行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※
※ ※
※ 個人化與群體化電子圖書館服務之分析與設計 ※
※ ※
※※※※※※※※※※※※※※※※※※※※※※※※*※※※※※※※※※※※※※※

計畫類別：■個別型計畫 □整合型計畫

計畫編號：90-2213-E-009-083-

執行期間： 2001 年 8 月 1 日至 2002 年 7 月 31 日

計畫主持人：楊維邦

執行單位：國立交通大學資訊科學系

本成果報告包括以下應繳交之附件：

□ 赴國外出差或研習心得報告一份

□ 赴大陸地區出差或研習心得報告一份

□ 出席國際學術會議心得報告及發表之論文各一份

□ 國際合作研究計畫國外研究報告書一份

個人化與群體化電子圖書館服務之分析與設計

計畫編號：NSC 89-2218-E-009-009

執行期限：89 年 8 月 1 日至 90 年 7 月 31 日

主持人：楊維邦　　交通大學資訊科學系

計畫參與人員：柯皓仁　交通大學圖書館

計畫參與人員：楊雅雯　交通大學資訊科學系

計畫參與人員：吳安琪　交通大學資訊科學系

計畫參與人員：戴玉旻　交通大學資訊科學系

計畫參與人員：陳莉君　交通大學資訊科學系

計畫參與人員：嚴文亨　交通大學圖書館

摘要

近年來，電子圖書館已成為圖書館界努力追尋的目標。電子圖書館的主要元素有三：電子化館藏、電子化作業和電子化服務。由於個人化和社群化將是未來圖書館電子化服務的發展趨勢，因此在本研究計畫中我們將探討以資訊擷取(Information Retrieval)和資料探勘(Data Mining)的技術來達成個人化、社群化圖書館電子服務的目標。本計畫的主要工作項目有三：

1. 研究動態追蹤讀者興趣的相關理論，並將其應用在個人化電子服務。個人化電子服務必須根據讀者的需求提供其必要的資訊，因此如何了解讀者的興趣並追蹤其興趣的更迭乃是一個非常重要的議題。本計畫將提出個人興趣關連圖(Key-Phrase Relationship Graph, KRG)的概念來動態追蹤讀者興趣。
2. 探討資料探勘相關理論，發展適用於社群化電子圖書館的資料探勘演算法。擁有共同興趣的讀者通常會借閱類似的書籍，反之亦然。本計畫擬由交大圖書館讀者借閱與預約歷史記錄中發掘具有共同興趣的讀者群組，了解其借閱的共通特性及順序，並根據資料探勘的結果達成社群化電子圖書館服務。
3. 根據以上的理論實作一套交通大學個人化數位圖書資訊環境系統雛形(PIE@NCTU)。

關鍵詞：電子圖書館，電子化服務，資訊擷取，讀者興趣追蹤，資料探勘，相關規則探勘，個人化，社群，個人資訊環境

Abstract

In recent years, libraries around the world have set the goal to move toward “electronic library”. Electronic collection, electronic operation, and electronic service are the three dimensions of electronic library. Because the tendency of electronic service is to facilitate personalization and grouping users into communities for sharing common interests and knowledge, in this project we will attempt to develop such kind of electronic service by using the techniques of information retrieval.

In this project, we will focus on three research topics:

1. Study how to dynamically track user interests, and apply the research achievement to personalized electronic service. Because a well-designed personalized electronic-service system has to provide services to a user according to his/her specific interests, it is essential for the system to know users interests and track the evolution of user interests. In this project, we will propose the idea of Key-Phrase Relationship Graph (KRG) to dynamically track the interests of users.
2. Study the data mining technique, and apply data mining to the electronic service that enable the formation of user communities. We believe users having common interests will borrow similar materials from library, and vice versa. Data mining is suitable for mining patterns for user communities. In this project, we will attempt to discover user communities with common reading interests by mining the

NCTU Library patron's borrowing and reserving history log.

3. Implement a prototype system, called *Personalized Information System at NCTU* (*PIE@NCTU*), to prove the feasibility of the research achievements.

1. 緣由與目的

近年來，電腦與網路科技的蓬勃發展，帶動了電子商務 (Electronic Commerce)的熱潮。電子商務的經營者皆期望經由網際網路接觸更廣大的客戶，更無遠弗屆地深入所有消費族群。然而，若想在眾多電子商務經營者中脫穎而出，其關鍵就在於是否能了解客戶的需求，提供每位客戶量身訂作之個人化服務，亦即採用客戶關係管理 (Customer Relationship Management – CRM)的理念，並利用電腦系統紀錄並分析客戶的需求，從而建立一對一之個人化服務，以期增加客戶的滿意度與忠誠度，並確保在激烈商業競爭的環境裡獲利[1][2]。

儘管圖書館並非以營利為目的之商業機構，然而圖書館的最終目的，乃在於使讀者有效地利用圖書館的資料，協助讀者獲取資訊、運用資訊，從而產生知識。因為每位讀者都有其特別的資訊需求，圖書館的服務應該把每位讀者視為不同的個體，儘量去滿足每一讀者個別的資訊需求。從這個觀點來看，若能採用 CRM 的理念並推動個人化服務，相信必能提昇圖書館對讀者的服務。

與個人化相對的概念是群體化。有時知識的產生不能光靠單一個體，而是得藉由具有相同興趣、專長的個體組成社群(Community)，彼此激發靈感與分享心得，方能促成知識的產生。表面上看來，個人化和群體化似乎是互斥的，但實際上對圖書館的服務而言，卻是一體兩面。

由於圖書館自動化系統(以下簡稱自動化系統)中儲存了圖書館館員專業知識的成果(編目資料)與讀者背景資料，且記錄了館藏的流通狀況及流通歷史，因此當圖書館界欲實施 CRM 理念並推動個人化和群體化服務之時，若能以自動化系統為基礎，將可收事半功倍之效。觀察現有圖書館自動化系統所提供的功能，大多侷限於圖書館本身作業的自動化，極少部分的功能是著眼於讀者服務，遑論個人化與群體化服務。以讀者最常接觸的線上公用目錄 (Online Public Access Catalog, OPAC，或稱館藏查詢系統)而言，除館藏查詢之外，僅具備基本的個人化服務，如館藏續借、館藏預約、借閱狀況查詢、讀者基本資料查詢與修改等。至於群體化服務更是付之闕如。

本研究計畫的目標在於運用資訊擷取(Information Retrieval)和資料探勘(Data Mining)的技術達成個人化和社群化的電子圖書館服務。我們將進行相關的理論研究並實作系統：

