利用群組發掘書籍最適性之推薦

陳垂呈

副教授
南台科技大學資訊管理系
E-mail: ccchen@mail.stut.edu.tw

黃俊榮

研究生
南台科技大學資訊管理研究所
E-mail: n9090020@webmail.stut.edu.tw

摘要

本篇論文藉由讀者之借閱資料為探勘的資料來源，每一筆借閱資料包含有讀者曾借閱過的書籍項目，利用群組(clusters)。從以下兩方面來發掘書籍最適性的推薦：一是以某一讀者為探勘的目標，並設定此一讀者之借閱資料為一群組的中心點，提出一個分群化方法，將與中心點滿足最小借閱相似度的借閱資料，歸屬於同一群組中。根據群組所顯示的借閱傾向特徵，可發掘此一讀者最適性的書籍推薦；二是以某一書籍為探勘的目標，並設定此一書籍為一群組的中心點，提出一個分群化方法，將包含有此一書籍的借閱資料，歸屬於同一群組中。在此一群組中，計算出此一中心點的關聯因子，根據借閱資料與關聯因子間的借閱相似度，可發掘此一書籍做適性借閱的讀者。根據所提出的方法，設計與建置一個發掘書籍最適性的推薦系統。此探勘結果，對圖書館在規畫最適性的書籍推薦服務時，可提供非常有用的參考資訊。

關鍵詞：資料探勘，群組，借閱資料，書籍推薦

簡　介

由於資訊技術及網際網路(internet)的支援，促進了圖書館提供數位化的資訊服務，電子圖書、儲存光碟、多媒體等科技的出現，也帶動了資訊儲存及檢索的新紀元。圖書館改變以往傳統搜尋書籍資料的方式，利用國內外網路上豐富的資源，配合各種電子媒介的輔助，使得讀者能以最少的時間及最便利的方式，即可享受到圖

書館最大的服務效益。但如何將傳統圖書館被動的服務方式，轉變成以主動積極、個人化及適性化的方式來吸引讀者到館借閱，進而提昇讀者的借閱率及圖書館的利用率，是圖書館管理者必須探討的課題之一。

在圖書館服務的項目中，以讀者個人化為中心的服務理念，是針對讀者在需求資訊的差異性下，善用個人化服務技術來調整圖書館的資訊服務，把最貼切的館藏資訊主動傳達給讀者個人，進而提昇館藏資料的利用率和圖書館的經營績效，是圖書館服務行銷重要的方式之一（卜小蝶，1998；Ou, Lin and Li, 2001）。辜曼蓉（1999）曾指出圖書館經營的核心價值在於顧客（讀者）和服務人員之間的互動，而館藏和資訊的傳播則扮演提昇圖書館服務品質的輔助角色。

目前，由國立交通大學所建置的個人化數位圖書館資訊服務，是國內各大專院校中，提供圖書館個人化資訊服務最具有代表性的系統之一（湯春枝，2002; http://mylibrary.e-lib.nctu.edu.tw/）。在系統所提供的「個人化推薦」功能中，其分別利用資料探勘技術（data mining techniques）中的關聯規則（association rules）及次序相關（sequences），來提供讀者個人化書籍推薦的服務。因此，將最貼切的書籍資訊推薦給讀者個人，以減少讀者搜尋書籍資料的時間，及提昇書籍的借閱率與利用率，是圖書館在規畫讀者個人化資訊服務中最重要的功能之一。

圖書館中每天均有相當大量的書籍被借閱，在讀者曾經借閱過的書籍資料中，往往隱藏著書籍之間的關聯性，也反映出讀者閱讀的傾向性。例如，讀者曾借閱「電子商務經營策略」的書籍，則往往也會有借閱「網路行銷」書籍，或一些與「網際網路」相關書籍的傾向特徵。因此，如何從累積數量龐大的借閱資料裡，找出對讀者有用的資訊或其他知識，是圖書館管理者必須思考的問題之一。

資料探勘是從大量資料中挖掘潛在有用的資訊與知識，以提供決策者最適當的參考資訊。目前資料探勘已應用於許多領域（Han and Kamber, 2000）。在本篇論文，我們以讀者之借閱資料為探勘的資料來源，每一筆借閱資料包含有讀者曾借閱過的書籍項目，並分別以某一讀者及某一書籍為探勘目標，利用群組（clusters）從以下兩方面來發掘書籍最適性的推薦服務：

㈠發掘某一讀者最適性的書籍推薦：假設目前欲探勘之讀者R的借閱資料為X；X為包含有一個或以上書籍的項目組，並設定借閱資料X為一群組的中心點。我們比對其他借閱資料與X之間的借閱相似度，將滿足「最小借閱相似度」的借閱資料，歸屬於R-群組中。然後，從R-群組中找出所指定之出現次數最大的前k個書籍項目，$k \geq 1$，且找出的書籍項目必須不包含於X中（由於只推薦未曾借閱的書籍項目），藉此做為發掘此一讀者最適性之書籍推薦的依據。

㈡發掘某一書籍最適性借閱的讀者：假設目前欲探勘之書籍為A，並設定書籍A為一群組的中心點。我們檢查其他借閱資料中是否包含有書籍A，若有則將借閱資料歸屬於A-群組中。然後，從A-群組中找出所指定之出現次數最大的前k個書籍項目，$k \geq 1$，且找出的書籍項目必須不包含A，以找出那些書籍項目與書籍A之間

具有同時出現的關聯性，藉此做為發掘此一書籍最適性借閱之讀者的依據。

根據所提出的方法，以南部某一科技大學圖書館的借閱資料為例，設計與建置一個書籍最適性推薦系統，並評估探勘結果的效益。此探勘結果，對圖書館在發掘讀者個人化最適性的書籍推薦，或發掘書籍最適性借閱的讀者，都可提供非常有用的參考資訊。

本篇論文的架構如下：下一節中，將介紹資料探勘技術，及其在圖書館服務之應用的相關研究；第三節中，以某一讀者之借閱資料為探勘的目標，設計一個分群化方法來發掘此一讀者最適性的書籍推薦；第四節中，則以某一書籍為探勘的目標，設計一個分群化方法來發掘此一書籍最適性借閱的讀者；第五節中，再根據所提出的方法，設計與建置一個書籍最適性推薦系統，並以探勘某一科技大學圖書館的借閱資料為例，評估所設計之方法的效益；最後於第六節中做一結論。

