数据管理的数据库组织

〔美〕S.P.高西 著

文 卿 译

国防工業出版社

数据管理的数据库组织

〔美〕S. P. 高　西　著
文　卿　译
戎　行
方　正　校
王东泉

国防工业出版社

内 容 简 介

本书主要论述数据管理的基本理论和基本方法。全书共分八章，目次如下：数据结构；查询及查询语言；对一个字段的查找；关键码到地址的变换；代数文件编排模式；连续检索特性；磁鼓存贮器上的组织；存取路径的检索。

本书可供高等院校师生和从事数据管理的研究、使用人员阅读。

DATA BASE ORGANIZATION FOR DATA MANAGEMENT

Sakti P. Ghosh

Academic Press 1977

*

数据管理的数据库组织

〔美〕S.P.高西 著

文 卿 译

戎 行

方 正 校

王东泉

*

国防工业出版社 出版

新华书店北京发行所发行　　各地新华书店经售

国防工业出版社印刷厂印装

*

850×1168 1/32　印张 12 1/2　316千字

1982年8月第一版　1982年8月第一次印刷　印数：0,001—4,800册

统一书号：15034·2367　　定价：1.55元

前　言

在过去的不多几年中，数据管理这一领域已成为计算机科学中的一个极为重要的分支。大数据库中存放着容量巨大的非数值数据，随着具体的需要，时时要根据复杂的查询对它们进行处理。由于非数值信息所固有的复杂逻辑结构，数值信息处理法并不能轻而易举地应用到非数值信息中去。非数值信息处理法大都是由实践家研制的，而见之于文献的形式描述却为量甚少。著名学府的学者和大计算机制造公司的高级研究人员曾为了从根本上理解这些方法而作了深入的研究。他们的著述散见于许多期刊、技术报导、会议记录，等等。这对于教师编写数据管理组织方面的基本教材来说，收集和整理资料成为一个十分困难的问题。对于学生学习这门课程来说，这个问题就显得更为困难。因此，学生不断向我索取资料和同事屡次邀我进行演讲，已几乎成为无暇应接之事了。社会上根本没有论述数据库组织的大学教科书出售，也根本不见涉及诸如数据描述模型、查询逻辑结构、组合查询集、二次剩余变换、对称文件编排模式、连续检索特性和磁鼓上数据组织等基本概念和理论的系统著作。这些论题能在各类期刊中找到，但就其内容的深广度来说，却只有高级研究人员才能懂得。它们不是为学生的学习撰写的。鉴于这一领域的知识在计算机科学中占有极为重要的地位，我深感不应该再让这种“真空”状态继续存在下去了。

作者以前编写的关于数据管理的教科书适用于初年级或高年级大学生，其中主要论述数据管理的算法和系统。在本书中，作者则力求将数据管理的基本理论和基本方法置于前沿而略去信息管理的系统细节。第一～四章可以供高年级大学生学习之用，第

五～八章可以供研究生学习之用。

本书共分八章，目次如下：（1）数据结构，（2）查询及查询语言，（3）对一个字段的查找，（4）关键码到地址的变换，（5）代数文件编排模式，（6）连续检索特性，（7）磁鼓上的组织，（8）存取路径的检索。

第一章介绍跟数据描述有关的一些基元概念和跟本书内容有关的一些基本数学知识，以及介绍如实体集模型、关系模型、和图论模型等一些数据描述模型。

第二章叙述各种类型的查询及它们的参数表示法，如简单查询、复杂查询、逻辑结构查询、基于关系代数集合表示法的查询语言上的查询等。

第三章涉及对一个字段的各种查找法；即串行查找法、排序、二叉树查找法、索引顺序查找法、和分层存取法等。

第四章综述关键码到地址的各种变换法、溢出问题、和关键码到地址的各种变换法之间的比较。

第五章分析应答多属性多值的查询所用的各种文件编排模式。在这一章中，为了构造文件编排模式，广泛地使用了有限几何的种种性质。

第六章研究无需冗余存贮而存取时间又最小的各种文件组织。

第七章论述磁鼓上的各种记录组织法；即磁鼓上的对分查找法、磁鼓上的连续存贮、磁鼓上的调度等。

第八章概述检索数据和定义逻辑结构用的各型存取路径。涉及线性表、逻辑存取路径、和查找路径算法等。

此为试读，需要完整PDF请访问：www.ertongbook.com

目　录

第一章 数据结构

§1.0 引　言

社会接受计算机是需要时间的。许多团体迄今仍将计算机拒之门外，但计算机毕竟经受了时间的考验而以更快的步伐为社会生活的各个方面所接受。计算机已经成为美国各项业务活动的一部分。它正迅速进入教育、货物盘存、零售、行政事务等领域；它正精查法典以张正义和实施法律，正敏锐判断各项金融活动，正调度空中航线等等。这些不过是计算机的极少一部分应用而已，它们的潜在发展被人们的想象力所束缚。

