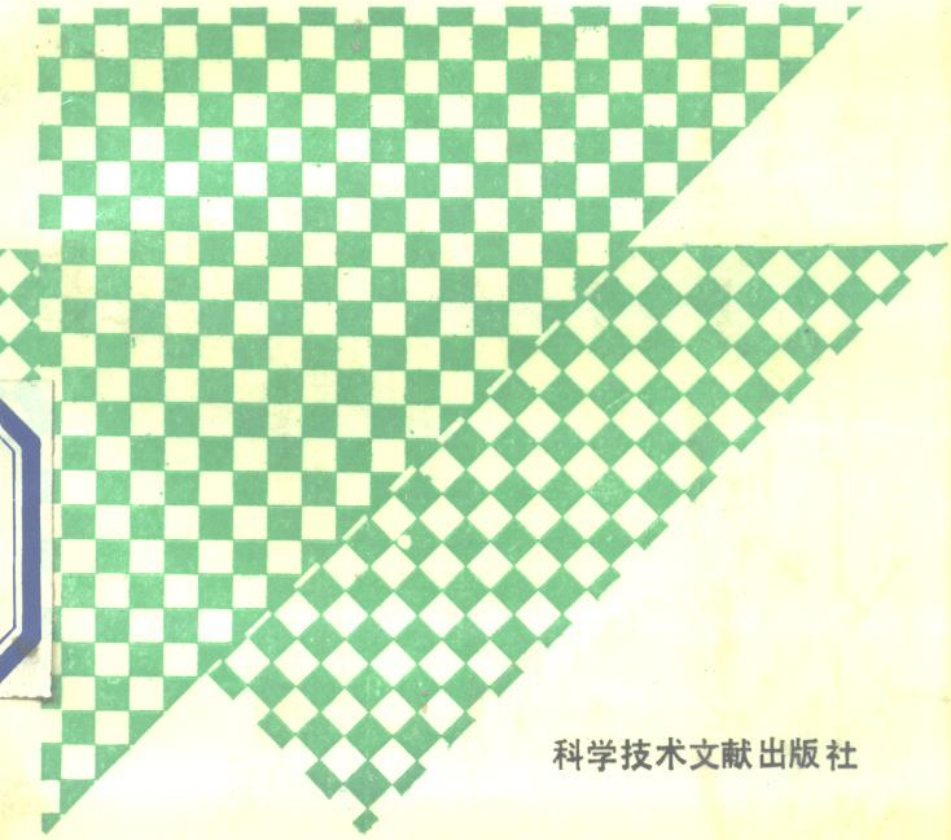
现代情报检索理论

康耀红 著



科学技术文献出版社

现代情报检索理论

康耀红　著

科学技术文献出版社

内 容 简 介

本书主要对已应用于商用系统的情报检索理论作较详尽的介绍和进一步讨论。另一方面，对那些在不久的将来可能影响到情报检索领域的理论也作深入的探讨。本书对情报检索理论的重要部分作了综合阐述，对情报检索的新领域，如模糊检索、代数检索、概率检索、多值相关性判定检索、布尔检索的新发展等，都作了全面的介绍，对与情报检索有关的文献内容的自动分析和自动标引、检索策略、查询加权、系统评价等作了比较充分的论述。本文还探讨了情报检索的哲学问题。

本书的读者对象是情报专业的教师及高年级学生，图书馆专业、计算机专业，以及某些应用数学专业的研究人员。

现代情报检索理论

康耀红 著

科学技术文献出版社出版

（北京复兴路15号）

航天部五院印刷厂印刷

新华书店科技发行所发行　　各地新华书店经售

*

787×1092毫米　32开本　10.125印张　217千字

1990年3月第1版　　1990年3月第1次印刷

印数：1—4100 册

科技新书目：210—136

ISBN 7-5023-0973-X/Z·138

定　价：5.30元

序　言

近三十年来，情报检索已发展成为一门具有丰富理论的学科，其中某些理论和方法（如布尔检索等）已发展得相当成熟，并已在实际中广泛应用。国外关于情报检索理论的研究仍然方兴未艾，因为对新的情报检索理论的探索，将有助于在新的起点上建立新型情报检索系统，对适应信息化社会更加广泛的情报检索应用具有重要的现实意义。了解情报检索的理论基础以及众多的新的检索理论，对我国情报检索学科建设和未来新型情报检索系统的设计，更具有重要的战略和现实意义。

情报检索的理论研究已越来越呈现出多样化的趋势。就检索策略而言，在商用系统中得到广泛应用的布尔检索方法及理论上的缺陷越来越受到挑战，人们一方面希图建立新型理论来代替布尔检索，提出了代数检索、概率检索等新型检索模型和方法，另一方面也致力于对布尔检索理论本身的推广和拓宽。近年来国内国外涌现了大批的文献和研究成果，可以肯定，随着科学技术的进一步发展和情报检索理论的进一步完善，一些先进的理论将会实用化和普及，并进入商用系统中产生更高的效益。