在理論研究方面：針對個人化電子圖書館服務，我們將進行研究動態追蹤讀者興趣(Tracking User Interests)的相關演算法；在社群化電子圖書館服務方面，我們首先採取將採取 Apriori[3] [4]、GSP [5]、H-Mine[6]等現有資料探勘的演算法，將之應用在社群化電子圖書館服務上，希望能夠找出讀者的特性以提升圖書館的業績，進而改良 GSP 與 H-Mine 演算法，使其更能適用於社群化電子圖書館服務。

在系統實作方面：實作兼具個人化和社群化能力的 *PIE@NCTU* 系統(Personalized Information Environment at NCTU)，以驗證所發展之理論與演算法的可行性。

2. 文獻探討

本節針對個人化以及群體化服務的相關文獻加以探討。

2.1. 個人化服務

目前許多網際網路服務都提供了個人化服務，如我的Yahoo!奇摩(http://tw.mykimo.yahoo.com)等；與圖書館相關的個人化服務則有 MyLibrary[7]、MyLirary@NCState[8]等。

French 與 Viles 二位學者在 1999 年提出個人化服務環境的架構[9]。概括來說，個人化服務環境應該要具備以下條件： (1) 個人化的使用者界面 (Customizable User Interface)，讓使用者依自己的喜好組織使用環境； (2) 有效的檢索 (Effective Search)，能提高檢索結果的正確性，引導使用者尋找資料，提高查全率 (Recall)；(3) 確保使用者的隱私權 (Privacy)。我們認為現階段的個人化服務系統應該具備以下條件：

2.1.1 個人化使用界面

個人化使用界面讓使用者依據個人的喜好來規劃其使用界面，前述幾種個人化服務均具備讓使用者設定個人化使用界面的功能。

2.1.2 個人資料紀錄 (User Profile)

個人化服務的每位使用者都有其個人的資料記錄，儲存背景、興趣、學科專長、檢索歷史等資料，做為個人化服務的主要依據，其中尤以個人興趣紀錄最為重要。個人興趣紀錄的主要來源有二：(1) 由讀者人工填寫，(2)由系統自動推導產生。

若個人興趣記錄是由人工填寫產生的，則此類系統均提供一個興趣關鍵字詞的輸入界面，讓使用者能自行設定感興趣的關鍵字。像 MyLibrary@NCState 即是根據讀者自行設定的關鍵字詞提供資訊選粹服務。這種作法雖然很直覺(因為是由使用者輸入興趣資料)，但是使用者往往會選擇過於普遍的字詞來描述自己的興趣

[10]，導致對於興趣的描述不夠精確。再者我們亦不能期望使用者都是勤勞、有足夠耐心，且總是能正確輸入關鍵字詞的。

至於自動產生個人興趣紀錄的系統，通常是根據使用者的使用歷程或特定行為來推導興趣，常見的方法有：(1) 將使用者瀏覽過的網頁中所含的關鍵字詞記錄下來當成使用者的興趣[10]；(2) 利用電子郵件通信紀錄來抽取關鍵字詞當成使用者興趣[10]；(3) 以交易行為推導興趣 (例如 Amazon 會以顧客買過的書之關鍵字為興趣)。

2.1.3 資訊選粹服務

有了個人興趣記錄之後，便可提供資訊選粹 (Selective Dissemination of Information， SDI) 服務。資訊選粹利用資訊過濾 (Information Filtering) 技術分析出個別使用者感興趣的新進資訊。資訊過濾技術可分為內容式資訊過濾 (Content-based Information Filtering) 以及協力式資訊過濾 (Collaborative Information Filtering) 二種[11]。內容式資訊過濾主要是以資訊內容作為過濾的依據並加以分析比較，使用者在興趣檔中只要紀錄感興趣的關鍵字詞，系統便會將新進資訊和興趣檔做比對達成資訊過濾。諸如 LA Times Custom News Services (http://www.latimes.com/) 、 MyLibrary @NCState 均採用內容式資訊過濾來提供資訊選粹服務。

協力式資訊過濾不直接分析資訊內容，而是找出與使用者背景、知識、興趣接近的同好或社群，再針對使用者的查詢主題，從這些同好或社群成員感興趣的資訊中，分析並選取最可能相關的資訊提供參考[12]。因為協力式資訊過濾較內容式資訊過濾複雜許多，因此實際採用協力式資訊過濾的系統較少。

2.1.4 個人化檢索

所謂個人化檢索是指能依個別使用者的背景、興趣或需求，幫助使用者尋找資訊。個人化檢索的可能應用有：(1) 提供適合使用者的背景、年齡等因素的資訊， 例如同樣是檢索"網路多媒體"這個主題，提供給小學生和提供給資訊系所博士生的資訊就應該要加以區別[12]；(2) 將檢索結果根據使用者的個人興趣加以排序(Ranking)，把使用者感興趣的檢索結果排列在較明顯的位置。

根據個人興趣紀錄提供資訊選粹服務的系統頗多，然而將其應用在個人化檢索上卻很少見，這對建置個人化資訊環境是一個很大的遺憾。有鑑於此，我們在本計畫中提出一個利用個人興趣紀錄及資訊過濾技術達成檢索個人化的方法。

2.2. 資料探勘與群體化服務

資料探勘(Data Mining)[13]也叫做資料庫探勘(Database Mining)或資料庫知識發掘(Knowledge Discovery in Database)。簡單地說，資料探勘是從儲存於資料庫(Database)、資料倉儲(Data Warehouse)或其他資訊儲存器(Information Repository)的大量資料中，發掘出具有價值的知識之過程。資料探勘在近年來廣泛地運用在各種領域或行業，例如行銷、財務、銀行、製造、通訊、保險等，用以發掘潛在客戶、管理異常狀況、管理客戶關係、或作為企業決策的參考。例如超級市場能將資料探勘運用於發掘顧客的消費模式，並利用所發掘的消費模式研擬促銷或貨物排架策略，以提昇超級市場的業績。

資料探勘主要功能有分類規則歸納(Classification)、推估分析(Estimation)、預測分析(Prediction)、相關規則探勘(Association Rules Mining)、同質分群(Clustering) 等五種。在本計畫中，我們將資料探勘中的相關規則探勘應用以及循序規則探勘(Sequential Pattern Mining)在圖書館的群體化服務，以自動化系統中的書目、館藏、借閱與預約歷史記錄為來源資料，探索讀者的社群特性。本計畫擬探索的讀者社群關係包含館藏借閱的共同性及順序性。發掘出讀者社群關係後，我們期望能運用這些社群關係來吸引讀者到館借閱、提昇館藏借閱率、提昇讀者忠誠度、促進館藏流通率。以下我們簡單介紹相關規則探勘的基本概念。

