

信息检索与数据库

〔日〕 中原啓一 三次衛 主编

张在浩 译

煤炭工业出版社

信息检索与数据库

〔日〕中原啓一 三次衛 主编

张在浩 译

煤炭工业出版社

(京)新登字042号

内 容 提 要

当今的时代是一个充满信息的时代,任何一个部门或行业都离不开信息。本书介绍了信息的收集、分类、管理及其数据库的建立方法;依据各部门、各行业的特点列举了多种系统与数据库之间的关系,以便充分利用数据库和检索技术而获取集萃于一库、检索于一瞬的效果和达到资源共享的目的。

中原啓一 三次衛 監修

情報の検索とデータベース

取次販売所 株式会社 コロナ社 1986

●

信息检索与数据库

〔日〕中原啓一 三次衛 主编

张在浩 译

责任编辑:陈日元

*

煤炭工业出版社 出版

(北京安定门外和平里北街21号)

煤炭工业出版社印刷厂 印刷

新华书店北京发行所 发行

✱

开本 850×1168mm¹/₃₂ 印张11³/₄, 插页2

字数300千字 印数1—1,254

1993年12月第1版 1993年12月第1次印刷

ISBN 7-5020-0783-0/TD·723

书号 3551 H0060 定价 15.00元

目 录

第一章 什么是信息检索	1
1.1 信息与信息检索系统	1
1.2 信息检索系统的发展	3
1.3 信息检索系统的基本功能	7
1.4 信息检索系统的分类	11
第二章 信息结构	14
2.1 信息的分析方法	14
2.2 信息结构	16
2.3 信息结构与数据库	24
第三章 软件技术	35
3.1 数据库管理系统	36
3.2 联机控制系统	85
3.3 运行管理系统	94
3.4 分散网络系统	113
第四章 文件装置	121
4.1 文件装置所处的地位及其结构	121
4.2 磁盘装置	123
4.3 磁盘系统的处理能力	128
4.4 半导体存储元件的利用	130
4.5 磁带装置	133
4.6 超大容量存储装置	135
4.7 光盘装置	137
第五章 业务领域中的适用例子	140
5.1 银行系统与数据库	140
5.2 制造系统与数据库	150
5.3 流通系统与数据库	160
5.4 行政系统与数据库	171

5.5	新闻系统与数据库	182
5.6	医疗系统与数据库	188
5.7	秘书系统与数据库	199
第六章	工程领域中的适用例子	211
6.1	计算机辅助设计系统与数据库	211
6.2	实验室自动化系统与数据库	225
6.3	图象处理系统与数据库	234
第七章	文献查询领域中的适用例子	249
7.1	文献查询系统与数据库	249
7.2	专利管理系统与数据库	261
7.3	图书馆系统与数据库	280
第八章	经营领域中的适用例子	288
8.1	经营领域中信息管理系统与数据库	291
8.2	经营决策支援系统与数据库	302
8.3	人事信息系统与数据库	313
第九章	数据库产业	327
9.1	数据库的分类及其利用方式	328
9.2	数据库产业结构	330
9.3	数据库服务的发展史	337
9.4	数据库产业的现状	341
9.5	数据库产业的今后课题	344
第十章	信息检索与数据库领域的今后课题	347
10.1	分散数据库	347
10.2	智能数据库	349
10.3	数据库处理机	352
10.4	私人信息的保护	355

第一章 什么是信息检索

1.1 信息与信息检索系统

人们在日常工作中，极其自然不加任何思索地使用着“信息”这个字眼。然而，到底什么是信息呢？

为“信息”这一概念下一个确切的定义，究竟有无意义，其本身就值得怀疑。但是，非要下一个定义的话，则定义可为：“所谓信息，就是对构成对象的发生源，通过耳闻、目睹、理解所产生的事物。”然而，仔细推敲这一定义，便会发现其中包含着很多非常抽象的要素。例如，所谓构成对象的发生源到底指的是什么，耳闻、目睹、理解之类的概念也非常抽象，耳闻、目睹者是何人，又是何物。诸如此类，没有一样是有具体规定的。

