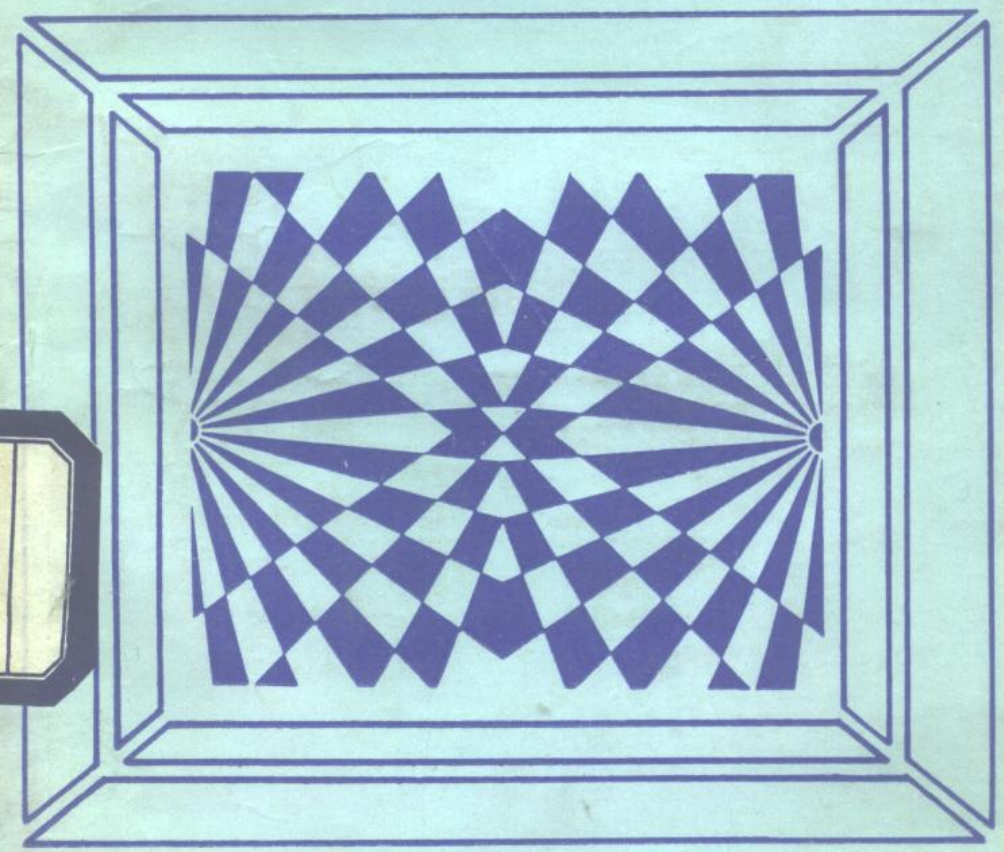
新一代的情报检索系统

[英] F . N . 德斯基著



新一代的情报检索系统

〔英〕F．N．德斯基 著

陈光祚 吴跃进 译

曾民族 校

科学技术文献出版社

1989

内 容 简 介

本书第一部分叙述了全文文本情报检索系统的基本功能，并提出了两个基本要求：一是建立更广义的情报模型化语言，一是发展简单的用户接口，以便用户能够进行复杂的特定提问。第二部分叙述了实现全文文本情报检索系统现有的硬件和软件方法，重点是探讨系统总体功能性的改进和提高性能的方法。第三部分提出了新型情报检索系统的设计方案及其主要特点。本书适合于从事各类情报检索系统的工作人员以及大专院校情报检索专业的师生参考。

F. N. Teskey
Information retrieval system for the future
The British Library Board
London 1984

新一代的情报检索系统

〔英〕F. N. 德斯基 著
陈光祚 吴跃进 译 曾民族 校
科学技术文献出版社出版
中国科学技术情报研究所印刷厂印刷
新华书店北京发行所发行　　各地新华书店经售

787×1092毫米 32开本 2.375印张 46千字
1989年6月北京第一版第一次印刷
印数：1—4000册
科技新书目：194—103
ISBN 7-5023-0811-3/Z·116
定价：0.95元

前 言

本文目的是研究全文文本情报检索系统所要求的基本功能。人们已经看到，情报检索各种类型的系统日益多样化，部分原因来自传统硬件的局限性；每一种特定的应用，为了能在现有的机器上达到实用的性能，不得不自己编写专门软件。随着新的硬件的发展，看来由于性能限制而带来的局限性可能不再如此严重了，因此，可以采取一种更为一致的方法来建立情报检索系统。本文目的在于研究对于各种类型情报系统共同的功能要求，并研究采用新的硬件实现这些功能的方法。