我们之所以能利用电子特性来处理信息，硬件工艺的进展实属关键条件，但软件和数据处理技术却跟不上硬件的发展速度。其所以如此，主要原因之一是软件和数据处理均属于新异而复杂的事物。计算机应用之迅速波及种色各异而又毫无关联的领域，对计算机使用者和科研人员都提出了严峻的挑战。计算机原来只用于处理数值信息，但我们要处理的信息大多数毕竟是非数值信息，而有关它们的描述法和处理法我们却知之甚少。如果我们用若干基元来描述跟某一业务有关的数据，则我们并不清楚这同一组基元是否也足以用来描述跟另一项业务有关的数据。会不会有一种什么方法适于描述所有业务的有关数据呢？用于描述实业领域数据的方法是否适于描述教育领域或法律系统中的数据呢？等等。跟非数值信息处理关联的这类基本问题至今很少作出回答。

信息量的爆发之势对此情况确实更无帮助。我们的知识量每10年大体要翻一番，这些新知识大部分正被记入书籍中。图书馆首当其冲地面向上述问题，因而也是自动信息处理领域中的先行

者。许多科研人员在图书馆自动化领域中所做的开拓工作为非数值信息处理奠定了基础。他们的工作大都涉及检索有关文件，索引和回答书中内容所包含的某些甚为简单的问题等等。最后这个问题实际上就是开始尝试解决非数值数据处理领域中的复杂问题。

甚至在发展出高度可靠的自动化图书馆体系之前，就已经用计算机来存贮由多种应用领域得到的结构复杂的数据。于是，不同区域的计算机连结为计算机网络，以便不同区域的计算机用户可以分享此数据。因此，研制高效描述数据结构的方法这一要求，就越发变得比以前更为重要。

最近的十年内，已研制了许多用于存贮和检索非数值信息的信息体系。这些信息体系都是从处理数值信息所用的操作系统扩张而来的。为了处理现有的信息检索系统，曾研制过多种查询语言，且有广泛应用，但对于它们的前途有何作为，则实在知之甚少。为了发展良好的查询语言，就必须研制良好的数据描述。大多数查询语言并不完全依赖数据描述；有些取决于计算机系统的其他方面。真正独立于数据的查询语言只取决于描述数据的基元本身。在这类情况中，就有可能基于相同的数据描述来评价不同查询语言的能力。于是，发展良好的数据描述理论就是为研制强力的查询语言和信息检索系统、以及将计算机的应用扩张到新而富有挑战性的领域去铺平道路。

我们在本书中将用到许多基本数学概念，以描述数据，表征存贮数据用的硬件，以及发展存贮、检索和修改数据用的方法。下节内容就述说全书所用的基本数学概念。

§1.1 基本数学概念

本节讨论跟书中内容有关的一些基本数学概念。而将单章中所涉及的较专用的概念放在各章本身讨论。本章所述的基本内容散见于多种教科书（参考 Stoll, 1963；Riordan, 1968；Harary, 1969; Peterson, 1972); 因此，除有助于读者的理解而必须之外，定理的证明或结论之推导一律不给。

1.1.1 集合论

一个集合是由叫做元或元素的对象所组成。给出某一元素，我们可以判明它是否是某个集合的元素。如果 a 是集 A 的一个元素，我们就写成如下记法：

$$a \in A$$

如果 a 不是 A 的元素，我们写成：

$$a \notin A$$

如果 a_1，a_2，…，a_n 是集 A 的全部元素，则将集 A 记作：

$$\{a_1,\ a_2,\ \cdots,\ a_n\} \equiv A$$

如果语句 $P(x)$ 定义了集 A，也就是 A 的元素都是使 $P(x)$ 为真的那些对象，则将 A 记作：

$$\{x \mid P(x)\}$$

当且仅当两个集合的元素相同，我们才称这两个集合相等。两个集合 A 和 B 的相等记为：

$$A = B$$

两者的不相等记为：

$$A \neq B$$

如果 A 和 B 是两个集合，当且仅当 A 的各个元素都是 B 的元素，才称 A **包含于** B 之内，记作：

$$A \subseteq B$$

在这种情况下，习惯上称 A 是 B 的一个**子集**。上列记法（$A \subseteq B$）也可以说成是 B **包含** A。如果集 A 包含于 B 但不等于 B，就称之为**真包含**，且记作：