反过来说，信息又是一种非常丰富多彩的无限存在的事物。即使对同一个发生源，随着耳闻者、目睹者、理解者的不同，即随着接收源的不同，所产生的信息也会有各种不同的变化。

现在，让我们设想一下“汽车在行驶”这一信息的发生源。

对这一发生源，假设有三个接收源：甲、乙和速度监测器。甲对汽车没有多大兴趣，因而对他来说只是产生和认识了“汽车在行驶”这一信息。而乙这时正盘算着要购置一辆汽车，因而从同一发生源中获取的是汽车的外形、性能、价格等信息。如图1.1。

这样，随着接收源的不同，所产生的信息也各不相同。这是由于不同的接收源，它的信息意识以及业已具备的知识等条件各不相同。

乙之所以能够产生汽车外形及价格等信息，是因为他以过去积累的信息为基础，进行了某种处理的结果。可见，“人”可以认为是一种复杂的信息系统。

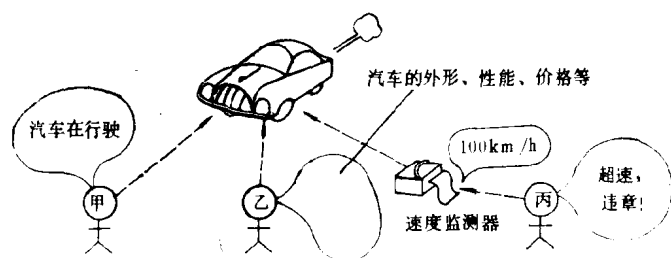


图 1.1 信息的产生

另一个接收源——速度监测器，也可以看做是一种信息系统，它具有将对象物的速度作为信息予以反应的功能。但是，速度监测器同人相比则简单得多，它的对象只局限于速度。

将速度监测器所产生并输出的信息作为发生源，警察从这一信息接收源又产生“此车超速，违章！”这一新的信息。这时，我们可以认为速度监测器即是信息的接收源，同时又是信息的发生源。

如此看来，所谓信息一定是由信息系统产生的，没有信息系统就不可能存在什么信息。这里所说的信息系统，是一个非常广义的概念。对这一概念的解释一般是这样的：将某一事件或信息做为输入对象，然后以自己原来已积累的信息为基础，使之产生新的信息。这种对信息进行积累和加工的过程，就是人们所说的信息系统。

图1.2表示了信息与信息系统之间的关系。可见，平常所说的信息，与信息系统之间存在着不可分割的极其密切的关系。特别是把占有重要位置的人作为信息系统考虑的时候，这一系统就变得非常复杂。

本书把这些概念包括人和计算机在内的信息系统，称为信息检索系统，并着重探讨以信息为中心的信息检索系统。同时，在本书里所说的信息检索系统，不是狭义的情报检索，即IR(Infor-

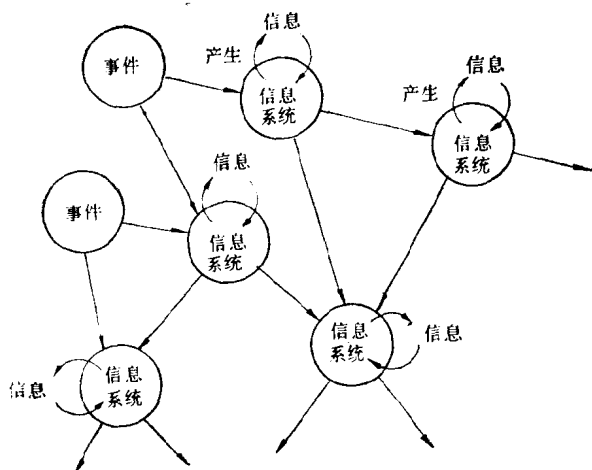


图 1.2 信息与信息系统

mation Retrieval) ，而是包括数据库管理系统在内的广义信息检索系统。另外，在探讨信息检索系统时，有各种不同的方法，有的从用户角度出发，有的从设计者角度出发，有的从运行管理员角度出发。本书则以设计者的立场为主，同时考虑到办公室自动化 (OA) 的需要，将介绍各个领域内的信息检索系统的适用例子。