二、相關研究

資料探勘可以用來進行大量資料的處理，完成如下的任務：關聯規則、分群、分類、次序相關分析（sequential pattern analysis）及預測等（Chen, Han and Yu, 1996）。其探勘結果對企業在從事行銷決策及市場預測等活動時，已被證實可以有效地提供非常有價值的參考資訊（Berry and Linoff, 2004）。對於圖書館的書籍借閱而言，讀者往往必須在龐大的書籍資料中，找尋其有興趣或想借閱的書籍資料，而圖書館卻只能被動地等待讀者來借閱書籍。如此結果，不僅造成讀者搜尋書籍資料的困擾，書籍資訊無法即時的傳達，也造成書籍的借閱率不佳。

目前，在探討利用資料探勘技術於圖書館經營服務中，已有許多相關研究相繼被提出。其中陳慶瑄（2000）曾探討以k-means的方法來形成學習社群，對電子圖書館之個人化服務可提供輔助的功能；鄭玉玲（2003）曾利用資料探勘的關聯規則，來建構一個在數位圖書館上提供讀者個人化的檢索與推薦系統；汪軒楷（2003）曾利用階層分析法（analytic hierarchy process）來建構個人化書籍推薦的網站系統，並對受測者進行系統功能的評估，有不錯的整體滿意度；杜逸寧（2005）曾結合分群法（clustering）與案例式推論（case-based reasoning）作為分類的機制來建構一個論文推薦系統，對使用者在檢索論文時，可提高檢索資訊的正確度。孫冠華（2003）曾利用關聯規則作為建構數位圖書館之個人化服務及管理的方法依據；吳安琪（2001）曾利用資料探勘技術來找出讀者間的社群關係，進而提昇圖書館的借閱率及讀者的忠誠度；洪志淵（2001）曾利用資料探勘技術來找出讀者與書籍之間一般化的關聯規則，藉此做為新書之讀者推薦的依據。

分群化是將物件根據相似度來進行分群，關於分群化的研究，主要可分以下幾種：分割式（partitioning）、階層式（hierarchical）、格子基礎（grid-based）、密度基礎（density-based）與模型基礎（model-based）等幾種（Han and Kamber, 2000）。在本篇

論文中，我們將修改分割式分群化的方法，做為分群化借閱資料的方法依據。

在眾多分割式分群化演算法中，較著名的有PAM（Kaufman and Rousseeuw, 1990）、k-means（Alsabti, Ranka and Singh, 1998; Dubes and Jain, 1988）及CLARANS（Ng and Han, 1994）等，其目的是分群成使用者所指定的k個群組，$k \geq 1$，此分割方式可將每一物件歸屬於最相似的群組中。以下介紹PAM（Partitioning Around Medoids）演算法的分群化步驟。

PAM演算法由Kaufman and Rousseeuw（1990）所提出，為了將全部物件分群成k個群組，PAM的方法是先為每個群組決定一個代表物件（representative objects）。此代表物件稱之為*medoid*，一旦把k個medoids選定之後，就依據相似度來決定非medoid物件是屬於那一個群組，其相似度是以物件彼此之間的距離（Euclidean distance）來表示，$d(O_a, O_b)$表示物件O_a與O_b之間的距離。例如O_i為medoid，而O_j為非medoid物件，如果$d(O_j, O_i) = min\{d(O_j, O_e)\}$，$O_e$表示所有的medoids，則$O_j$歸屬於$O_i$群組。

對任一個非medoid物件O_j而言，當一個medoid O_i被一個非medoid物件O_h取代時，所造成的改變成本C_{jih}定義如下：

$$C_{jih} = d(O_j, O_p) - d(O_j, O_q)$$

O_p表示以O_h取代O_i之後，與O_j有最大相似度（最短距離）的medoid；

O_q表示以O_h取代O_i之前，與O_j有最大相似度（最短距離）的medoid。

以O_h取代O_i成為medoid之後，所造成的總改變成本為：

$$TC_{ih} = \sum_j C_{jih}$$

若$TC_{ih} > 0$時，表示以O_h取代O_i之後的總距離比取代前大，則O_i將不會被O_h所取代。以TC_{ih}為衡量依據，PAM演算法說明如下：

Algorithm PAM

1. 任意選取k個物件做為medoids。

2. 對所有O_i與O_h之組合，計算出其TC_{ih}，其中O_i表示任一個的medoid，O_h表示任一個非medoid物件。

3. 選擇出TC_{ih}為最小值的O_i與O_h配對；假如$TC_{ih} < 0$，則以O_h取代O_i成為medoid，並跳至2.。

4. 否則停止執行，已完成分群。

在以上分群化的探勘過程，假設共有m個物件，欲分群化成k個群組，由於須要考量每一個物件都有機會成為medoid，因此共有C_k^m種medoids的組合須做考量。而計算每一種medoids組合的改變成本就須掃瞄全部物件一次，因此共須要掃瞄全部物件C_k^m次。

在考量探勘的執行效率及推薦的差異性，則將利用群組分別從以下兩方面來探討如何發掘書籍最適性的推薦服務：一是以某一讀者之借閱資料為群組中心點，設

計一個分群化方法，將與中心點具有滿足所設定之最小相似度的借閱資料，歸屬於同一群組中，並藉由群組所顯示出的傾向特徵，來發掘此一讀者最適性的書籍推薦；二是以某一書籍為群組中心點，設計一個分群化方法，將包含有此一書籍的借閱資料，歸屬於同一群組中，並藉由群組所顯示出的傾向特徵，來發掘此一書籍最適性借閱的讀者。

三、發掘讀者個人化最適性之書籍推薦

在此一章節，以讀者之借閱資料為探勘的資料來源，每一筆借閱資料包含有讀者曾借閱過的書籍項目，並以某一讀者之借閱資料為探勘目標。我們設計一個簡單，但快速的方法來分群化讀者的借閱資料，並從分群化後之群組所顯示出的傾向特徵，做為發掘此一讀者個人化最適性之書籍推薦的依據。此章節共分為兩小節：㈠小節說明發掘讀者個人化最適性之書籍推薦的探勘過程；㈡小節以一實例做說明。

(一)分群化之探勘方法

在分群化借閱資料的過程中，我們以欲探勘之讀者的借閱資料為群組的中心點，然後計算借閱資料與中心點之間的借閱相似度，將具有滿足「最小借閱相似度」的借閱資料，歸屬於同一群組中。假設Y及Z為兩筆借閱資料，Y、Z分別為包含有一個或以上書籍項目所形成的項目組，其中Z為群組的中心點，並定義此兩筆借閱資料之間的借閱相似度d為：

