自動資訊查尋的新方法

張 系 國*

一、引 言

資料基(database)或**資料庫**(databank)是近幾年來在**資訊科學**(Information Science)的研究以及**資訊系統**(information system)的設計方面的一個新觀念。英格斯氏曾對資料基做了如下的定義：

「資料基就是儲存的各種資料的總和，可供某一機構應用上需要的。」[1]

例如，一所醫院有關病人的病歷、診斷結果和檢查報告；一所大學有關學生的成績資料、學生入學統計資料和課程提要；一家圖書館有關書籍的編號、分類和要目；一個政府機構有關人力資源的統計、分析和普查報告；都構成每個機構的資料基。資料基既然包括了一個機構的所有資料，因此，有關這個機構的各種資訊，都可以從資料基裏設法查尋。

在過去，資料基泛指一個機構利用各種方法儲存的資料。例如卡片檔案、顯微膠片、文書記載、文件檔案等等。自從電子計算機的廣泛採用以來，一個機構的各種資料，都可以經過重新整理以後，存入電子計算機的各型記憶器裏。例如**磁蕊記憶器**(magnetic core memory)，可以容納數千字到數萬字的資料；[2]**磁碟記憶器**(magnetic disk memory)，可以容納數十萬字到數百萬字的資料；**磁帶記憶器**(magnetic tape memory)，可以容納數百萬字到億萬字的資料；更龐大的**磁帶庫**，可以容納百億字的資料。因此，近年來逐漸出現不少以電子計算機爲主體的電化訊息系統，儲存了有關一個機構所有的重要資料，構成一個由電子計算機控制的資料基。本文所討論的資料基，所指的就是儲存在各型電化記憶器內，由電子計算機來控制儲存及提取資料的電化資料基。

電化資料基具有以下的優點：

㈠由於電子計算機能在極短時間內完成複雜的運算，因此資料的分析統計並不必依賴人工，可以由機器自動處理。這不僅節省人力，而且可以爭取時效，對現代機構的管理和決策，有很大的幫助。

---

* 著者現任美國伊利諾大學芝加哥分校資訊工程系副教授。

㈡由於電子計算機在軟體系統的控制下，可以迅速找尋到儲存在各型記憶器內的資料，因此原來依賴人力難以查獲的資料，現在都可以迅速查獲。

㈢由於所有的資料都存在各型記憶器內，構成一個完整的資料基，因此各種資料的**互查**（cross referencing）和**互相驗證**（cross checking），都非常方便。這是在以往的人力檔案系統裏，不容易辦到的事。例如稅務局調查逃稅事件，如果沒有電化資料基和電子計算機的幫助，會十分困難。又如圖書館查尋某一方面的資料，如果有電化資料基，便可以縮短查尋的時間，節省查尋者所花費的精力。

在設計以電子計算機爲主體的電化資訊系統時，一個最重要的問題，便是如何從電化資料基裏查尋訊息。本文的目的，在介紹一種設計資料基的新觀念，以及**自動訊息查尋**（automated information retrieval）的一些新方法。

二、關係資料基

關係資料基（relational database）是設計資料基的一個新觀念。這個構想，首先由柯德氏提出3，由於觀念簡單而清晰，很易爲一般人接受，所以受到研究電化訊息系統的學者的廣泛注意。在本節裏，我們以一個圖書館的假想資料基爲例，介紹關係資料基的基本觀念。

關係資料基的基本構想，就是將資料基裏的資料，組織成爲大大小小的許多個表（table）。從關係資料基裏查尋資訊，便可以查表而得。這種查表的觀念，很容易爲一般人接受。我們平常許多資訊查尋的工作，其實都是查表。例如到車站買火車票，旅客可以查票價表以找出票價的多少；到學校查成績，學生可以查成績公佈欄以找出自己的成績多少。這樣的例子，不勝枚舉。現在以假想的圖書館資料基爲例。這個資料基，包括以下三個表：