1.2 信息检索系统的发展

可以说，目前普遍使用的“信息检索”这一词，是以往的狭义情报检索 (IR) 和文献情报服务的总称。确实，至今为止狭义情报检索 (IR) 一直被视为以文献情报为中心的一个完全独立的利用计算机管理的专门领域。

过去把狭义情报检索 (IR) 作为专用系统予以对待，可能出于如下的原因：文献情报与日常事务处理相比较，后者是一种非常难以规范化的信息，而且要求具备多种信息查询形式。但是，

这种分析比较，在批处理中心对日常事务处理的接受能力尚受到限制的时代，也许是正确的。然而，时至今日，计算机技术已经发展到在日常事务处理领域中也能够处理诸如顾客信息、不动产信息等各种各样的数据，而且出现了能够满足各种要求的系统。这时还象从前那样做比较，就不见得正确。

换句话说，过去把狭义情报检索（IR）作为专用系统予以对待，在很大程度上是由于受到计算机技术的限制。像近几年这样，联机数据库等硬件和软件技术业已得到长足进展的情况下，将日常事务处理和情报检索加以区分，已经变得没有实际意义。基于这样一种观点，我们准备结合计算机技术的发展史，简述一下信息检索系统的发展史。

研究计算机技术的发展史，即可从电路技术角度入手，也可从软件技术角度入手，其方法是多种多样的。这里我们打算从构成信息检索系统核心的信息存储介质和存储装置入手。

信息存储装置的历史可分为以磁带为基础连续处理中心时代和以磁盘为前提的随机处理中心时代。信息检索系统的基本功能，就是如何存储信息，又如何将已存储的信息查找出来。

这时，连续处理中心系统和随机处理中心系统，从用户角度看，其系统的利用形式是大不相同的。也就是说，对特定的信息需求来说，不需一秒钟就能将信息检索出来的系统和花几分钟以上才能检索出来的系统，两者之间已经不是单纯的速度方面的差异问题，而是意味着系统本身的类型存在着差别。

因此，我们一方面把计算机技术的发展阶段，分为连续处理中心时代和随机处理中心时代，另一方面参考操作系统、联机系统和数据库管理系统等软件技术的发展过程，来论述信息检索系统的发展史。

1.2.1 连续处理中心时代

这个时期，几乎所有的信息都存储在磁带上，而计算机的利用形式，完全是采用批处理方式。

这一时期的系统，可以认为是以业务工作尽量省力为其主要

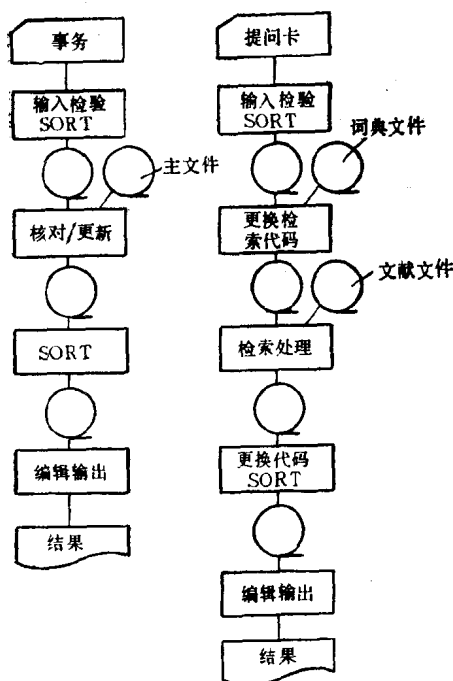


图 1.3 连续处理形式

目的。其中一半以上是按着业务要求进行逐日的库存管理或逐月的工资计算等代表性应用领域。这些处理形式，如果从内部结构进行分解的话，如图1.3所示，可分为事务处理、同主文件的核对处理、输出明细表格处理和编辑输出处理。当然，根据业务的复杂程度，有时也需要同若干个主文件进行核对，输出多种结果。但基本上是采用以分类（SORT）为中心的处理形式。

1.2.2 随机处理中心时代

这个时期，由于用户需求的特定信息能够在瞬时之间检索出来的随机处理特性和联机技术的发展，联机信息检索系统得到了迅速的普及。在初期阶段，因受磁盘容量和速度的限制，应用领域只局限于那些依靠联机检索能够得到较大效益的银行业和订货

