

南 華 大 學

資訊管理學系碩士論文

運用資料探勘技術實作數位圖書館上個人化之檢索與推  
薦服務-以南華大學圖書館為例

Apply Data Mining Techniques to Implement Personalized Retrieval and  
Recommendation on Digital Library - the Library of NHU for example



研 究 生：鄭玉玲

指導教授：邱宏彬 博士

中 華 民 國 九 十 二 年 六 月

運用資料探勘技術實作數位圖書館上個人化之檢索與推薦服務  
-以南華大學圖書館為例

Apply Data Mining Techniques to Implement Personalized Retrieval and  
Recommendation on Digital Library - the Library of NHU for example

研 究 生 : 鄭 玉 玲

Student : Yu-Ling Cheng

指 導 教 授 : 邱 宏 彬

Advisor : Hung-Pin Chiu

南 華 大 學

資 訊 管 理 學 系 ( 所 )

碩 士 論 文

A Thesis

Submitted to Department of Information Management  
College of Management

Nan-Hua University

in partial Fulfillment of the Requirements

for the Degree of

Master of Business Administration

in

Information Management

June 2003

Chaiyi Taiwan, Republic of China.

中 華 民 國 九 十 二 年 六 月

# 博碩士論文授權書

本授權書所授權之論文為本人在南華大學(學院)資訊管理系所  
九十一學年度第二學期取得碩士學位之論文。

論文名稱：運用資料探勘技術實作數位圖書館上個人化之檢索與推薦服務

-以南華大學圖書館為例

同意  不同意 (政府機關重製上網)

本人具有著作財產權之論文全文資料，授予行政院國家科學委員會科學技術資料中心、國家圖書館及本人畢業學校圖書館，得不限地域、時間與次數以微縮、光碟或數位化等各種方式重製後散布發行或上載網路。

本論文為本人向經濟部智慧財產局申請專利(未申請者本條款請不予理會)的附件之一，申請文號為：\_\_\_\_\_，註明文號者請將全文資料延後半年再公開。

同意  不同意 (圖書館影印)

本人具有著作財產權之論文全文資料，授予教育部指定送繳之圖書館及本人畢業學校圖書館，為學術研究之目的以各種方法重製，或為上述目的再授權他人以各種方法重製，不限地域與時間，惟每人以一份為限。

上述授權內容均無須訂立讓與及授權契約書。依本授權之發行權為非專屬性發行權利。依本授權所為之收錄、重製、發行及學術研發利用均為無償。上述同意與不同意之欄位若未鉤選，本人同意視同授權。

指導教授姓名：邱震彬

研究生簽名：鄭玄玲  
(親筆正楷)

學號：90112513  
(務必填寫)

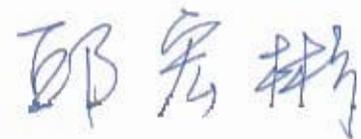
日期：民國 92年 6月 30日

1. 本授權書 (得自 <http://nr.stic.gov.tw/theses/html/authorize.html> 下載) 請以黑筆撰寫並影印裝訂於書名頁之次頁。
  2. 授權第一項者，請確認學校是否代收，若無者，請個別再寄論文一本至台北市(106-36)和平東路二段 106 號 1702 室 國科會科學技術資料中心 王淑貞。(本授權書諮詢電話:02-27377746)
  3. 本授權書於民國 85 年 4 月 10 日送請內政部著作權委員會(現為經濟部智慧財產局)修正定稿，89.11.21 部份修正。
- 本案依據教育部國家圖書館 85.4.19 台(85)圖編字第 712 號函辦理。

南華大學碩士班研究生  
論文指導教授推薦函

資訊管理系碩士班鄭玉玲君所提之論文  
運用資料探勘技術實作數位圖書館上個人化之  
檢索與推薦服務-以南華大學圖書館為例  
係由本人指導撰述，同意提付審查。

指導教授



九十二年六月二日

# 南 華 大 學

## 碩 士 學 位 論 文

### 資 訊 管 理 學 系

運用資料探勘技術實作數位圖書館上個人化之檢索與推薦服務

-以南華大學圖書館為例

研 究 生：鄭 玉 玲

經考試合格特此證明

口試委員：

吳光閔

賴智錦

\_\_\_\_\_

\_\_\_\_\_

指導教授：

邱宏彬

所 長：

資訊管理學系  
系主任  
吳光閔

口試日期：中華民國

92 年 6 月 14 日

## 誌 謝

未進入研究所之前，對研究所是一種憧憬，所以在此論文完成之際，便覺得自己是幸運地。猶記得剛踏入校門時，深深為這個清新的校園所吸引，而今已到了離別時刻。感謝兩年來教導過我的資管所老師，您們開闊我的視野，豐富我的領域知識，培養我解決問題的能力。

感謝我的指導老師邱宏彬博士，從論文方向的确立、題目擬定、研究架構的擬定、文句修擅，到論文的完成，多蒙邱老師不辭辛勞的撥冗指正與引導，才能使本論文能夠完成。老師的耐心和包容，使我由懵懂中成長，為人處事更是我學習的地方。在論文提案及口試期間，感謝吳光閔博士及賴智錦博士所提出的建議與指教，使本論文更臻完善。

很幸運能與慧君 惠玲還有家扶的所有同工；Andy, Snop, Samper, Sely, Grace 在台林的相互照顧；信旭、佳昌、胤良、清涼．．相遇而沒有失之交臂，謝謝你們在課業、生活的關心指導及鼓勵，使我的研究所生活更加充實，你們是我難得的朋友。

走過了這麼一段工作和唸書的日子，深深感覺到兩邊要取其平衡的壓力，尤其是工作夥伴的協助與體諒，謝謝你們大夥兒。

最後，感謝母親為家庭無私的奉獻，也感謝家瑜和家瑜爸媽和我親愛的家人，玉玫、玉英、俊傑、阿志、最可愛的昕昀，在我經歷挫折時不斷給我打氣。有你們當我的精神支柱，真好！

碩士論文的完成，要感謝的人實在太多了，在此無法一一詳列，謹以此論文獻給我所愛的人，以及我最敬重的邱宏彬老師。

鄭玉玲 謹誌於 嘉義

南華大學資管所

中華民國九十二年六月

# 運用資料探勘技術實作數位圖書館上個人化之檢索與推薦服務 - 以南華大學圖書館為例

研究生：鄭玉玲

指導教授：邱宏彬 博士

南華大學資訊管理學系碩士班

## 摘 要

拜無遠弗屆網際網路與電腦硬體的快速發展所賜，現行圖書館系統皆運用電子化的書籍資料庫、借閱紀錄資料庫，以及運用兩者間關聯性來加強圖書館之各樣服務。在此，本研究將提出一「數位圖書館上個人化之檢索與推薦系統」，以中國圖書分類法的明確分類優勢協助讀者找出確實符合的資料。另一方面，運用資料探勘的技術，從書目、交易歷史、及讀者身分屬性資料庫中，找出身分背景相同的分群，並找出同類群中最感興趣的書目，作為圖書館推薦服務的參考。結果顯示，事先運算同類身分背景的分群，讓讀者進入系統後，直接將已分類完成且關聯度高的書籍推薦之，將增加個人化的服務速度。

關鍵字：資料探勘、個人化資訊環境、自動化推薦系統、關聯法則

Apply Data Mining Techniques to Implement Personalized Retrieval and  
Recommendation on Digital Library - the Library of NHU for example

Student : Yu-Ling Cheng

Advisor : Dr. Hung-Pin Chiu

Department of Information Management

The M.B.A Program

Nan-Hua University

**ABSTRACT**

Due to the rapid advance of computer-related technology, the large quantities of electronic data produced everyday has been explored and analyzed to discover meaningful information so as to enhance the service quality of digital library. The objective of this study is to enable personalized retrieval and recommendation services on digital library. We utilize the well-defined “New Classification Scheme For Chinese Libraries” to support audience to retrieve the books that they really want. Besides, memory-based reasoning was applied to assign the unlabeled user to the cluster of its nearest labeled neighbors based on some predefined measures of users’ characteristics. Association rules discovered from the books borrowed by the readers in the same cluster are used as the basis of book recommendation. A simple clustering algorithm was exploited to speed up the processing time of recommendation. The experimental results show that the proposed approaches are effective in promoting the searching efficiency and accuracy of the system.

**Keyword** : Data mining、Personal Information Environment、  
Automatic Recommender System、 Association rule

# 目 錄

書名頁 .....	I
授權書 .....	II
論文指導教授推薦書 .....	III
論文口試合格證明 .....	IV
誌謝 .....	V
中文提要 .....	VI
英文提要 .....	VII
目錄 .....	VIII
表目錄 .....	IX
圖目錄 .....	X
第一章導論 .....	1
第一節研究背景 .....	1
第二節研究動機 .....	1
第三節研究目標 .....	2
第四節論文架構 .....	2
第二章相關研究工作 .....	4
第一節個人化服務的意義與發展 .....	4
第二節中文圖書分類法 .....	5
第三節資料探勘意義與發展 .....	7
第四節關聯法則 .....	11
第五節記憶基礎理解 .....	14
第六節資料探勘模組建立互動流程 .....	20
第三章系統架構 .....	24
第一節圖書推薦檢索系統整體架構 .....	24
第二節系統模組 .....	26
第三節資料探勘模組建立互動流程 .....	35
第四章系統實作與討論 .....	39
第一節系統建置 .....	39
第二節系統展示 .....	51
第三節模組優點 .....	59
第五章結論與未來研究方向 .....	61
參考文獻 .....	63
附錄一 .....	65
附錄二 .....	66
附錄三 .....	67

# 表 目 錄

表 1	中國圖書分類法細目說明表.....	6
表 2	中國圖書分類法簡表.....	7
表 3	MBR 同類群組合函數例子.....	19
表 4	MBR 組合函數判斷是否推薦.....	19
表 5	種子樣本點和其他樣本點的距離矩陣表.....	31
表 6	編號 1 子分群借閱書目.....	31
表 7	編號 1 子分群關聯法則推薦書籍.....	31
表 8	經由 Improved MBR 分類後大分群.....	33
表 9	探勘資料表(lib_mining).....	37
表 10	書籍資料表 schema( lib_book ).....	39
表 11	中國圖書分類索引之 schema( lib_index ) .....	39
表 12	借閱紀錄表之 schema( lib_loanred ).....	42
表 13	身分屬性資料表.....	42
表 14	探勘資料表.....	42
表 15	屬性轉換數值分布表(變數：身分).....	44
表 16	屬性轉換數值分布表(變數：單位).....	44
表 17	單位距離矩陣表.....	45
表 18	身分距離矩陣表.....	45
表 19	探勘資料表之簡例.....	46
表 20	子群矩陣表.....	48
表 21	同一子群借閱書籍表.....	48
表 22	同一子群推薦資料表.....	48
表 23	系統合併子分群資料表.....	50

# 圖 目 錄

圖 1	資料庫知識發掘流程圖.....	9
圖 2	Apriori 之處理程序.....	13
圖 3	Apriori 之演算法.....	13
圖 4	apriori-gen()產生 $C_k$ 的兩個步驟.....	14
圖 5	類別變數 - 性別的距離函數例子.....	16
圖 6	類別變數 - 郵遞區號的距離函數例子.....	16
圖 7	類別變數 - 郵遞區號的改良式距離函數例子.....	17
圖 8	鄰近資料演算法(Simple nearest neighbor algorithm).....	18
圖 9	鄰近資料演算法距離公式.....	18
圖 10	資料探勘模組互動流程圖.....	20
圖 11	圖書館推薦檢索系統架構圖.....	24
圖 12	主題式同類書目推薦模組.....	25
圖 13	互助式資訊過濾推薦模組圖.....	26
圖 14	中國圖書分類法示意圖與勾選例子.....	28
圖 15	單一種子樣本點分群示意圖.....	30
圖 16	多個種子樣本點分群示意圖.....	33
圖 17	SCSA Algorithm.....	34
圖 18	系統需求資料表欄位說明.....	35
圖 19	南華大學原始資料表.....	35
圖 20	身分距離函數.....	44
圖 21	單位距離函數.....	45
圖 22	分群合併程式.....	49
圖 23	系統畫面一.....	52
圖 24	系統畫面二.....	52
圖 25	系統畫面三.....	53
圖 26	系統畫面四.....	54
圖 27	系統畫面五.....	54
圖 28	系統畫面六.....	55
圖 29	系統準確性評估例證 (以單位讀者為例).....	56
圖 30	隨機抽取讀者之訓練和測試資料重疊程度圖一.....	57
圖 31	隨機抽取讀者之訓練和測試資料重疊程度圖二.....	58
圖 32	改良式分群探勘法和直接關聯法則探勘法比較圖.....	59

# 第一章、導論

## 第一節 研究背景

網際網路蓬勃發展，近年來圖書館也漸漸朝向數位化的腳步邁進，在資訊提供者主義(Provider-ism, PRVM)和使用者的主義(User-ism, USRM)的兩種觀念之下，將使用者主義的觀念落實在數位圖書館的做法是化被動為主動的貼心服務[6]。此外，拜網路及硬體快速發展所賜，多數圖書館運用電子化的書籍資料庫、借閱者資料庫、借閱記錄檔來加強圖書館之各項服務，本研究利用這些現有資料庫來協助讀者快速且準確地找到欲借閱的資料。

## 第二節 研究動機

數位圖書館的建置是利用資訊數位化以及網際網路的技術來達成傳統圖書館服務品質的升級。但是在此資料繁多且容易取得的環境下，資訊氾濫的結果往往會令使用者無法隨心所欲地尋找需要的資訊，因此，探究如何讓使用者能正確且快速地取得所需資料，並發展一個以人為中心之智慧型使用者介面是當今數位圖書館發展中的重要議題[20]。因此，本論文將個人化資訊服務導入數位圖書館服務之中，以圖書館現有館藏資源為基礎，針對讀者的借閱模式及特性，採因人而異的介面，提供更多元化的資訊增值服務以提高資訊服務品質及使用者滿意度。本論文希望藉由資料探勘的技術與應用，架構一「數位圖書館上個人化之檢索與推薦系統」，運用現有資料庫中的歷史紀錄檔，藉以了解讀者的閱讀習性

及興趣，來建構個人化圖書檢索介面，協助讀者找到符合其自身需求的書籍與資料。

### 第三節 研究目標

由於現今資訊的快速成長，各種資料量變得龐大且分散，使用者利用傳統關鍵字檢索搜尋資料的方式，變得相當費時且不容易聚焦。因此，為了要降低使用者在網路上的搜尋時間，提供使用者更適切的内容資訊，自動化推薦系統(Automatic Recommender System)便因應而生。

自動化推薦系統有許多種不同的推薦方式，如 Content-based，Filtering-based 等方式[5]，而本研究主要針對實作一圖書館線上推薦系統及探討如何有效地運用資料探勘的技術從大量的資料庫中挖掘出符合需求的知識，以推薦適當的資訊給使用者。

本研究結合了資料探勘理論中的關聯法則(Association Rule)法及記憶基礎理解(Memory Based Reasoning，MBR)技術來建構一個事先聚類讀者特徵且能挖掘出借閱關聯特質的個人化之檢索與推薦系統。運用方法的特色在於將特徵事先聚類、且能自動定期更新並和關聯法則做結合。

本研究所欲達到的目標如下：

- 一、提供適用於圖書館之個人化線上推薦檢索模組。
- 二、事先聚類讀者特徵、且能自動定期更新讀者資料。
- 三、實作一圖書館個人化線上推薦與檢索系統。
- 四、提供更符合個人化要求的線上推薦結果。

### 第四節 論文架構

本研究分析南華大學圖書館現行館藏查詢系統，並探討目前在資訊

搜尋所遇到的問題，且針對推薦系統的讀者分群特性提出一個加速分類檢索方法，最後則實際應用資料探勘技術於處理圖書推薦的程序上，以此驗證本研究的可行性及應用性。

第一章說明本論文之研究背景、研究動機、研究目標；第二章是敘述個人化發展、中國圖書分類法、資料探勘觀念，包含關聯法則及 MBR、資料探勘模組建立互動流程等相關文獻之探討；第三章提出系統架構，分別就圖書推薦檢索系統整體架構、系統模組、資料探勘模組建立互動流程說明；第四章是說明系統實作與討論系統建置、系統展示、模組評估與優點；第五章歸納結論與未來研究方向。