通过用户调查，本文提出了对情报系统的两个基本功能要求，并对这些功能的实现方法作了研究。这些方法已用于研究新型情报检索系统的一般设计原理。作者希望这项未来的研究工作能完善一个更详细的设计方案并实现一个原型系统。

目　录

第一部分 用户需求

1 全文文本检索系统导论

1.1 情报检索系统

随着我们的社会结构日益复杂化，个人和集团需要交流信息不断增多。许多这样的信息是作为文本传播的，但大量文本的累积使得检索相关资料的问题变得尖锐起来。文本情报检索系统（text information retrieval system）的目标，是组织与存储大量的文本，以便这些文本能够用于有效的信息交流。同时由于文本数量的增加以及需要存取这种文本的人数的增加，简单的手工处理方式已变得无能为力，因而需要更为有效的自动化文本检索系统。

传统的情报检索系统的基础是对每一篇文献指定一个或多个索引词，这些索引词通常是从某种叙词表中选出的。这种方式造成了一些问题：首先，必须付出相当的努力来维护叙词表，以便保证当文献中出现了新的概念时，在叙词表中就有描述这些概念的新词可供使用。其次，对文献指定索引词的过程是很费时间的并且很难实现自动化。最后，对文献的检索将要受叙词表范围的限制。然而，在全文文本检索系统（free-text retrieval system）中，原文本中的每一个单词都可以作为一个索引词使用。因此，在维护叙词表或指定索引词方面，都不需要有智力性的劳动，虽然可能需要相当

大的计算机资源来支持。此外，检索将不再受叙词表专业范围的限制，而完全采用原文本的词语。

一个全文文本情报检索系统必须能够支持以下功能：首先，系统必须能从各种来源收集文本；这些来源可能是跨越广泛的地理区域分布的，并且可能是以不同的介质、不同的格式产生文本的。要在一个地方把所有这些信息整理成标准的形式，通常决不是很简单的事。其次，系统必须提供存贮所有这类文本的空间。存储大量的文本，本身固然是一个问题，但是如果我们不能对文本进行存取，那么存储的文本也就毫无意义。这给我们提供了第三个要求，即必须能够从存储的文本中检索特定的信息的能力。最后，全文文本检索系统应该具有对被检索出的文本进行处理的能力，并且以用户乐于接受的形式向用户提供检索出的文本。在本书中，我们将较详细地考察这些基本要求，讨论现有的系统在何种程度上可满足这些要求以及新一代的情报检索系统如何能满足这些要求和其它要求。

1.2 结构化数据库

情报检索系统满足这些要求的一个途径，是采用结构化数据库。这一要求导致了IMS和 CODASYL 这类层次的和网状数据库管理系统的发展。这些系统允许(在一种主语言中)抽取和处理单个的数据单元。但是这些系统有一定的技术局限性，并且需要相当的专业技能来建立和使用。关系型数据库系统（Codd，1970），试图克服某些这样的局限性。所有这些系统都限于处理结构化的、固定格式的数据，因而不适合于作为非结构化的、或全文文本的大量信息的存储和传

输。正是为了满足这种要求，非结构化的、全文文本数据库得到了发展。

1.3 全文文本数据库

全文文本数据库最初是在书目检索的领域内发展的(Luhn, 1975)。书目型自由文本检索系统的目标是通过提供可以找到相关情报的文献的参考书目来满足用户的情报需求。这种系统通常依靠给每篇文献指定一组索引词并利用这些索引词来识别相关的文献。最简单的方法是，由用户指定一个相关的索引词并由系统来检索利用该索引词标引的文献。标准的布尔型检索系统就是由这种简单的过程加上索引词的布尔组配所组成的。例如，让我们假设在系统中文献编号为：

D1,D2,D3……

并且假设索引词为：

T1,T2,T3……

如果这些文献被标引如下：

D1:T2,T4,T5

D2:T1,T2,T5

D3:T2,T3,T4

D4:T1,T5

那么一个要求检索T2 AND T5的提问将检出文献 D1和D2。这种检索通常是通过建立一个索引文件来实现的，这个索引文件为每一个索引词列出它所标引的文献。在上面的例子中，索引文件的形式如下：

T1:D2,D4

T2:D1,D2,D3

T3:D3

T4:D1,D3 。

T5:D1,D2,D4

通过相关的索引表的组配，文献可以容易地检索出来。因此，要求检索T2A ND T5的提问将检索对T2和T5两个表共同的D1、D2文献。

这种利用倒排文件结构来检索出全文文本情报的方法已经相当成熟。目前英国已出现为数不少的商业性全文文本情报检索系统，附录1列出其中部分系统。本书的第一部分，将考察这些系统所提供的功能、它们的局限性以及所需的附加设备，附录2列出这些系统提供的各种功能。