相關規則探勘最常應用在商店交易記錄資料庫，用以分析並發掘顧客的交易模式，並根據發掘出來的交易模式採取適當的行銷策略，以提昇商品的銷售率。例如：20%買牙刷的顧客也會買牙膏、毛巾、和香皂就是一個典型的相關規則。相關規則探勘的正規敘述([3])如下：

令 $I=\{i_1,i_2,\ldots,i_m\}$ 是由交易項目(Item)組成的集合，由一個或一個以上的項目所組成的集合稱為項目集 (Itemset)。令資料庫 $\boldsymbol{D}$ 是由一群交易(Transaction) T 所組成的集合，每個 T 為一項目集，代表交易記錄，$T \subseteq I$ ，每個交易記錄有其唯一的識別碼，稱為 *TID*。如果 $X \subseteq I$ 且 $X \subseteq T$，則稱 T 包含(Contain) X。以商店的應用來看，每一種商品就是一個交易項目，一個顧客在某次來訪時中所購買的商品所成的集合即為一交易。

一個相關規則 (Association Rule) 表示成 $X \Rightarrow Y$，其中 $X \subseteq I$ ，$Y \subseteq I$，$X \cap Y = \varnothing$ 。若 $\boldsymbol{D}$ 中包含 X 的交易裡有 c%也同時包含了 Y，我們就說規則 $X \Rightarrow Y$ 的確信值 (Confidence) 為 c%；如果 $\boldsymbol{D}$ 中包含 $X \cup Y$ 的交易記錄有 s%，我們就說規則 $X \Rightarrow Y$ 的支持度 (Support) 為 s%。相關規則探勘定義為：給定交易記錄資料庫 $\boldsymbol{D}$，在當中找出所有確信值和支持度大於最小確信值跟最小支持度的規則。

Agrawal 等人[3][4]將相關規則探勘分為二個子問題：

1. 子問題 1：找到所有支持度大於最小支持度的項目集。為了探勘方便起見，有時把某一項目集的支持度定義為包含此項目集的交易個數，而不是原來的交易百分率。支持度大於最小支持度的項目集稱為大項目集 (Large Itemset)。
2. 子問題 2：用子問題 1 中所找到的大項目集來產生所期望的規則。此步驟的演算法非常直接，即：對於任一大項目集 L，找出其所有非空子集合。對於每個非空子集合 a，如果規則 a⇒(L-a)的確信值(也就是 support(L)/support(a))大於最小確信值，則此規則即符合所求。

表 1 為一個交易資料庫範例。在此交易資料庫中共包含四筆交易，其交易紀錄識別碼分別為 100、200、300、400。此交易資料庫中共有五種交易的項目(或商品)：牙刷、牙膏、毛巾、手錶、香皂，為方便電腦運算起見，我們將這五種商品依序賦予 1、2、3、4、5 的代碼。假設最小支持度為 2，則只要出現在交易資料庫二次或以上的項目集便為大項目集，因此如表 2 所示，本範例中的大項目集包含{1}、{2}、{3}、{5}、{1, 2}、{2, 3}、{2, 5}、{3, 5}、{2, 3, 5}；本範例中可能出現的相關規則列於表 3，以{2, 5}⇒{3}此一相關規則來看，其確信值為 67%，這代表有 67%買 2 和 5 的客人也會時買 3(在 200, 300, 400 這三筆交易中都有 2 和 5，但是只有 200 和 300 這兩筆交易有 3)，而同時購買 2, 3, 5 的交易有二次(即 200 和 300 這兩筆交易)。

若設定之最小確信值為 100%(在實際狀況下不可能如此設定)，則探勘得到的相關規則有：{1}⇒{3}、{2}⇒{5}、{5}⇒{2}、{2, 3}⇒{5}、{3, 5}⇒{2}等五個規則，分別代表{牙刷}⇒{毛巾}、{牙膏}⇒{香皂}、{香皂}⇒{牙膏}、{牙膏, 毛巾}⇒{香皂}、{毛巾, 香皂}⇒{牙膏}。

表 1 交易資料庫範例

資料庫 ***D***	
交易紀錄識別碼(TID)	項目集(Itemset)
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

表 2 大項目集(最小支持度為 2)

大項目集	支持度
{1}	2
{2}	3
{3}	3
{5}	3
{1, 3}	2
{2, 3}	2
{2, 5}	3
{3, 5}	2
{2, 3, 5}	2

表 3 相關規則

相關規則	確信值
{1}⇒{3}	100%
{3}⇒{1}	67%
{2}⇒{3}	67%
{3}⇒{2}	67%
{2}⇒{5}	100%
{5}⇒{2}	100%
{3}⇒{5}	67%
{5}⇒{3}	67%
{2}⇒{3, 5}	67%
{3}⇒{2, 5}	67%
{5}⇒{2, 3}	67%
{2, 3}⇒{5}	100%
{2, 5}⇒{3}	67%
{3,5}⇒{2}	100%

從相關規則探勘延伸出各式各樣的問題，其解決方法和基本的問題息息相關，且可應用在更多領域。以下簡述目前的相關研究[14]：

1. 有些研究考慮到顧客多次交易的狀況，同一位顧客會有先後時間關係的交易記錄。此類研究的目的是想要在這些交易記錄中找出顧客最常買的商品序列(Sequence)，此序列是由項目集所組成，而每一個項目集內的項目是不考慮個數且沒有順序關係的。此種相關規則探勘又稱為循序規則探勘(Sequential Pattern Mining)。
2. 支持度和確信值是用來控制產生規則多寡的兩個限制，除了這二個限制外，有些研究還定義更多的限制，繼而利用這些限制的特性來加速探勘。
3. 在實際的應用環境中，資料庫內的交易記錄筆數會隨著時間而增加或減少，而資料探勘要耗費龐大的計算時間，若是每次資料庫一更新就重新探勘將會浪費很多的時間，相對地若只探勘資料庫更新的部份便可節省很多時間，因此有些研究希望只探勘資料庫增加或減少的部份就能得到所有的規則，此即漸進更新(Incremental Update)問題。
4. 有些研究除了考慮交易記錄之間的順序關係外，也探討交易記錄之項目集內的項目順序關係。
5. 有些研究將交易記錄中每一個交易項目的個數考慮進去，或者將交易項目分類，以分類來當相關規則的項目。

3. 動態學習使用者興趣

個人化服務的關鍵技術在於如何準確得知使用者的興趣。在圖書館自動化系統(如線上公用目錄)中，使用者的資訊需求經常是透過檢索來滿足的，因此，我們認為使用者的興趣或資訊需求能夠從其曾經用過的檢索策略來解讀。基於此一概念，在本論文中，我們考慮個別使用者曾經在檢索策略中用過的關鍵字詞的頻率、各關鍵字詞間的相關性，以及時間對興趣的影響，來動態學習其興趣。