## 第二章、相關研究工作

### 第一節 個人化服務的意義與發展

現今個人化資訊環境(Personal Information Environment, PIE)的研究主要著重於提供可程式化(Programmable)的個人使用介面，並且利用資訊過濾的相關技術提供使用者有興趣的相關資訊。其中的主要幾個探討主題包含個人化的使用者介面(Customizable User Interface)、資訊過濾技術(Information Filtering)、個人化瀏覽(Personalized Browsing)、非文字資訊檢索(Non-textual Information Retrieval)等方面[4]。

#### 壹、個人化的使用介面

個人化的使用介面，是讓使用者在進入資訊系統後能根據自己的喜好來設定使用環境。如現有的奇摩網站，能夠依照個人喜好自訂背景。

#### 貳、資訊過濾技術

資訊過濾技術一般可以區分為內容式資訊過濾 (Content-based Information Filtering, 或稱為 Feature-based Information Filtering)以及互助式資訊過濾(Collaborative Information Filtering)兩類[1]。內容式資訊過濾通常利用關鍵字詞比對方式，對文件內容(Content)加以分析比對，著名的商用系統如 Verity、CompassWare、IBM InfoSage 及 InfoSeek Personalized News 等皆是採用此類技術。至於互助式資訊過濾技術，則不直接分析文件內容，而改以分析使用者的個人背景、知識、興趣等，先找出興趣接近的同好，之後針對使用者的查詢主

題，從這些同好所感興趣的資訊中，再分析選取最可能相關的資訊提供參考。相關服務包括資訊選粹服務(Selective Dissemination of Information, SDI)、個人化推薦(Personalized Recommendation)以及個人化檢索(Personal Search)。本論文所著重的主題在於個人化推薦及個人化檢索兩項服務。

### 參、個人化瀏覽

對一般使用者而言，瀏覽是最容易親近的資訊服務。將圖書分類目錄加以簡化，允許使用者依不同的興趣及背景，提供個人化的分類目錄瀏覽，節省使用者時間[11][21]。

本研究提出一「個人化圖書資訊環境」的模組架構，達到下列目的：

- 一、提供使用者易用、親和力高的介面(User-friendly)。
- 二、偵測使用者的興趣、幫助使用者適性地找尋資料，以切合使用的的需求，提高搜尋結果。
- 三、依個人欲查詢的關鍵字為主，將圖書分類目錄加以簡化，提供使用者更簡潔明瞭的個人化的分類目錄瀏覽，節省搜尋時間。

## 第二節 中文圖書分類法

「中國圖書分類法」是以學術性質與知識體系為基本架構，並輔以體裁、地域、時代、語文、用途、版本等作為分類的方式，分成總類、哲學、宗教、自然科學、應用科學、社會科學、中國史地、世界史地、文學、美術十大類[9]。

這個分類法的優點在於使用年代久遠，從民國初年使用迄今，廣泛

應用，且分類主題詳盡、明瞭，讓人容易一眼就明白，並隨著時代不同而持續更新，不會流於過時的窠臼之中。

國內現在約有百分之八十以上的圖書館在使用這種方法，南華大學圖書館亦是採用中國圖書分類法為中文書目分類原則，本文就地取材，依現有資源的分類規劃，加強圖書館個人化服務。

中國圖書分類法的分類方法，是採階層式分布，依種類分為十大類，每一大類以一個百位數整數代表，如 400 代表「應用科學類」，其下一個十進位代表更細的分類如 410 代表醫藥、420 代表家事、...，490 代表商學等十個應用科學的分類；以 410 來說，其下一個十進位又以醫學為主再細分為十類，411 代表衛生學、412 代表公共衛生，以此類推下去，階層式的表示法。在圖書館的編制當中，每一個中國圖書分類編號都會有一些書籍隸屬之。表 1 是說明圖書分類方法的例子，412 代表「公共衛生類」，則 412.4 代表「公共衛生類」中的「防疫類」，412.1 412.2 412.3 ...，412.9 都等子類別都是歸屬於「公共衛生類」之下。表 1 與表 2 將對中國圖書分類法細目及簡表加以說明。

表 1 中國圖書分類法細目說明表

412	為「公共衛生」
412.4	為「公共衛生」中的「防疫」
412.48	為「公共衛生」中的「防疫」中的「其它傳染病預法」
412.481	為「公共衛生」中的「防疫」中的「其它傳染病預法」中的「隔離」
412.483	為「公共衛生」中的「防疫」中的「其它傳染病預法」中的「消毒」

表 2 中國圖書分類法簡表

總類	宗教類	應用科學類	史地類中國	語文類
000 特藏	200 總論	400 總論	600 史地總論	800 語言文存學
010 目錄學	210 比較宗教學	410 醫藥	610 通史	810 文學
020 圖書館學	220 佛教	420 家事	620 斷代史	820 中國文學
030 國學	230 道教	430 農業	630 文化史	830 總集
040 百科全書	240 基督教	440 工程	640 外交史	840 別集
050 普通雜誌	250 回教	450 礦冶	650 史料	850 特種文學
060 普通社會	260 猶太教	460 應用化學	660 地理	860 東洋文學
070 普通論叢	270 其他各教	470 製造	670 方志	870 西洋文學
080 普通叢書	280 神話	480 商業;各種營業	680 類志	880 西方諸國文學
090 群經	290 術數;迷信	490 商學;經營學	690 遊記	890 新聞學
哲學類	自然科學類	社會科學類	史地類世界	美術類
100 總論	300 總論	500 總論		900 總論
110 思想	310 數學	510 統計	710 世界史地	910 音樂
120 中國哲學	320 天文	520 教育	720 海洋	920 建築
130 東方哲學	330 物理	530 禮俗	730 東洋;亞洲	930 雕塑
140 西洋哲學	340 化學	540 社會	740 西洋;歐洲	940 書畫
150 論理學	350 地質	550 經濟	750 美洲	950 攝影
160 形而上學;玄學	360 生物;博物	560 財政	760 非洲	960 圖案;裝飾
170 心理學	370 植物	570 政治	770 澳洲及其他	970 技藝
180 美學	380 動物	580 法律	780 傳記	980 戲劇
190 倫理學	390 人類學	590 軍事	790 古物;考古	990 遊戲;娛樂;休閒

### 第三節 資料探勘意義與發展

根據 Frawley, Piatetsky-Shapiro 及 Matheus 的定義，資料探勘即是從資料庫中，發掘出隱含、未知的、且可能有用的資訊[23]。

而 Cabena 等人更進一步區分資料探勘與資料庫知識發掘 (Knowledge Discovery in Database; KDD) 的定義 [8][10]。整個過程包括對應用領域的認知、具備相關的專業知識，確認分析目標的資料來源，接著進行資料淨化 (Data Cleaning)、資料整合 (Data integration)、資料選擇 (Data Selection)、資料轉換 (Data transformation)、資料探勘 (Data mining)、型樣評估 (Pattern evaluation)，以及知識呈現 (Knowledge representation) 等七個步驟 [17]。

- 一、資料淨化：實際應用的資料庫，通常會有不完整，或是有雜訊，及不一致的情形產生。資料淨化的工作主要是要將不完整的資料 (missing value) 填上、去除資料中的離群值，以及修正資料的不一致性等等。
- 二、資料整合：將與目標相關的各種資料來源加以整合，例如將不同類型的資料庫整合成資料倉儲。
- 三、資料選擇：自資料庫或資料倉儲中取得與分析目標相關的資料。
- 四、資料轉換：執行加總 (summary) 及整合 (aggregation) 的運算，將資料轉換成適合探勘的形式。
- 五、資料探勘：是一種不斷循環的決策分析過程，運用智慧型的方法，從資料中粹取出型樣 (Patterns)、關係 (Relations) 等等。
- 六、型樣評估：資料探勘所產生的型樣並非完全都是使用者所期待的資訊，因此必須利用一個較客觀的評估方式，確認出真正有意義的知識。
- 七、知識呈現：使用視覺化的技術或是知識表達的方法，將知識傳達給使用者。

# Data Mining: A KDD Process

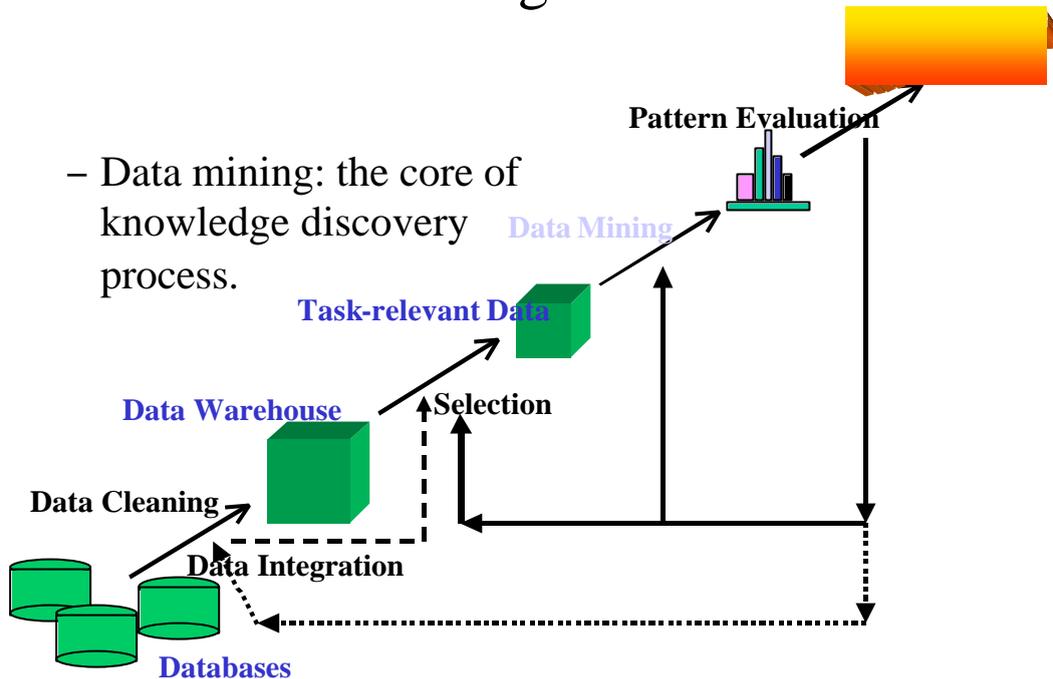


圖1：資料庫知識發掘流程圖[8]

經過多年的發展，已有許多資料探勘的技術及方法被提出，分別針對不同的應用領域或需求而有不同的演算法。

資料探勘可發揮之功能包括關聯分組(Affinity Grouping)、資料分類(Classification)、資料分群分析(Data Clustering Analysis)、推理(Estimation)、預測(Prediction)。依序說明如下：

## 一、關聯分組：

又稱購物籃分析。針對一個銷售的交易資料庫，我們感興趣的是所有商品項目之間的關聯，利用項目的相關性來決定其分類。通常會以關聯法則來表示，例如「購買麵包的顧客，則同時也會購買牛奶，即麵包 牛奶」。運用這樣的方法可以用來分析顧客的消費行為，找出相關的商品集合間彼此的關聯性，以改善貨物架的擺設，確認交叉銷售(Cross Selling)的機會，並

設計出吸引人的產品組合[19]。

## 二、資料分類：

資料分類是在資料庫的物件集合中，按照分析對象的屬性分門別類加以定義，找尋共同性質，並建立類組(Class)。例如將客戶支付訂單的償債能力分為高、中及低三種等級。使用的技巧有決策樹(Decision Tree)、記憶基礎推理(Memory-Based Reasoning)等。

## 三、資料分群分析：

資料分群的意思就是根據資料分群的原則，把資料庫中的資料叢聚成群。所謂資料分群的原則，就是「讓群組內的資料相似度最高，讓群組跟群組間的資料相似度最低」，將異質母體中區隔為較具同質性之群組(Clusters)。例如股票分析法。所使用的工具包括K-means法及Agglomeration法等。

## 四、推理：

與資料分類的功能相當，在資料庫的物件集合中，按照分析對象的屬性分門別類加以定義，找尋共同性質並建立類別(Class)的過程。唯一不同的是資料分類的方法，大多建立離散資料(Discrete)的類別，而推理的方法大多建立連續型態(Continuously Valued)的類別，例如按照客戶的資本額來推估其訂貨量。所使用的工具包括相關分析、迴歸分析及類神經網路等。

## 五、預測：

根據對象屬性之過去觀察值來推估該屬性未來之值，例如由顧客過去之刷卡消費量預測其未來之刷卡消費量、由客戶過

去的訂購量推估未來的訂購量。所使用的工具包括迴歸分析、時間數列分析及類神經網路等。

#### 第四節 關聯法則

在資料挖掘的技術中，關聯法則(Association Rule)是最常被使用之方法[18]。關聯法則指出在資料庫中某些資料間彼此的關聯性，此法被廣泛應用於各領域。

在商業上的應用，如超市購物車分析(Market Basket Analysis)，例如在購買烤肉炭的客戶中，75%也會購買打火機燃料，由此可以幫助超級市場或百貨店規畫如何擺設貨品。另外如信用卡公司、保險公司以及政府單位對於詐欺行為的偵測(Fraud Detection)，這些行業每年因為詐欺行為而造成的損失都非常可觀。資料挖掘中的關聯法則可以找出可能的詐欺交易，減少損失。在學術上的應用，如網路教學，可藉由分析學生的學習情形，改善相關課程的設計提升教學品質。

關聯式法則的表示形式為： $X \rightarrow Y$ ；其中  $X$  及  $Y$  為資料庫中的任意物項集合，且  $X \cap Y = \emptyset$ 。關聯式法則的產生可以藉由兩個參數來決定：支持度(Support)及可靠度(Confidence)；支持度定義物項在資料庫中所出現的比例，而可靠度定義此關聯式法則可信的程度。一般而言，關聯式法則的支持度及可靠度皆必須分別大於使用者訂定的最小限制，才能判定其為有意義的關聯式法則。

找出關聯法則有以下二個步驟：

- 一、首先必須找出高頻項目集合(Large Itemsets)，即此集合之交易(Transaction)要能夠滿足其支持度(Support)大於最小支持度(Minimum Support)的條件。

二、再使用第一步驟所產生的 Large Itemsets 來產生 Association Rule。

截至目前為止已有很多演算法可以找出 Large Itemsets，例如著名的 Apriori、DHP 等[18]，最為人所知且最具代表性的 Apriori 演算法之處理程序如圖 2，其處理步驟說明如下：

- 一、定義最小支持度及最小信賴度。
- 二、Apriori 演算法使用了候選項目集合(Candidate Itemset)的觀念，若候選項目集合的支持度大於或等於最小支持度(Minimum Support)，則該候選項目集合為高頻項目集合(Large Itemset)。
- 三、首先由資料庫讀入所有的交易，得出候選 1 項目集合(Candidate 1-Itemset)的支持度，再找出高頻單項目集合(Large 1-Itemset)，並利用這些高頻單項目集合的結合，產生候選 2 項目集合(Candidate 2-itemset)。
- 四、再掃描資料庫，得出候選 2 項目集合的支持度以後，再找出高頻 2 項目集合，並利用這些高頻 2 項目集合的結合，產生候選 3 項目集合。
- 五、重覆掃描資料庫，與最小支持度比較，產生高頻項目集合，再結合產生下一級候選項目集合，直到不再結合產生出新的候選項目集合為止。

以下我們利用簡單的例子，分別來看 Apriori 演算法的處理過程，若資料庫中有四筆交易，每筆交易都具有不同的 TID 作代表，而交易中都包含了有數種物品，如下所示：