借閱相似度d=（借閱資料Y與Z之間有相同書籍項目的個數）／（借閱資料Z的書籍項目個數），當借閱相似度愈大，表示借閱資料Y包含有愈多與Z相同的書籍項目。

從計算兩筆借閱資料的借閱相似度，可得知兩位讀者是否具有相近的借閱傾向，並且也反映出在借閱資料Y中不包含於Z的書籍項目，其與群組中心點之讀者具有關聯性的潛在傾向。我們將兩筆借閱資料中的書籍項目直接進行比較計算，可以很有效率地得到兩筆借閱資料Y與Z之間的借閱相似度。我們定義一函數*get-item*(Y, i_1)表示擷取借閱資料Y中第i_1個的書籍項目。例如，Y={ABC}，A、B、C∈ 書籍項目，則*get-item*$(Y, 2)$ = B。假設每一筆借閱資料中所包含的書籍項目都已事先由小到大的排序過，例如A < B < C，因此計算兩筆借閱資料Y與Z之間的借閱相似度d，可表示成以下的演算法：

Float ***Per_Same-Item(Y, Z)*** {
int *same_item* = 0; /* 相同書籍項目的數量變數 */
int $i_1 = i_2 = 1$; /* 表示借閱資料Y中第i_1個書籍項目，及Z中第i_2個書籍項目 */
while (*get-item*(Y, i_1) <> ∅) and (*get-item*(Z, i_2) <> ∅) {

```
if (get-item(Y, i1) == (get-item(Z, i2) {
    same_item++;
    i1++;
    i2++;
  }
  elseif (get-item(Y, i1) > (get-item(Z, i2)
    i2++;
  else i1++;
}
return same_item/|Z|; /*|Z|為借閱資料Z的書籍項目個數*/
}
```

例如，Y = {BCE}及Z = {ABD}，A、B、C、D、E∈ 書籍項目，經由上述演算法的計算，其借閱相似度d = 1/3 = 33%。

假設目前欲探勘之讀者R的借閱資料為X，X為包含有一個或以上書籍項目所形成的項目組，設定X為一群組的中心點，並設定一個「最小借閱相似度」值。依據之前所定義的借閱相似度d，分別計算其他借閱資料Y_j與中心點X之間的借閱相似度，並將具有滿足「最小借閱相似度」的借閱資料Y_j，歸屬於同一群組中，稱之為R-群組，$1 \le j \le m$，表示共有m筆的借閱資料，分群化的過程可表示成以下演算法：

```
Clustering-1(X) {
  for (j = 1; j≤ m ; j++)
    if Per_Same-Item(Yj, X) ≥ 最小借閱相似度
      Yj ∈ R-群組;
}
```

例如，假設所設定的「最小借閱相似度」為70%，則與中心點X具有70%或以上借閱相似度的借閱資料，就歸屬於R-群組中。經由上述演算法的分群化步驟，即可將具有滿足最小借閱相似度的借閱資料，歸屬於R-群組中。借閱資料經由以上分群化之後，可在R-群組中找出群組的借閱傾向特徵，藉此做為發掘此一讀者最適性之書籍推薦的依據，其定義如下：

讀者最適性的書籍推薦=max{計算R-群組中各書籍項目出現的次數，並且為此一讀者未曾借閱過}。

從R-群組中，可以藉由群組所顯示出的傾向特徵，找出此一讀者未曾借閱過且出現次數最大的書籍項目，以做為發掘此一讀者最適性之書籍推薦的依據。由於在R-群組中的借閱資料與中心點X具有滿足最小借閱相似度的特徵，也就是說，借閱資料中其他非包含於X的書籍項目與中心點X之間具有潛在的關聯性。因此，若能在R-群組中找出非包含於X且出現次數最大的書籍項目，其可顯示出此書籍項目與中心

點X之間具有最大的關聯性，中心點之讀者借閱此書籍項目的傾向也最大。此外，也可依需求來調整書籍推薦的個數，即在R-群組中，找出此一讀者未曾借閱過且出現次數最大的前k個書籍項目，$k \geq 1$。從以上探勘演算法的計算過程中，在最差的情況下，只需要掃瞄借閱資料庫2次的計算。

(二)實例說明

茲以一實例來說明發掘某一讀者最適性之書籍推薦的探勘過程，表1為一借閱資料庫，其包含有5筆的借閱資料，{A, B, C, D, E}表示全部書籍項目的集合，{R_1, R_2, R_3, R_4, R_5}表示全部借閱資料的集合。假設目前欲探勘之讀者的借閱資料編號為R_5，其包含的書籍項目為CE，設定「最小借閱相似度」為50%，發掘此一讀者最適性之書籍推薦的探勘過程說明如下。

表1　借閱資料庫

借閱資料編號	書籍項目
R_1	ABD
R_2	BE
R_3	ACE
R_4	BCE
R_5	CE

以讀者R_5之借閱資料{CE}為一群組的中心點，經由*Clustering*-1演算法的計算，可得到以下的R_5-群組：

R_5-群組 = {R_2, R_3, R_4}。

在R_5-群組中非包含於{CE}且出現次數最大的書籍項目= max{A=1, B=2} = B。因此，從分群化之後的R_5-群組中，可發掘B為此一讀者最適性的書籍推薦。

四、發掘書籍最適性借閱之讀者

在此一章節，仍以讀者之借閱資料為探勘的資料來源，並以某一書籍項目為探勘的目標。我們設計一個簡單、但快速的方法來分群化讀者的借閱資料，並從分群化後之群組所顯示出的傾向特徵，做為發掘此一書籍項目最適性借閱之讀者的依據。此章節共分為兩小節：㈠小節，說明發掘某一書籍項目最適性借閱之讀者的探勘過程；㈡小節則以一實例做說明。

(一)分群化之探勘方法

在分群化借閱資料的過程中，以欲探勘之書籍項目為群組的中心點，然後檢查借閱資料是否包含有此一書籍項目，若有則將借閱資料歸屬於同一群組中。假設目

前欲探勘之書籍項目為A，並設定書籍A為一群組的中心點，共有m筆的借閱資料$Y_1, Y_2, \cdots, Y_m$，分群化的過程可表示成以下演算法：

***Clustering*-2(A)** {
for (j=1; $j \leq$ m; j++)
if $A \subseteq Yj$
$Yj \in A$-群組;
}

經由上述演算法的分群化步驟，即可將包含有書籍A的借閱資料歸屬於A-群組中，然後找出A-群組所顯示出的借閱傾向特徵。在找出書籍A最適性借閱的讀者之前，必須從A-群組中找出與書籍A具有最大關聯性的書籍項目，稱之為「關聯因子」，其定義如下：