A表

編號	書　名	作　者	出版年份	出版書局	分類號
01	中國通史	張得功	1950	A	M1
02	西洋通史	李　四	1972	B	M1
03	天文學入門	王　五	1973	C	M4
04	中國地理	張得功	1952	A	M2
05	莊子哲學	周志堯	1950	B	M3

B表

分類號	分類名稱	總類別
M1	歷史	文史哲學
M2	地理	文史哲學
M3	哲學	文史哲學
M4	天文	科學

C表

出版書局	書局名	地點
A	中華	台北
B	東方	台北
C	科學新知社	香港

A表儲存了有關書籍的編號、書名、作者、出版年份、出版書局，以及分類號等資料。B表儲存了有關書籍分類的分類號、分類名稱及總類別等資料。C表儲存了有關出版書局的代號、書局名以及地點等資料。

一個關係資料基，就包括了類似上例的大小各種表格。「關係」一詞，意指每一個表都代表了各個**資料單位**（ data item ）間的某種**關係**（ relation ）。例如C表的每一**行**（ row ），都表示了「出版書局」，「書局名」以及「地點」間的關係。我們平常要查尋的資訊，正是這些**資料單位**之間的關係。例如我們想要知道一本書的作者是誰，便是找尋「書名」和「作者」兩個資料單位間的關係。

關係資料基裏每一個**表**（table），如果儲存在計算機的記憶器裏，就構成一個**檔案**（ file ）。如何將檔案儲存在記憶器，而能够使得提取資料方便，是資訊系統設計者面對的一個重要問題。簡單的說，每個表都必有一個**主鍵**（ primary key ）。例如A表的主鍵是「編號」，B表的主鍵是「分類號」，C表的主鍵是「出版書局」。在任何一個表裏，每一行資料的主鍵值都不相同。因此我們可以根據主鍵的值，找尋到所需要的各行資料。各個檔案，也常依照主鍵值的大小順序排列後，存到計算機的記憶器裏。例如A表，B表和C表的各行，都依照主鍵值，由小排到大。這可說是便利查尋的基本方法。我們如果要查有關一本書的資料，便可依照「編號」，很快在A表查出它的書名，作者等資料。其他更複雜的方法，就不在此細述了。4

三、關係資料基的幾種運作

如果要從關係資料基表查尋訊息，可以使用一些基本的**運作**（operation），來找尋所要的資料。下面舉七個例子，來說明這些基本運作。

例一：所要查尋的是

「找尋所有的出版書局名稱」

這個**查尋訊問**（retrieval query），可以用一個**投影運作**來表示。投影運作（projection operation）就是從一個大表內拿出有用的**列**（column），來構成一個小表。例如以上的查尋訊問，相當於如上的投影運作。

U＝C（書局名）

在上述的投影運作之後，U表便包括下述的資料。

U表

書局名
中華
東方
科學新知社

投影運作很像是幾何學上的投影，因此得名。上例說明，投影運作可用來從大表找出小表。

例二：所要查尋的是

「找尋所有的書名和作者名稱」

這個查尋詢問相當的投影運作如下。

U＝A（書名，作者）

在上述投影運作後，U表便包括下述的資料。

U表

書名	作者
中國通史	張得功
西洋通史	李四
天文學入門	王五
中國地理	張得功
莊子哲學	周志堯

例三：查尋詢問如下

「找尋所有的總類別」

相當的投影運作如下。

U＝B（總類別）

在上述投影運作後，U表便包括下述的資料。

U表

總　類　別
文　史　哲　學
科　　學

注意B表原有四行資料，U表則只有兩行，換句話說，重複的行都已合併成一行。我們只需知道總類包括「文史哲學」及「科學」，所以「文史哲學」並不需要重複三次。

例四：查尋詢問如下

「找尋張得功所有的著作」

這個查尋詢問，不能只用投影運作來解答。我們只對張得功的著作有興趣，並不想知道別人的著作。因此，我們用一個**限制運作**（restriction operation）來檢出張得功的著作，方法如下。

U＝A〈作者＝張得功〉

上述的運作，尖括弧內是一個**限制條件**（restriction condition），表示「作者」必須等於「張得功」。在上述限制運作之後，U表便包括下述的資料。