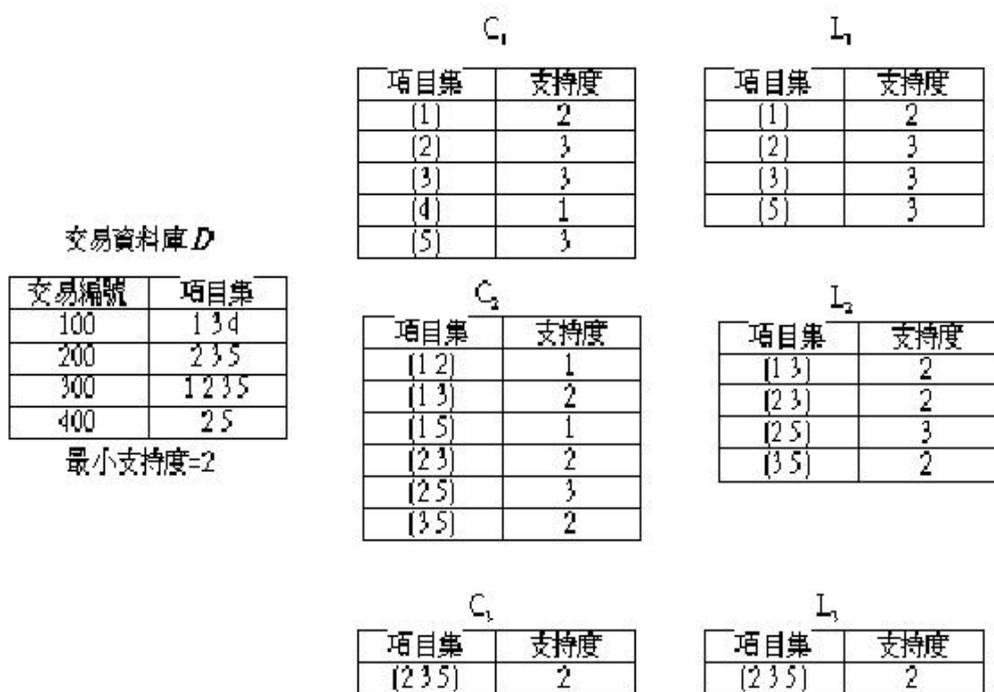


圖 2：Apriori 之處理程序

```

L1={large1-itemsets};
for(k=2;Lk-1 ≠ ∅;k++) do begin
  Ck=apriori-gen(Lk-1);
  for all transactions t∈D do begin
    Ct = subset(Ck,t);
    for all candidates c∈Ct do
      c.count++;
  end

  Lk = {c∈Ck |c.count ≥ minsup};
end
Answer = ∪k(Lk);

```

圖3：Apriori之演算法

在圖 3 的 Apriori 演算法中，apriori\_gen() 函數是用來產生候選集。由於任何大項目集的子集合也都是大項目集，因此 apriori\_gen() 以所有大 (k-1)-項目集為參數，傳回大 k-項目集的候選集(C<sub>k</sub>)。apriori\_gen() 分為兩個步驟來產生 C<sub>k</sub>：

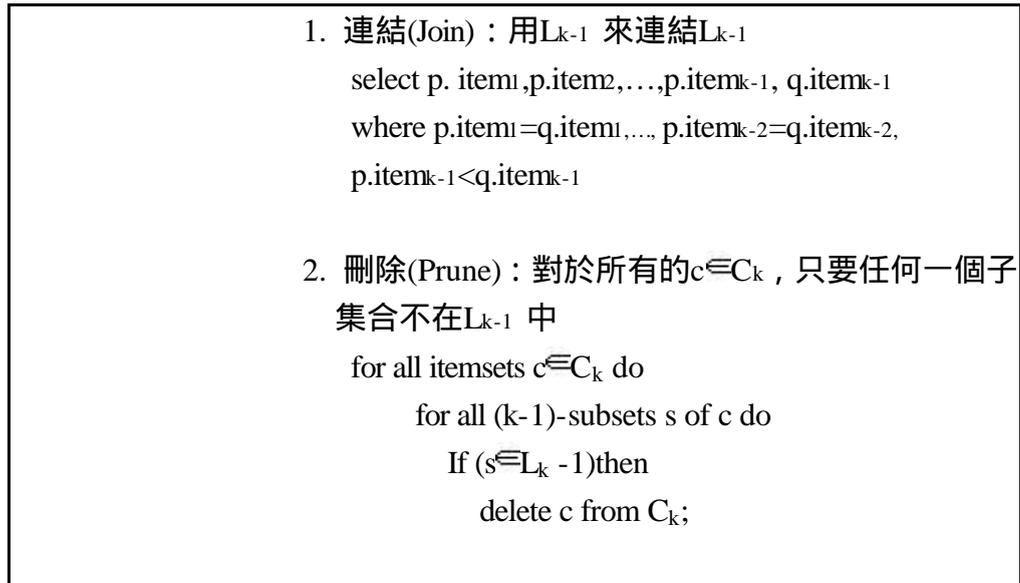


圖 4：apriori-gen()產生  $C_k$  的兩個步驟

而 subset()函數用來找出一個交易紀錄中所有包含於  $C_k$  的子集合。如前所述，由於 Apriori 在每一個程序中使用  $L_{k-1}$  的連結來形成 Candidate large itemsets  $C_k$  (以  $L_{k-1} \times L_{k-1}$  表示)。然後掃描資料庫且計算  $C_k$  中每一個 item 的 support 數量，來決定  $L_k$ 。因此  $C_k$  的數量越多，要產生  $L_k$  的成本也就越高。

## 第五節 記憶基礎理解 (Memory-based reasoning , MBR)

人們總是憑藉著過去的經驗下判斷：當身處人群，我們有時會突然看見熟悉的面孔，事實上我們是將每一張臉和之前的所看過的所有臉作對照。這也就是記憶基礎理解(MBR)[15][22]的原理，只不過是用人工智慧的方式來進行比對，依據現有的資料庫，「MBR」先找出新資料的「鄰近資料(Neighbor)」，然後根據鄰近資料，對新資料進行分類和預測。MBR 的優點在於它不在乎資料的格式、複雜的資料型態也有效，如影像、純文字、地理區域、可產生新的分群。只要設定好「距離函數」(Distance Function)和「組合函數」(Combination Function)兩項機制，即可將資料作

分類和預測。它的缺點是會佔用大量硬體資源，主因是因為它需要儲存大量的歷史資料以方便比對[14]。

MBR 的主要實作精神包含以下幾步驟：選擇歷史資料、轉換歷史資料、設定距離函數、組合函數、和鄰近資料的數目、最後是分類、預測並評估。

#### 一、選擇歷史資料：

在做分類預測之前需要有分類正確且分布平均的歷史資料來作為後續新進資料的分類基礎。選擇歷史資料又稱為選擇訓練資料組(Training Set)，是現有資料的子群組，而訓練組資料的範圍必須能夠涵蓋原始資料，並且在各個分類項中的樣本數須接近，如此未知紀錄的鄰近資料才能作為良好的預測基礎。

#### 二、轉換歷史資料：

將良好的訓練資料組轉換到電腦上，會決定 MBR 的表現。明確了解已知變數的數值範圍，並為每個數字做正規化，才不會造成數字範圍較大的變數完全主宰距離函數的情形。增強系統效能的方法是減少歷史資料庫的紀錄數量，但會造成記錄涵蓋不完整的情形發生。

#### 三、設定距離函數：

對於 MBR 來說，每一個點代表的是資料庫中的一筆資料，也就是一位具有身分、單位、借閱數量等特徵變數的借閱者，點與點之間的距離用來決定是否屬於同一分群，距離越近代表背景屬性越相近。

變數又分為類別變數和數值變數兩種，類別變數如性別等變數之間並無距離上的意義；此時必須依靠資料值及探勘需求

訂定距離函數，最直接的類別距離函數就是 1 代表不同，0 代表相同。而數值變數則是在資料值之間有距離上的關係，通常是直覺式的將資料經過正規化後即定出距離函數；如年齡、薪水...等。而類別變數的距離函數定義通常是依資料探勘的實際需求訂定，分別舉性別和郵遞區號二個例子說明：

(一)性別的距離變數表示：

$d_{\text{性別}}(\text{女性}, \text{女性})=0$
$d_{\text{性別}}(\text{男性}, \text{女性})=1$
$d_{\text{性別}}(\text{女性}, \text{男性})=1$
$d_{\text{性別}}(\text{男性}, \text{男性})=0$

圖 5：類別變數 - 性別的距離函數例子

(二)五位數的郵遞區號通常用來當作簡單的數值形式，隨意抽出兩個郵遞區號計算標準距離函數的差值是不具意義的。因此，郵遞區號比較接近類別變數的型態。必須視資料特性來定義距離函數。例如，郵遞區號「100」前三位數字是代表曼哈頓區的。因此，處理的方式是去掉後面兩位數字，然後將前三位數當程式類別變數，所以「10014」、「10001」和「10016」都被當成「100」。而這三者之間的差異值是 0，但是和不同分類項的郵遞區號(如「101」和「905」的差異值是 1)，距離函數表示法如下：

$d_{\text{郵遞區號}}(A,B)=0$ ，當前三位數相等時
$d_{\text{郵遞區號}}(A,B)=1$ ，當前三位數不等時

圖 6：類別變數 - 郵遞區號的距離函數例子

設定特殊的距離函數最大的優點是不需改變原始資料，而就函數本身進行調整成適合的資料型態。如郵遞區號進一步改良成為：

$d_{\text{改良後}}(A,B)=0.0$ 如果郵遞區號相同
$d_{\text{改良後}}(A,B)=0.1$ 如果郵遞區號前三位數相同(「20008」和「20015」)
$d_{\text{改良後}}(A,B)=0.5$ 如果郵遞區號第一個數字相同(「95050」和「98125」)
$d_{\text{改良後}}(A,B)=1.0$ 如果郵遞區號第一個數字不同(「02138」和「90024」)

圖 7：類別變數 - 郵遞區號的改良式距離函數例子

這樣的改良可進一步從郵遞區號中獲得更多的資訊。其他如電話號碼和產品編號亦是如此，我們能透過特殊距離函數找出隱藏在他們之間的結構。建立這些結構化的距離矩陣能夠增進 MBR 的效能。此外，也可以採用建構距離矩陣的方式來建立距離函數。

距離函數具備四個關鍵特性：明確界定性(Well-defined)、區辨性(Identity)、可互換性(Commutativity)、三角不等式(Triangle Inequality)。在做 MBR 時因為要計算距離，通常都將變數轉換成數值來表示，數值變數有三種最常見的距離函數可供選擇：

1. 絕對值： $|A - B|$
2. 平方差： $(A - B)^2$
3. 標準化絕對值(Normalized Absolute Value)： $|A - B|$ (最大差值)，其優點在於數值都介於 0 - 1 之間，上述選擇權也因系統特性而異。

在圖 8 中，Marko 提到鄰近資料演算法[13](Simple

nearest neighbor algorithm), 分別計算兩個點間單一變數 (feature) 的值, 比較兩個值大小並計算符合程度, 再乘上權重係數 (weight coefficient) 來表示此變數的重要性, 將所有得到變數的權重值加總起來超過 threshold 的則加到所需要的結果之中, 參閱圖 8 的演算法。而圖 9 是將鄰近資料演算法用距離函數來呈現。

for each feature in the input case,  
 find the corresponding feature in the stored case  
 compare the two values to each other and compute the degree of match  
 multiply by a weight coefficient representing the importance of the feature to the match  
 add the results to derive match score

圖 8：鄰近資料演算法 (Simple nearest neighbor algorithm)

$$\frac{\sum_{i=1}^n W_i * sim(f_i^I, f_i^R)}{\sum_{i=1}^n W_i}$$

$W_i$  是變數(feature)的權重  
 $n$  表變數(feature)個數  
 $sim(f_i^I, f_i^R)$  代表原始資料的相似度函數  
 $f_i^I, f_i^R$  are the values for feature  $f_i$  in the input and retrieved cases

圖 9：鄰近資料演算法距離公式

#### 四、組合函數：

組合函數就是將來自不同鄰近資料的特性結合起來, 用以判定預測新進點某種特性的的動向。舉例說明, 同一群組中的借閱者, 有距離遠近之分, 越近的代表借閱者背景屬性越接近,

由不同個數的借閱者組成之組合函數代表某種書籍的借閱比例。最常用的方式是民主方式(democracy)，也就是讓最近似的 K 個鄰近資料「投票」決定選出答案。所採用的方法是先由決定鄰近資料的數目開始，而新資料的類別就是最多鄰近資料所屬的類別。為了避免同票的情形，當類別只有兩個的情形下，K 必須是奇數；一般來說，當類別數為 C 時，鄰近資料應該要有 C+1 個。其他如加權投票、或迴歸分析來依不同特性而做決定採用。舉例說明：表 3 中紀錄編號 1-5 表示經由距離函數運算過後由紀錄編號(新)的成員為中心點計算出和此中心點距離最近的 5 個點，表示為同一分群，並由同一分群中是否借閱「孟子」一書來投票決定是否要推薦給新的讀者，如表 4 所示。

表 3：MBR 同類群組合函數例子

同一分群紀錄			
紀錄編號	身分	單位	借閱「孟子」
1	專任教授	南華通識教學中心(O)	是
2	兼任教授	南華哲學研究所	是
3	大四	南華宗教學研究所	否
4	大二	南華社會學系所(O)	是
5	大二	南華哲學研究所	否
新	專任教授	南華傳播學研究所	?

表 4：MBR 組合函數判斷是否推薦

使用 MBR 判斷新進讀者是否要推薦「孟子」							
	相似者	相似者流失狀態	k=1	k=2	k=3	k=4	k=5
d <sub>加總</sub>	4,5,2,1,3	是,否,是,是,否	是	?	是	是	是

## 五、設定鄰近資料的數目：

要決定這個數目常是要看測試實際結果而定，一般來說，鄰近資料數量越多結果越好[17]，但是 Marko 在[13]中提到一般



建立資料探勘模組步驟：

#### 一、定義資料需求(Identify Data Requirements)及資料取得

模組發展過程中的第一步驟就是定義及取得正確的資料。正確資料的來源通常只是從資料庫中取得所需用到的欄位，這些資料可能會經過合理化的清理(Clean)過程得到，並不是資料越多越好，而是要找出符合我們需求的資料。另外，在正確資料的前提下，資料樣本數量及樣本期間越長，則系統的所達到的正確率會越高。

#### 二、資料驗證、探勘、清理(Validate, Explore and Clean Data)

不正確的資料會導致錯誤從不同的點產生。一般企業的資料庫中，對於企業有重要意義的欄位通常都保持資料百分百正確，但若不重要的資料，則不注重其正確性。以資料探勘的觀點來看，所包含進來的資料並不一定是重要的欄位，所以會有不正確資料產生的情況，此時，資料清理及驗證就顯得格外重要。

#### 三、資料結構轉換(Transpose Data to the Right Granularity)

我們在架構資料探勘模組時，有時會用到階層式的概念；且由於資料探勘的演算法在實作時所作用到的資料記錄都是每一行獨立的，若有階層式的資料記錄，都要做轉換的工作。所以，在模組中的所有紀錄，在放進資料表時，都須將之轉換成單一行的結構。

#### 四、增加衍生變數(Add Derived Variables)

衍生變數的來源是從現有資料庫中的資料依資料探勘的資訊需求做加總、乘、除等動作形成另一類資料，只要是對探勘模組有幫助的皆可加入。如：金額的加總。

## 五、建立模組資料集(Prepare Model Set)

要如何確定從模組中所探勘粹取出來的資料是隱含的、有用的。使用 Model Set 方法來認證。模組資料集(Model Set)、就是真正在資料探勘系統模組中執行的資料。通常都分成訓練資料集(Train Data Set)、測試資料集(Testing Data Set)、評估資料集(Evaluation Data Set)，運用訓練資料集來將資料帶入模組中，並用測試資料集來驗證資料的正確性，最後再帶入評估資料集作為實際上的測試及預測，以達到資料探勘的目的。

- 1.訓練資料集(Train Data Set)：這個模組是採用資料庫中預先分類好的部分資料作為訓練資料集，在這個步驟當中，資料探勘的演算法將會找出型樣及預測值。
- 2.測試資料集(Testing Data Set)：用測試資料集的資料來讓模組更加精確的，使用資料庫中的另一個子集合來作為測試的資料。為什麼要讓模組更加精確呢？是因為避免模組落入用訓練資料集所產生的預測資料型樣(Pattern)之中，因為預測資料集只是所有資料庫中的一部分資料，並不能完全代表資料庫，故採用測試資料集來調校教及確認。並能更加確定模組的一般化(general)和未來在做評估時更加準確的依據。
- 3.評估資料集(Evaluation Data Set)：評估系統模組效能(performance)及計算採用評估資料集所產生的結果和測試及訓練資料集所產稱的結果差異性有多大。

上述三個 Model Set 的資料都是互斥的，如此才不會產生資料交互重疊的影響。

## 六、選擇探勘技術(Choose Modeling Technique)及訓練模組(Train Model)

Choose 用什麼 Mining 的技術,如類神經、MBR 等,現行有許多 DataMiningTools 內建好不同的 Modeling Technique,只要將資料輸入即會自動將資料放入不同的模組中執行,並選擇一個較符合的模組供使用。