關聯因子= max{計算A-群組中各書籍項目出現的次數，並且非書籍A}。

從A-群組中，可藉由群組所顯示出的傾向特徵，找出非書籍A之出現次數最大的書籍項目，以做為發掘書籍A之關聯因子的依據。由於在A-群組中的借閱資料都具有包含書籍A的借閱特徵，也就是說，借閱資料中的其他書籍項目與中心點之間具有潛在的關聯性。因此，若能在A-群組中找出非書籍A且出現次數最大的書籍項目，其可顯示出此書籍項目與中心點A之間具有最大的關聯性，即中心點A與此書籍項目會同時出現在借閱資料中的傾向也最大。我們也可依需求來調整「關聯因子」的個數，即在A-群組中，找出非書籍A且出現次數最大的前k個書籍項目，$k \geq 1$。接下來，根據「關聯因子」與書籍A之間具有關聯性的特徵，定義借閱資料與「關聯因子」之間的借閱相似度為：

書籍A的借閱相似度=(借閱資料∩關聯因子)的書籍項目個數／關聯因子的書籍項目個數。

根據以上的定義，設定一個「最小借閱相似度」，做為發掘書籍A最適性借閱之讀者的依據，其定義如下：

書籍A最適性借閱的讀者= 在借閱資料中包含的「關聯因子」滿足最小借閱相似度、並且未曾借閱過書籍A的讀者。

根據以上定義所顯示出的借閱傾向特徵，可做為發掘書籍A最適性借閱之讀者的依據。從以上探勘演算法的計算過程中，在最差的情況下，仍只需要掃瞄借閱資料庫2次的計算。

(二)實例說明

仍以表1之借閱資料庫為例，來說明發掘某一書籍最適性借閱之讀者的探勘過程。假設目前欲探勘之書籍項目為B，設定「關聯因子」的個數為1，借閱相似度為100%，發掘此一書籍最適性借閱之讀者的探勘過程說明如下。

以書籍B為一群組的中心點，經由*Clustering-2*演算法的計算，可得到以下的B-群組：

B-群組 = $\{R_1, R_2, R_4\}$。

在B-群組中書籍B的關聯因子= max\{A=1, C=1, D=1, E=2\}=E。因此，在借閱資料中包含有書籍E且未曾借閱過書籍B的讀者有：R_3、R_5。

五、發掘書籍最適性推薦系統之設計與實作

茲將前面章節所描述的探勘方法，設計與建置一個發掘書籍最適性推薦的探勘系統，系統探勘過程的模型如圖1，表2為探勘系統的開發平台。

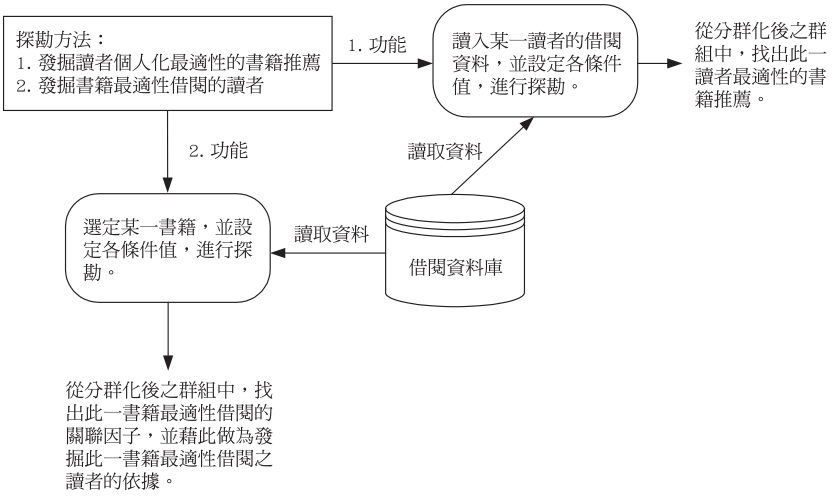

圖1　系統探勘過程模型

表2　系統開發平台

作業系統	Windows XP Professional Edit
CPU	Intel Pentium-4 1.7GHz
主記憶體	512M SDRAM
程式語言	C#
網頁設計	ASP.NET
資料庫	Access 2003

且以南部某一科技大學圖書館之讀者的借閱資料為例，共有2000、2001、2002、2003、2004，及2005等6年的借閱資料，各年份曾借閱過書籍之讀者人數分別為967、2172、4424、7050、9350，及8666，做為所設計探勘方法的資料來源。以前5年（2000-2004）讀者之借閱資料做為探勘計算的訓練資料，若去除重複的讀者，則共有16033位不同的讀者。而以最後一年（2005）讀者之借閱資料做為探勘計

算的驗證資料，其中在前5年有出現的讀者共有6532位。在書籍借閱方面，前5年（2000-2004）曾被借閱過之不同書籍的數量分別為6494、15196、23638、34232，及40255，若去除重複的書籍，則共有73305本不同的書籍。而最後一年（2005）曾被借閱過之書籍的數量為37850，其中在前5年有出現的書籍共有26368本。

圖2為借閱資料的原始資料，包含有書籍的「條碼號」、「讀者編號」、「借閱日期」、「借閱時間」、「歸還日期」、「歸還時間」、及「書名」等欄位資料，這些原始資料是以每一本書籍為一個記錄來儲存。因此，在探勘計算之前，須先將屬於同一讀者的記錄彙整成一筆借閱資料。