康耀红同志是一位从数学专业转向计算机情报检索理论研究的青年教师。《现代情报检索理论》一书是他以近五年时间悉心阅读数百篇国内外有关论著的读书心得和综合分析的研究成果，对情报检索的众多研究成果进行了精心推敲并予以系统化。本书对情报检索理论的新领域如模糊检索、代

数检索、概率检索、多值相关性检索、布尔检索的新发展等，从基本内容到当前进展作了比较系统的介绍，对与情报检索有关的文献内容的自动分析与自动标引、检索策略、查询加权、系统评价、查全-查准互逆关系的数学解释等作了比较充分的论述，其中对国外某些理论和方法的介绍在我国还是第一次见到，无疑对我国学者了解情报检索的基础理论及国际国内的研究现状具有重要参考价值。

我国已出版了不少关于情报检索的论著，其中大都侧重于情报检索系统的组成和软件介绍，而类似《现代情报检索理论》这一理论化的专著，在我国是第一次出版。作者在书中结合国际国内的现状研究，还融合了本人的一些观点和研究成果。例如，布尔检索中的双曲模型，概率检索中对于词相依性的不相关化处理，第八章中的词关联性测度，第十二章中多值相关性判定检索理论，第十三章关于情报检索的哲学问题等，都含有作者独立研究或具有独到见解的一些研究成果。本书写作注意语言表述的通俗性和数学论证的严密性，书中涉及的重要结果，都指出了出处和有关的参考文献，写作作风和态度是严谨的。

总之，本书是一部具有一定学术价值、对我国情报检索学科建设具有一定意义并能反映我国有关情报检索理论研究现状的著作。我相信它的出版将会对我们的情报检索理论研究产生有益的影响，因此很乐于向我国计算机情报检索同仁和同行推荐这一学术著作。

曾民族谨识

1989年6月27日

前　言

作为一本试图系统介绍和全面研究现代情报检索理论的书，又不致使篇幅太长，在其形成过程中，作者面临的最大困难是对数百篇文献的取舍和加工。本书的作法是，对于已应用于商用系统或实验系统的理论作较详尽的介绍和进一步的讨论，另一方面对那些不久的将来可能影响到情报检索领域的理论也作了深入的探讨。对于所有的重要结果，作者都指出出处及其它必要的参考文献。

除第一章、第十一章和第十二章外，其余各章最后都附有补记。这主要是为了对该章内容作补充的说明，特别是对有关的理论作简要的介绍和评价，以满足部分研究者的特殊需要。

书中没有用专门的篇章介绍有关的数学知识，除个别的理论（§4—1，§11—1）作了概略的介绍外，大部分都是直接引用的。这些知识可以从现有的许多数学教科书中找到。

我衷心感谢北京文献服务处曾民族老师，正是在他的支持和帮助下，我才得以坚持始终地完成了本书的写作。我还要感谢汤兆魁、周智佑、曹天顺几位老师，他们对本书的出版提供了许多有益的帮助。**本书的出版受到了西安电子科技大学科研基金的资助。**

虽然作者已尽了一切力所能及的努力，但书中还难免存在一些错误或不足之处，恳请广大读者予以批评指正。

康耀红

1989年3月

目　录

第一章 引 论

本章阐述情报检索的特点，介绍情报检索系统以及现代情报检索理论的发展状况。

§1—1 情报检索及其分类

情报检索一词同“情报”一样，目前仍是一个没有统一定义的术语。但从人们在“情报检索”范围内所进行的研究或活动来看，它主要涉及情报资料的表示、组织和存取，通过检索系统为用户提供与用户查询主题相关的情报资料。

按照检索对象的性质不同，情报检索可分为数据检索、事实检索和文献检索三种类型。数据检索的对象是数据，检索就是搜索数据文档，并针对查询提供答案。事实检索的对象也是数据，但要针对查询，由检索系统对数据文档进行分析、推理后，将最终结果输出。文献检索的对象是文献或文献的某种表示形式，通过检索系统为用户提供与查询主题相关的文摘索引、文摘内容或文献全文。

数据检索和事实检索是一种确定型检索，它提供与用户要求相关的情报；文献检索是一种不确定型检索，它不仅要提供与用户要求相关的情报，而且要提供相关的程度。这样的差异导致了必须用不同的方法去处理文献检索、数据检索和事实检索。Van Rijsbergen 在〔1〕（指每章后所列参考文献编号，下同）中详细地阐述了处理数据检索和文献检索

