目 录

张庆国

第八章　数据检索系统

8.1 概论

在以上的章节中，我们讨论的都是文献检索。尽管我们也称其为情报检索。但是我们还应注意到这样一个问题；大多数情报检索系统感兴趣的是事实或数据，而不是在从中抽取出所需信息前还要对其进行分析研究的文献。因此，数据检索和事实检索是情报检索的重要任务。

数据检索的一种方法是进行所谓的“细节检索”。细节检索的原理是：将每篇文献分解成小的部分——单个的章节、页、段，甚至句子——并且只检索这些个别的部分而不检索整篇文献。为达到这个目的，我们需要对文献内容进行足够深度的分析，以便能够刻划文献各个部分的特性并且指出它们之间的区别和联系。还要把文献的原文也存贮起来，以提供需要的细节检索能力。细节检索的文档组织、提问公式和用户——系统交互方法所用的基本检索处理过程与整篇文献的检索并无太大的不同。

但是，目前的数据检索基本上都是用比较成熟的数据库技术实现的。并且，除了包含在文献中的数据之外，还有大量的以独立形式存在的数据，它们通常也是由数据库技术管理（包括检索）的。本章简单介绍与情报检索有关的数据库技术。

数据管理系统和标准的文献检索系统有许多相似之处：生成和维护一个存贮数据库；信息查询的结果是检索存贮数据中可回答用户提问的一部分，等等。数据管理系统最显著的特性是它存贮的数据的确定结构，这一点不同于我们已介绍过的文献检索，也不同于我们以后将要介绍的事实检索。在数据管理系统中，只使用结构完

全限制和完全确定的信息项目，清除了不确定的和含混的成份。实际上，数据管理系统通过处理小的、预先指定的属性集合来处理数据文件。例如，一个由个人记录组成的文件可以通过以下项目确认：所涉及的人员姓名、地址、每人的年龄、工作类别、及每人的年工资。每个属性可期望只带有许多特定值中的一个——例如，对上述问题中的工作人员，其年龄能被限制在18岁到65岁之间。对于文件中的一个记录，可以通过引用这个特定记录的属性所带的特定值来描述。所以，一个名为Smith的人可被指定为

（姓名=Smith；地址=110Main st；

年龄=25；工作类别=123；

工资=19，500）

显然，在这样的环境中，许多困难的检索问题可以避免。尤其是索引语言的选择、项目本身的内容分析和标引操作。更进一步地，如果假设每个项目通过属性值的选择都已得到完全的和确定的描述，那么需要进行的处理工作的种类也是有限的。特别地，我们可以比较用户提问中指定的属性值和存贮数据的属性值，以期得到属性值满足一个给定属性范围的全部记录（如年龄在25岁和35岁之间的全部工作人员）。再者，数据库管理系统可设计用来进行简单的数值处理，如按照特定的规则对记录数量进行计算——例如，在某年出版的图书数量；或者某特定出版者的出书数；或者得到平均数，例如在某年中出版的图书的平均费用。

能够完成包括上述各种功能的系统即为我们在这里讨论的数据库管理系统（DBMS Data Base management System）

一个数据库管理系统由三个主要部分组成：首先是一个由各种文档构成的数据库，文档被分解成单个的记录。这一部分还包括在

文档中运行的访问和文件维护操作；其次，要有一个通信系统，提供系统用户和自动化系统之间的界面，还包括报文处理、编辑、输出显示功能；最后一个是事务处理管理系统，负责对来自各种用户的作业进行调度，对文件的存取控制，对可能同时进行的并发作业的管理，和在系统出现故障之后系统重启和恢复过程的实现。

数据库管理系统一般具有以下特点：

1、用户在编写应用程序时，可以不考虑表示和组织存贮中数据方法的细节，不考虑访问这些数据的特定过程，换言之，用户程序有一种“物理数据独立性”，因为它们并不依赖于数据库管理系统使用的实际物理实现方法。

2、类似地还有“逻辑数据独立性”，它的意义是，用户程序可以有效地与数据库实体的内部结构和各实体之间定义的各种关系独立开来。

3、高级语言在使用上的便利经常用来帮助用户对系统提交查询和指定需要的文件处理操作。

4、通过设置用户提出的与文件和数据项特性有关的合法性语句以保持数据的质量。例如，对于特定属性，其值的指定范围是可访问的，还可以让系统自动提供错误检验以查证合法性断言。

5、提供重启能力，当系统出现故障时，通过适当的中断，使系统记录当前的状态，以保持正确的处理顺序，并且在对这些记录下的状态信息进行校正后，据此恢复系统的运行。

6、通过保密变换和存取控制提供安全性，保证存贮中的数据不被误用和独占。

为了将高级处理说明书转换成特定机器的操作，并且提供上文中提到的数据独立性，需要存贮实际使用的文件结构的详细描述。

和每个文件中的单个记录的结构的详细描述。这种数据描述和文件表示法被称为模式。模式有几种类型：个别用户用来指定他们自己的数据的外模式，或称用户模式；概念模式。对于所有用户都是一样的，并且被用来提供需要的数据独立性；机器模式或称内模式，用来表示实际的物理数据结构。这个结构是系统需要用来完成文件处理操作的。