登记等用户，但是已经对社会和企业产生了很大的影响。

不久，对广泛应用计算机的期待越来越强烈。人们开始呼吁：不要仅仅停留在对特定业务的高度计算机化，要求建立用于经营性信息管理和制订经营计划决策的战略性系统以及计算机辅助设计系统。要想满足这些要求，就有必要把以往分散的业务予以统一起来。但到目前为止，大家都在以业务处理为中心考虑问题。由于存在着若干种按业务系统编制的专用文件，所以做为业务统一化的基础工作——信息统一化就很难实现，无形中为向高度系统化方向发展设置了障碍。

由于这个缘故，后来的研究工作就不象过去那样分别为不同的业务编制若干个不同的专用文件，然后通过应用软件对这些文件之间的相关关系进行管理；而是研究那种能够把各种信息之间依存的相关关系也做为信息来进行存储并能扩大应用领域、统一业务的数据库。这一研究得到了人们的重视，并得到了很快的发展。此外，随着办公室自动化（OA）而带来的事务处理范围的扩大和信息处理设备的发展，使得构成信息对象的范围更加广泛。以此为背景，数据库同网络技术和微型计算机技术的关系变得更加密切，而做为数据库内容被贮存的信息质量也发生了很大的变化。

做为信息检索系统，系统的联机化正在得到实施，检索需求的服务正在得到质的改进。也就是说，将过去那种向用户单方面提供信息的服务方式，改进为查询用户试探性提出的要求，从而使用户能够获得自己所需求的信息。此外，构成对象的信息，也不仅仅局限于文献情报，逐渐呈现出向化合物和医疗信息等领域扩大的趋势；更为重要的是服务方式的转变，即利用个人计算机通过联机检索能够获得上述外部数据库的信息。

在这种随机处理中心系统中，为了实现富于灵活性的利用方式，并能够进行多种目的检索和多种条件检索，所面临的重要课题，就是如何存储信息和显示信息，即所谓的信息结构问题。

除此之外，同局限于连续处理的时代相比，对软件功能的要

求也大幅度增加。诸如，如何将信息存储在磁盘上进行管理，如何控制联机处理所必需的终端机和电路，如何有效地利用光盘等大容量存储介质，以及如何同各种办公室自动化（OA）设备等智能装置进行分散处理等等。

图1.4表示了连续处理时代和随机处理时代的软件系统。从图中可以看出，系统所需要的整个软件中，应用软件所占的比例相对地减少，而操作系统等系统软件所占的比例却逐渐增加。

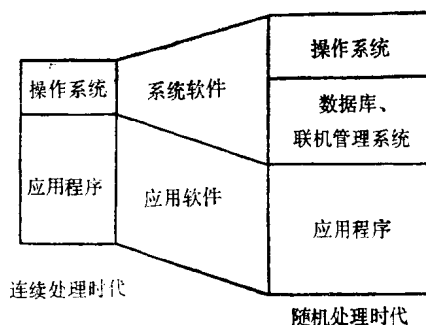


图 1.4 软件系统

在系统软件中，数据库管理系统的重要性正在逐步提高，并且伴随着关系数据库的出现，有可能建成使用起来更加方便、更加灵活的信息检索系统。

从上述观点出发，难道我们不应该感到有必要彻底地重新认识对信息的分析方法吗？关于信息的分析方法，将在第二章信息结构中详细论述。

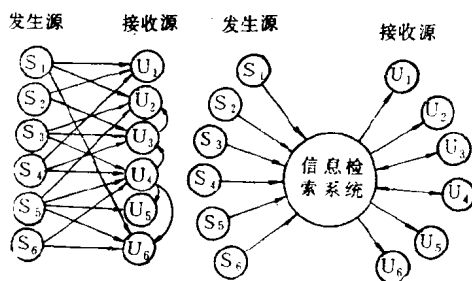
1.3 信息检索系统的基本功能

1.3.1 任务

信息检索系统的任务，无非是从信息发生源收集信息，并向信息接收源即向用户提供所需求的信息。也就是说，对信息发生源来说，它起着一个接收源的作用，而对信息的用户（接收源）