的不同方法。文献检索较数据检索和事实检索内容更为丰富，处理方法更为一般，因此 Van Rijsbergen 指出用“文献”来代替“情报”就足以论述情报检索了。从这个意义上讲，情报检索主要就是文献检索。实际上情报检索研究的中心问题也就是以计算机在文献检索中的应用为对象，研究适用于计算机处理的文献的表示、存贮、组织和检索的理论、技术和方法。

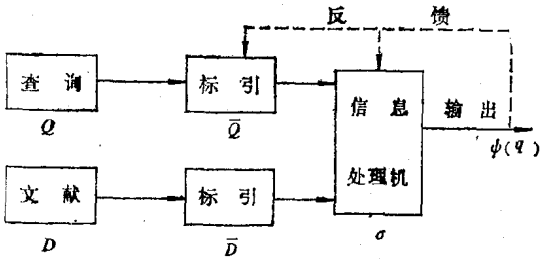
本书研究的对象是文献检索。

§ 1—2 情报检索系统

检索系统，狭义地说就是利用一定的检索设备从整理好的，存贮在某种载体（如卡片、书本、缩微胶卷或磁带等）上的文献集合中找到所需文献的系统。这里所说的检索设备是指卡片机、选卡机、电子计算机和缩微品检索装置等。所谓整理好的文献集合，可以是用自然语言描述的正文，或经过标引之后给出检索标志并按规定的顺序排列而成的文摘或题录的集合〔2〕。情报检索系统响应一个查询请求的输出是由一组参考文献所构成，这些参考文献倾向于向系统用户提供可能有兴趣的资料。

我们采用下图说明情报检索系统的主要功能。标引就是将来到系统的文献和查询要求经过概念分析，转换成某些词或“标引语言”。然后，文献和查询的标引被输入信息处理机。由图示看出信息处理机处理的不是原始文献和查询需求，而是文献和查询的一种表示。显然，如果不能构造出准确反映文献和查询的主题内容的标引，检索结果是很难想象

的。所以，标引是影响检索效益的一个重要因素。信息处理机的功能包括通过某些适当的途径将情报结构化，例如对情



情报检索系统示意图

报进行分类、排档等；同样包括实际的检索功能，即执行响应查询的检索策略。检索策略是建立在对文献和查询的比较之上的，是文献表示和查询表示的一种连接。对于给定的文献表示和查询表示，通过检索策略确定与查询表示相关的文献表示。检索策略的优劣直接影响最终检索结果，它也是影响检索效益的一个重要因素。由于标引和检索策略往往出现误差，信息处理机往往不能为用户提供满意的情报，因而需从用户对输出结果的分析所得的情报进行反馈，以得到修正的标引和检索策略。

采用数学语言，情报检索系统可以定义为一个四维数组：

$$S=\langle D,\ Q,\ T,\ \sigma\rangle .$$

其中 D 表示原始文献集，Q 表示用户查询等，T 是标引词集合。而 σ 为匹配函数：

$$\sigma:\ \overline{D}\times\overline{Q}\longrightarrow R.$$

此处，$\overline{D}$是经过标引的文献集合，$\overline{Q}$是经过标引的查询集合，R是函数值集合。值$\sigma(\bar{d}, \bar{q})\epsilon R(\bar{d}\epsilon\overline{D}, \bar{q}\epsilon\overline{Q})$表示文献$d$关于查询$q$的相关程度。设$k$为检索状态值，则关于查询$q$的检出文献集为：

$$\Psi(q)=\{d|\sigma(\bar{d}, \bar{q})\geqslant k, d\epsilon D\}.$$

要建立一个能够有效利用情报资料的情报组织系统是困难的。原因至少有两个：首先，对于不同的主题范围来说，情报量的增长是不平衡的。某些领域，例如计算机科学，以极快的速度在发展。然而其它一些学科，例如外语研究方面的资料，就可能根本不增加。将来情报增长的模式是难以预测的，因而对各个学科的任何预言，实际上都有很大的悬殊。为了仔细研究情报未来增长的情况，人们可以就每一学科和每一主题领域内，力求事先提出某些规定的扩充结构。最后，在这些扩充结构中，某些领域的资料可能是容纳不下的，而在另外的一些领域中却可能根本就没有用到。

在创造有效情报组织方面的第二个困难，是如何把相关条目放置得相近一些。例如，关于代数，图论和拓朴学的书籍，在一个文献集合中应是相互靠近的。第一眼看上去，这个问题似乎很容易，特别是当人们把这些主题都清清楚楚地放在更广泛的一般数学主题下面的时候，则更为简便。然而，会出现特殊的问题，例如对于某一交叉学科，这一特定学科可能与几个主题有关。要把分散在所有相关主题类目中的系统分析方面的资料，都放在相近的位置上，仅用把资料按顺序放在书架上的方法（这是一种一维的组织方法），是不能实现的，而必须采用多维的组织方法。