U表

編號	書　名	作　者	出版年份	出版書局	分類號
01	中國通史	張得功	1950	A	M1
04	中國地理	張得功	1952	A	M2

如果我們只需要知道張得功著作的書名，那麼便可以先用一個限制運作選出U表如上述，再從U表做一投影運作，選出V表如下。

U＝A〈作者＝張得功〉

V＝U（書名）

在上述兩組運作之後，V表包括下述的資料。

V表

書　　名
中 國 通 史
中 國 地 理

例五：查尋詢問如下

「找尋張得功在一九五〇年以後出版的著作」

這個查尋詢問相當的運作如下。

U＝A〈（作者＝張得功）∧（出版年份＞1950）〉

V＝U（書名）

上述的限制運作，有一個較複雜的限制條件，意思是「作者」等於「張得功」和「出版年份」大於「1950」。∧是一個**邏輯符號**，表示邏輯的「和」。其他可用的邏輯符號有∨（邏輯的「或」）和～（邏輯的「非」）。**比較符號**可用的則有＝（等於），≠（不等於），＞（大於），≥（大於或者等於），＜（小於），≤（小於或者等於）。限制運作表的限制條件，是由邏輯符號、比較符號以及資料單元所構成的邏輯式子，因此可以表達十分複雜的邏輯上的限制條件。

在上述的兩個運作之後，V表包括下述的資料。

V表

書　　名
中 國 地 理

上述的書是張得功在一九五〇年以後出版的唯一著作。

例六：查尋詢問如下

「找尋在台北出版的所有書籍」

這個查尋詢問，不能只用投影運作和限制運作來解答。我們可以先在C表上做一限制運作如下。

U＝C〈地點＝台北〉

在上述運作之後，U表包括下述的資料。

U表

出版書局	書局名	地名
A	中華	台北
B	東方	台北

如果再在U表上做一投影運作，

V＝U（出版書局）

便可以得到V表如下。

V表

出版書局
A
B

在V表裏，我們得到所有在台北的書局代號。如果我們找出這些書局所出版所有的書籍，那麼這些書籍，就是在台北出版的所有書籍。因此，我們可以在A表再做如下的限制運作。

W＝A〈（出版書局＝A）∨（出版書局＝B）〉

在上述運作之後，W表包括下述的資料。

W表

編號	書名	作者	出版年份	出版書局	分類號
01	中國通史	張得功	1950	A	M1
02	西洋通史	李四	1972	B	M1
04	中國地理	張得功	1952	A	M2
05	莊子哲學	周志堯	1950	B	M3

上述的限制運作，可以用一個**結合運作**（ join operation ）來代替，得到完全相同的效果。

W＝A（※出版書局）V

「結合運作」的主要功用，是用來結合兩個表內的資料。例如V表包括了

所有在台北的「出版書局」。因此，我們選擇「出版書局」來結合兩個表。※是結合符號，「出版書局」是用來結合的資料單位。如果V表內有一個出版書局A，那麼W表內所有出版書局也是A的各行就可以保留。因此，上述結合運作，就保留了A表內所有出版書局是A或B的各行，造成了新表W。

本例的查尋詢問，因此相當於下列四個運作。

U＝C〈地點＝台北〉　　（限制運作）
V＝U（出版書局）　　（投影運作）
W＝A（※出版書局）V　　（結合運作）
X＝W（書名）　　（投影運作）

在這四個運作之後，W表內包括下述的資料。

W表

書　　名
中　國　通　史
西　洋　通　史
中　國　地　理
莊　子　哲　學

上述四本書就是所有在台北出版的書籍。

例七：查尋詢問如下

「找尋所有文史哲學類，在一九六〇年前出版的書籍」

這查尋詢問相當於下列五個運作。

U＝B〈總類別＝文史哲學〉　　（限制運作）
V＝U（分類號）　　（投影運作）
W＝A（※分類號）V　　（結合運作）
X＝W〈出版年份＜1960〉　　（限制運作）
Y＝X（書名）　　（投影運作）