来说，它又起着发生源的作用。这样把两者连接起来，最终起着信息发生源与信息接收源之间的媒介作用。这就是信息检索系统本来的目的。

但是，正象图1.5所示，当存在若干个发生源和接收源时，信息的收集与提供就变得复杂起来。特别是信息的接收源，为了获得它所需求的信息，不得不跟多个发生源打交道；有时发生源所产生的信息，也随着时间的变化而变化。这时，信息的发生源与接收源之间的连接就变得非常复杂。另外，有时还可能出现这种情况，即某个信息接收源对其它接收源又构成它的发生源。可见，把能够妥善而恰当地收集信息与提供信息做为信息检索系统的本来目的是正确的。



S—发生源，U—用户，即接收源

图 1.5 信息查询系统的任务

这时，一个重要的问题就是不能象联机通信系统那样，只是把信息妥善地传递过去，从而把发生源产生的信息原封不动地提供给用户；而是使信息检索系统具有将信息加工成适合于用户需求的形式，并能够补偿信息发生与信息需求之间的时差的功能。

1.3.2 系统管理员

如上所述，信息检索系统的任务，不单起着发生源与接收源之间的媒介作用，还根据某种目的，从发生源收集信息，然后将

它积累存储，最后向用户提供适合其需求的信息。因此，不能象单纯进行信息交换的联机通信系统那样，把信息检索系统交给用户管理；而必须按照一定的规程使之运行。已经存储在信息检索系统中的信息，绝不可有意无意地受到破坏或被窃。

为此，信息检索系统必须配备系统管理员或专门机构。系统管理员的职责，是建立必要的规程，以保证该系统恰当地收集信息，正确地保存信息，并向正当的用户提供准确的信息，进而管理好系统，使之有效地运行，如图1.6。

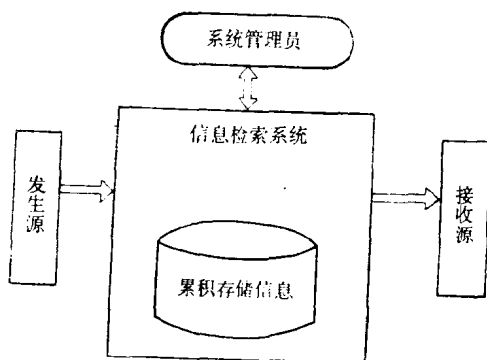


图 1.6 系统管理员

因此，信息检索系统必须具备能够按照系统管理员制定的各种规程得以运行的功能；同时还要具备能够向系统管理员提供表示该系统运行状态的各种信息的功能，以保证系统管理员能够实施最优化操作。

1.3.3 基本功能

如图1.7所示，从外观上看信息检索系统，其基本功能可分为三大类：信息收集功能、信息提供功能和系统管理功能。有关这些功能的软件系统，将在第三章详细论述，这里只对这些功能的概况分别进行叙述。