## 七、檢查模組效能(Check Performance of Model)

資料探勘的工具會提供實驗數據及參數,協助我們了解模組的效能。

### 第三章 系統架構

本系統建置一個人化導向的查詢介面，協助找出真正符合讀者個人化需求的書籍，並推薦之。本研究提出兩個模組應用於系統的功能之中。第一個模組是「主題式同類書目推薦模組」，第二個模組是「互助式資訊過濾推薦模組」。本章中分別就圖書推薦檢索系統的兩個系統模組架構流程進行說明。

#### 第一節 圖書推薦檢索系統整體架構

圖書推薦檢索系統主要分成兩個模組進行：「主題式同類書目推薦模組」利用中國圖書分類法中詳盡分類的特性，為讀者推薦適當的書籍。在「互助式資訊過濾模組」之中，結合使用者屬性資料表和借閱紀錄表，探勘出背景屬性相近的讀者，並運用改良式 MBR 分群方法加速搜尋的腳步，透過此機制，盡可能達到個人化環境的目標。

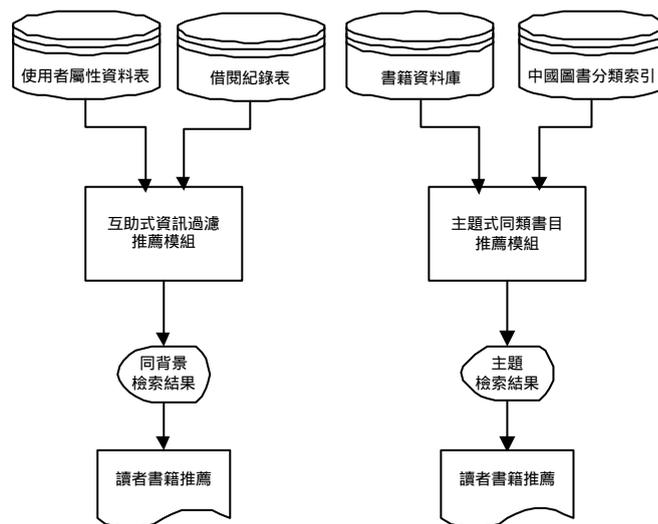


圖 11：圖書館推薦檢索系統架構圖

二、「主題式同類書目推薦模組」：

以中國圖書分類法的分類主題為主，依讀者所鍵入的字串為

搜尋關鍵字，從資料庫中找出和關鍵字相關的資料，並將這些資料依中國圖書分類法的主題分門別類，讓讀者進階勾選主題，最後再將屬於此主題的書籍推薦給讀者，這樣不會造成所搜尋到的資料數量龐大且不符讀者需求；如此即使書名和讀者所搜尋的關鍵字不符合，但只要是讀者想要的類別，仍會推薦給讀者，達到推薦的書籍符合讀者的要求，主題式同類書目檢索推薦模組架構圖參見圖 12。

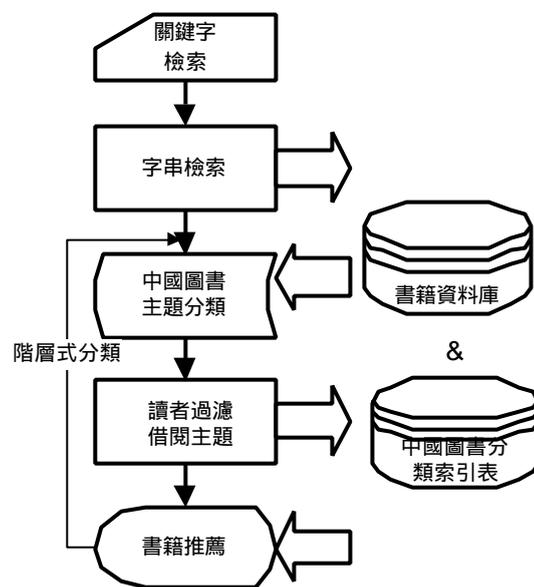


圖 12：主題式同類書目推薦模組

### 三、「互助式資訊過濾推薦模組」：

個人化資訊環境強調以人為本，因此，本研究找出同一群聚中某些相同的特徵，將背景屬性屬於同一分類者群聚在一起，作為新進入者的書籍推薦參考。因為 MBR 不在乎資料格式，因此我們利用 MBR 從資料庫中找尋和讀者身分屬性相近的其它所有讀者的分群，並用關聯規則技術找出相關性高的推薦書籍。另一方面，MBR 在資料量大時相當費時，為了達到即時線上推薦的目的，我們提出改良式 MBR，利用一些預處理與相關資訊的儲存，

來快速取得推薦的書籍。關於互助式資訊過濾推薦模組架構圖，參見圖 13。

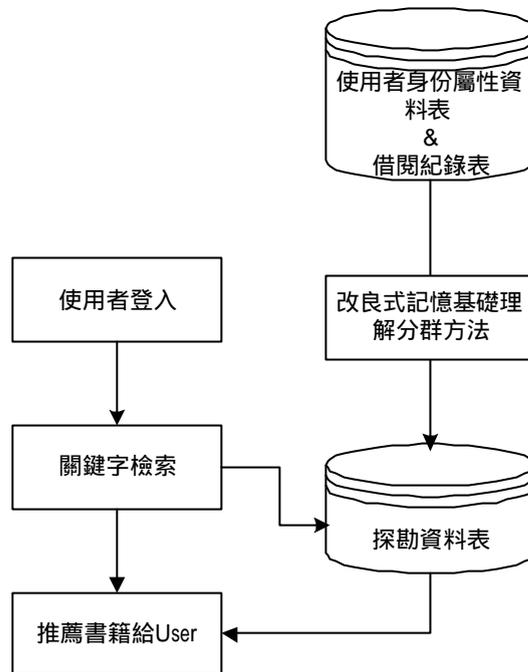


圖 13：互助式資訊過濾推薦模組圖

## 第二節 系統模組

### 壹、主題式同類書目推薦模組

以中國圖書分類法的分類階層為主，依讀者所鍵入的字串為搜尋關鍵字，從資料庫中找出和關鍵字相關的資料，並將這些資料依中國圖書分類法的主題分門別類，讓讀者進階勾選，再將屬於此主題選項下更詳細的分類主題顯示出來，亦可再讓讀者選取，如此逐層進行，直到讀者確定所要的分類，並將屬於讀者需求的分類書籍依照熱門借閱率排序出來做為推薦的書目；另一方面，系統所推薦的書籍是位於讀者所確定過的主題之下，可能包含多本書籍，而且書名和讀者所輸入的關鍵字亦不盡相符，所以推薦的書籍經過熱門借閱率的排序，期望此主題式同類書目推薦模組能盡量達成符合個人化環境的效益。

本文以圖 14 舉例說明中國圖書分類法的配置流程及在於主題式同類書目推薦模組的應用；在第二章第二節中曾經提到，中國圖書分類法的分類方法是採階層式分布，依種類分為十大類，每一大類以一個百位數整數為代表，每往下一位數即代表下一階層的分類。圖 14 中，310、320 ...，390 是屬於 300 下的分類；再往下一層來看，321、322、...，329 是屬於 320 下的分類，此階層架構在圖中可明顯的表示出來。另一方面，每一個主題分類中都會包含屬於此分類的書籍，並不限定於最底層的主題才有書籍隸屬之。如 323(記述天文學)主題下有我們的太陽、冥王星、海王星、月亮..等書籍，323.2(記述天文學,太陽系)主題下有太陽、我們的太陽系、土星、小行星群等書籍。323.3(記述天文學,行星)主題下有星的奧秘、探測行星、移民火星..等書籍。

當讀者輸入「太陽」關鍵字時，系統會從資料庫中找出所有書名包含有「太陽」的紀錄，並將主題按階層呈現出來，讓讀者勾選主題，若讀者選的主題是「科學總論,天文,記述天文學」，則將我們的太陽、冥王星、海王星、月亮..等書籍按熱門借閱的比率推薦給讀者；若讀者選的主題是「科學總論,天文,記述天文學,太陽系」，則將太陽、我們的太陽系、土星、小行星群等書籍按熱門借閱的比率推薦給讀者。

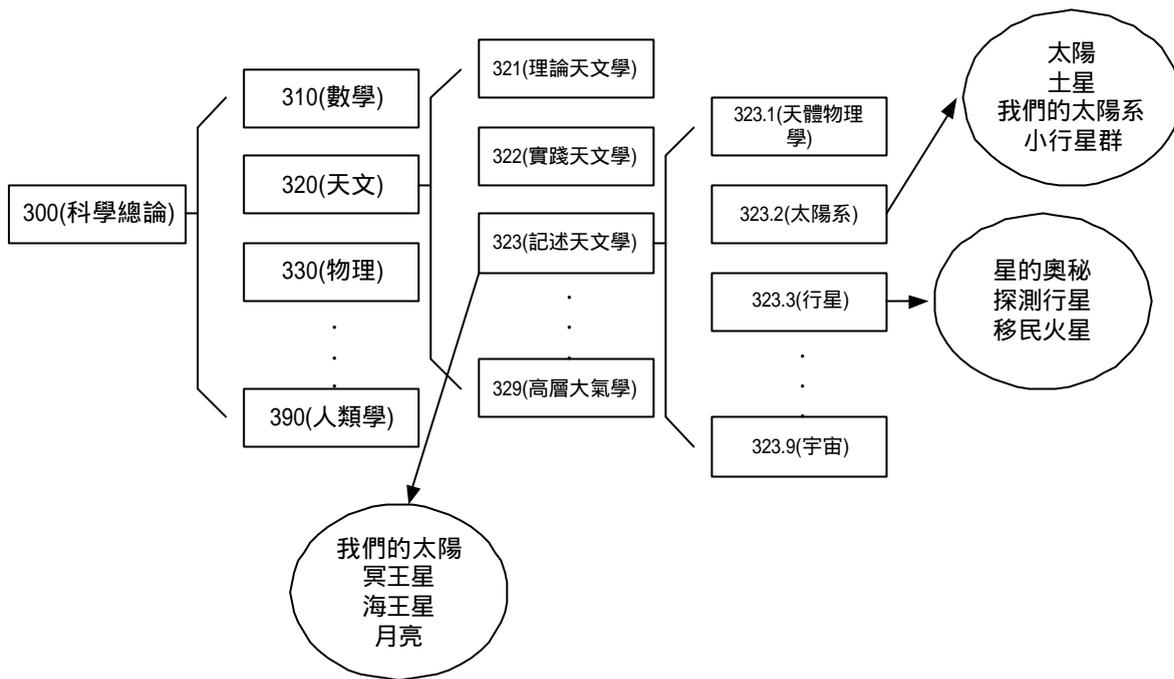


圖 14：中國圖書分類法示意圖

## 貳、互助式資訊過濾推薦模組

MBR 方法中，主要的精神在於利用「距離函數」計算新進樣本點和樣本空間中其它所有樣本點的距離，並將距離符合鄰近資料門檻值個數的樣本歸類為同一分群，再由同一分群內的所有樣本點投票決定是否要推薦書目。

在圖書館的個人化系統中，在讀者按下搜尋鈕後，所強調的是系統能夠立即的給予回應，若搜尋時間過久，會導致讀者喪失耐心。以探勘資料表中有 10,000 紀錄為例，要計算新進樣本點與其它一萬筆記錄(10,000)間的距離，再將這些距離由小到大作排序，並以距離最短的幾個鄰近資料來投票決定預測的情形。如此一來，每有一個讀者登入系統時，系統都要重新計算一次。隨著探勘資料表中資料記錄的增加，搜尋時間會增加，降低系統效能，將無法符合個人化需求。

為了達到即時推薦的目的，本研究採用 MBR 的分類方法，事先找出每個樣本點的鄰近資料(Neighbor)，歸類為同一分群，並使用

(Simple Cluster - Seeking Algorithm, SCSA)演算法，將群與群之間做歸類合併的動作，減少分群的數目，以加速個人推薦書目搜尋的速度，將個人化環境圖書資料推薦的系統效能提高。

#### 一、互助式資訊過濾模組流程大綱：

(一)、分別以每位讀者做為空間樣本點中的種子樣本點，以 MBR 的方法找出每個種子樣本點的鄰近資料當做同一分群，這樣的分群稱為子分群(subgroup)，每一個子分群都有一個編號來代表；換言之，我們以空間中的每個樣本點為主做分群的動作，因此，會有  $n$  個子分群。每一子分群中內包含的是一個質心(種子樣本點)，以及和距離質心最近且符合鄰近資料數目的樣本點。

(二)、運用 SCSA 演算法，將子群和其他子群做重疊程度比對，當相似程度超過我們所定義的相似度門檻值(Similar threshold)時，則將兩子群合併成為一大分群，若重疊程度低於相似度門檻值時，則歸類為另一新的大分群。如此，重疊程度高的子群即可歸於同一群，以減少子群的數量。本文中相似度門檻值定義為兩個子分群內相同樣本點的重複程度；舉例說明，兩個子分群內各有 10 個樣本點，而其中有三個樣本點一樣，則 Similar threshold=30%。在本論文中，我們定義 Similar threshold 為 75%，也就是兩個子分群之間，必須要有 8 個以上的點重複才會歸類為同一大分群。

(三)、有一讀者進入系統搜尋時，只要計算這位讀者樣本點和所有大分群代表點的距離，找出距離最近的大分群；再計算這位讀者所在大分群內所有子分群質心的距離，找出距離最近的子分群，將(一)所事先計算好子分群關聯規則所推薦的資料，推薦給

讀者。如此一來，讀者搜尋的線上等待時間將會從計算所有空間樣本點的時間降低到只計算一次大分群代表點的最短距離和計算大分群下子分群代表點的最短距離的時間。

## 二、互助式資訊過濾模組流程說明：

### (一)、以讀者為中心探勘子分群及推薦書籍：

樣本空間中所有點都依序為種子樣本點，計算每個種子樣本點與其他樣本點的距離，並找出每個種子樣本點的最近鄰近資料(要預先定義最近鄰近資料個數 nearest threshold)，形成一筆「分群子紀錄」。

以圖 15 舉例說明，樣本空間點中有  $n$  位讀者，分別以編號 1 ~ 編號  $n$  代表。

1. 設定鄰近資料門檻個數(nearest threshold)值  $k=4$ 。
2. 以編號 1 樣本點為中心點所做的分群，灰色部分代表以編號 1 為質心且和編號 1 距離最短的 4 個樣本點，歸類為子群 1，此群組內包含(1,3,12,14,7)。

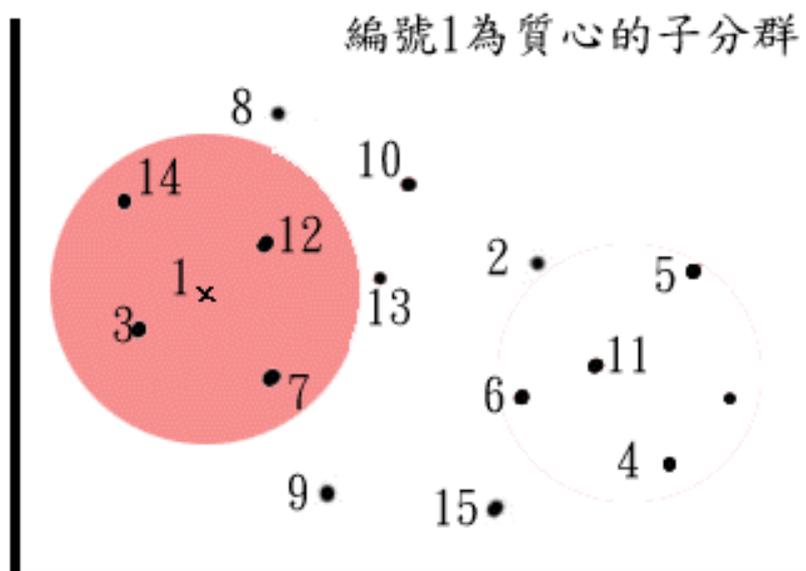


圖 15：單一種子樣本點分群示意圖

將編號 1 的分群概念用資料記錄的角度來看，編號 1 的種子樣本點為起點，分別計算編號 1 和其他所有點的距離，並依距離大小將所有樣本點編號由小到大排序，形成如表 5 的矩陣，小括弧內所代表的為距離編號 1 質心的距離長度。

表 5：種子樣本點和其他樣本點的距離矩陣表

種子樣本點	依新樣本點和其他樣本點的距離由小到大排序							
1	1(0.0)	3(0.8)	12(1.0)	14(1.2)	7(1.5)	13(2)	...	dist(n)