第一和第二個運作，找出文史哲學總類的分類號。然後，第三個運作根據「分類號」結合A表和V表，找出A表內所有文史哲學類的各行。第四個運作找到出版年份小於一九六〇的各行。最後一個運作，投影X表到「書名」這一列上面。因此，在五個運作之後，Y表包括下述的資料。

Y表

書　　名
中　國　通　史
中　國　地　理

上述兩本書是文史哲學類，在一九六〇年前出版的書籍。

四、自動資訊查尋

以上說明了如何使用投影運作、限制運作和結合運作三個基本運作，從關係資料基裏查尋資訊。一般的查尋詢問，幾乎都可以用這三種運作來找尋答案。還有一些更複雜的運作，不在此細述。這些運作，根據柯德氏的理論，可以用來回答「**一度繫詞邏輯**」(First-Order Predicate Calculus) 表所有的邏輯命題。因此，它們構成一個完整的資訊查尋語言 (a complete information retrieval language)。柯德氏稱這些運作構成的語言爲**關係代數** (Relational Algebra)。使用關係代數，一般的資訊查尋詢問，都可以妥善的處理了。

現在的電化資訊系統，利用上述的資料基的新觀念和自動資訊查尋的新方法，可以解答使用者的資訊查尋詢問，滿足使用者找尋資料的各種需求。它的作業程序，大致如下。

㈠使用者利用**電傳打字機** (teletypewriter)、讀卡機 (card reader) 或其他型式的計算機終端機 (computer terminal)，向電化資訊系統提出查尋詢問，例如：

「找尋所有王五在一九七三年出版的著作」

「找尋地理類所有在一九六〇年到一九七〇年間在香港出版的著作」

「找尋中華書局出版的哲學類的著作」等等。使用者亦可提出一些**鍵語** (keyword)，要求資訊系統做一**鍵語查尋** (keyword in context search)，找到某一類的資料。

㈡資訊系統將上述的查尋詢問，翻譯成爲查尋指令，到資料基內找尋所需資料，經過處理後，再通過電傳打字機，報表印出機 (line printer) 或其他型式的計算機終端機，將查獲的資料提供給使用者。

上述的關係資料基的新觀念，幫助解決了自動資訊查尋上面許多問題。使用者可以將查尋詢問，改寫成一串簡單的查尋運作，資訊系統便可以據此找到所需的資料。因爲這些運作構成的「關係代數」，是一嚴謹的邏輯體系，許多複雜的邏輯命題，都可以清楚而毫不含混的表達出來。

目前資訊科學家研究的一個課題，是如何能利用自然的語言來表達查尋詢問，並自動將自然語的查尋詢問，翻譯成精確的查尋指令。如果這一方面能有更大的進展，資訊系統的應用，就會更加廣泛而普遍了

〔附　註〕

1. Engles, R. W."A Tutorial on Date Base Organization", *Annual Review in Automatic Programming*", Vol. 7, Part 1. p. 111. (1972)
2. 在本文中，一個「字」指由八個數元（bit）構成的譯碼單位。一個數元是0或是1。因此，一個「字」可以代表二的八次方，或是二百五十六種不同譯碼。英文字母加上數字、標點符號、特殊運算符號，可以用一個「字」表示。中文字則至少需要兩個「字」來表示。例如電報號碼1791，可以譯成000 00110 11111111的譯碼。
3. 有關「關係資料基」設計的原理，請參看 Date, C. J. "*An Introduction to Database Systems*" Reading, Mass.: Addison Wesley, 1975.
4. 有關「檔案設計」的原理，請參看張系國著「檔案設計原理」，一九七三年中央研究院數學研究所出版。

A New Approach to Automated Information Retrieval

Dr. Shi-Kuo Chang

Department of Information Engineering, University of Illinois at Chicago Circle, Chicago, Illinois, U. S. A.

ABSTRACT: The concept of a relational database as a collection of tables is introduced. A new approach to automated information retrieval based upon the relational algebra is then described. Examples on information retrieval using relational operations, such as projection, restriction, and join, illustrate how retrieval queries can be formulated in relational algebra and applied to the relational database.