(1) 信息收集功能 该功能将信息发生源产生的信息输入到系统里,然后把符合系统要求的信息进行选择、加工,最后把信息存储到数据库里。

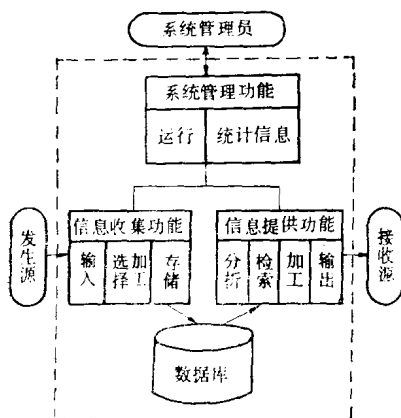


图 1.7 查询系统的基本功能

理工作,例如检索用关键词的代码化以及数据压缩等。

③ 存储 为了把已经选择加工的信息变成累积存储信息,或者对数据库进行追加存储,或者对原数据库的信息进行更新。

(2) 信息提供功能 该功能首先对用户提出的信息需求进行分析,然后检索用户所需求的信息,并将检索结果按用户要求的形式加工,最后编辑输出。

① 分析 因为用户提出的要求是用某种语言描述的,所以为了满足用户的要求,就要事先分析查询哪一类信息,并通过什么样的处理,打印出何种形式的信息。

② 检索 根据对用户需求的分析结果,从数据库中有效地检索出所需要的信息。检索条件从简到繁各式各样,检索的信息种类和数量也依用户的需求而各不相同。

③ 加工 将检索结果按用户要求进行加工,但根据不同的要求有各种不同的加工处理。例如,从简单的分类、合并统计,一直到管理科学技巧等。

① 输入 根据发生源的性质和收集信息的种类,可以是脱机输入,也可以是联机输入,收集方法是多种多样的。输入信息的检验方法,也要依据信息性质来确定。

② 选择加工 从已输入的信息中选择符合系统要求的信息,并根据需要进行存储前处理

④ 输出 将检索结果和加工处理结果,按用户要求的形式进行编辑,然后予以输出。依据输出的信息种类,其输出形态也各式各样。例如,从简单的报表一直到设计图案。

(3) 系统管理功能 该功能包括按照系统管理员制定的规程运行系统的功能,掌握并向系统管理员提供运行状态的功能。

① 运行 该功能包括数据保护功能(例如为防止系统的信息被破坏或盗窃而对数据进行保密措施等)、发生故障时的系统修复功能、数据库修复功能以及对低效数据库的再组织功能等。

② 运行信息 该功能对系统的效率和数据的利用情况,提供各种有关运行信息。运行信息中还包括有关系统的统计信息、为修复故障所必需的滞后信息以及为计算费用所需要的信息。

1.4 信息检索系统的分类

信息检索系统在不同领域有各种用途,人们正在广泛地使用和开发信息检索系统。由于构成系统特征的要素本身就有繁多的种类,所以为众多的信息检索系统进行分类,是一件非常困难的事情。例如:

- 系统是联机方式还是批处理方式。
- 系统所要求的实时性如何。
- 信息的提供方式,是由系统定期地提供,还是允许用户采用试探法来提供用户所需求的信息。
- 系统服务的用户,是局限于企业内部的用户,还是向社会开放。
- 系统所处理的信息,是以文字信息为主,还是以数值信息为主;是处理图像信息,还是多种混合信息。
- 具有多大的存储容量。
- 对已存储的信息,其加工处理工作是通过主机进行,还是通过个人机进行。

所有这些要素,在设计信息检索系统时都很重要,此外还有其它很多要素。但是,对信息检索系统,即使把这些要素组合起

来进行详细的分类,恐怕也很难掌握其本质。

本书着眼于掌握信息检索系统的本质,我们认为最主要的两个要素是需求模式的多样性和信息加工处理的复杂性,根据这两个要素的组合情况,对信息检索系统进行分类。如果把这两个要素与人做个类比的话,其中需求模式的多样性相当于知识面的宽度。但是,即使一个人的知识再丰富,仅仅靠这些知识还不能说他是一个具有智慧的人,他还需要具备运用这些知识的能力,诸如管理方面的决策能力或判断能力。而这种应用能力,恰似信息检索系统中的加工处理的复杂性。

1.4.1 需求模式的多样性

一个信息检索系统究竟存有什么样的信息,能够满足何种形式的检索,这就是需求模式的多样性。

从理想的角度讲,最好有一个对任何信息需求都能进行检索的系统。但是要做到这一点,就需要把信息按照实际社会所存在的形式原封不动地储存起来。这就不得不储存庞大的信息,同时不可忽视的问题是,系统的效率将下降,数据的冗余度将增大。

在联机存款系统中,所需求的信息是当时的存款额,并规定必须以户头号码做为关键词,因此能够进行非常有效的检索。

与此相反,在文献情报系统中,很显然那种只能通过文献题名进行检索的系统,是没有使用价值的。

1.4.2 加工处理的复杂性

一个信息检索系统根据其检索结果,能够进行什么样的加工处理,这就是加工处理的复杂性。当一个用户在管理方面想要制订某项决策时,他不仅要查询以往的实际信息,还想预测当这一决策付诸实施后会产生什么样的结果。因此,重要的是能不能预测这种结果,并在多大程度上满足运用管理科学方法的处理要求和复杂的技术计算要求。

当然,每一个信息检索系统不一定都需要具备所有的功能。象文献情报系统那样,用户所期待的功能,只是要求将检索结果加工到编辑输出为止。