3. 在編號 1 為質心的子分群之中，每一個點代表一位讀者，每一位讀者都有若干筆借閱記錄，再用關聯規則的方法從同一子群中找出關聯性高(符合門檻值 support=2)的書籍做為這一子群的推薦書目，參閱表 6。從表 6 中經由關聯規則所得到的結果推薦書目是(管理數學 組織行為)和(記憶術 88 法則 組織行為)。
4. 重複上述 1-3 流程得到表 7，n 個子分群中，分別為每個子分群預先計算好要推薦的書籍，儲存於資料表之中。

表 6：編號 1 子分群借閱書目

紀錄編號	書名
1	管理數學,組織行為,教學模式,鬧中取靜
3	記憶術 88 法則,組織行為,創業的智慧
12	管理數學,記憶術 88 法則,組織行為,創業的智慧
14	記憶術 88 法則,鬧中取靜
7	競爭優勢,企劃實務

表 7：編號 1 子分群關聯法則推薦書籍

子分群編號	推薦書籍
1	(管理數學 組織行為)(記憶術 88 法則 組織行為)
2	(灌籃高手 名偵探柯南)
...	.....
n	(統計學 財務管理)(社會心理學 當代社會理論)

## (二)、用 SCSCA 演算法合併子分群：

(一)所介紹的是分別為單一種子樣本點作分群的情況。為了加速線上處理的速度，我們打算將子分群做可能的合併，當為每個樣本點都分群後，會產生分群區域重疊(overlap)的狀況，這些重疊部分的程度有高有低，因此我們使用重疊程度做為合併的依據，將重疊程度高的子分群做合併。我們思考要如何將子分群做合併，將重疊程度低的子分群不合併，我們採用 SCSCA 演算法的快速分群能力，期望將重疊程度高的子分群合併成同一分群，以減少分群的數目。

資料探勘表中有  $n$  個樣本點，則距離矩陣表中即會有  $n$  筆記錄，每一筆記錄代表一個子分群，我們要計算第一個子分群和其他子分群重疊的部分，若重疊程度大於 Similar threshold，則將之併入第一個子分群，並給予一個大分群編號；若小於 Similar threshold，則將之標示為另一個大分群。以此類推，直到找出所有的大分群為止，每個大分群都有一中心樣本點作為代表。

圖 16 中是以(二)中用 SCSCA 演算法合併子分群的步驟來做說明，圖 16 所示為樣本空間中的所有點作分群，同一圓圈內的點代表同一分群，圖中以質心 1,2,3 做子群代表。

質心 1 所包含的同一分群有(1,12,6,2,14)、質心 2 所包含的同一分群有(2,1,12,6,13)、質心 3 所包含的分群有(3,11,7,15,13)。依 SCSCA 演算法的流程，首先先將子群 1(質心是 1)視為大子群 1，和子群 2 的點作比較，對於大子群 1 來說，重疊的點有(1,12,6,2)，5 個點中有 4 個點重複，重疊

程度為  $4/5=80\%$ ，大於  $75\%$ (Similar threshold)，將子群 2 併入大子群 1。接著和子群 3 比較，重複點只有 1 個(點 13)，重複率為  $1/5=20\%$ ，低於 similar threshold，將子群三做為另一個大分群。依此類推，直到所有子群處理完畢，最後會形成若干個大分群。每個大分群都會有一個子分群作為質心。從表 8 中「大分群」欄位中可看出系統分為幾個大分群。大分群 1 的質心為「子分群 1」。SCSA 的演算法如圖 17 所示。

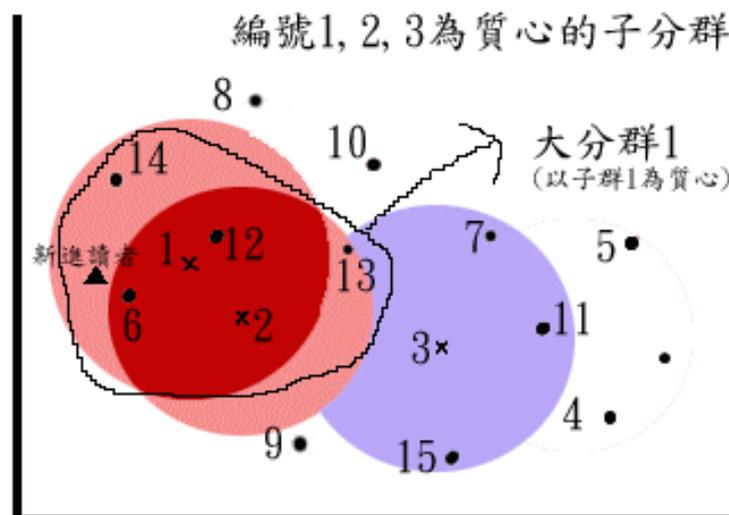


圖 16：多個種子樣本點分群示意圖

表 8：經由 Improved MBR 分類後大分群

子分群	樣本點和其他樣本點的距離由小到大排序					大分群
1	1	12	6	2	14	1
2	2	1	12	6	13	1
3	3	11	7	15	13	2
...	...	...	...	...	...	...
...	...	...	...	...	...	...

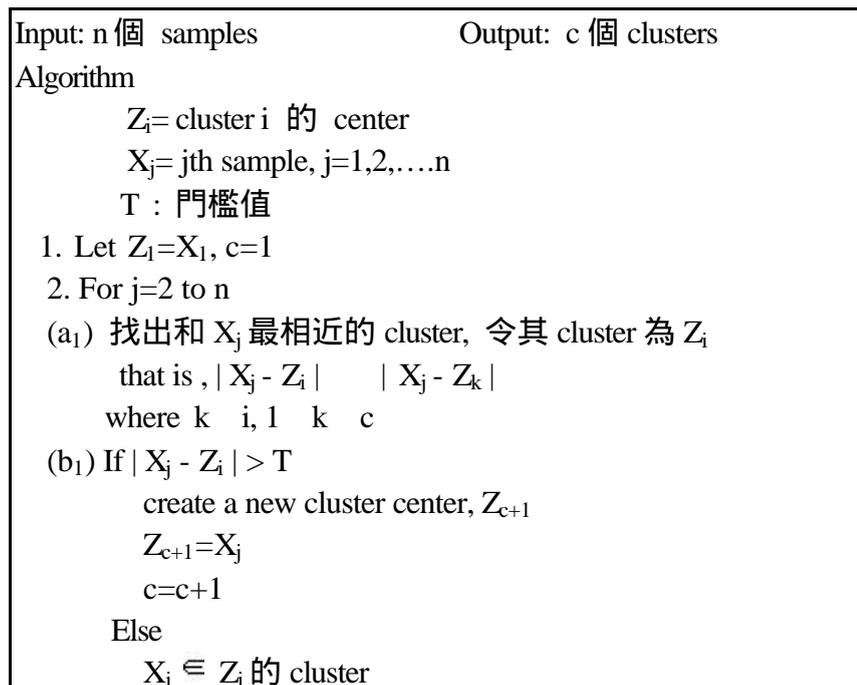


圖 17 : SCSA Algorithm

(三)、新進讀者書籍推薦：

以上的處理程序都在系統背景先執行完畢，當新讀者進入時，只要計算讀者與所有大分群質心樣本點的距離，距離最短的即為讀者所屬大分群；再計算讀者與此大分群內所有子分群質心點的最短距離，找出讀者屬於哪個大分群下的子分群，並由事先計算好子分群的相關性高的書籍推薦給讀者。圖 16 中用三角形所代表的是新進讀者，先計算出他是屬於大分群 1，在和在大分群 1 內的所有子分群計算距離，找出他是最接近子分群 6，於是便將子分群 6 的相關書籍推薦給此新進讀者。

在記憶基礎理解(MBR)的方法中，每有一筆新進資料進入，則須重新計算此新樣本點和樣本空間中所有點的距離，計算的時間會超出讀者所能等待的時間，對於提供個人化服務的系統來說無法達到其效能。

改良式記憶基礎理解 (Improved MBR)方法，先依樣本空間中點的分布特性做好分群，分群大小不限，而是一群聚特性決定，符合實際資料分群特性。將樣本空間內分成若干個大分群，降低新進資料比對時間，提昇系統效率。

### 第三節 資料探勘模組建立互動流程

#### 一、定義資料需求及資料取得

##### 1. 定義資料需求：

由第三章第一節的兩個模組架構圖中可得知本研究所需要的資料需求包括了中國圖書分類索引表、書籍資料表、讀者屬性資料表、借閱紀錄資料表。依據系統需求定義資料屬性及格式，參閱圖 18。

書籍資料表(lib_book)：登錄號、書名、作者、出版者、出版年、分類號
借閱紀錄資料表(lib_loanred)：讀者證號、書名、借閱紀錄編號
讀者屬性資料表(lib_property)：讀者證號、所屬單位、身分、借閱紀錄
中國圖書分類索引表(lib_index)：中國圖書分類編號、中國圖書分類主題

圖 18：系統需求資料表欄位說明

##### 2. 資料取得：

南華大學取得的資料表包含書籍資料表和借閱紀錄資料表，詳細資料格式參見圖 19。

書籍資料表(dbo_allbook)：作者、出版、出版年、ISBN、ISSN、出版地、分類號、作者號、年代號、卷冊號、複本號
借閱紀錄資料表(dbo_hua)：單位名稱、身分別、交易名稱、書名、登錄號、特藏名稱、交易日期、交易時間

圖 19：南華大學原始資料表

本系統從南華大學圖書館獲取資料，書籍資料表計 197,951 本，借閱紀錄表計 554,876 筆，有效的取樣年限從 1999/10/20 到 2003/1/24 日合計 3 年 4 個月。

## 二、資料驗證、探勘、清理(Validate, Explore, Clean Data)

以資料探勘的角度還看，所要做的是去除多餘的、不符合探勘需求(unpopular、illegal、unreasonable)的資料，但對於圖書館資料來說，我們將不符合需求的資料，如不需要的欄位、重複的書籍、空白欄位的紀錄、不符合時限的資料做清理的工作；符合需求但不符合格式的資料做轉換的工作。

從南華大學圖書館所取得的資料中，將無須使用到的欄位刪除掉，並按照系統模組的需求加入所需用到的欄位，形成如圖 18 的資料表。

在書籍資料表中，在去除複本後所得資料計為 169,018 本；而在借閱紀錄表中，借書、續借、還書，本系統視為同一筆資料，經過資料轉換後有效筆數為 58,858 筆。

## 三、將原始資料轉換成適合探勘的資料格式(Transpose Data to the Right Granularity)

由於資料探勘的演算法在實作時所作用到的資料記錄都是每一行獨立的，以資料探勘的角度來看有時需要轉換成線或面的需要，所以有必要做轉換的工作。本系統中，單一位讀者連續兩天到圖書館去分別一次借閱五筆記錄，在借閱紀錄表中會有十筆記錄，而在探勘資料表中只視為單一筆記錄，也就是一位讀者借了十本書，可參閱表 9。

## 四、增加衍生變數(Add Derived Variables)

由於南華大學圖書館所給予的資料當中並無讀者證號一

欄，在紀錄中我們無法判別哪一本書為何人所借閱，所以本研究採用以現有資料中日期、時間、身分和單位的特性模擬單一借閱者，並給予讀者證號。我們將借閱紀錄資料表和讀者屬性資料表做整合的動作，形成探勘資料表，如表 9，並在其中增加了讀者證號、借閱數量等衍生變數，以達到系統施行可能性的目的。

讀者證號	所屬單位	身分	借閱數量	借閱紀錄
90112513	資管所	在職專班	3	系統分析、漂亮家居、資料結構

表 9：探勘資料表(lib\_mining)

#### 五、建立模組資料集(Create Model Set)

建立模組資料集主要分為二部分，訓練資料集(Train Data Set)、測試資料集(Testing Data Set)，沒有採用 Evaluation Data Set 的原因在於探勘資料表中的讀者證號及借閱數量是屬於虛擬的資料，在系統評估上的信度低，所以我們只選擇 Train Data Set 和 Testing Data Set 兩組資料集。系統的模組資料按照時間區分為二份，並分別代表這二項模組的資料集；從 1999/10/20 到 2001/12/31 的資料為 Train Data Set、2002/1/1~2003/1/24 為 Testing Data Set。運用 Train Data Set 來將資料帶入模組中，找出型樣及預測值，並用 Testing Data Set 來驗證系統的正確性，亦以達到資料探勘的目的。

#### 六、選擇探勘技術(Choose Modeling Technique)

個人化圖書館著重於個人化資訊，通常是由興趣、習慣組成個人化環境，而對現有電子化圖書資料庫中的資料而言，個

人的背景屬性是一項很好的探勘分析的資料，要將這些資料依背景屬性做分類，MBR 是一很適合的方法，因為 MBR 能運用距離函數的觀念將複雜的資料型態做分群的優點。另一方面，圖書館的書籍資料屬於 Items 的型態，非常適合採用關聯法則的方法來做探勘，所以我們希望結合這兩種技術來加強圖書館之個人化環境服務。

#### 七、檢查模組效能(Check Performance of Model)

本研究運用 Train Data Set、Testing Data Set 二部分來測試系統效能。「主題式書目推薦模組」流程是很直覺的，本研究運用中國圖書分類法將書目預先詳盡分類的特性，加上讀者自行選取符合需求的主題，進行書籍推薦的動作；雖未請讀者進行測試評估的工作，但符合個人化環境的觀念。而「互助式資訊過濾模組」將資料分為訓練資料集和測試資料集兩群，分別執行這兩群資料集，所得到的結果以相同背景身分和類似借閱數量的讀者所推薦的書籍相互比較，會發現推薦的書籍相似度較高，由此可知我們所提出的系統模組有其效能上之可行性。

## 第四章 系統實作與討論

### 第一節 系統建置

#### 壹、主題式同類書目推薦模組

以中國圖書分類法的分類階層為主，依讀者所鍵入的字串為搜尋關鍵字，從資料庫中找出和關鍵字相關的資料，並將這些資料依中國圖書分類法的主題分門別類，讓讀者進階勾選主題，再將屬於此主題的書籍中屬於熱門借閱的書籍推薦給讀者。這樣不會造成所搜尋到的資料數量龐大且由於主題是由讀者自行勾選，系統所推薦的書籍會是更加符合讀者個人需求。

##### 一、檔案架構

本模組所採用到的表格包含書籍資料表、中國圖書分類索引表，我們將書籍資料表的分類號和中國圖書分類索引表的編號做關聯，便於系統使用。

表 10：書籍資料表 schema ( lib\_book )

欄位說明	Field	Type	Length
登錄號	acc_no	Char	12
書名	title	Char	150
作者	author	Char	100
出版者	publisher	Char	100
出版年	publish_year	Char	4
分類號	class_no	Char	50

表 11：中國圖書分類索引之 schema ( lib\_index )

欄位說明	Field	Type	Length
編號	indnum	Char	10
主題	indtitle	Char	32

## 二、模組說明

台灣現行圖書館中文書目分類中，皆是採用賴永祥教授所編訂的「中文圖書分類法」；在一般數位圖書館中所採用之資料檢索的「索書號」檢索，也是以此種分類法為主。雖然「中文圖書分類法」有分類詳盡明瞭的優點，但因為它的檢索方式是以索書號的編號來查詢書籍，一般讀者在尋找書籍時，無法得知中文圖書分類編號及作者編號，因此雖然有以中文圖書分類法編號來做書籍檢索的功能，卻因編號記憶不易，而喪失的數位圖書館中中文圖書分類法檢索的美意。

基於能夠保持中國圖書分類法明確詳盡的優點，又能改善檢索介面友善性（friendly）的立足點，系統中提出以反向的思考模式，讓圖書檢索能夠以讀者為本的思考方式，以字串搜尋的方式，配合中國圖書分類法，找出符合讀者真正需要的熱門主題書籍並推薦之。

## 三、模組流程：

- 1.讀者登入個人化環境，若為第一次登錄，將個人紀錄登錄到讀者屬性資料表當中
- 2.讀者輸入欲搜尋之書籍字串 search\_key
- 3.系統依書名字串，運用 SQL 語法在書籍資料書中搜尋所有紀錄，並將所有比對到的資料存入暫存資料表（temp table）中。

```
“select acc_no,title,class_no into Temp.acc_no,Temp.title,Temp.class_no  
from lib_book where title like search_key;”
```

```
“select class_no from Temp group by class_no”
```

- 4.以中國圖書分類法索引表中的編號和主題為主，將暫存資料表中的資料做主題分類，多本書籍可能隸屬單一主題，系統會只

將單一主題列出

```
“SELECT indtitle FROM lib_index INNER JOIN temp ON  
lib_index.indnum= temp.class_no;”
```