$$A \subset B$$

在这种情况中，也称 A 是 B 的一个**真子集**。下面列出包含关系的几个基本特性：

此为试读，需要完整PDF请访问：www.ertongbook.com

$A \subseteq B$

$A \subseteq B$ 与 $B \subseteq C \Rightarrow A \subseteq C$ （$\Rightarrow$表示蕴涵）

$A \subseteq B$ 与 $B \subseteq A \Rightarrow A = B$

仅有一个集合允许其没有元素，称之为**空集**。并记作ϕ。如果A有n个元素，则包括空子集在内，A的子集个数共有2^n个。

或属于集A或属于集B的所有元素的集合称之为A和B的**并集**。记作：

$$A \cup B = \{x \mid x \in A \text{或} x \in B\}$$

例如，$\{a_1, b_1, a_2\} \cup \{a_1, b_2\} = \{a_1, b_1, a_2, b_2\}$。既属于集$A$又属于集$B$的所有元素的集合称之为$A$和$B$之**交集**。记作：

$$A \cap B = \{x \mid x \in A \text{与} x \in B\}$$

例如，

$$\{a_1, b_1, a_2\} \cap \{a_1, b_2\} = \{a_1\}$$

如果两个集合的交集是空集，即$A \cap B = \phi$，则称这两个集合是**不相交**的。当且仅当集的集合内两两相异的集都不相交，此集合才称为**分离集**。集A的一个**分割**是指一个非空的分离集，这分离集内的各个集合都是A的相异子集，且这些子集的并集等于A。例如，$\{a_1, a_2\}$，$\{a_3\}$，和 $\{b_1, b_2\}$ 三者就是 $\{a_1, a_2, a_3, b_1, b_2\}$ 的一个分割。

凡是不属于集A的所有元素的集合，称为集A的 **绝对 补集**，记作 $\overline{A}$。因此：

$$\overline{A} = \{x \mid x \notin A\}$$

所谓集B关于集A的**相对补集**，是指A内不属于B的元素的集合。记作：

$$A - B = \{x \in A \mid x \notin B\}$$

集A和集B的**对称差集**记作$A + B$，定义为：

$$A + B = (A - B) \cup (B - A)$$

如果在某一论述中所考虑的所有集合都是一个集合如Ω集的子集，则称Ω是**泛集**。因此 $\overline{A} = \Omega - A$。

而对于$\cup$和$\cap$这两种运算，结合律、交换律和分配律都成立。考虑任何三个子集A，B和C；下列各式均成立：

结合律

$$A\cup(B\cup C)=(A\cup B)\cup C$$
$$A\cap(B\cap C)=(A\cap B)\cap C$$

交换律

$$A\cup B=B\cup A,\quad A\cap B=B\cap A$$

分配律

$$A\cup(B\cap C)=(A\cup B)\cap(A\cup C)$$
$$A\cap(B\cup C)=(A\cap B)\cup(A\cap C)$$

下列各等式是直接可得的：

$$A\cup\phi=A,\quad A\cap\Omega=A$$
$$A\cup\overline{A}=\Omega,\quad A\cap\overline{A}=\phi$$

上述的结合律、交换律和分配律都可以容易地推广到任何有限个数的集合上去。结合和交换二律的推广留作读者练习。分配律的推广如下：

$$\begin{aligned}&A\cup(B_1\cap B_2\cap\cdots\cap B_n)\\=&(A\cup B_1)\cap(A\cup B_2)\cap\cdots\cap(A\cup B_n)\\&A\cap(B_1\cup B_2\cup\cdots\cup B_n)\\=&(A\cap B_1)\cup(A\cap B_2)\cup\cdots\cup(A\cap B_n)\end{aligned}$$

还有如下三个定律，在集合论中也很重要：

幂等律

$$A\cup A=A,\quad A\cap A=A$$

吸收律

$$A\cup(A\cap B)=A,\quad A\cap(A\cup B)=A$$

德·摩根律

$$\overline{A\cup B}=\overline{A}\cap\overline{B},\quad \overline{A\cap B}=\overline{A}\cup\overline{B}$$

自对偶的例子如：$\overline{\overline{A}}=A$，以及如果$A\cup B=\Omega$，且$A\cap B=\phi$，则$B=\overline{A}$。