通常，高级数据描述语言是由数据库管理系统提供的，用来帮助用户指定个别的用户模式。一旦用户指定了一种模式，将外模式转换成程序使用的内模式的工作便完全留给了系统。在执行程序建立和转换工作时，系统通常使用一种存贮的“数据字典”。这个字典中记录了系统中所有对象的描述，包括文件和相应的文件模式，用户终端及其确认和特点，单个用户及其状态和文件存取权限，实际需要完成的特殊的事务处理和文件操纵等等。

8.2 数据库管理系统的结构

8.2.1 信息项的结构

为了描述DBMS的单个操作，需要较详细地考察对其进行操作的信息项的结构。事实上，对特殊信息结构的利用构成了常规书目检索和数据库管理的基本区别。特别地，书目信息系统中使用的信息通常被认为是无结构的，并且数据元素经常是“自描述”的。后者的意义在于：我们很容易区分作者名和出版者或文献简名。在数据管理环境中，单个记录不是自描述的，一个确定的记录结构或者是假定的，或者是隐含的。

作为一个例子，我们考察INSPEC检索服务中心使用的一个典型的书目记录的成份，并且将其表示在表8.1中。这个记录有

三种类型的标识符：目标词，如作者、期刊名、或者页号；内容标识符，如自由词和从受控词汇表中抽取的词；篇名和文献的文摘。除了这个表中隐含的结构，在篇名和文摘中还有大量固有的句法和语义结构。更进一步地，这个项目中的实际文章和书目参照之间的牵连关系也可用来定义文章之间的明确的层次关系。

表8.1 INSPEC的一个记录的成份

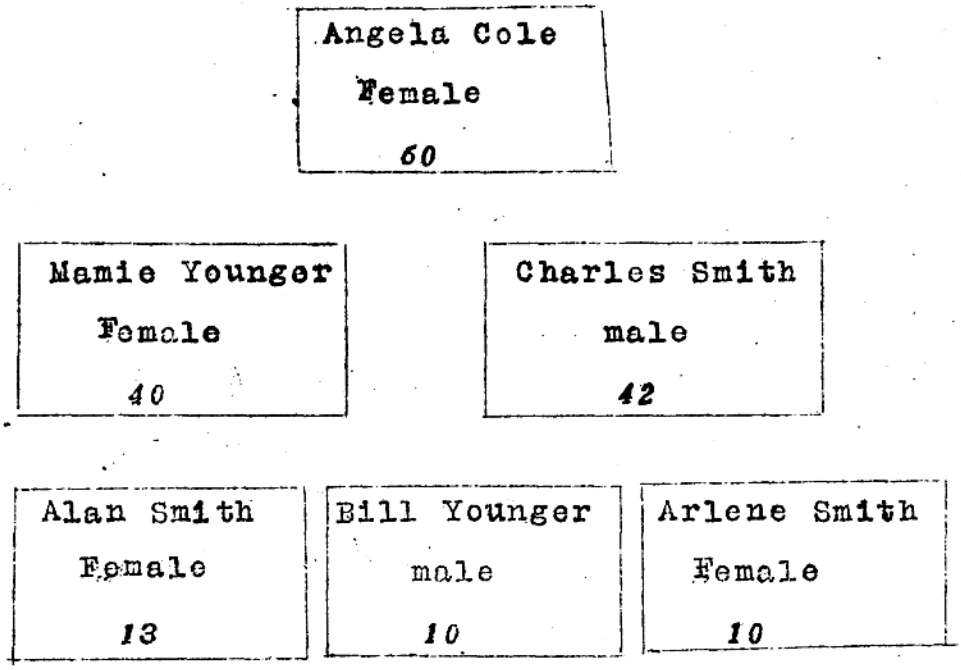
篇　名	参考号
作者名	分类号
出版商	受控标引词
卷号、期号	自由标引词
出版日期	文章类型和处理类型
页号	（例如：书目或是文献考察、
出版的语种	一般性或是综述性文章）
文摘的全文	

在实际上，置入一个书目记录中的语言学结构通常并不用来进行信息检索。检索词或关键字几乎总被认为是相互独立的，并且为显示篇名和文摘的语言学结构而需要的语言分析方法的完成是昂贵的，对实际使用的可靠性也不足。在数据库系统中，情况则完全不同，因为结构是经过精心设计的，并且用于检索。