本節先提出一個動態學習使用者興趣的演算法，並透過一個範例來說明演算法；接著說明如何根據此演算法實作資訊選粹服務和個人化檢索。

3.1. 個人興趣關連圖

若要提出一個能有效偵測使用者興趣的方法，首先必須對"興趣"的特性有所了解。我們認為每個人的興趣通常會維持一段時間，且隨著歲月流逝而有所改變，因此，偵測使用者興趣的演算法必須要能隨著時間修正使用者的興趣 (興趣加強或減弱)。如前所述，本計畫假設使用者的興趣可由其曾使用過的檢索策略來解讀，在綜合考量時間因素和檢索行為的情形下，我們針對每位使用者曾用過的檢索策略，建立個人興趣關連圖 (Personal Keyphrase-Relationship Graph，簡稱 PKRG)。個人興趣關連圖的主要用途在於計算單一使用者曾用過的各關鍵字詞的權重，以及該使用者認為各關鍵字詞間的關連性，從而推演使用者的興趣。

3.1.1 個人興趣關連圖的建立

對於每一位使用者，我們會為其建立專用的個人興趣關連圖(PKRG)。PKRG 為一有向圖(Directed Graph)，圖 1 為 PKRG 的範例。如圖 1 (a)所示，PKRG 中的每一個端點 (Vertex) V_i 代表使用者曾在檢索策略中用過的關鍵字詞；每一條邊線 (Edge) E_{ij} 表示使用者曾用過以該邊線兩端點 (V_i 及 V_j) 代表之關鍵字詞做 "且 (AND) " 運算所產生的檢索策略 (例如：'詩 AND 古典')。

我們為 PKRG 中的每個端點和邊線都賦予權重(Weight)：端點 V_i 的權重代表使用者對關鍵字詞 V_i 感興趣的程度，邊線 E_{ij} 的權重則代表使用者對'V_i AND V_j'此檢索策略感興趣的程度，同時也代表使用者認為 V_i 和 V_j 這兩個關鍵字詞的關連性。當系統發覺一位使用者執行了新的檢索策略'x AND y'時，系統會自動在該使用者的 PKRG 中產生二條邊線：E_{xy} 和 E_{yx}。我們假設在一檢索策略中，使用者對第一個關鍵字詞會比第二個關鍵字詞有興趣，所以 E_{xy} 的權重會高於 E_{yx} 的權重。

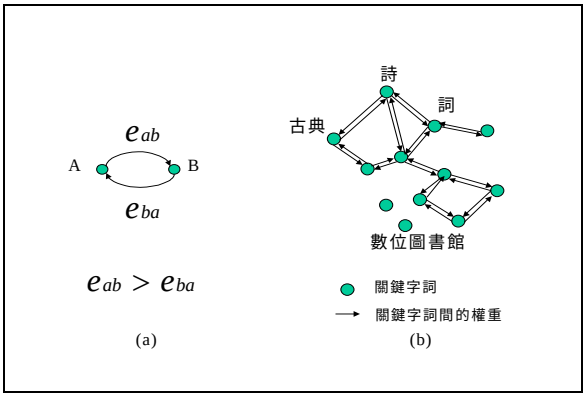

圖 1 個人興趣關連圖 (Personal Keyphrase-Relationship Graph, PKRG)

3.1.2 權重計算 (Weight)

PKRG 中的每個端點和邊線都具有一個權重：端點的權重代表使用者對相對應的關鍵字詞感興趣的程度；邊線的權重則代表使用者對相對應的檢索策略感興趣的程度，同時也代表相對應的二關鍵字詞的關連性。計算權重時需考量三個因素：(1)單一關鍵字詞出現在檢索策略中的次數，(2)單一檢索策略曾使用的次數，以及(3)時間對使用者興趣的影響。我們假設時間對使用者興趣的影響可依固定時間劃分成若干區段，在此假設下，我們採用下列步驟來計算 PKRG 中每一個端點(代表關鍵字詞)和邊線(代表檢索策略)的權重。

- ◆ 計算個人關鍵字詞的權重
- 計算單一時間區段內曾使用過的關鍵字詞之權重；
- 計算各個時間區段內時間對興趣的影響係數；
- 結合各個時間區段內求得的個人關鍵字詞權重，以及時間對興趣的影響係數，求得整體的關鍵字詞權重。

- ◆ 計算個人檢索策略的權重
- 計算單一時間區段內曾使用過的檢索策略之權重；
- 計算各個時間區段內時間對興趣的影響係數；
- 結合各個時間區段內求得的個人檢索策略權重，以及時間對興趣的影響係數，求得整體的檢索策略權重。

接下來我們將詳細介紹計算個人關鍵字詞和檢索策略權重的方法。

3.1.2.1 單一時間區段內關鍵字詞的權重計算

我們將單一時間區段內某關鍵字詞的使用頻率加以正規化之後所得的值, 作為單一時間區段內該關鍵字詞的權重。計算公式如下：

$$W_{ik} = \frac{TF_{ik}}{\sum_{j=1}^{n_k} TF_{jk}}$$

- Wik：第 k 個時間區段中，關鍵字詞 i 的權重。
- TFik：第 k 個時間區段中，關鍵字詞 i 的使用頻率。
- nk：第 k 個時間區段中，用過的關鍵字詞總數。

3.1.2.2 單一時間區段內的檢索策略權重計算

我們依據關鍵字詞和檢索策略的頻率來計算單一時間區段內檢索策略的權重。公式如下：

$$Sim_{ijk} = \frac{\frac{EF_{ijk}}{\sum_{j=1}^{m_k} EF_{jk}}}{\sqrt{\frac{TF_{ik}}{\sum_{j=1}^{m_k} EF_{jk}} * \frac{TF_{jk}}{\sum_{j=1}^{m_k} EF_{jk}}}} = \frac{EF_{ijk}}{\sqrt{TF_{ik} * TF_{jk}}}$$

- Sim_{ijk}：在第 k 個時間區段中，‘i AND j’這個檢索策略的權重。
- TF_{ik}：在第 k 個時間區段中，檢索策略中含有關鍵字詞 i 的正規化頻率。
- TF_{jk}：在第 k 個時間區段中，檢索策略中含有關鍵字詞 j 的正規化頻率。
- EF_{ijk}：在第 k 個時間區段中，檢索策略中含有‘i AND j’這個檢索策略的正規化頻率。
- m_k：在第 k 個時間區段中，用過的檢索策略總數。