5. 列出所有分類好之主題，讓讀者階層式勾選真正所需要的主題，將讀者選定主題之中國圖書分類編號儲存於陣列 list\_ary[] 之中，

6. 依讀者所選定之主題，用中國圖書分類編號到書籍資料庫中擷取符合讀者需要主題的所有書籍資料，並推薦給讀者，

```
For count=0 to max_count(list_ary) -1  
    rec=Select title from lib_book where list_ary[count] like  
    title;  
    for cnt=0 to max_count(rec)-1  
        select count(title) from lib_book where indtitle=rec(cnt)  
        order by rec(cnt)  
    end for  
    Count++;
```

End For

7. 將讀者所借閱的書籍資料記錄下來，存入借閱紀錄表中，作為日後互助式資訊過濾模組推薦的參考依據

8. 結束

## 貳、互助式資訊過濾推薦模組 Collaborative Information Filtering

「互助式資訊過濾書目推薦」是依讀者的身分屬性(身份、科系、借閱數量)，以 MBR 的技術做群集化 ( clustering ) 的分析，紀錄相同群集間所有借閱的資料記錄，運用關聯規則的技術過濾出同群讀者間高頻項目集 ( Large Itemsets ) 的書籍，日後針對讀者的查詢主題，推薦給讀者。

## 一、檔案架構

本模組中將借閱紀錄表和身分屬性資料表做清理結合的動作而成探勘資料表，參閱表 12、表 13、表 14。

表 12：借閱紀錄表之 schema ( lib\_loanred )

欄位說明	Field	Type	Length
讀者證號	reader_ID	Char	10
書名	title	Char	150
借閱紀錄編號	recorder_ID	Char	12

表 13：身分屬性資料表

欄位說明	Field	Type	Length
讀者證號	reader_ID	Char	9
所屬單位	dept_name	Num	2
身分	type_name	Num	2
借閱數量	amount	Num	2

表 14：探勘資料表

欄位說明	Field	Type	Length
讀者證號	reader_ID	Char	9
身分	Type_name	Num	2
單位	dept_name	Num	2
借閱數量	amount	Num	2
借書紀錄 ( 書名 )	acc_no	Char	65536

## 二、模組說明

互助式資訊過濾模組採用了關聯規則、MBR 及 SCSSA 演算法三項技術，故資料需事先經過轉換和選擇訓練資料兩個步驟。

MBR 是一個監督式資料採礦技術，利用已知的範例形成推斷未知的模型，它會找尋已知部分與最接近的案例，並且將它們結合。MBR 技術包含選擇歷史資料、轉換歷史資料、設定距離函數、組合函數、和鄰近資料的數目、最後是分類、預測並評估等幾個步驟。本研究根據系統需求將上述步驟稍作修改，先將資料做清除轉換，再運用 MBR 及關聯規則找出推薦書籍，最後運用 SCSSA 演算法加速分群的觀念，協助讀者進入系統時能快速獲得所需的資料。

#### (一) 資料清理和資料轉換：

身分屬性資料表中欄位內的資料都是文字形式的，為了適用於 MBR，須將文字資料轉換成數值型態，以方便距離計算。在 MBR 技術中，通常會視變數的型態不同而設定距離函數。在圖書館的環境之中，可做為變數的資料還很多，如興趣、查詢關鍵字、將上時段限制的讀者借閱紀錄等...，由於單位和身分是屬於類別變數，而借閱數量是屬於數值變數，故本系統在類別變數上依實際需求訂定距離函數，以矩陣方式表示，而數值變數則依實際取得資料後做正規化的動作。

本研究由資料取得的難易程度選取了身分、單位、借閱數量三項變數作為本系統 MBR 的分類依據，以下分就三項變數說明：

1. 身分：表 15 中的分類標準是南華大學現階段所採用的身分分類，是屬於分類明確的類別變數，分為老師、學生、職員三大類，本研究定義出身分距離函數（圖 20），並用距離矩陣的方式表達，參閱表 16。

$d_{\text{身分}}(A,B) = 0.0$ 如果身分相同 (老師、學生、職員)
$d_{\text{身分}}(A,B) = 1.0$ 如果身份不同 (老師、學生、職員)

圖 20：身分距離函數

表 15：屬性轉換數值分布表 (變數：身分)

老師	學生	職員
兼任教授	二年制學士專班	本校職員
專任教授	大學部先修生	行政教學單位專用
	大學部學生(日)	
	在職碩士專班	
	進修學士班(夜)	
	碩士先修班	
	碩士班學生(日)	

表 16：身分距離矩陣表

	老師\兼任教師	老師\教授	...	學生\大學生	學生\研究生	...	職員\行政職員
老師\兼任教師	0	0	...	1	1	...	1
老師\教授	0	0	...	1	1	...	1
...	...	...	...	...	...	...	...
學生\大學生	1	1	...	0	0	...	1
學生\研究生	1	1	...	0	0	...	1
...	...	...	...	...	...	...	...
職員\行政職員	1	1	...	1	1	...	0

2. 單位：表 17 中為南華大學現階段所有的單位屬性，不同科系的人，所使用的書籍會有所不同，由此來作為分群的依據，距離函數參閱圖 21、距離矩陣參閱表 18。

$d_{\text{單位}}(A,B) = 0.0$ 如果相同學院、相同科系
$d_{\text{單位}}(A,B) = 0.5$ 如果相同學院、不同科系
$d_{\text{單位}}(A,B) = 1.0$ 如果不同學院、不同科系

圖 21：單位距離函數

表 17：屬性轉換數值分布表（變數：單位）

文學院	社會科學院	管理學院
哲學系所	社會學研究所	管理研究所
生死學系所	亞太研究所	環境管理研究所
環境與藝術研究所	公共行政與政策研究所	出版學研究所
美學與藝術管理研究所	經濟學研究所	財務管理研究所
宗教學研究所	歐洲研究所	旅遊事業管理學研究所
文學系所	教育社會學研究所	非營利事業管理研究所
應用藝術與設計系	國際關係學系	資訊管理學系所
民族音樂系	應用社會學系	傳播管理學系所
		企業管理學系所
		電子商務管理學系

表 18：單位距離矩陣表

	文學院 \哲學系	文學院 \生死學	...	管理學院 \資管系	管理學院 \企管系	...	社科院 \歐洲所
文學院\哲學系	0	0.5	...	1	1	...	1
文學院\生死學	0.5	0	...	1	1	...	1
...	...	...	...	...	...	...	...
管理學院\資管系	1	1	...	0	0.5	...	1
管理學院\企管系	1	1	...	0.5	0	...	1
...	...	...	...	...	...	...	...
社科院\歐洲所	1	1	...	1	1	...	0

3. 借閱數量：系統評估每位讀者的歷史借閱紀錄來計算借閱數量，發現大多數者的借閱數量分配大多在 10-100 之間，

刪除逾越此範圍的資料，留下的做為借閱數量一欄的資料，可瞭解借閱數量相近的讀者之借閱習性；此外，顧慮到數值分布大小的情形，若不同欄位的資料變化差異太大易導致數值大的資料牽引小的資料情況產生，故須將這些值做資料正規化的處理，由於身分、單位兩個變數的資料值都界於 0-1 之間，我們也將借閱數量這個變數資料正規化成 0-1 之間；另一個問題是借閱數量資料的範圍界於 0-100 之間，故計算之後需除以 100，所得的範圍才會界於 0-1 之間，也才符合三項變數的值都界於 0-1 的資料範圍。

(2) 設定距離函數：

距離函數的計算方法是採用空間向量中的距離公式為主。計算每一個點和空間中其他所有點的距離，並兩兩比較，找出符合 k-nearest neighbor (k 值一般都設定為 10-15 之間，本文設定為 9 個，加上自己計 10 個)，之後再將所找到的點存入資料表之中，表 19 為說明探勘資料表的簡例。

表 19：探勘資料表之簡例

讀者證號	單位	身分	借閱數量	書名
90112501	資管系所	在職生	24	XML.....
90101111	傳播系所	大學生	31	.....
90101112	企管系	職員	19	.....
...				.....
92110000	文學系所	教授	59	中國思想,蛇,盡情做自己.....

本研究採用最直覺的距離公式的觀念 (公式 1),  $x, y$  代表空間上的兩個點,  $x_i, y_i$  系統採行的第  $i$  個變數,  $i=1 \dots n$  代表有

n 個變數。

距離公式： $d(\mathbf{x}, \mathbf{y}) = W_1 d_{v_1}(x_1, y_1) + W_2 d_{v_2}(x_2, y_2) + \dots + W_n d_{v_n}(x_n, y_n)$

$w_1, w_2, \dots, w_n$  代表每個變數(feature)的權重

$v_1, v_2, \dots, v_n$  代表變數(feature)

$x = [x_1, x_2, \dots, x_n]$

$y = [y_1, y_2, \dots, y_n]$

每一點都有一筆以它自身為中心點的紀錄。MBR 技術中的一項缺點即是分類的距離計算上運算作業繁複，要計算每個點和其他所有點的距離，需耗費大量運算時間及記憶體空間；舉例說明，若有 n 個點，要計算一和其他 n-1 個點的距離，並找出符合 k 值的點將之取出做為分群的依據，如此，全部 n 筆記錄總共要計算  $n*(n-1)$  次；經過上面步驟，可得到以每個點為中心點的子群，子群之間有可能是相互重疊的，也有可能是互不相干的。經過距離函數的計算後，我們可得到以每個中心點為準的子分群。

表 19 舉例說明，計算 90112501 和 90101111 的距離，身分分別是  $d_{\text{單位}}(90112501_{\text{資管系所}}, 90101111_{\text{傳播系所}}) = 0.5$ 、單位分別是  $d_{\text{身分}}(90112501_{\text{在職生}}, 90101111_{\text{大學生}}) = 0$ 、借閱數量分別是  $d_{\text{借閱數量}}(90112501_{24}, 90101111_{31}) = \sqrt{(-7)^2} / 100$ ，距離函數可由查表得知，而借閱數量除以 100 的原因是為了配合身分、單位兩項類別變數值的距離都借閱 0 - 1 之間，將之加總  $0.5 + 0 + 0.07 = 0.57$ 。以此類推，計算子群一和其他點的距離，比較距離，由小至大作排序，即成為表 20，子群一代表點 90112501 距離讀者 90101111 最近， $d = 0.57$ ，第二接近的是讀者 90101112， $d = 1.55$ ，如此找出距離子群一最接近的 9 個點(灰

色部分), 子群二...子群 n 也都有距離自身最接近的 9 個點。

以子群一來說, 包含了 90112501 自身和距離最近的 9 個讀者, 每位讀者在探勘資料表中都會有多本曾經借閱過的書目(表 21), 再從這個表中用關聯規則找出相關性高的書籍(表 22); 如此可以得到每個子群都已預先計算好要推薦的書籍。

表 20：子群矩陣表

	讀者證號	距離最接近者 1	距離最接近者 2	...	距離最接近者 9	...	距離最接近者 n
子群 1	90112501	0.57(90101111)	1.55(90101112)	....	6.7(90112521)	....	126.9(87121311)
子群 2	90112503	1.2(90112505)	2(90112501)	....	8.8(90112515)	....	110.3(86112211)
子群 3	90112505	2(90112501)	2.1(90112503)	....	10(90112533)	....	152.1(85142214)
.....	.....	.....	.....	....	.....	....	.....
子群 n	92110000	3(91212313)	3.2(89121341)	....	15(82131111)	....	144.6(87111245)

表 21：同一子群借閱書籍表

	讀者證號	借閱書籍
子群 1	90112501	管理數學,組織行為,教學模式,鬧中取靜
	90101111	記憶術 88 法則,組織行為,創業的智慧
	90101112	管理數學,記憶術 88 法則,組織行為,創業的智慧
	.....	記憶術 88 法則,鬧中取靜
	92112521	競爭優勢,企劃實務

表 22：同一子群推薦資料表

	讀者證號	推薦書籍
子群 1	90112501	( 管理數學 組織行為 )( 記憶術 88 法則 組織行為 )
	90101111	( 灌籃高手 名偵探柯南 )
	90101112	
	.....	
	92110033	( 統計學 財務管理 )( 社會心理學 當代社會理論 )

### 三、模組流程：

互助式資訊過濾模組的流程共分為三大部分，分別說明之：

#### 1.以讀者為中心探勘子分群及推薦書籍

分別以每位讀者做為空間樣本點中的種子樣本點，從探勘資料表中以 MBR 的方法找出每個種子樣本點的鄰近資料做為同一子分群(subgroup)，每筆子分群中都會包含種子樣本點(id)、及距離種子最近的 9 位讀者，然後每一個子分群都後用關聯規則探勘出要推薦的書籍，參考表 22，分群合併程式部分參見圖 22。

```
reclib_mining.MoveFisrt
Do Until reclib_mining.EOF
    reader_id=reclib_mining("reader_id")
    typename=reclib_mining("type_name")
    deptname=reclib_mining("dept_name")
    amount=reclib_mining("amount")
reclib_mining.MoveNext
    recunit(id)=reader_id
    recunit(distance)=abs(typename-reclib_mining("type_name"))
        +abs(deptname-reclib_mining("dept_name"))
        +abs(amount-reclib_mining("amount"))
    For
        Select id,* from recunit order by distance
    End For
Loop
```

圖 22：分群合併程式

#### 2.用 SCSA 演算法合併子分群

運用 SCSA 演算法，將子群和其他子群做重疊程度的比對，當相似程度超過 75%時，我們將兩子分群合併成為一大分群，若相似程度低於 75%，則歸類為另一新的大分群，並給予大分群一

個識別號。如此，相似程度重疊程度高的子群即可歸於同一群，以減少子群的數量。

表 23：系統合併子分群資料表

	讀者證號	距離 1	距離 2	....	距離 10	大分群編號
子群 1	90112501	0.57(90101111)	1.55(90101112)	....	6.7(90112521)	(90112501)
子群 2	90112503	1.2(90112505)	2(90112501)	....	8.8(90112515)	(90112501)
子群 3	90112505	2(90112501)	2.1(90112503)	....	10(90112533)	(92110000)
.....	.....	.....	.....	....	.....	
子群 n	92110000	3(91212313)	3.2(89121341)	....	15(82131111)	(92110000)

從表 23 中可看到縱軸，分為 n 個子群，以橫軸來看，每筆記錄代表一個子群。以「子群一」開始，將子群一內所包含的成員分別和所有 n 個子群內（包含自己）所包含的成員做比對，計算重疊程度，子群一和子群二成員比對重疊程度用  $\text{sim}(90112503/90112501)=90\%$ ，若  $90\% > 75\%$  ( Similar Threshold )，則將子群一和子群二合併成為一個大分群，並給予大分群編號，若是小於 75%，則將子群一和子群二標上不同的大分群編號。接著，子群一和子群三比較，以此類推，直到子群一和子群 n 比較，完成一輪的比對。SCSA 的優點在於只要掃描資料庫一次 ( one pass ) 即完成分群的動作。經過此相似度分類計算，每個子分群都會隸屬於一個大分群編號。

### 3.新進讀者書籍推薦

上述這些動作都在使用者登入前即計算完成，一但有使用者登入系統，只要計算這位讀者樣本點和所有大分群代表點的距離，找出距離最近的大分群，讓此讀者屬於這個大分群；再計算

這位讀者所在大分群內所有子分群質心的距離，找出距離最近的子分群，將模組流程一中所事先計算好子分群中經由關聯規則所探勘出的資料，推薦給讀者。如此一來，讀者搜尋的線上等待時間將會從計算所有空間樣本點的時間降低到只計算一次大分群代表點的最短距離加上和計算大分群下子分群代表點的最短距離的時間。

4.將書籍推薦給讀者

5.結束

## 第二節 系統展示

### 壹、系統建置需求

#### 一、硬體設備

個人電腦 ( PC ) 一部

中央處理器(CPU)：Intel Pentium III 600

記憶體：128MB

硬碟：10GB

#### 二、軟體設備

作業系統：Microsoft Windows 2000 Professiona l

資料庫：Microsoft Access 97 ( 資料連接方式 ADO )

開發軟體：Microsoft Visual Basic 5.0 Enterprise

### 貳、系統功能

本研究的資料來源為南華大學圖書館現行採用電子化資料表，經過系統需求分析後，從中粹取出我們所需要的資料欄位，並分別存入資料探勘所需要的資料表格之中。

## 一、主題式同類書目推薦系統功能說明：

圖書推薦檢索系統是一個人化的檢索系統，主要服務成員為學校內所有教職員及學生。每位成員之個人基本資料檔皆已建構於讀者屬性資料表之中，我們以每位成員的基本資料及過去的借閱紀錄架構個人化環境。在登入系統時都須輸入帳號及密碼，作為個人化環境的區隔。如圖 23：