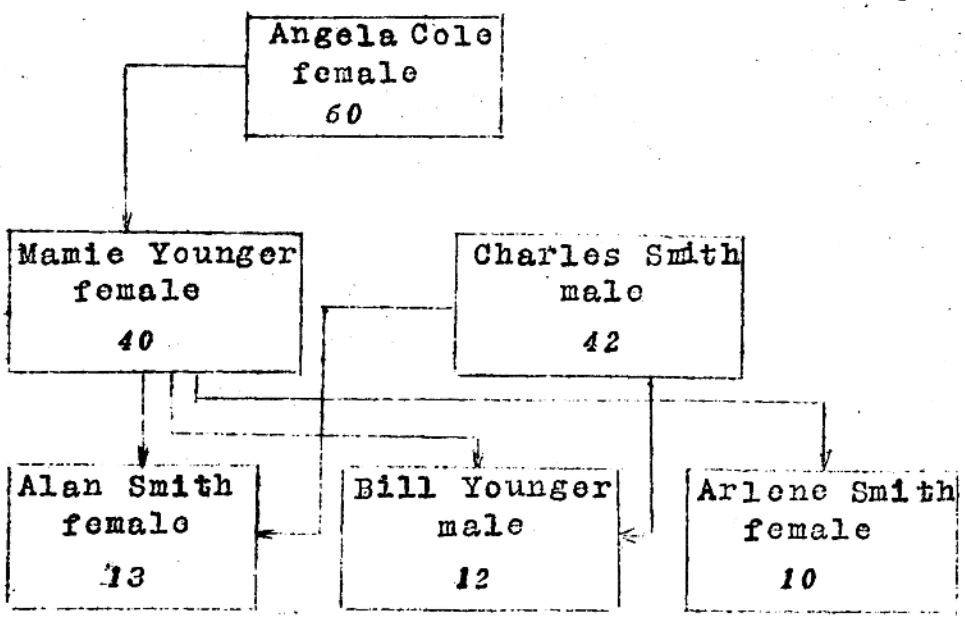
为了描述数据库系统中信息项的结构，比较方便的方法是区分要操纵的实体和刻划实体的属性。这二者通常的不同点依赖于它们在一个特定检索环境中所扮演的角色。实体是数据对象，有其自己的独立生命，并且因此，它们构成了对于检索系统的用户有主要兴趣的元素。另一方面，属性之所以存在，只是因为它们作为标识

管被赋予了有定义的实体。在一个DBMS中，同一个数据在不同的环境下可以作为实体和属性分别出现。例如，在一个对学生记录的描述中，可把教师作为属性（教了这个学生的某门课，或者说，“任课教师”这个属性的值是该教师的姓名）；在另外的情况下，当然可以把这个教师当成单独的实体。

图 8.1 示出了一个“个人”数据库的例子。在这个例子中，人是实体，属性包括姓名、性别和年龄。图 8.1.a 给出了六个这样的个人记录，图 8.1.b 则给出了这六个记录（六个实体）之间的父——子关系。



a



b

图 8.1　一个“个人”数据库

在DBMS环境中，习惯于在实体之间定义两种类型的关系。一种是普通关系，或称“层次包含”关系。在这里，某些实体是一般性的，支配若干类实体；其它的实体则较小，依赖于较一般的实体。通常用树型结构来表示实体之间的层次包含关系。实体之间的另一种关系是非层次关系。非层次关系一般用网状结构表示。

为了有效地实现实体之间的关系，通常按照一种实体可以联系多少其他类型的实体来对这些关系分类。所以，有1对1联系，象飞机和航班，每架飞机只能指派给一个航班，每个航班只能由一架飞机完成；有多对1关系，例如学生和教室，每个教室可容纳多个学生，每个学生只能在一个教室里；有多对多关系，例如工人和工程项目，每个工人可参加多个工程项目，每个工程项目可由多个工

人参加。

在数据库中，实体之间的联系通常是用指针表示的。图 8.1 中已经给出了指针的示例。在实体关系参预作用的情况下，物理地址指针的使用简化了对信息要求的处理工作。例如，对于图 8.1 的数据库，我们很容易满足“给出 Marie Younger 的全部后代的姓名和年龄”的提问。

除了实体与实体之间的关系外，属性之间也存在关系。在属性之间，可以定义许多种语义关系。例如，在人的年龄和职业之间存在关系（初等学校的学生通常小于 15 岁，而管子工或焊工则超过 15岁）。属性之间的另一种关系取决于某些属性值能不能在同一个集体的不同记录中是一样的。例如，考虑一个人员文件。假定社会保险号（SS）唯一地确定了文件中的每一个人。在这种情况下，显然，如果同样的SS号碰巧确认了两个记录，那么它们的性别年龄等其他属性值应该是一样的。因为这两个记录代表同一个人。从技术上说，我们认为在一个人的社会保险号和这个人的年龄之间存在一种函数相关性，或者说SS号功能性地决定年龄。