條碼號	讀者證號	借閱日期	借閱時間	歸還日期	歸還時間	書名
A158804	98990003	20000221	150332	20030527	144154	布瓜的世界=Pourquoi/幾米著
A143428	T0500114	20000221	150042	20030527	152334	WINDOWS2000Professional使用手冊/施威銘研究室著
A103192	08110001	20000301	200515	20000419	142540	NT Server 4.0網路資源手冊/浦勒斯(Microsoft Press)原著;劉剛樑翻譯
A112357	02000277	20000301	200643	20000303	193935	SQL Server 7.0設計實務/施威銘研究室著
A121542	08110001	20000301	201526	20000330	243256	Visual InterDev 6.0Web應用系統輕鬆學習手冊/陳宗興編著
A115844	02000349	20000301	195753	20000306	102525	Engineering problem solving with MATLAB/Delores M. Etter
A074070	02000175	20000301	202018	20000316	151712	九六文錄:中國人文探索/許倬雲著
A105148	08110001	20000301	200735	20000313	130926	PC DIY網路自己裝/施威銘研究室著
A103233	08109010	20000301	201601	20000425	163358	WEB 的設計藝術/吳耀光，黃輝凱，黃冠諭著
A109562	02000277	20000301	200620	20000428	91917	Outlook 98全新出擊:通往資訊世界的一扇窗/王仲麒編著
A115982	02000349	20000301	200043	20000306	102525	Introduction to applied fuzzy electronics/Ahmad M. Ibrahim
A117732	08408015	20000301	201120	20000315	140535	Optical physics/S. G. Lipson;H. Lipson;D. S. Tannhauser
A120716	08110001	20000301	200411	20000313	130926	MCSE微軟資格認證指導手冊:Windows NT Server 4.0/Joe Casad,Thomas
A115367	02000934	20000301	201056	20000317	164336	Production & inventory management/Donald W. Fogarty;John H. Blackstor
A103555	02000349	20000301	200013	20000306	102525	Genetic algorithms + data structures=evolution programs/Zbigniew Michale
A112400	08110001	20000301	201449	20000419	142540	UML精華:應用標準物件模式語言/Martin Fowler, Kendall Scott原著;許鈞
A099548	08205001	20000301	201727	20000302	90416	一分鐘時間管理/布蘭查(Kenneth Blanchard)等著;官如玉譯
A111293	02000277	20000301	194951	20000302	174151	1999 NT 架站實務/施威銘研究室著
A103595	02000934	20000301	195123	20000306	103803	Balancing and sequencing of assembly lines/薛爾(Armin Scholl)
A021735	02000200	20000301	201843	20010418	142459	工程材料學/金重勳著
A074633	02000200	20000301	202057	20000322	154222	二二八學術研討會文集(1991)/二二八民間研究小組等編著
A119131	08110001	20000301	201510	20000419	142540	Visio 5.0中文版輕鬆學習/馬孝瑀,蔣益寶著
A121357	07208002	20000301	201014	20000306	102210	SPSS For Windows統計分析:初等統計與高等統計/張紹勳,林秀娟著
A023536	02000175	20000301	201750	20000301	201759	概率萬花筒/會蘭英發行
A023536	02000175	20000301	201806	20000301	201822	概率萬花筒/會蘭英發行
A112363	08110001	20000301	200533	20000313	130926	NT網路安全詳論.實務篇/Matthew Strebe, Charles Perkins, Michael G. Mc
A067645	02000175	20000301	202128	20000323	133856	人生哲學/張永鳶等合著
A120394	02000277	20000301	201627	20000303	194354	Windows NT Workstation 4.X中文版使用手冊/林龍震編著

圖2 借閱資料

我們以前5年的借閱資料做為探勘的訓練資料，找出讀者借閱書籍的傾向特徵。接下來，說明所建置之發掘書籍最適性推薦的探勘系統，其在探勘訓練資料的執行過程。圖3為點選「書籍推薦作業」→「發掘讀者最適性的書籍推薦」功能的探勘畫面，分別在「讀者編號」欄位填入欲探勘的讀者編號，「最小借閱相似度」欄位填入數值，及在「群組出現次數最大前幾項」欄位填入數值。由第三節所描述之演算法的探勘過程，可在「書籍推薦」欄位顯示出探勘的結果，如圖4。