3.1.2.3 時間對興趣的影響係數

儘管使用者通常會隨著時間的流逝改變其興趣, 但我們假設使用者的興趣在某一段固定時間內都不會有所改變, 且越久之前的興趣紀錄對使用者而言越不重要。為了表現時間流逝對興趣的影響，我們將使用者用過的檢索策略依固定時間間隔劃分成若干區段, 每一時間區段對使用者整體興趣的影響呈半衰期 (Half-Life) 遞減。假設將時間劃分成 n 個區段, 則在第 k 區段中(k 值越大時間越遠), 時間對使用者整體興趣的影響率為：

$$Hl_k = \frac{2^{n-k}}{2^n - 1} \quad \text{where} 1 \le k \le n$$

- Hl_k：第 k 時間區段的半衰期時間係數。

例如當 n=5 時，五個時間區段的半衰期時間係數為$(Hl_1, Hl_2, Hl_3, Hl_4, Hl_5) = (16/31, 8/31, 4/31, 2/31, 1/31)$。

3.1.2.4 個人關鍵字詞權重計算

在將使用者檢索歷程劃分成若干時間區段 個別計算單一時間區段內關鍵字詞的權重後, 我們可將之合併計算，求得整體的個人關鍵字詞權重。計算公式如下：

$$W_i = \sum_{k=1}^{n} (W_{ik} * Hl_k)$$

- Wi：關鍵字詞 i 的權重。
- W_{ik}：第 k 個時間區段中，關鍵字詞 i 的權重。
- Hl_k：第 k 時段中時間對興趣的影響係數。
- n：劃分的時間區段數目。

3.1.2.5 個人檢索策略權重計算

在將使用者檢索歷程劃分成若干時間區段 個別計算單一時間區段內檢索策略的權重後, 我們可將之合併計算，求得整體的個人檢索策略權重。計算公式如下：

$$W_{ij} = \sum_{k=1}^{n} (Sim_{ijk} * Hl_k)$$

- W_{ij}：‘i AND j’這個檢索策略的權重。
- Sim_{ijk}：在第 k 個時間區間中，‘i AND j’這個檢索策略的權重。
- Hl_k：第 k 時區中時間對興趣的影響係數。
- n：劃分的時間區段數目。

3.2. PKRG 與資訊選粹服務、個人化檢索的關係

運用 PKRG 計算出每個關鍵字詞以及檢索策略的權重之後, 可用以推演出使用者的興趣並應用在資訊選粹服務與個人化檢索。運用 PKRG 進行資訊選粹服務的方法如下：

1. 透過 PKRG 計算出使用者用過的關鍵字詞的權重；
2. 將關鍵字詞依照權重排列, 挑出權重最大的前幾個關鍵字詞，當作該使用者的興趣，並存入個人興趣紀錄中；
3. 比對新進資訊所含關鍵字詞是否和使用者興趣記錄檔中的興趣關鍵字詞相符。若相符，則視該資訊為使用者有興趣的新進資訊。

本計畫中所實現的個人化檢索，乃指系統能考慮各關鍵字詞間的相關性，將檢索結果依使用者興趣重新排列。在概念上，我們認為經常共同出現的關鍵字詞對使用者而言代表這些關鍵字詞的關連性很強，因此當使用者使用某一檢索策略搜尋資料時，若檢索結果中含有和該檢索策略經常共同出現的關鍵字詞，便會是使用者比較有興趣的。例如：若使用者常使用 '*古典* AND *詩詞*' 這個檢索策略，代表使用者認為*古典*和*詩詞*間的關連性很強，因此當使用者用*詩詞*來尋找資料時，我們便認為包含*古典*這個關鍵字詞的檢索結果會是使用者較有興趣的。運用 PKRG 進行個人化檢索的方法如下：

1. 透過 PKRG 計算出使用者用過的檢索策略的權重，此權重亦代表關鍵字詞間的關連性，並將檢索策略權重資訊存入個人興趣紀錄中；
2. 當使用者輸入一檢索策略時，根據個人興趣記錄中的檢索策略權重資訊篩選出和當次檢索策略最有關的前幾個關鍵字詞；
3. 比對檢索結果和篩選出的關鍵字詞，含有愈多關鍵字詞的檢索結果代表使用者愈有興趣。

4. 資料探勘與群體化服務

我們將資料探勘技術中的相關規則探勘與循序規則探勘應用在圖書館，期能發掘出讀者借閱館藏之社群性，以作為實施群體化服務的依據，我們欲探索的讀者社群關係包含：

1. **館藏借閱的共同性**：興趣相同的讀者們往往會借閱類似的館藏，若我們能發掘出館藏借閱的共同性，當有某位讀者借閱某館藏時，我們便可推薦給他借過此館藏的讀者亦曾借閱的其他館藏。
2. **館藏借閱的順序性**：對於某些館藏，讀者可能會依據一定的順序來閱讀(例如先借入門的再借進階的)，若我們發現許多讀者都按照一定的順序來閱讀某些館藏，那麼當有某位讀者借閱這些館藏中的某一本時，我們便可建議他按照順序來閱讀相關書籍。

4.1. 前置作業

在在此步驟採取的動作如下：

1. 確定資料來源：資料來源為圖書館自動化系統中的書目檔、館藏檔、讀者檔、及交易歷史檔。
2. 資料選取：
   - ✧ 確定所要分析紀錄的時間範圍：交易歷史檔中通常包含許多年份的歷史紀錄，因此首先必須確定要分析的時間範圍。本研究的時間範圍是從 1998 年 1 月 1 日至 2000 年 8 月 31 日。
   - ✧ 確定所要分析紀錄的類別：交易歷史檔中通常包含許多類型的交易紀錄，例如借閱、歸還、預約、聲明歸還…等，由於本計畫所要探討的是館藏借閱的共同性和順序性，因此只需要用到交易歷史檔中的借閱和預約歷史紀錄。
     經過本步驟的處理，我們在交易歷史檔中選取了 1998 年 1 月 1 日至 2000 年 8 月 31 日間有關借閱及預約的交易歷史資料，共有 487,786 筆。
3. 資料的前置處理及轉換：針對發掘館藏借閱共同性和順序性的資料處理需求，進行必要的資料前置處理及轉換。
   - ✧ **館藏借閱的共同性**：採用相關規則探勘發掘館藏借閱的共同性。我們將每一筆書目資料視為一個項目(Item)，以書目號作為項目代碼，而每位讀者在一段時間內(如一學期或一學年)所借閱或預約的書目所成之集合即為一筆交易(Transaction)，資料庫 ***D*** 便是由所有交易組成的集合。舉例來說，若圖書館中有二位讀者 A 及 B，A 在一段時間內借了「1343」及「253」這二本書，B 在一段時間內借了「3423」、「34636」及「9689」三本書，資料庫 ***D*** 中就會有 {1343, 253} 和 {3423, 34636, 9689}二筆交易。
   - ✧ **館藏借閱的順序性**：採用循序規則探勘發掘館藏借閱的順序性。我們先將讀者借閱或預約的交易歷史紀錄依時間排序，並將每一筆書目資料視為一個項目，以書目號作為項目代碼，讀者同時間借閱或預約的書目所成之集合即為項目集，同一位讀者所借閱或預約的項目集依時間排序而成的序列即為讀者的交易序列。例如：圖書館中有一位讀者 A 在 10 月 1 日借了「3425」，在 10 月 15 日又借了「9823」及「4875」，則其交易序列即為 { (3425) (9823, 4875) }。