Lancaster 在〔3〕中对情报检索系统的特性、试验及评价作了定性的论述，Salton 在〔4〕中也用专门一章介绍了实际运行的检索系统。本书将对已应用于商用系统或实验系统的理论作较为详尽的介绍和进一步的讨论，另一方面对那些在不久的将来可能影响到情报检索领域的理论也作深入的探讨。

§1—3 现代的检索理论

我们已经知道文献的标引和检索策略是文献检索中的两个重要因素。如何对文献进行标引，如何制定高效益的检索策略，这是情报检索理论研究的两个重要课题。

在传统的情报检索中，一方面标引是由学科专家或职业标引人员手工完成，其固有的缺点是需要占用大量的人力物力；另一方面，当文献量激增，文献的专业化程度愈来愈高的同时，我们不可能把各方面的学科专家聚集起来用于文献标引，而职业标引人员要就不同内容的文献给予精确的标引又越来越困难，从而使手工标引常常无法与实际需求相适应。

在过去，人们建立了关于布尔检索策略的系统理论。这种检索策略的优点是运算程序简单，查询描述准确。因为这些原因，布尔检索在实际中得到了广泛的应用。但这种传统的检索方法还存在不少缺陷，例如不能控制系统输出量的大小，不能对系统输出按与用户的目的相关程度进行排序等。随着科学技术的飞速发展，情报用户队伍的结构愈来愈复杂，对情报检索系统的要求也愈来愈高，传统的布尔检索将

无法与新的形势相适应，必须有新的理论和更先进的方法。

科学技术的大发展不仅需要发展情报检索理论，而且也给现代情报检索理论的发展准备了两个重要条件——现代数学和电子计算机。现代数学，如概率论、模糊数学等，为现代情报检索理论提供了多种多样的研究工具，而电子计算机的发展更具有决定性的作用，可以说情报检索的现代理论是和电子计算机平行发展起来的。50年代末Luhn提出了自动抽词的思想，奠定了自动标引的理论基础；60年代Salton提供了矢量检索理论，并成功地应用于SMART实验系统。S.K.M.Wong建立了广义矢量模型，考虑了词与词的相依性。Z.W.Ras利用格与布尔代数理论建立了代数模型。Cooper和Bookstein利用集合论建立了情报检索的一般模型。Maron、Roberton和Sparck Jones于60—80年代期间也已先后建立了三个概率检索模型。Van Rijsbergen还在词相依情形下讨论了概率检索模型。荷兰的Raedcki在模糊检索理论方面作了出色的工作。最近几年，Salton又建立了扩展布尔检索模型，都小健、康耀红在多值相关判定的前提下提出了广义检索理论。关于自然语言的正文文献的输入和输出还是一个尚待解决的问题，但已越来越引起人们的重视。可以想象，未来硬件以及现代检索理论的发展，完全有可能使自然语言的输入输出成为现实。

现代情报检索理论的特点是标引、分类、反馈等过程的自动化和理论的多样化。人们可以根据文献和查询的相关程度将输出的文献进行排序，以便使用户可以首先得到最相关的文献，还可以控制输出量的大小。

应该指出，尽管现代情报检索理论有很多优点，但传统

情报检索理论也有其长处。例如查询的布尔表示比向量表示就准确一些。所以布尔检索的思想是我们在以后的研究中应该予以借鉴的。

§1—4 数学在情报检索中的重要性

周智佑在[5]中谈到情报科学的性质时指出："情报科学是一门交叉学科，但它不是自然科学内一门学科与另一门学科间的交叉，而很大程度上是社会科学与自然科学这两大门学科之间的交叉。"许多交叉学科如管理学、未来学、科学学都已发展到相当的程度，其原因就是成功地应用了数学。

情报检索的实际问题涉及到大量的数据。一个高效的检索系统要求的复杂程度越高，处理情报的工作量就越大。在这种情况下，数学工具是必不可少的。运用数学不仅能使被研究的对象和现象的质的概念精确化和深化，并能预见新现象的产生。在情报检索中引入数学模型，巧妙地运用数学方法能够事先预见在这些参数或那些参数变化条件下，某种假设的接近于实际的情报体系的发展过程。

随着计算机的发展，使用计算机可以建立各种既可快速又有智能的情报检索系统。早些时候，图书馆的编目和一般管理都为计算机所完成，随之而来的是众所周知的"图书馆联机革命"。但现在仅这些已无法与科技文献迅猛增长的形势相适应，于是进一步出现了计算机自动抽取文摘、自动标引、自动分类、自动收集书目数据等处理方法。尽管许多自动化的情报检索系统都还停留在试验阶段，但它们已越来越