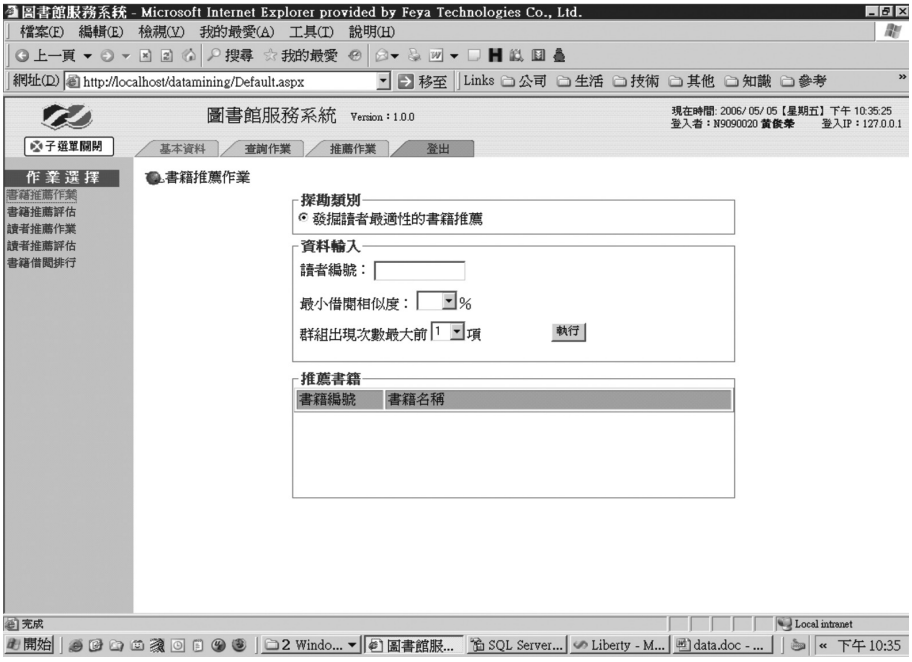

圖3　發掘讀者最適性之書籍推薦的探勘執行畫面

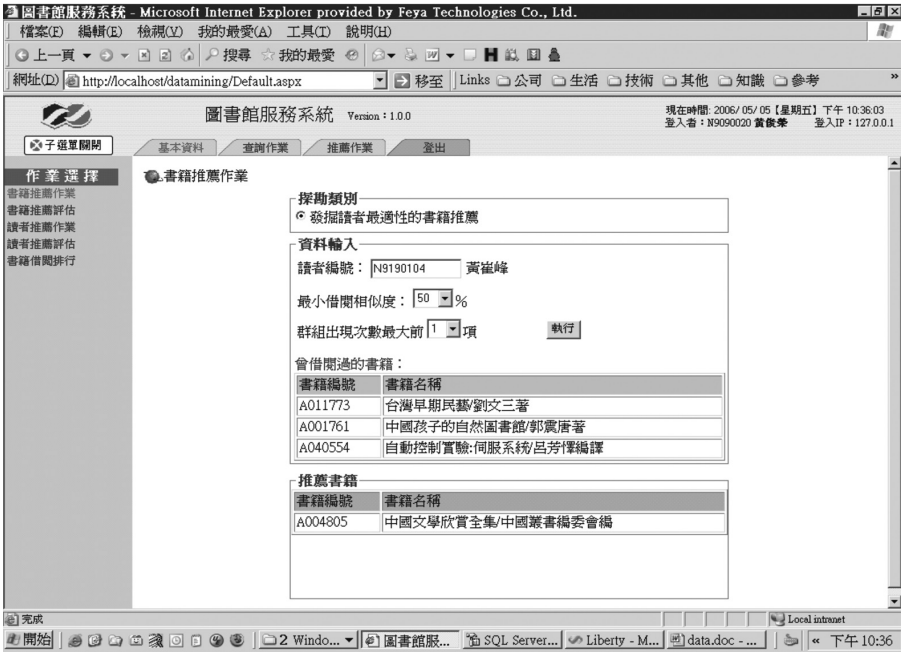

圖4　發掘讀者最適性之書籍推薦的探勘結果畫面

圖5為點選「讀者推薦作業」→「發掘書籍最適性的讀者推薦」功能的探勘畫面，分別在「書籍編號」欄位填入欲探勘的書籍編號，「最小借閱相似度」欄位填入

數值，及在「群組出現次數最大前幾項」欄位填入數值。由第四節所描述之演算法的探勘過程，可在「推薦讀者」欄位顯示出探勘的結果，如圖6。

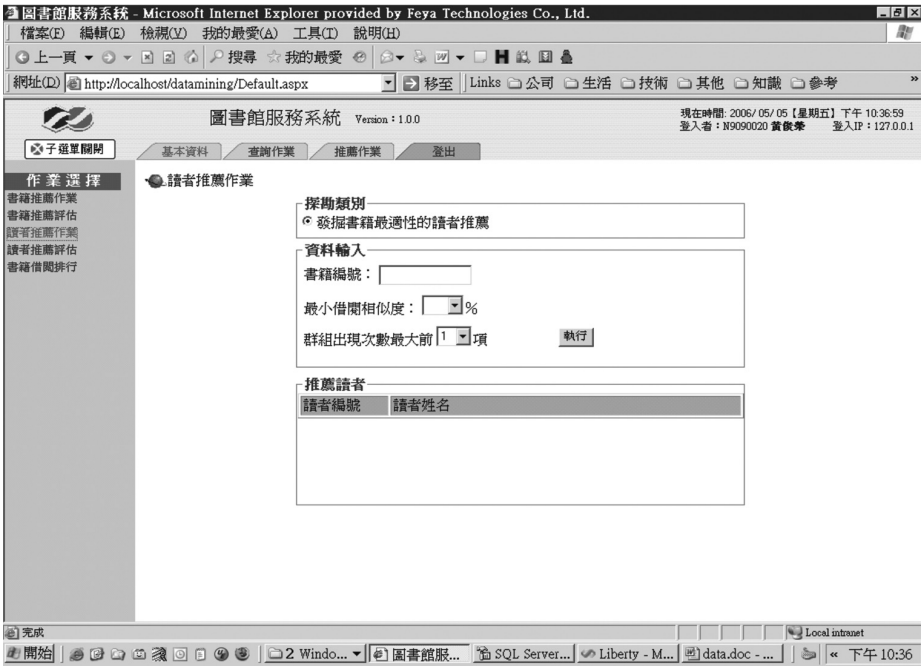

圖5 發掘書籍最適性借閱之讀者的探勘執行畫面

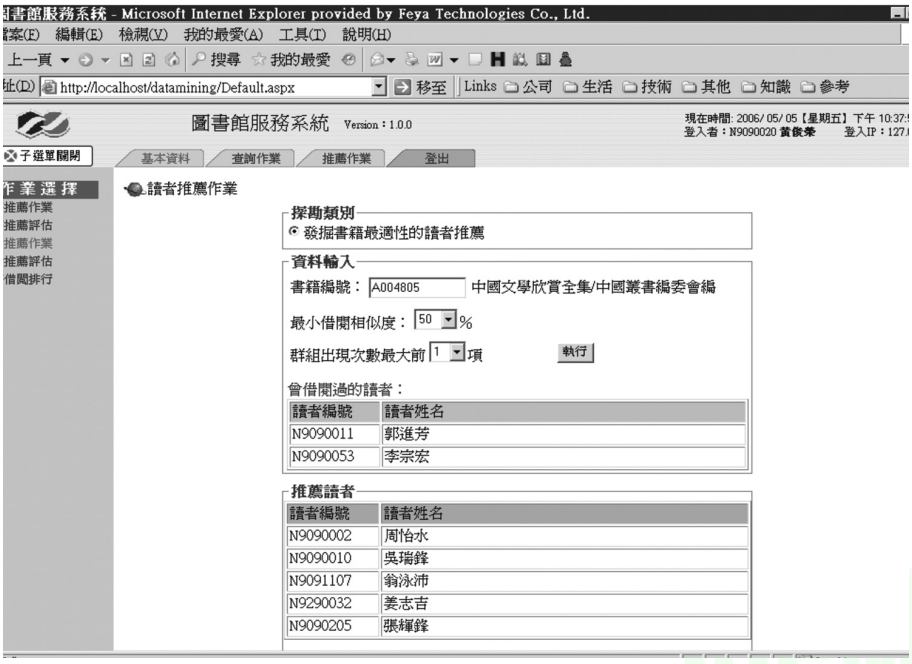

圖6 發掘書籍最適性借閱之讀者的探勘結果畫面

我們以最後一年（2005）的借閱資料做為探勘的驗證資料，做為評估在前面訓練資料所推薦之結果的成效。茲說明此探勘系統在評估驗證資料的執行過程如下。圖7為點選「書籍推薦評估」功能的執行畫面，在「書籍推薦評估」欄位中，除了顯示讀者與書籍的借閱資料，在最後一欄中會註明是否為推薦給讀者的書籍，若是則須以符號標示之。在6年的借閱資料中，會同時出現在前5年及最後一年之借閱資料中的讀者共有6532位，便以此6532位讀者做為評估的對象。在這6532位讀者中，共有1076位讀者在最後一年的借閱書籍中，包含有系統以讀者為目標來推薦書籍之方式的書籍，其推薦出現率為16.5%（1076/6532）。