1.1.2 关系

如果一个对象 x 由**关系** ρ 跟另一对象 y 有关系，我们就写成 $x\rho y$ 或 $\langle x, y\rangle\in\rho$。关系的例子如 $x=y$，$x<y$，$x\equiv y$ 等等。x 和 y 都称为这关系的自变量。由于只有两个自变量，这种关系称为二元关系。一个二元关系也可以认为是一个有序对。设 $x\in X$ 和 $y\in Y$，则 ρ 就是**笛卡儿积空间** $X\times Y$ 的一个子集（参考 Stoll，1963）。关系 ρ 的元素都是满足 $X\times Y$ 中 $x\rho y$ 的有序对。我们称有序对 $\langle x, y\rangle$ 中的 x 为第一坐标，称 y 为第二坐标。如果 $x\rho y$ 和 $y\rho z$ 两者蕴涵 $x\rho z$，则称这关系是**可递的**。

正如可以将有序对的概念扩张为有序 n 元组 $\langle x_1, x_2, \cdots, x_n\rangle$ 一样，也可以将二元关系的概念扩张为 n 元关系。n 元关系 R 是积空间 $X_1\times X_2\times\cdots\times X_n$ 的一个子集，式中 X_i 都是集，$i=1, 2, \cdots, n$。R 的元素都是满足关系 R 的有序 n 元组。X_j 称为 R 的第 j 个**定义域**。例如，如果 $R\equiv$ 整数域上一个五元空间中的相等的自变量，则 R 的各元素为（1，1，1，1，1），（2，2，2，2，2），（3，3，3，3，3），…。

由于关系也是集合，故集合论的所有一般运算都适用于关系。但是，关系在集合论的所有运算下都是不封闭的；例如，一个二元关系与一个三元关系之并不是一个关系，它们之间的交也不是。

有些运算适用于关系，但不适用于一般集合。设给定一个关系 R，在交换任何两个定义域后，就能得到一个新关系。这新关系称为原关系的**置换**。一般说来，从 n 个自变量的一个关系可以得出 $n!$ 个置换关系。

如果从关系 R 中去掉定义域 X_n，并去掉重复元素，所得到的集合是有 $n-1$ 个自变量的一个关系。称之为 R 在 $X_1\times X_2\times\cdots\times X_{n-1}$ 上的**射影**。

对 $k\leqslant n$，令 $I=\{i_1, i_2, \cdots, i_k\}$ 是 k 个指标的一个列表；则当选择运算符 Π_I 作用于 R 上时，给出一个 k 元关系 $\Pi_I(R)$，它的第 j 个定义域是 X_{i_j}。选择运算符 Π_I 既可用于置换，也可用于

射影。

将两个关系联合为另一关系这个概念是较难建立的。可以绕着圈子定义为：给定 $X_1 \times X_2 \times \cdots \times X_k$ 上的一个关系 R_1，及 $X_k \times X_{k+1} \times \cdots \times X_n$ 上的另一关系 R_2，如果在 $X_1 \times X_2 \times \cdots \times X_n$ 上有一个 n 元关系 R_3，使得 $\Pi_{12\cdots k}(R_3)=R_1$ 且 $\Pi_{k\cdots n}(R_3)=R_2$，则称 R_1 和 R_2 是可联合的。R_3 就称为 R_1 和 R_2 的联合。联合也可以定义在多重定义域上。

给定两个关系 R_1 和 R_2，如果有 R_1 和 R_2 的一个联合 R_3，使得 $\Pi_{12\cdots k-1,k+1\cdots n}(R_3)=X_k$，则$X_k$ 称为 R_1 和 R_2 的**合成关系**。

1.1.3 函数

函数是一种 n 元关系，其任何两个相异元素的前 $n-1$ 个坐标都相异。因此，当且仅当满足下列两个条件，f 才是一个函数：

（i） f 的元素都是有序元组。

(ii) 如果$\langle x_1, x_2, \cdots, x_{n-1}, x_n\rangle$和$\langle x_1, x_2, \cdots, x_{n-1}, x_n'\rangle$是 f 的两个元素，则 $x_n=x_n'$。

如果$\langle x_1, x_2, \cdots, x_n\rangle \in f$，我们就写为$\langle x_1, x_2, \cdots, x_{n-1}\rangle f x_n$。如果 f 是一个函数，而$\langle x_1, x_2, \cdots, x_{n-1}, x_n\rangle \in f$，使得$\langle x_1, x_2, \cdots, x_{n-1}\rangle f x_n$，则称$\langle x_1, x_2, \cdots, x_{n-1}\rangle$是 f 的自变量。称 x_n 为 f 在$\langle x_1, x_2, \cdots, x_{n-1}\rangle$的值，或称之为$\langle x_1, x_2, \cdots, x_{n-1}\rangle$在 f 作用下的象，等等。词“映射”和“变换”都是词“函数”的同义语。上述函数也可记作$\langle x_1, x_2, \cdots, x_{n-1}\rangle \to x_n$。以后用符号$X$表示 f 的定义域，用Y表示x_n的值域。函数的例子如⟨1，3，7⟩、⟨狄克，亨利，汤姆⟩、⟨书，钢笔，学生⟩等。一个函数的象空间不必只包含一个定义域，也可包含多个定义域。一个 n 元组的象仍可以是一个 n 元组，只要多个 n 元组不映射为同一个 n 元组。因此，当 $X=\{\langle x_1, x_2, \cdots, x_{n_1}\rangle\}$ 而 $Y=\{\langle y_1, y_2, \cdots, y_{n_2}\rangle\}$ 时，$X\to Y$ 是一个函数。