2 全文文本检索的基本功能

2.1 文献和索引词

在建立一个全文文本情报检索系统之前，必须确定存入系统的文献、所使用的索引词以及对每篇文献指定索引词的方法。系统的这一初步要求应该保证每篇文献能够容易被识别。例如，在一个存贮与检索判例法律的系统中，每个判例报告都构成一个独立的文献，而在一个商业信件情报检索系统中，每一封信件就是一篇文献。确定和指定索引词有多种多样的方法，从完全手工的方法到完全自动化的方法，大致可分为如下几类：

1．以手工方式指定一组固定的索引词；

2．人工指定关键词；

3．依据事先编制的词典进行自动标引；

4．全文标引；

5．词干一级的自动标引。

第一种方法是依据一个固定的索引（如杜威十进分类法）进行手工编目的传统方法。第二种方法相当于在文献的摘要或全文中手工指定关键词和短语。而其它的几种方法均以机器可读文本为前提，其中第三种方法是从文本中抽取预定义索引词。第四种方法是把每一个单词作为一个索引词来处理。最后一种方法则力图通过去掉前缀和后缀并用所得到的词干作为索引词的方法，来控制索引词集合。在最后的两种方法中，都需要从索引词集合中排除某些普通的单词，如 and、of 和 the等。

早期的某些研究（Cleverdon，1967），已经作了大量的实验来评价上述几种方法的相对优缺点，本文不拟重述这些实验，但必须指出，每一种方法都有其最适用的具体条件，而任何全文文本检索系统的一个基本功能是对上述各种标引方法都适用。的确，某些应用可能需要在同一文献的不同部分内进行不同类型的标引。例如，在一个技术报告数据库中，在词干上的自动标引可能是最合适的，而当研制的数据库具有特定用途时，人工指定索引词则可能是较有用的。LexiBOSS系统，采用它的参考词（reference word）、关键词（key-word）和特征词（tagword）的概念，在某种程度上可满足这种要求。

2.2 布尔检索

如果一个文献集合已经对每篇文献指定了各自的索引词，那么这个文献集合的布尔检索功能，就是索引词的逻辑

组配（用and,or和not）功能。这种功能用于检索出带有特定索引词组合的那些文献。例如，检索式：

（情报and检索）not计算机

目的是想查出标引了“情报”和“检索”但不标引“计算机”的文献集合。人们普遍认为，所有全文文本检索系统都应该具备这一基本功能。我们所考察的所有系统也都提供这种功能。

但在某些系统中，要用一条指令完成一个复杂的布尔检索是不可能的。例如，在DIALOG和ESA系统中，必须检索各个单个的词，然后组配所得到的各个结果集合。在这些系统中，上述检索需要如下四个指令来表述：

检索指令1　select　情报

检索指令2　select　检索

检索指令3　select　计算机

检索指令4　Combine（1 and 2）not3

尽管这种对话对于简单检索来说不算缺点，但无论对学习该系统的初学者，还是构造复杂检索提问式的专家来说，都会增添很多麻烦。

就现实情况来说，布尔检索的问题之一是用户对检出的文献数量不能进行直接控制。系统可能检索出一千多篇文献，让用户去识别相关的文献，或者系统可能根本没检索到任何文献。在许多情况下，用户想要知道的是，比如说十篇同他的提问最为相关的文献。这个问题，通过使用查全和查准的方法能够部分地得到解决。前者是扩大检索的范围，从而检索出更多的文献；而后者是使检索更为精确，从而检索出更少量的文献。查全方法旨在用更为广义的词来取代不常用的

和非常专指的词，而查准方法不仅仅旨在用更专指的词来代替广义的词，而且可将检索限制在数据库的特定部分。下面是一些查全查准的方法，在下面的各节中，我们将进一步讨论这些方法。

截断和字符串检索；

同义词表和叙词表；

字段和节；

相邻和组配。

2.3 截断和字符串检索

如果在建库时采用全文标引，那么在检索时，通过截词功能能够在某种程度上实现词干标引的效果。截词检索将检索出在检索式中包含有该词干的文献，而不限于那些包含有确切检索词的文献。例如，如果符号‘＊’用于表示截断，那么检索要求 inform ＊将检索出含有以‘inform’这些字符开头的文献，例如含有 inform，informs，information 等文献。我们考察的所有系统都提供这种功能。的确，在标准的索引文件（1.3节）中，这种功能是易于实现的。因为索引按字母顺序排列，将保证具有同一词干的所有检索词存贮在邻接的位置。英文中的词干通过去掉后缀的处理可以很容易找得到，所以截词检索的方法正好把全文检索和和词干标引两者的优点很好结合在一起。