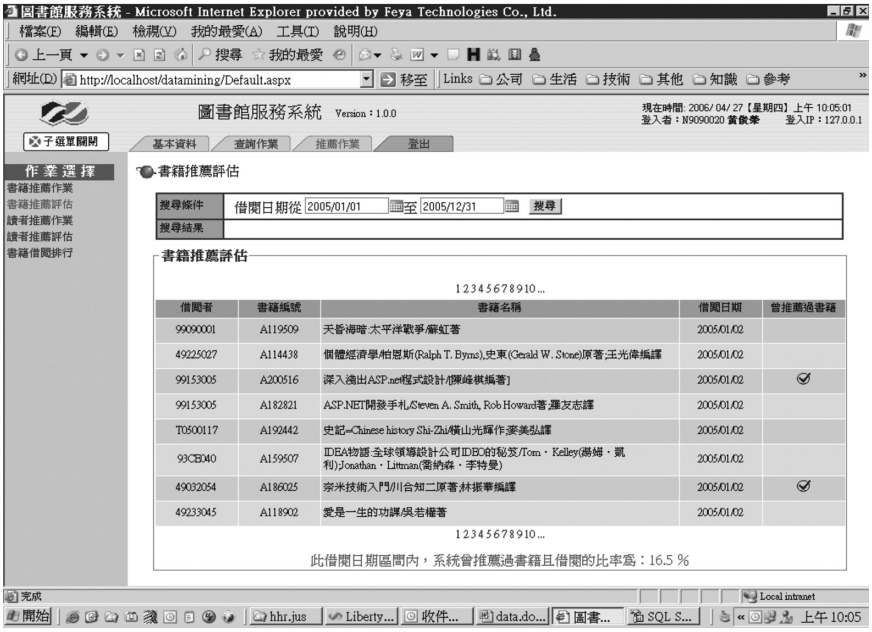

圖7　以讀者為目標來推薦書籍之評估的執行畫面

圖8為點選「讀者推薦評估」功能的執行畫面，可在「讀者推薦評估」欄位，除了顯示書籍與讀者的借閱資料，在最後一欄會註明讀者是否為書籍推薦借閱的讀者，若是則須以符號標示之。我們仍以在前5年中有出現的讀者共有6532位為評估的對象，其中共有6472位在探勘訓練資料時列為書籍最適性借閱的讀者，檢視最後一年的借閱資料，發現共有872位讀者的借閱書籍中至少包含有一本，是讓探勘系統視讀者為此一書籍最適性的借閱者，其推薦出現率為13.5%（872/6472）。

我們所建置的探勘系統中，分別以讀者個人及書籍做為探勘的目標，來找出讀者個人化最適性的書籍推薦，及找出書籍最適性借閱的讀者，若能對訓練資料所找出之借閱傾向特徵做實際的書籍推薦，讓讀者接受到書籍推薦的訊息，及延長驗證的期間，其推薦的出現率應可更為提高。在探勘的過程中，由於不同參數的設定

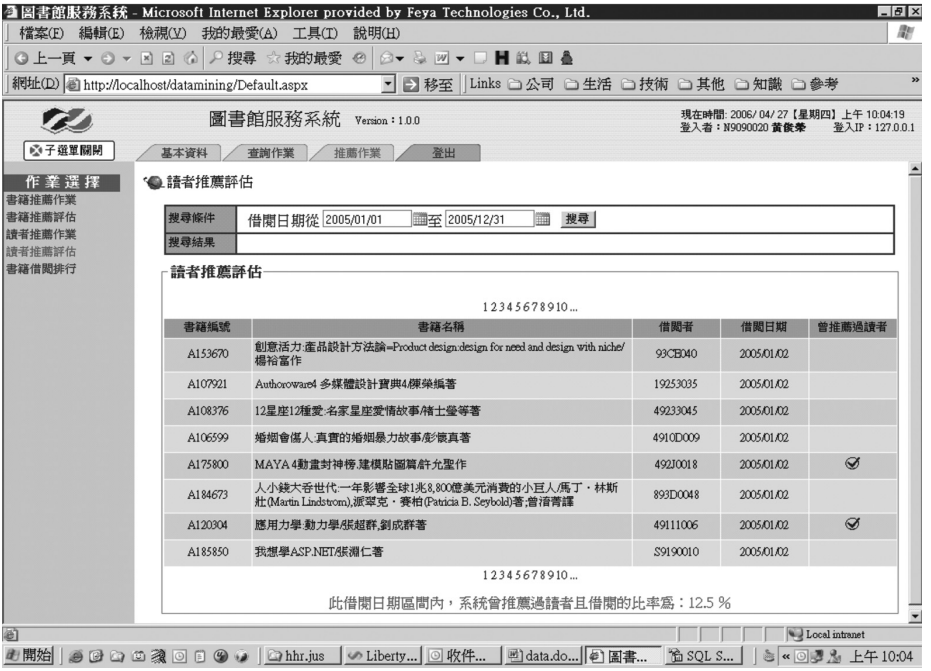

圖8 以書籍為目標來推薦借閱讀者之評估的執行畫面

值，往往會影響所找出的借閱傾向規則，對推薦時所導致的準確度也會有所影響。因此，圖書館管理者可依其經驗設定參數值，而探勘出更令人滿意的推薦結果。

六、結論與未來研究

圖書館蘊藏豐富的書籍與其他多樣的資料，如何將這些大量的資料有效地、貼切地推薦及傳達給讀者，進而吸引讀者到館借閱及利用，以提升圖書館的利用率及效益，是圖書館管理者必須思考的問題。讀者依其需求及興趣到圖書館借閱書籍，從這些被借閱過的書籍中，可反映出讀者對書籍的偏好傾向，及書籍之間的關聯性，若能從讀者之借閱資料中找出書籍彼此間的關聯性，對圖書館管理者在擬訂書籍最適性的推薦時，必可提供相當有用的參考資訊。在本論文中，分別以某一讀者及某一書籍為探勘的目標，利用群組來找出此一讀者最適性的書籍推薦、及此一書籍最適性借閱的讀者。從資料的蒐集、分析、方法的設計、訓練資料中所推導出的書籍推薦、系統設計與實作、及驗證資料中所得結果的準確性，顯示出所提出之探勘方法具有實務應用的有用性及創新。

在表3中，說明一些利用分群化技術做為個人化推薦的研究，及本論文中所設計之方法間的差異。目前已有許多相關的研究，探討如何利用分群化技術以提昇圖書館的經營服務（杜逸寧，2005；陳慶瑄，2000），但鮮少直接針對讀者個人，及

書籍本身進行做分析與計算。在本論文中，已設計兩個分群化方法，分別從讀者個人，及書籍本身來進行探勘，對書籍的推薦勢必能找出更貼切與適性的結果。

表3　以分群化為基礎之推薦方式的比較

	Weng and Liu（2004）；Wu, Chen and Chen（2001）；陳慶瑄（2000）；杜逸寧（2005）；邵秀梅（2004）；劉仲原（2002）等演算法	文中發掘讀者最適性之書籍推薦的演算法	文中發掘書籍最適性借閱之讀者的演算法
中心點選定方式	必須計算所有可能成為較佳解的中心點組合	直接以欲探勘之讀者的借閱資料為中心點	直接以欲探勘之書籍為中心點
分群化方法	採用傳統分群化演算法	根據與中心點之間的借閱相似度，是否滿足「最小借閱相似度」來進行分群化計算	根據是否包含有中心點之書籍來進行分群化計算
推薦方式	視欲推薦之個人包含於那一群組中，藉由群組所顯示出的傾向特徵，做為個人化推薦的依據	由專屬於讀者個人的群組中，藉由群組所顯示出的傾向特徵，做為發掘讀者最適性之書籍推薦的依據	由專屬於書籍本身的群組中，藉由群組所顯示出的傾向特徵，做為發掘書籍最適性借閱之讀者的依據
執行效率	較差	較好	較好
群組特徵	與欲推薦之個人的關聯性較差	與欲推薦之讀者的關聯性較好	與欲發掘之書籍的關聯性較好