4.2. 探勘成果及討論

在經過前置處理之後，我們進行相關規則[16][17]以及循序規則的探勘[16]。相關規則探勘主要採用 Aprori[4]與 H-Mine[6]，循序規則探勘則是採用 GSP[5]。除此之外，我們還改良 H-Mine 演算法，提出 H-Mine(Gen-MMS)[17]演算法，針對書籍

類別進行探勘，並可在書籍類別的不同階層，設定不同的支持度(Support)。

我們將 1998 年 1 月 1 日至 2000 年 8 月 31 日間的借閱及預約歷史資料納入探勘，每位讀者在這一段時間內借閱和預約過的館藏都視為同一筆交易；然後，我們將只含一本書的交易刪除(因為不可能探勘出共同性和順序性)，以減少交易記錄的筆數。經過這些處理後，探勘的資料包括 11,398 筆交易，且最長的交易有 542 個項目。由於讀者借閱館藏的期限大多為一個月，因此，若假設館藏只有一本，且讀者一借就是一個月，則二年半內最多只有 30 位讀者可借閱，因此我們將最小支持度設為 0.21%(即 23 人，約是二年內可借閱的讀者人數)，並針對書籍館藏探勘，分析出讀者借閱館藏的共同性，部分成果如表 4。書籍借閱順序性的部分成果則如表 5。

表 4 館藏借閱的共同性 (部分成果)

書　　名	借閱人數
● 精通 Borland C++ Builder:視覺化 C/C++程式設計.基礎篇 ● Borland C++ Builder 視窗程式設計經典	52
● MPEG video:compression standard ● Digital video:an introduction to MPEG-2	51
● CMOS circuit design, layout, and simulation ● Low-power cmos wireless communications:a wideband CDMA system design	42
● MATLAB 入門引導 ● PC MATLAB 入門與實例應用	38
● 親蜜心事 ● 是誰拿走了那一雙雪靴	37
● CDMA systems engineering handbook ● CDMA techniques for third generation mobile systems	36
● FreeBSD 抓得住 INTERNET:伺服器架設與管理 ● FreeBSD 網路應用	34
● 精通 Borland C++ Builder:視覺化 C/C++程式設計.基礎篇 ● Borland C++ Builder 完全征服手冊	33
● JPEG still image data compression standard ● Win 32 多緒程式設計:執行緒完全手冊 =Multithreading Applications in Win 32	30
● 仙河飲馬 ● 淨土之春	30
● RF power amplifiers for wireless communications ● Microwave circuit design using linear and nonlinear techniques	28
● Win 32 多緒程式設計:執行緒完全手冊 =Multithreading Applications in Win 32 ● PC 影像處理技術.(一),圖檔壓縮篇	28
● FreeBSD 抓得住 INTERNET:伺服器架設與管理 ● 抓住你的 PhotoImpact 4.2 中文版	28
● 線性代數 ● 通訊系統	27
● Numerical Recipes in C:The Art of Scientific Computing ● An introduction to wavelets	24
● Visual C++ 6.0 程式開發手冊 ● Visual C++ 6.0 程式設計指南,應用程式架構篇	24
● 麵包樹上的女人 ● 賣海豚的女孩	24
● Delphi 4.0 徹底研究 ● 煞死你的網頁設計絕招	24

表 5 館藏借閱的順序性(部分成果)

書　　名	借閱人數
● MATLAB 入門引導 ● PC MATLAB 入門與實例應用	37
● Wideband CDMA for third generation mobile communications ● CDMA systems engineering handbook	32
● CDMA systems engineering handbook ● CDMA techniques for third generation mobile systems	29
● 仙河飲馬 ● 淨土之春	29

在分析出館藏借閱的共同性和順序性之後，我們可以將分析的結果供個人化檢索與推薦時使用。舉例而言，當使用者經由個人化檢索檢索到*麵包樹上的女人*這本書時，在使用者檢視這本書的詳細資料時，系統可以將*賣海豚的女孩*推薦給這位使用者(在二年半的時間裡有 24 位讀者看過*麵包樹上的女人*與*賣海豚的女孩*這二本書)；類似的，若使用者檢索到 *PC MATLAB 入門與實例應用*這本書時，我們可以建議他先閱讀 *MATLAB 入門引導*(在二年半的時間裡有 37 位讀者是依照 *MATLAB 入門引導*→*PC MATLAB 入門與實例應用*的順序來閱讀這二本書)。

在此必須特別指出，並非資料探勘的所有結果都是有價值的；例如我們在探索館藏借閱的順序性時就發掘出以下的順序性：名流劍客沒羽箭第一部 → 名流劍客沒羽箭第二部 → 名流劍客沒羽箭第三部 → 名流劍客沒羽箭第四部 → 名流劍客沒羽箭第五部。很明顯地，這是一個無用的循序規則。

5. PIE@NCTU 系統實作

根據前述系統架構，並輔以動態偵測讀者興趣的演算法[15]、以及資料探勘實施社群化服務的技術[16][17]，我們實作了一套適用於圖書館之個人化圖書資訊系統：PIE@NCTU。PIE@NCTU 的發展目的為：

- 提供一個 User-friendly、Web-based 的界面以便於讀者使用圖書館的圖書資源；
- 能針對使用者需求給予個別的使用環境；
- 具有社群化功能，促進讀者間的知識分享；
- 動態偵測讀者興趣，協助讀者尋找圖書館館藏；
- 提供資訊選粹服務，根據讀者興趣，定期通知讀者有興趣的新進資訊；
- 提供跨平台書籤功能，方便讀者利用館藏

資源；

- 提供讀者與圖書館互動的園地，促進讀者與圖書館的交流，並協助讀者解決利用館藏資源時遭遇到的困難。

從使用者的觀點來看，PIE@NCTU 是一個提供個人化和群體化服務的 WWW 網站。使用者可透過任何 WWW 瀏覽器進入。經由連接圖書館自動化系統的讀者資料檔，各館的讀者都能夠使用 PIE@NCTU 所提供的各項服務。讀者以其在所屬圖書館自動化系統中既有的帳號、密碼成功登入系統後，便能享用 PIE@NCTU 的所有服務。PIE@NCTU 的服務可分為共同使用環境、個人化使用者界面、管理者界面等三部分。圖 2 PIE@NCTU 首頁(http://pie.e-lib.nctu.edu.tw/pie)。