当且仅当 f 的值域如 R_f 是 Y 的一个子集，才称函数 f **映入** Y，而当且仅当 $R_f=Y$，才称 f **映到**Y。符号 $f(x)$ 也表示 x

在 f 作用下的象。

如果一个函数将相异元素映到相异元素，称这个函数是**一对一函数**，即：

$x \neq x' \Rightarrow f(x) \neq f(x')$ 且 $f(x)=f(x') \Rightarrow x=x'$。

例子如 $f(x)=x+1$，式中 x 是实数。

两个函数 f 和 g 的**合成函数**写作 $g \circ f$，定义为如下的集：

$\{\langle x,\ z\rangle |$ 存在一个 y 使得 xfy 且 $ygz\}$。

函数的合成运算是不可交换的，但遵循结合律。函数 f_1，$f_2, \cdots$，f_n 在合成定义下所得到的函数是唯一的，记作：

$$f_1 \circ f_2 \circ \cdots \circ f_n$$

如果 f 是一个一对一函数，则将 f 的各个元素的自变量跟值坐标交换后所得到的函数叫做 f 的**反函数**。记作 f^{-1}。如果 f^{-1} 存在，则它的定义域就是 f 的值域，它的值域就是 f 的定义域，而当且仅当 $y=f(x)$，才有 $x=f^{-1}(y)$。

1.1.4　二项式系数

牛顿所发现的二项展开式如下：

$$(1+x)^n=\sum_{k=0}^{n}\binom{n}{k}x^k$$

式中$\binom{n}{k}$是从 n 个对象中选出 k 个对象所能得到的不同款式的个数，称之为二项式系数。可以将它表为积或阶乘之比：

$$\binom{n}{k}=\frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots(1)}$$

$$=\frac{n!}{k!(n-k)!}\text{ 对整数 } n \geqslant \text{整数 } k \geqslant 0$$

二项式系数可以分解为：

$$\binom{n}{k}=\binom{n-1}{k}+\binom{n-1}{k-1}$$

本书内容要用到的二项式系数的几个基本公式及二项式系数之和的公式为：

$$\binom{n}{k}=\binom{n}{n-k},\quad \binom{n}{k}=\frac{n}{k}\binom{n-1}{k-1}$$

$$\binom{n}{k}\binom{k}{l}=\binom{n}{l}\binom{n-l}{k-l}$$

$$=\binom{n}{k-l}\binom{n-k+l}{l}\quad \text{对整数 } l\geqslant 0$$

$$\sum_{i=0}^{n}\binom{n}{i}=2^n$$

$$\sum_{i=0}^{m}\binom{n+i}{i}=\binom{n+m+1}{m}\quad \text{对整数 } m\geqslant 0$$

$$\binom{n}{k}=\sum_{i=0}^{K}\binom{n-1-i}{k-i}\quad \text{式中 } K=\min(k,n-1)$$

$$\sum_{i=0}^{m}\binom{i}{k}=\binom{m+1}{k+1}$$

$$\sum_{i=m}^{\min(n,l+m)}\binom{n}{i}\binom{l}{m-i}=\binom{n+l}{m}$$

$$\sum_{i=0}^{\min(n,l-m)}\binom{n}{i}\binom{l}{m+i}=\binom{n+l}{n+m}$$

1.1.5 图论

图论是数学的一个分支，是研究平面上用点和线段画成的图形特性的。图形中的点称为**顶点**或**结点**或**点**。两个结点之间用一条线段连接。一条线段的两个端点如果连接在同一结点上，就称这线段为一个**圈**。如果线段本身无任何方向性，就称它为一条**棱**。如果线段本身有方向性，就称它为一条**弧**。用结点和棱构成的图形叫**无向图**。一个图形可以用它的结点集$\{v_i\}$和棱集$\{e_i\}$指定。图1.1.5.1就是无向图一例。

此为试读，需要完整PDF请访问：www.ertongbook.com