圖 23：系統畫面一

系統登入之後，主要分為二大功能：主題式同類書目推薦、互助式資訊過濾推薦，參見圖 24。使用者可點選分別進入各個功能。



圖 24：系統畫面二

點選「主題式同類書目推薦」會進入如圖 25 畫面，使用者輸入關鍵字串，系統會在書籍資料庫中做關鍵字搜尋，以書名比對的方式，將所搜尋到的資料依中國圖書分類法做主題分類，並顯示主題讓使用者勾選，如圖 26 中系統會依照「網路應用」關鍵字到資料庫中搜尋，並將所包含的主題顯示出來，使用者可以用多選的方式選擇主題；「進階選取」是讓讀者自行決定要往下一階層主題搜尋，並將下一層中包含有「網路應用」書名的子主題顯示在螢幕上讓讀者選擇要繼續往下一類別搜尋或是直接推薦屬於此主題的書籍，若要往下一層主題搜尋則在勾選之後按下「進階搜尋」；若只要單一主題的推薦書籍，則只要勾選主題，按下「確定」，系統會將屬於讀者勾選的主題下的書籍依熱門借閱程度推薦給讀者，參考圖 27。

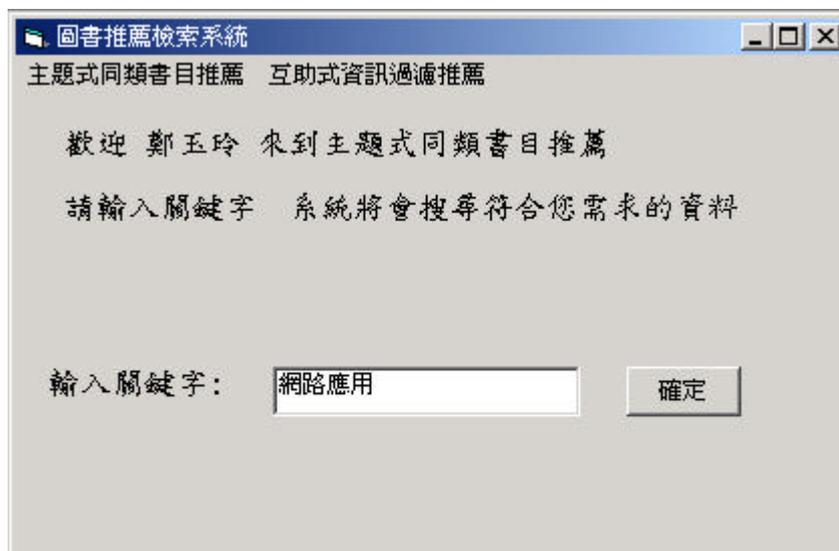


圖 25：系統畫面三

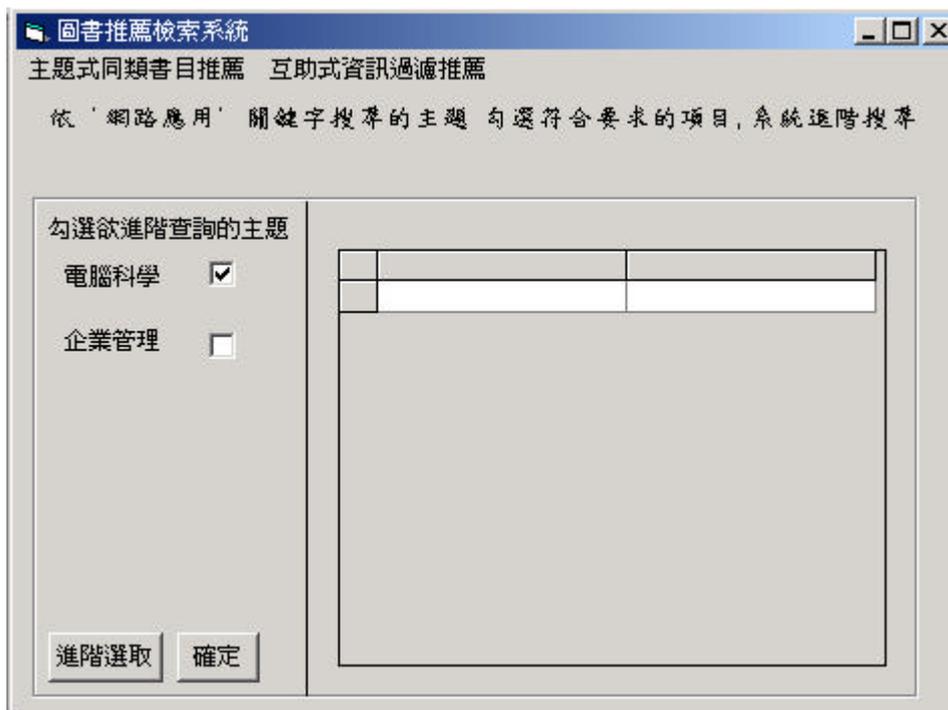


圖 26：系統畫面四

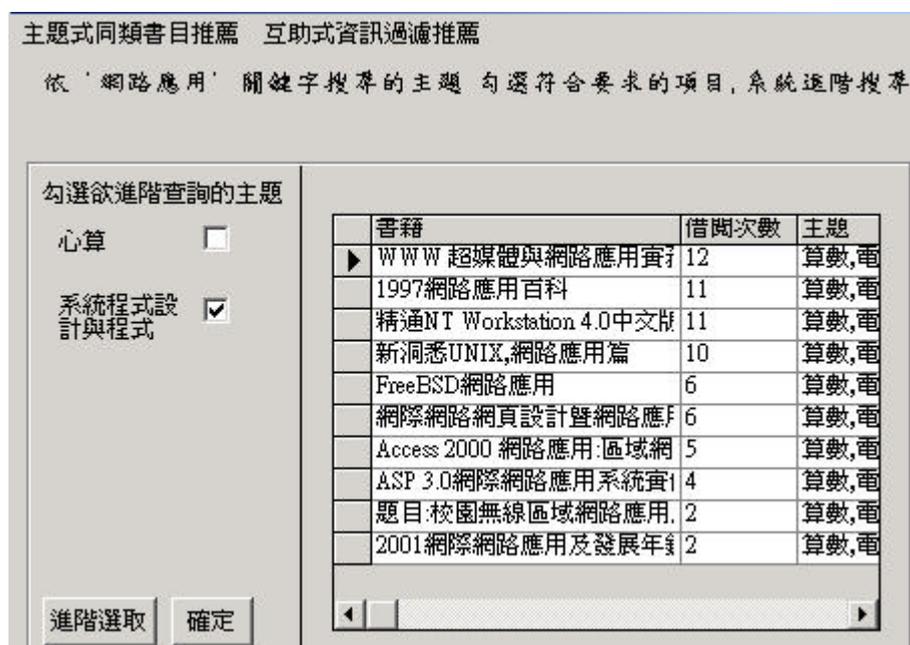


圖 27：系統畫面五

## 二、互助式資訊過濾書目推薦系統說明：

首先，系統會按照本研究所提出的互助式分群方法為每位讀者找出相關性高的書籍，並為資料庫中的所有讀者做大分群歸類，將分群資訊儲存於資料庫中。當讀者進入如圖 23 系統登入畫面後，並點選「互助式資訊過濾推薦」，系統會根據登入時各人的背景屬性，到資料庫中找尋所屬大分群，並找出最接近的人，將其所借閱過的書籍，並推薦給新進讀者，如此，推薦給新讀者的書籍將會是一逼近解，也就是相似度最高，畫面參見圖 28。

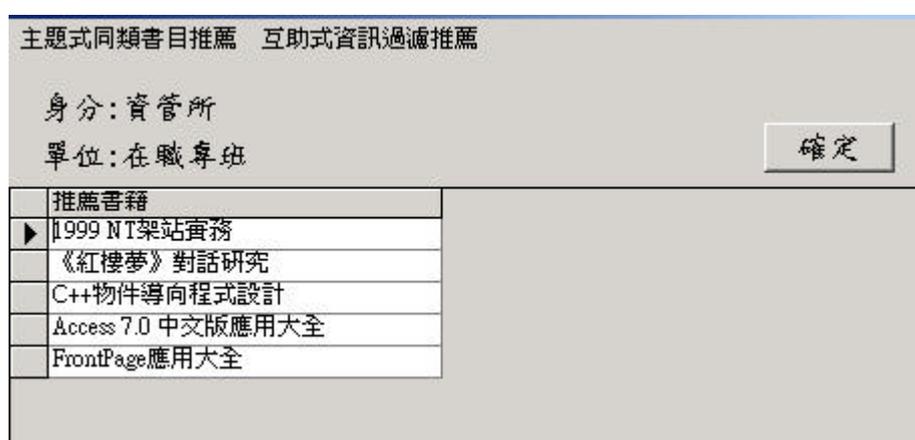


圖 28：系統畫面七

## 參、資料評估

本系統資料評估分為三類，系統準確性評估，比較不分群和分群之間的推薦準確性；模組正確性評估，利用訓練資料集和測試資料集評估模組正確性；系統執行速度評估，比較不分群和分群之後的搜尋速度。

### 一、系統準確性評估：

本研究將不同的讀者資料分別輸入「互助式資訊過濾模組」和「MBR 及關聯法則推薦模組」中做比較。所謂「MBR 及關聯法則推薦模組」是當一新進讀者進入系統後，直接計算新讀者和資料庫中所

有其他讀者的距離，找出符合鄰近資料的群組，同一群組中再找出關聯規則高的書籍，當作推薦的依據。從圖 29 是以一位讀者為例證的表示，其中可以看出同一個讀者證號經過兩個模組的計算並產生推薦書籍中，4 本書中有三本是相同的，代表相似度達到 75%。我們以這樣的方法為 1,000 位讀者做評估，發現所達到的相似度平均達到 83%，代表本研究的方法所找到的逼近解的可靠度相當高，且也改良了 MBR 中耗時的缺點。準確性評估的資料參考附件三。

「互助式資訊過濾模組」	
讀者證號	推薦書籍
90112513	管理資訊系統,管理數學,SQL Server 7.0 設計實務,資料庫系統概論

「MBR 及關聯法則推薦模組」	
讀者證號	推薦書籍
90112513	管理資訊系統,管理數學,SQL Server 7.0 設計實務,鹿鼎記

圖 29：系統準確性評估例證（以單位讀者為例）

## 二、 模組正確性評估：

經過資料清理及轉換的步驟後，借閱紀錄表中的紀錄剩下 169,325 筆，並由系統統計後得知每個科系的借閱紀錄筆數由於背景不同而有所不同；附錄一為以身分和單位統計出的紀錄表數，從表中我們可以清楚的知道不同科系不同身分中借閱的紀錄多寡不同，多如資管系大學部達 19,074 筆，少如只有數十筆，如此系統在設定關聯規則支持度門檻值時容易產生紀錄少之背景無法找出關聯法則，紀錄多之背景有找出太多的推薦書籍，在系統實作上，為避免紀錄數量太少無法探勘出正確資料，我們將筆數低於 500 筆的記錄刪除；驗證方法為將借

閱紀錄依日期分為訓練資料集（1999/10/20 到 2001/12/31 的資料）和測試資料集（2002/1/1~2003/1/24）兩部分，並將訓練集及測試集的內每個背景屬性相同的數量控制在相近的前提中，從訓練資料集中探勘出每位讀者的關聯推薦書籍，也從測試資料集中探勘每位讀者的關聯推薦書籍，並在每個不同身分單位的讀者中隨機抽取 100 位，比對每位讀者在訓練資料集及測試資料集中的差異性，藉以瞭解系統模組之正確性。圖 30 和圖 31 為不同科系單位中，隨機抽 100 位讀者的訓練資料集和測試資料集的重疊程度折線圖(部分)，圖 30 為資料記錄筆數界於 10,000 筆和 20,000 之間，圖 31 為界於 3,000 和 4,000 筆之間，由圖 30 中可以得知重疊程度最低的為社會科學系大學部的學生，只有 52%，而最高為傳播系大學日間部學生，高達 83%；而圖 31 中可以得知重疊程度最低的為美學與藝術系日間部大學生，只有 38%，而最高為社會教育系日間部碩士生，達 69%大體上來看，圖 30 的四個科系的借閱符合程度都界於 50% ~ 75%之間；而圖 31 的借閱符合程度界於 38% ~ 58%，所以說，本模組在紀錄筆數越多的情形之下，訓練及和測試集的符合程度相關性越高，也就是系統的模組正確性越高。附錄二為訓練資料集和測試資料集之關聯法則探勘結果。

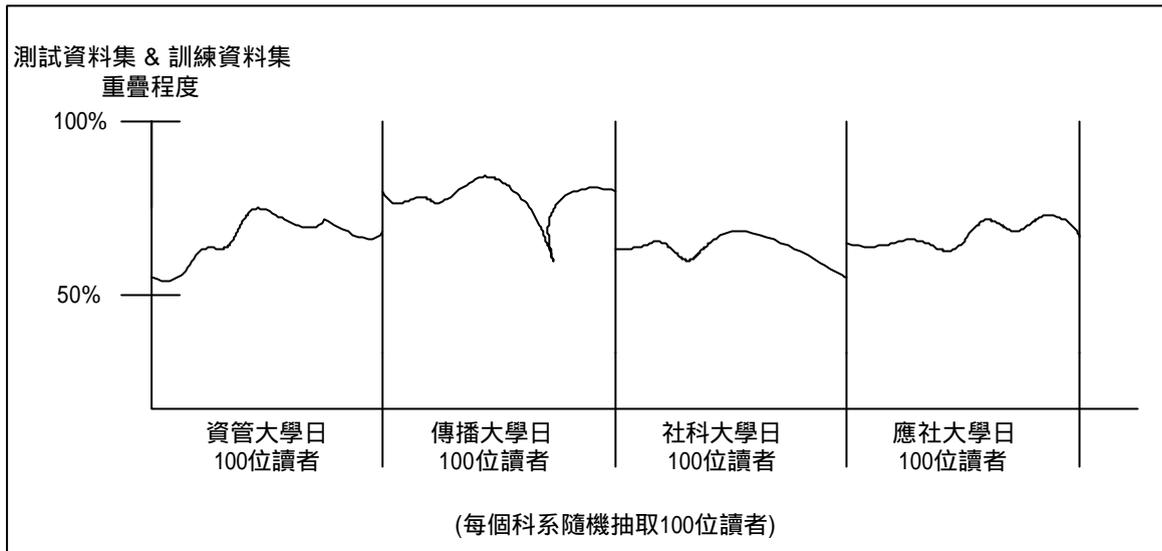


圖 30: 隨機抽取讀者之訓練和測試資料重疊程度圖一(一萬到兩萬筆紀錄)

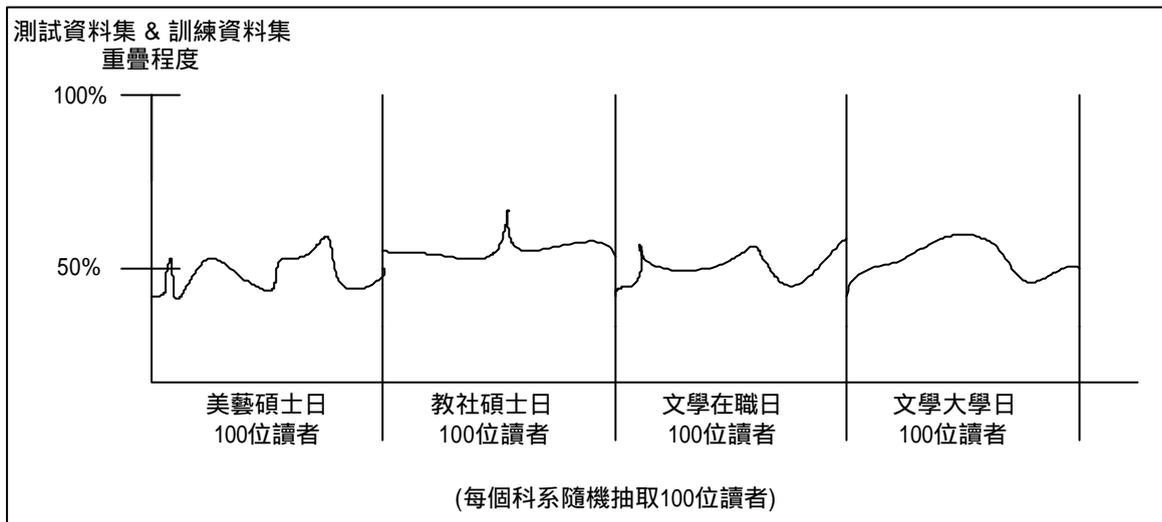


圖 31: 隨機抽取讀者之訓練和測試資料重疊程度圖二(三千到四千筆紀錄)