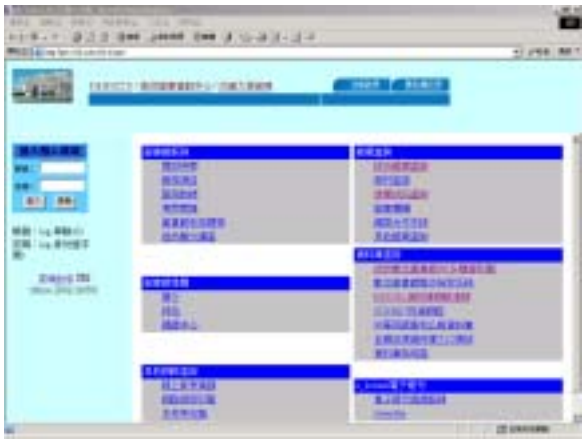

圖 2. PIE@NCTU 首頁

PIE@NCTU 所提供的個人化服務有： (1) 個人化環境設定， (2) 個人化搜尋引擎， (3) 個人書籤， (4) 資訊粹選服務， (5) 我要問問題, (6) 個人通告。由於我要問問題和個人通告與系統架構中的個人參考諮詢、個人資訊中心功能相近，不再贅述，以下僅針對前四項服務詳加說明。

- 個人化環境設定

  PIE@NCTU 提供三項個人化環境設定的功能：個人化桌面、個人服務設定、個人興趣設定。

  個人化桌面設定讓使用者根據其需要組織桌面環境。設計上是將整個桌面環境依不同功能劃分成數大類，各類別中含有子集合，然後提供選單讓使用者點選希望出現在其桌面上的項目與功能。個人化桌面提供的服務類別包括：圖書館服務、圖書館導覽、館藏查詢、資料庫查詢、新書通告、借閱狀況、檢索界面、系統公告、圖書館連結等。圖 2 右方畫面為系統預設的桌面環境，但在某使用者根據其需求設定之後的個人化桌面如圖 3。

  個人服務設定讓讀者選擇希望收到的資訊選粹服務，包含個人新書目錄、藝文活動通告、圖書館公告。

  至於個人興趣設定，顧名思義就是讓使用者設定其個人的興趣。儘管在 PIE@NCTU 中我們已經應用動態偵測興趣演算法來判斷讀者個人興趣，但是系統自動判斷出來的興趣可能沒有辦法完全符合使用者的興趣，為了輔助系統的不足，PIE@NCTU 提供讀者手動設定環境的功能；此外，使用者也可以設定感到興趣的書目類別。

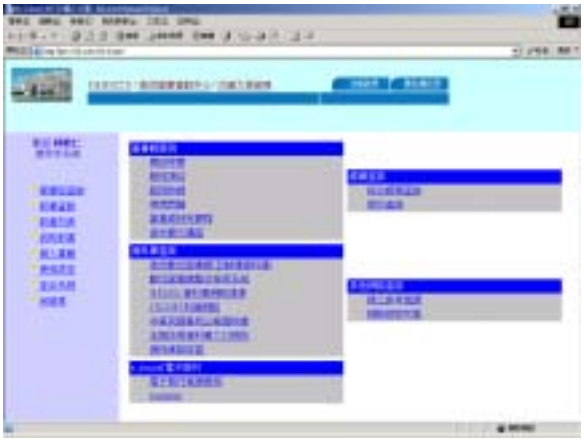

圖 3 個人化桌面

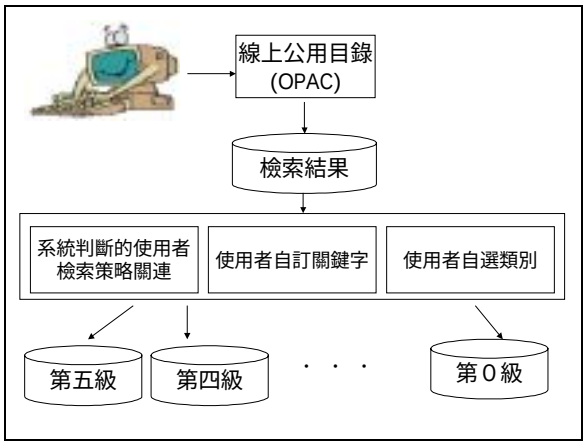

圖 4 個人化搜尋引擎示意圖

- 個人化搜尋引擎

  PIE@NCTU 提供二種個人化搜尋引擎服務：個人館藏查詢、個人新進館藏查詢。圖 4 為 PIE@NCTU 如何達成個人化搜尋引擎的示意圖。當使用者輸入一檢索策略時，系統首先會根據此檢索策略修正讀者個人興趣，然後將此檢索策略傳遞給線上公用目錄(OPAC)搜尋館藏，當線上公用目錄將檢索結果回傳給 PIE@NCTU 後，PIE@NCTU 根據系統判斷的使用者檢索策略關聯、使用者自訂興趣關鍵字詞、使用者自選興趣類別等三個條件，將檢索結果分成六個等級加以排序。愈符合上述三個條件的檢索結果等級越高，亦即系統認為該檢

索結果對使用者較有用，在結果呈現時，會將其排列在較明顯的位置。系統將檢索結果分成零至五共六個等級，其中皆不符合為第零等級，剩餘五個等級依檢索結果符合上述三個條件的程度排列。圖 5 為某使用者檢索*事件*這個關鍵字詞的結果，列出的六筆資料為等級三的結果。

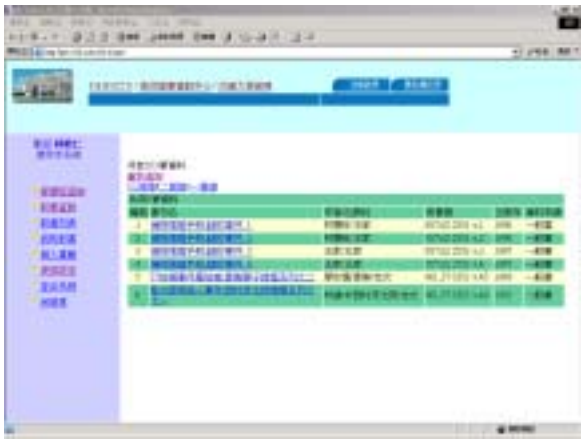

圖 5 以「事件」為檢索策略進行個人化檢索的結果(等級三)