依靠数据操纵中明确使用的关系类型，通常区分为三种抽象的数据库模式。它们分别称为关系模式、层次模式和网络模式。我们以下讨论这三种模式。

8.2.2 关系数据库模式

在关系数据库模式中，实体型之间并没有定义的明确关系，并且不实际存贮指针。实体之间的关系必须在它们用来回答查询之前抽取出来。

一个关系是一个表，它通过这样一种方法来表示文档中的记录：

每记录。亦称元组，是通过属性值的一个有序集合来确定的。每个记录对应于表中的一行。每一列则表示用以刻划记录的一个属性。表8.2是一个简单的关系：

姓名	号码		地址	城市
Brindle	01764	264	First St	New York
Cumind	25511	11	A St	Kansas City
Daniel	4759	1011	Main St	Kingston
Kaizer	77084	2146	Meadowrook	Philadelphia
McGill	5730	189	Edgemont	Washington
Boreault	1920	12	Ackerman	Syracust
Salton	3081	2365	Meadowlark	Cambridge

表8.2 一个简单的关系

一个关系数据库模式中通常包含多个关系。图8.2是一个有三个关系的关系数据库模式。

1. | 雇员号 | 雇员姓名 | 工资 | 地址 | 工作类别 |
2. | 雇员号 | 雇主 |
3. | 雇主 | 工程号 | 工程名 |

图8.2 由三个关系组成的关系数据库模式

在上图中我们看到，有重复存贮的属性。这种存贮上的冗余在关系数据库中通常是不可避免的。因为关系数据库中不设指针。

对于关系可以定义各种运算。它们称之为关系代数。设R和S分别是两个关系，它们各自包含若干行。常用的关系代数有

1. 关系并R∪S。R∪S是一个关系。其中包括R和S中的所有不同的行。

2. 关系差R－S。R－S是一个关系。其中包括在R中但不

在S中的行。

3．选择操作。在给出一定的条件后，可以对一个关系R施行选择操作，产生的关系中包括原关系R中满足该条件的行。这个新的关系用 SELECT(条件)(R) 来表示。

4．投影操作。对关系R的投影操作记为 $PROJECT_{i1,i2,\cdots it}(R)$。这也是一个关系，其中只包含原关系R中的t列，并且这t列的顺序为 i1，i2，…，it。投影后删去相同的行。

5．关系R和关系S的卡氏积R×S。设R中有n行，S中有m行，则R×S中共有n·m行。其中每一行由R中的一行和S中的一行连接而成。

图 8.3示出了这五种关系运算。其中(a)和(b)分别是关系R和关系S。(c)是关系RUS。(d)是关系R—S。(e)是选择操作 $SEL_{(NAME=ADAMS)}(R)$。(f)是投影操作 $PROJ_{3,1}(R)$。(g)是卡氏积R×S。

Adams	22	New	York
Adams	25	New	York
Brown	22	New	York

(a)

Adams	25	New	York
Smith	25	Chicago	

(b)

Adams	22	New	York
Adams	25	New	York
Brown	22	New	York
Smith	25	Chicago	

(c)

Adams	22	New	York
Brown	22	New	York

(d)

Adams	22	New	York
Adams	25	New	York

(e)

New	York	Adams
New	York	Brown

(f)

Adams	22	New York		Adams	25	New York	
Adams	22	New York		Smith	25	Chicago	
Adams	25	New York		Adams	25	New York	
Adams	25	New York		Smith	25	Chicago	
Brown	22	New York		Adams	25	New York	
Brown	22	New York		Smith	25	Chicago	

(g)

图 8.3 关系代数

在上述关系代数中，我们没有定义关系交R∩S，这是因为 $R \cap S = R - (R - S)$。

关系的卡氏运算是将两个关系合并成一个关系。它的两个特例是所谓的限制联接和非限制联接，或统称为联接操作。它是关系数据库中一类非常有用的操作：

1. 限制联接。设T和V是两个关系。限制联接 T JOIN V (米) 是一个关系。它由这样的元组组成：从关系T中抽取一个元组，再从关系V中抽取一个元组。这两个元组的关系要满足条件米。如米是B＜D，则表示T中元组中的属性B的值要小于V中属性D的值。限制联接相当于在卡氏积中进行一次选择操作。

非限制联接又称自然联接。对于非限制联接，我们总假定要联接的关系中的某些属性包括相同的值域，即两个关系中的某些列代表同样的属性的值。记这个相同的属性值为C，则非限制联接相当于限制联接 T JOIN (C=C) V ，非限制联接后，两个相同的属性值被删去一个。

图 8.4 给出了非限制联接和限制联接的例子。其中(a)是关系T，(b)是关系U，(c)是非限制联接 T JOIN U；(d)是关系T，(e)是关