目前，本研究僅就如何利用群組概念，分別發掘讀者個人化最適性之書籍推薦、及書籍最適性借閱之讀者的探勘過程作探討，並根據所提出的探勘方法，以某一科技大學之讀者的借閱資料為探勘的來源資料，設計與建置一個書籍最適性的推薦系統，以網頁界面來具體顯示書籍推薦的基本操作功能。對於未來繼續從事之相關研究有：

㈠對借閱資料中之書籍借閱時間的次序性，作有效的分析與運用。

㈡利用其他資料探勘技術，例如關聯規則，在此研究問題的可行性與有效性。

㈢考量讀者的族群特徵，例如就讀科系、性別、年級、學業成績等因素，來探勘書籍適性化的推薦。

參考文獻

卜小蝶（1998）。「淺析個人化服務技術的發展趨勢對圖書館的影響」。國立成功大學圖書館館刊，2，63-73。

汪軒楷（2003）。「策略式資料探勘在個人化推薦上之研究」。未出版碩士論文，真理大學管理科學研究所，台北縣。

杜逸寧（2005）。「結合叢集法與案例式推論於協同分類—以論文推薦系統為例」。未出版碩士論文，國立彰化師範大學資訊管理研究所，彰化市。

邵秀梅（2004）。「資料探勘應用於個人化網路學習導覽推薦之研究」。未出版碩士論文，銘傳大學資訊管理研究所，台北市。

吳安琪（2001）。「利用資料探勘的技術及統計的方法增強圖書館的經營與服務」。未出版碩士論文，國立交通大學資訊科學研究所，新竹市。

洪志淵（2001）。「圖書流通記錄之一般化相關規則找尋之研究」。未出版碩士論文，國立中山大學資訊管理研究所，高雄市。

湯春枝（2002）。「從個人化服務行銷的理念談交通大學個人化數位圖書資訊服務（PIE@NCTU）系統」。國立成功大學圖書館館刊，9，33-49。

陳慶瑄（2000）。「學習社群對電子圖書館個人化服務之影響」。未出版碩士論文，國立中正大學資訊管理研究所，嘉義縣。

劉仲原（2002）。「個人化網頁推薦系統之研究—以歷史博物館為例」。未出版碩士論文，國立雲林科技大學資訊管理研究所，雲林縣。

鄭玉玲（2003）。「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」。未出版碩士論文，南華大學資訊管理學研究所，嘉義縣。

辜曼蓉（1999）。「讀者資訊尋求行為與以讀者為中心的圖書館行銷」。書府，20，81-111。

孫冠華（2003）。「應用資料探勘技術於數位圖書館之個人化服務及管理」。未出版碩士論文，南華大學資訊管理學研究所，嘉義縣。

Alsabti, K., Ranka, S. and Singh, V. (1998). “An Efficient K-Means Clustering Algorithm”, *IPPS/SPDP Workshop on High Performance Data Mining*, Orlando

Berry, M. J. A. and Linoff, G. S. (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. (2nd Ed.). New York: John Wiley.

Chen, M. S., Han, J. and Yu, P. S. (1996). “Data Mining: An Overview from a Database Perspective”, *IEEE Trans. on Knowledge and Data Engineering, 8*(6), 866-883.

Dubes, R. C. and Jain, A. K. (1988). *Algorithms for Clustering Data*. New Jersey: Prentice Hall.

Han, J., Kamber, M. (2000). *Data Mining: Concepts and Techniques*. Morgan Kaufmann. Retrieved from http://mylibrary.e-lib.nctu.edu.tw/

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: an Introduction to Cluster Analysis*. New York: John Wiley & Sons.

Ng, R. T. and Han, J. (1994). “Efficient and Effective Clustering Methods for Spatial Data Mining”, *Proceedings of the 20th International Conference on Very Large Data Bases*, Santiago.

Ou, J., Lin, S. and Li, J. (2001). “The Personalized Index Service System in Digital Library”, *DW, DM and DL*. Third International Symposium Cooperative Database Systems for Advanced Applications, USA.

Weng, S. S. and Liu, M. J. (2004). “Personalized Product Recommendation in e-Commerce”, *EC Technologies*. 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service, USA.

Wu, Y. H., Chen, Y. C. and Chen, A. L. P. (2001). “Enabling Personalized Recommendation on the Web Based on User Interests and Behaviors”, *Eleventh International Workshop on Research Issues in Data Engineering Document Management for Data Intensive Business and Scientific Applications*, Germany.

Using Clusters to Find the Most Adaptive Recommendations of Books

Chui-Cheng Chen

Associate Professor
Department of Information Management
E-mail: ccchen@mail.stut.edu.tw

Jun-Rong Huang

Graduate Student
Institute of Information Management
Southern Taiwan University of Technology
Tainan, Taiwan, R.O.C.
E-mail: n9090020@webmail.stut.edu.tw

Abstract

In this paper, we use readers' borrowing history records as the source data of mining. Each borrowing history record contains a reader ever borrowed books, and use clusters to find the most adaptive recommendations of books from two aspects. One is to let one reader as the target of mining and assign his borrowing history record as the center of cluster. Then, we propose a clustering method to let each other borrowing history record is grouped with the center to which it contains the reader's borrowing history record for satisfying the threshold of the minimum borrowing similarity. We can find the most adaptive book recommendations for the reader according to the characteristics of borrowing tendency of the cluster. The other is to let one book as the target of mining and assign it as the center of cluster. Then, we propose a clustering method to let each other borrowing history record is grouped with the center to which it contains the book. We compute the association factors of the center in the cluster, and find the most adaptive readers of borrowing the book according to the borrowing similarity between the association factors and borrowing history records. We design and construct a mining system for fining the most adaptive recommendations of books according to we propose the both methods. The results of the mining can provide very useful information to plan the services of the most adaptive book recommendations for libraries.

Keywords: *Data mining; Cluster; Borrowing history record; Book recommendation*