- 個人書籤

  PIE@NCTU 提供跨平台的書籤功能：「個人書籤」，可讓使用者紀錄有興趣書目的超連接、及使用者的註解。圖 6 為個人書籤的範例。

- 資訊選粹服務

  PIE@NCTU 所提供的資訊選粹服務可定期比對新進資訊和使用者當時的興趣。資訊選粹服務的內容有個人新書通報、個人藝文活動通報，個人圖書館公告等，未來將繼續發展其他的個人資訊選粹服務。PIE@NCTU 係根據系統判斷的使用者興趣關鍵字詞、使用者自訂興趣關鍵字詞、使用者自選興趣類別，來作為資訊選粹服務的依據，由於做法與個人化檢索類似，在此不加贅述。

圖 6 個人書籤

除了前述個人化功能外，我們亦將資料探勘的成果與 PIE@NCTU 整合。我們所使用的資料探勘技術為相關規則探勘(Association Rule Mining) [16][17]，當使用者在查閱某一館藏的詳細資料時，PIE@NCTU 會顯示與該館藏經常被同一讀者借閱的其他館藏。範例如圖 7。

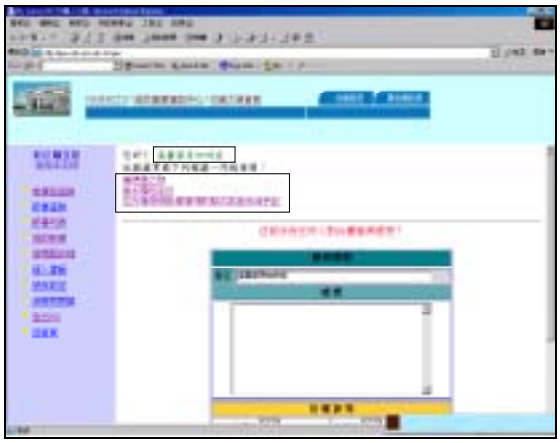

圖 7 整合資料探勘與 PIE@NCTUrary

6. 結論與未來發展

在資訊爆炸與網路科技的時代，圖書館在校園中所扮演的角色必須重新定位。圖書館除了提供使用者快速、精確、完整的查詢之外，還要能累積前人的智慧、綜合讀者的知識、提升校園閱讀文化，而建置一整合個人化與群體化的圖書館讀者服務，正是達到此一目標的必要之途。

本計畫提出一套兼具個人化與群體化服務的圖書資訊環境系統：PIE@NCTU。在未來，為了增加圖書館館藏資源的附加價值、增進讀者學習興趣，彼此激勵學習興趣、提升圖書館與讀者互動、讓讀者能充分利用圖書館的資源，我們將持續設計以下功能。

1. 投票系統：藉由讀者投票讓圖書館更加瞭解讀者的意見。
2. 討論區：討論區讓讀者針對各主題發表意見與分享別人的看法；圖書館亦能經由討論區瞭解讀者的意見及看法，作為業務參考的依據。
3. 知識分享與網路讀書會：網路讀書會讓讀者針對特定書目/主題發表意見、獲得其他讀者的看法；也可以透過資料探勘的成果將讀者組成社群，讓彼此分享社群成員的學習心得、搜尋策略...等，藉此促進知識分享，激發更多靈感。
4. 協力式資訊過濾：在本計畫中我們係以個人興趣來做資訊過濾，未來可加入群組興趣實現協力式資訊過濾。

5. 在資料探勘中考量讀者分類：目前我們是將圖書館的所有讀者視為一個大社群，從中了解成員在館藏借閱的共同性與順序性。然而，讀者背景與學科領域可能會影響到其借閱行為，因此若能先將讀者分群(如根據系所、學院、性別、年級等分群)，再針對每一群讀者探索其借閱的共同性和循序型，相信資料探勘的結果更能切合讀者的需求。

7. 參考文獻

[1] 蔣以仁，「一對一個人化服務機制」，電腦與通訊，95 期，(民國 90 年 3 月)，頁 88-93。

[2] 張進群，陳建良，「客戶關係管理」，機械工業雜誌，89 年 12 月號，頁 165-176。

[3] R. Agrawal, T. Imieliski, and A. Swami. “Mining Association Rules between Sets of Items in Large Databases”. Proceedings of the 1993 ACM SIGMOD Conference, pp.207-126.

[4] Agrawal, R., & Srikant R. (1994) Fast algorithms for mining association rules. In Proceedings of the International Conference on Very Large Data Bases (VLDB'94) (pp. 487 - 499), Santiago, Chile: VLDB Endowment.

[5] R. Srikant and R. Agrawal. “Mining Sequential Patterns: Generalizations and performance improvements”. *IBM Research Division Almaden Research Center*, 1995.

[6] Pei, J., Han, J., Lu, H., Nishio, S., Tang, S., & Yang, D. (2001). H-Mine: hyper-structure mining of frequent patterns in large databases. In Proceedings of International Conference on Data Mining (ICDM'01) (pp. 441 – 448), San Jose, CA: IEEE.

[7] S. Cohen, J. Fereira, A. Horne, B. Kibbee, H. Mistlebauer, and A. Smith, “PIE@NCTUrary: Personalized Electronic Services in the Cornell University Library”, *D-Lib Magazine*, April 2000.

[8] K. Morgan and T. Reade, “Pioneering Portals: PIE@NCTUrary@NCState,” Information Technology and Libraries, 19 (4): 191-198, 2000.

[9] J. C. French and C. L. Viles, “Personalized Information Environments: An Architecture for Customizable Access to Distributed Digital Libraries”, D-Lib Magazine, June 1999.

[10] I. B. Crabtree and S. J. Soltysiak, “Identifying and tracking changing interests”, International Journal on Digital Libraries, 2: 38-53, 1998.

[11] 卜小蝶，「網路資訊過濾技術與個人化資訊服務」，21 世紀資訊科學與技術國際研討會，台北市：世界新聞傳播學院圖書資訊學系，(民國 85 年 11 月 7-9 日)，頁 339-350。

[12] 卜小蝶，「提供個人化服務的線上公用目錄檢索系統初探」，中國圖書館學會會報，59 期，(民國 86 年 12 月)，頁 127-133。

[13] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, and A. Zanasi. “Discovering Data Mining: From Concept to Implementation”. Prentice Hall, Upper Saddle River, NJ, 1998.

[14] 林高煌，「一個有效率的大參考序列探勘方法及 其在全球資訊網上的應用」，國立交通大學資訊科學系碩士論文，(民國 89 年 6 月)。

[15] 楊雅雯，「個人化數位圖書資訊環境 - 以 PIE@NCTU 為例」，國立交通大學資訊科學系碩士論文，(民國 90 年 6 月)。

[16] 吳安琪，「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」，國立交通大學資訊科學系碩士論文，(民國 90 年 6 月)。

[17] 戴玉旻，「圖書館借閱記錄探勘系統」，國立交通大學資訊科學系碩士論文，(民國 91 年 6 月)。