### 三、系統執行速度評估：

在資料分群的評估之中，我們將本研究所提出的改良式分群探勘法和直接用關聯法則探勘法來做比較。從圖 32 中可得知，改良式分群法探勘法中所用到的時間為正方形的線加上圓形的線。正方形的線代表系統預先分群探勘所耗費的時間，圓形的線代表系統已經做完分群探勘後，讀者實際進入系統搜尋的時間。直接關聯法則探勘法是用中間倒三角形的線代表，也就是讀者進

入搜尋時，系統才從資料庫中探勘出讀者所需的資料。從圖中我們可以看出，採用改良式分群法探勘所花的總時間成本高於直接關聯法則探勘；雖然如此，在個人化推薦檢索系統中，讀者的搜尋等待時間是一個重要的目標，以此觀點來看，則只要比較到三角形和圓形的線，因為正方形的線早在讀者進入系統前已運算完成，由此可知，在資料筆數少的時候，搜尋時間相當接近，但隨著資料筆數的增加，直接關聯法則探勘的時間會隨之增加，而改良式分群法卻一直保持在 7 秒之內完成。

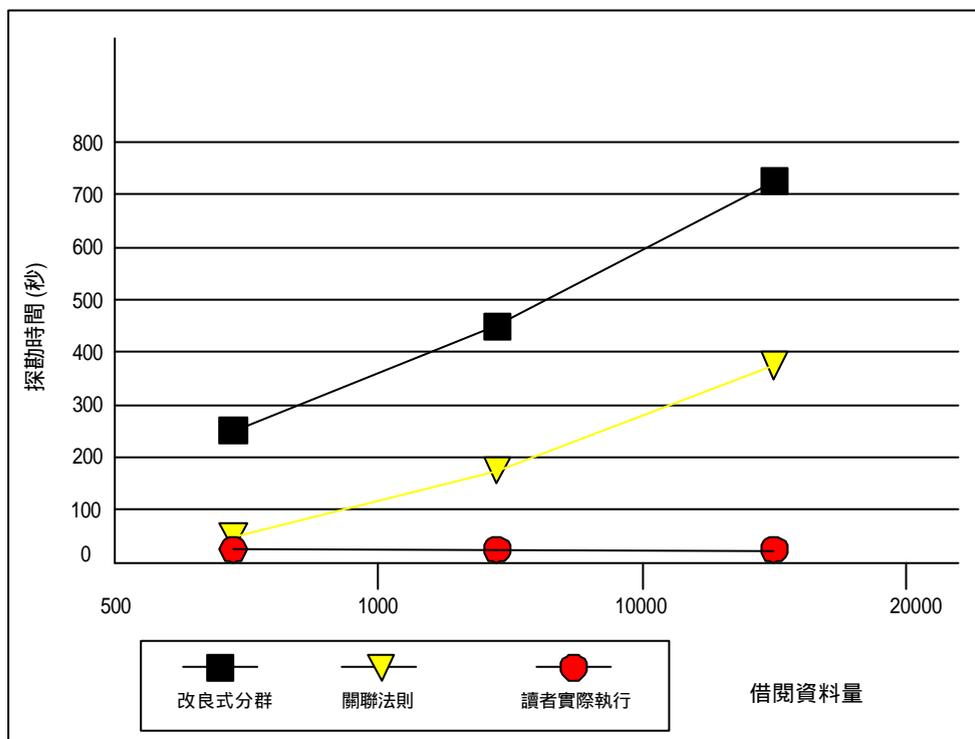


圖 32：改良式分群探勘法和直接關聯法則探勘法比較圖

### 第三節 模組優點

#### 壹、主題式同類書目推薦

1. 中國圖書分類法的分類明確。

2. 透過讀者自行選擇所需的主題，提昇讀者需求的正確性。
3. 系統並會將讀者每次進系統所選擇的資訊紀錄載於借閱記錄檔 (Log)中，此搜尋結果將會儲存於探勘資料表中，並成為系統探勘的一部分。
4. 符合讀者需求的主題書籍，經過熱門排行的篩選，推薦書目是更加符合大眾化興趣的資料。

## 貳、互助式資訊過濾推薦

1. 這個模組運用了個人化資訊環境的方法。
2. 確保使用者的隱私權 (Privacy)，在系統登入時即要求使用者作帳號密碼的確認動作。
3. 運用資料探勘技術結合個人化概念，為圖書館服務提供另一經營概念。
4. 將相近的背景屬性及興趣者曾經借閱過的資料推薦給個人，提高檢索結果的正確性 (Effective Search)。

## 第五章 結論與未來研究方向

本研究中，設計與實作一個人化圖書系統，依讀者身份屬性及借閱行為架構一推薦系統，透過中國圖書分類及資料探勘技術，及讀者個人化資訊，推薦最符合讀者需求的書目；未來，可藉由此系統的使用率及非熱門書目之借閱率來了解此網站的成效如何，以作為改進的參考，讓數位化圖書館品質更加提昇。如此，個人化環境的功能將更完整。

(一) 本研究所達到四項目的：

- 1.提供圖書館使用者一個個人化線上推薦與檢索模組
- 2.提出了一個適用於加速分群的分類方法
- 3.實作一圖書館線上推薦與檢索系統
- 4.提供更符合個人化要求的線上推薦結果

(二) 未來研究方向

- 1.此系統是使用於個人化之單機環境之上，未來可嘗試將單機環境移到Web介面上，或是和現有圖書館之搜尋機制做結合。
- 2.加入興趣分類，當圖書館資料庫成長到一定數量時，可以將興趣做更詳盡的分類，如此系統所推薦的書籍將會更符合讀者需求。
- 3.在個人化環境設定上面，可加上讓讀者自行建構一符合個人化風格及需求的環境。
- 4.運用Improved MBR的分群方法，可以找出同一興趣群中的讀者，那個身分及單位所佔的最多，做為圖書館管理者為不同身分屬性群購書及推薦的參考。
- 5.系統可朝向協助館方了解對於不同主題所感興趣的讀者群，作為行銷服務的參考。

6. 在系統上加上讀者使用評比，讓讀者自行設定變數權重，加強個人化推薦系統的準確度。

## 參考文獻

- [1]卜小蝶,「淺析個人化服務技術的發展趨勢對圖書館的影響」,成功大學圖書館館刊,第二期,八十七年十月。
- [2]曹健華,蘇建源,邱宏彬,「應用資料探勘技術於數位圖書館之個人化服務及管理」,第四屆網際網路應用與發展學術研討會,21-28頁,台南,92年5月。
- [3]彭文正,Data Mining 資料探礦-顧客關係管理暨電子行銷之應用,初版,維科圖書股份有限公司,台北,民國九十年。
- [4]黃夙賢、柯皓仁、楊維邦(2001),交通大學個人數位圖書館資訊環境,新世紀數位圖書館與數位博物館趨勢研討會論文集,交通大學圖書館編印。
- [5]黃智育(2002),「資料探勘於即時線上推薦系統之應用研究」,朝陽科技大學資訊管理系碩士論文。
- [6]湯春枝,從個人化服務行銷的理念談交通大學個人化數位圖書館資訊服務 PIE@NCTU 系統,國立成功大學圖書館館刊。
- [7]楊雅雯(2001),「個人化數位圖書館資訊環境—以 PIE@NCTU 為例」,交通大學資訊科學研究所碩士論文。
- [8]鄭玉玲,張晉赫,邱宏彬,「運用顧客行為智慧實作數位圖書館上個人化之檢索與推薦服務-以南華大學圖書館為例」,第四屆網際網路應用與發展學術研討會,393-398頁,台南,92年5月。
- [9]賴永祥,「中國圖書分類法」,2001 增訂八版,文華圖書管理資訊股份有限公司。
- [10]H. Liu, and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Kluwer Academic Publisher, 1998.
- [11]J. Han, and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000.
- [12]J. T. Tou, and R. C. Gonzalez, Pattern recognition principle, Addison-Wesley, pp. 90-92, 1974.
- [13]MIT media Lab WebSite  
(<http://web.media.mit.edu/~mtu/mab/>)
- [14]M. J. A. Berry, and G. S. Linoff, Mastering Data Mining – The Art and Science of Customer Relationship Management, John Wiley & Sons Inc., pp. 39-61, 2000.
- [15]M. J. A. Berry, and G. S. Linoff, Data Mining Techniques : for marketing, sales, and customer support, John Wiley & Sons Inc., pp. 163-196, 1997.
- [16]National Science Foundation WebSite

- (<http://www.nsf.gov/pubs/1998/nsf9863/nsf9863.htm>)
- [17]P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, and A. Zanasi. Discovering Data Mining From Concept to Implementation, Prentice Hall PTR, 1998.
- [18]R. Agrawal, and R. Srikant. "Fast Algorithms for Mining Association Rules", Proc. of the 20th VLDB Conference, pp. 487-499, 1994.
- [19]R. Agrawal, T. Imielinski, and A. Swami. "Mining association Rules between Sets of Items in Large Databases", Proc. of the 1993 ACW SIGMOD Conference, pp. 207-216, 1993.
- [20]S. Cohen, J. Fereira, A. Home, B. Kibbee, H. Mistlebauer, and A. Smith, "MyLibrary: Personalized Electronic Services in the Cornell University Library", D-Lib Magazine, April 2000.  
(<http://www.dlib.org/dlib/april00/mistlebauer/04mistlebauer.html>)
- [21]Tseng S. M., "Mining Association Rules with Interestingness Constraints in Large Databases," International Journal of Fuzzy Systems, Vol. 3, No. 2, pp. 415-421, June 2001.
- [22]S. Kasif, S. Salzberg, D. Waltz, J. Rachlin, and D. Aha. "A Probabilistic Framework for Memory-Based Reasoning", Artificial Intelligence, pp. 287-311, 1998.
- [23]W. J. Frawley, G. P. Shapiro, and C. J. Matheus, " Knowledge Discovery in Databases: an Overview", AI Magazine, pp. 1-27, 1992.

附 錄 一：不同身分單位紀錄筆數表

科系	身分	借閱紀錄筆數
南華資訊管理學系	本校大學部學生(日)	19074
南華傳播管理學系	本校大學部學生(日)	16955
南華社會科學學院	本校大學部學生(日)	10087
南華應用社會學系	本校大學部學生(日)	10043
南華文學研究所	本校碩士班學生(日)	8220
南華國際關係學系	本校大學部學生(日)	7720
南華生死學研究所	本校碩士班學生(日)	5426
南華歐洲研究所	本校碩士班學生(日)	5289
南華應用藝術與設計學系	本校大學部學生(日)	5099
南華出版學研究所	本校碩士班學生(日)	4945
南華美學與藝術管理研究所	本校碩士班學生(日)	4039
南華教育社會學研究所	本校碩士班學生(日)	3563
南華文學研究所	本校在職碩士專班	3372
南華文學系	本校大學部學生(日)	3318
南華傳播管理學研究所	本校碩士班學生(日)	3182
南華亞太研究所	本校碩士班學生(日)	2998
南華環境與藝術研究所	本校碩士班學生(日)	2745
南華旅遊事業管理研究所	本校碩士班學生(日)	2580
南華哲學研究所	本校碩士班學生(日)	2573
南華圖書館	本校職員	2335
南華資訊管理學研究所	本校碩士班學生(日)	2332
南華生死管理學系	本校大學部學生(日)	2298
南華佛學研究中心	本校碩士班學生(日)	1970
南華哲學研究所	本校在職碩士專班	1963
南華傳播管理學系	本校二年制學士專班	1879
南華非營利事業管理研究所	本校碩士班學生(日)	1817
南華企業管理學系	本校大學部學生(日)	1784
南華社會學研究所	本校碩士班學生(日)	1748
南華公共行政與政策研究所	本校碩士班學生(日)	1734
南華電子商務管理學系	本校大學部學生(日)	1712
南華宗教學研究所	本校碩士班學生(日)	1704
南華環境管理研究所	本校碩士班學生(日)	1671
南華生死學研究所	本校在職碩士專班	1360
南華民族音樂學系	本校大學部學生(日)	1142
南華大學部(不分系)	本校大學部先修生	1058
南華學務處	本校職員	1038

## 附錄二：身分類別相關法則探勘結果（部分）

借閱紀錄量：19,074

身分類別：資管系大學日間部學生

館藏書名	借閱次數
網頁設計快易通 網頁設計經典	32
福爾摩斯一 福爾摩斯二	30
福爾摩斯辦案記 福爾摩斯一 福爾摩斯二	29
管理資訊系統 管理資訊系統導論	28
微積分:原理及題解 微積分入門	28
網頁設計經典 網頁設計快易通 網頁設計寶典:FrontPage 2002	27
微積分入門 微積分 微積分:原理及題解	27
管理資訊系統 管理資訊系統實務 管理資訊系統導論	20
大地瑰寶:100 個您必遊的名景 優遊台灣導覽手冊	19
德川家康全傳 神雕俠侶	18
灌籃高手 中華小廚師	16

### 附錄三：系統準確性評估（部分）

#### 主題式同類書目推薦 ■ MBR + 關聯法則

讀者證號	推薦書籍
87101213	多情劍客無情劍, 白玉老虎, 不確定的年代, 浣花洗劍錄
87101213	多情劍客無情劍, 白玉老虎, 不確定的年代, 經濟學
87111417	Flash 3 躍動的網頁, 網路概論, 最新 HTML 入門與應用白皮書
87111417	Flash 3 躍動的網頁, 網路概論, 最新 HTML 入門與應用白皮書
87121321	會計學原理, 紅樓夢校注, 中國古代神話, 雍正皇帝: 九王奪嫡
87121321	會計學原理, 紅樓夢校注, 中國古代神話, 雍正皇帝: 九王奪嫡, 三國演義校注
88111312	碧血劍, 俠客行, 射雕英雄傳, 笑傲江湖
88111312	碧血劍, 俠客行, 射雕英雄傳, 笑傲江湖, 鹿鼎記, 連城訣
88121011	C 語言程式設計基礎, C & C++ 程式設計, C 語言程式設計入門: 使用 Turbo C
88121011	C 語言程式設計基礎, C & C++ 程式設計, C 程式語言設計
88121123	天魔刀, 離島醫生, 白色巨塔, 小王子, 音樂欣賞
88121123	天魔刀, 離島醫生, 白色巨塔, 小王子
88121025	全面優質管理, 窗邊的小荳荳, 幽靈山莊, 水滸英雄傳, 神話的智慧: 時空變遷中的神話
88121025	全面優質管理, 窗邊的小荳荳, 幽靈山莊,
89112513	第三謊言, 鹽田兒女, 蕭十一郎, 競爭優勢
89112513	第三謊言, 鹽田兒女, 蕭十一郎, 環境與人
89123312	管理經濟學, 專案管理, 你是 EQ 高手嗎, 原子空間
89123312	管理經濟學, 專案管理, 你是 EQ 高手嗎, 行政學
89132111	大自然的數學遊戲, 生產與作業管理, 愛回到最初, 音樂與心靈
89132111	大自然的數學遊戲, 生產與作業管理, 愛回到最初, 人文教育
90103125	蘇東坡全集, 宋詩話全編, 中國文學批評史, 中國詩學
90103125	蘇東坡全集, 宋詩話全編, 中國文學批評史, 沈從文文集
90231517	冷戰後美國的東亞政策(1989-1997), 後冷戰時期亞太集體安全, 大地紀行
90231517	冷戰後美國的東亞政策(1989-1997), 後冷戰時期亞太集體安全, 兩岸關係變遷史
90112519	放眼新世界, 中國關係的軌跡, 政治學, 新戰爭論, 國際組織法
90112519	放眼新世界, 中國關係的軌跡, 政治學, 新戰爭論, 亞洲的安全挑戰
90132423	社會學理論, 社會學, 當代社會學理論, 蘇菲的世界
90132423	社會學理論, 社會學, 當代社會學理論, 大地記行
91101234	哲學概論, 人類學導論, 社會學
91101234	哲學概論, 人類學導論, 死亡教育概論