在某些其它应用中，例如检索化学名称，有意义的词干常常出现在词的中间。在这些领域中就要求一种比在英文文本中所用的简单的词干右截断更为广义的字符串检索。这种扩充包括左截断，将允许检索提问如：

acetyl

这个检索提问式将检索出包含有字符串'acetyl'在内的任何化学名称。更为广义的字符串检索将允许多字符屏蔽（*）和单字符屏蔽（!）的结合，因此检索提问：

t!sk*y

将检索包含teskey，teskely，tiskey等文献。这种功能对于检索那些不知道其确切拼写形式的特殊名称时，显然是很有用的。许多现有的系统都提供这种字符串检索的功能。但是，要从一个按字母顺序排列的索引中找出所有可能的匹配是很花费时间的，因此，这种检索总是限于对数据库的一些小的、完全确定的部分内进行。（但以硬件支持的系统除外，如内容可寻址文档存贮（CAFS）和OFIS文档等。）

2.4 同义词表和叙词表

在传统的文献检索系统中，叙词表的作用是提供固定的标引词。在一个全文文本检索系统中，每一个除通用词以外的语词都是一个索引词，从而叙词表的作用从一种标引辅助工具变成了一种检索辅助工具。对于这种新的作用，叙词表对一个给定的词应该提供同义词、广义词或下位词，且这些词应能够包括在用户的检索提问中，以便用户根据需要来提高查全率或查准率。STATUS系统提供了这种检索手段。在某些应用中（如判例法律检索），组织得很好的同义词表看来也能够大大地改善检索性能。虽然一部具有结构完整的叙词表体系能够在全文文本检索系统中提高主题检索的效率，但在现在，在检索中使用叙词表词间层次关系的实践经验还很少。

2.5 字段和节

大家普遍认为，全文文本数据库需要一定的结构，这种结构可以从简单的书目数据的作者—题名—摘要结构到复杂的结构，如关于设备安全记录所要求的论述设备的类型、检验与维护工作等详细信息。我们所考察的所有系统都提供了把一单篇文献分解成带有标识的字段的手段。这样就可以把检索与显示限制在数据库的特定部分。例如，如果一个用户想要检索一个书目数据库以得到由 Mount 所著的所有书籍的书名；那么，对 Mount 的一个简单检索可能不仅仅检索出由Mount所著的书，而且也会检出有关 Mount St Helens的书或由Mount出版公司印刷的书等。如果这些文献采用结构化方法，分出著者、题名和出版者等字段，那么用户就能够进行更为精确的检索，例如：

mount in author

这里in用来指明在一定的字段内进行检索。当浏览检出的文献集合时，用户可以只浏览书名字段，这样就可以找到由Mount所著的所有的书名。在某些应用中，有的数据库希望把文献分解成若干节，这些节的名称比数据库的范围更小。例如一个技术报告数据库，每一份报告都是一篇独立的文献，但不同文献的章节之间可能需要相互参照。没有一个系统提供对此问题的解决办法，只有 STATUS 系统的段落(paragraph)功能，在解决这个问题方面取得了一定进展。

2.6 相邻和组配

以上我们考察的各种检索功能，都是以文献中索引词的

出现与否为基础的。然而，在全文文本检索中，各个词的相邻度和相对地址是很重要的。一个典型例子是blind Venetian（百页窗）和Venetian blind（威尼斯盲人）之间的差别。这种功能最普遍的应用，是要求检索词作为短语出现，而不是在文献中的任何地方出现，以提高检索的准确性。例如，假设在数据库中有两篇文献：

1．A user survey of information retrieval system（情报检索系统的用户调查）；

2．Some new information on the retrieval of nuclear by-product（关于核副产品回收的一些新情报）。

“情报 and检索”的布尔检索式，将同时检出上述两篇文献，因为“回收”与“检索”在英文中为同一个词retrieval，而这两篇文献是互不相关的。如果以短语“情报检索”进行更精确的检索，将只检出第一篇文献。在STATUS中，最一般的形式是以“连语”（collocation）这一算符提供这种检索手段。在大多数其它系统中，则通过字符串检索的形式来提供这手段，必须先采用更广义的检索词得出文献集合后才能进行字符串检索。

3 主要问题总结

3.1 数据库的建立

建立一个有用的全文文本情报检索系统所要付出的努力往往被低估。在确定数据库结构的第一阶段，不仅要决定合适的文献结构和索引方法，而且要估价这些决定对系统未来性能的影响。在许多情况下，在数据库设计中的小的变化，