或几种语言记录的文献正文中查找需要的事实,为了从已有的事实中通过对它们的逻辑输出获得这些事实,必须通过算法从这些正文中吸收其语义内容,对之进行分析和综合等等,即逻辑信息处理。这类任务只有借助于逻辑情报系统才能解决,关于逻辑情报系统将在最后一章末论述。

情报检索系统的功能见图1。从图中可以看出,向情报检索系统输入时,每篇文献必须标引。这里及以后,我们把

-
3. D. H. Barlow, *Designing a Commercial SDI Product awareness Service*.——载: Samuelson K. (ed.), *Mechanized Information storage, retrieval and dissemination*, (Proceedings of the F. I. D./I. F. I. P. Joint Conference, Rome, June 14—17, 1967), Amsterdam, North-Holland Publishing Co., 1968, pp. 349—363.