圖書與資訊學刊
50(Aug '04)76-91
ISSN 1023-2125

資料探勘應用於圖書館之探討

Applying Data Mining to the Library

彭 于 萍
Yu-Ping Peng
銘傳大學管理科學研究所博士生
Ph.D. Student, Ming Chuan Graduate Institute of Management Science
E-mail：sunniapg@ms43.hiner.net

【摘要 Abstract】

為達成圖書館滿足使用者資訊需求之目標，圖書館實有掌握使用者需求，並主動行銷服務之必要性。而近年極熱門應用於企業行銷決策之資料探勘技術，能否應用於圖書館環境協助決策，值得探討。本文說明資料探勘之意義、功用，並介紹一般性的實施過程與應用技術，接著針對圖書館服務內容及館藏特質，進一步剖析圖書館應用資料探勘技術之適用範圍。最後，將研究建議提供圖書館決策者參考。

In order to achieve the goal that library should satisfy the information needs of users, it is necessary to understand the needs of users and marketing. Data mining applied to the decisions of enterprises popularly in recent years. Whether data mining could apply to the library environment to assistance decision, it's worthful to discuss. This paper discusses the significance and function of data mining, and introduces the general processes and technologies at first. Then, I aim at the characteristics of library services and collections to analyze the appropriateness of applying data mining to the library. Finally, implications for decision makers of library were proposed.

關鍵詞 Keyword

資料探勘 圖書館
Data mining；Library

壹、研究動機

Aristotle Onassis 指出：「企業致勝關鍵在於得知沒人知道的東西」。(註 1)西元 1980 年後，各企業陸續建置資料庫，蒐集並累積大量顧客、競爭者以及產品之交易資料，其中隱藏許多記錄著企業決策者或消費者的決策過程及結果之重要資訊。因此，若能由資料寶山挖出隱藏其中不易發現的知識或訊息，進而發展成有效的消費行爲或模式，除了協助商業應用、後續制訂決策及預測之外，更重要的是此資訊可能成爲企業致勝的關鍵，此一過程稱爲「資料探勘(Data Mining)」。根據麻省理工學院(MIT)的科技評論中指出，資料探勘是未來改變世界的十大新興科技趨勢之一，時代雜誌更預測－資料探勘是二十一世紀最重要的五大新興行業之一，顯見資料探勘之重要性。

「資料探勘」最常應用在企業行銷及顧客關係管理決策之制訂。就現代圖書館而言，在網路發達、資訊激增、媒體多樣化等因素相互雜揉之影響下，讀者服務項目及內容複雜度亦隨之大增。爲提高圖書館資訊資源之利用率，達成「適時提供適當資訊給適當的使用者」之金科玉律，圖書館實有切實掌握使用者需求、進行顧客關係管理，並主動積極行銷及推廣適當服務之必要性。而資料探勘是否也能應用於圖書館環境，以協助達成滿足使用者資訊需求之目標呢？

針對「資料探勘應用於圖書館環境」之議題，國內外研究文獻爲數不多，國外文獻多由理論觀點探究其適用性，國內則以實際進行資料探勘之實證性學位論文爲主。總括來說，圖書館或可應用資料探勘支援決策，更加瞭解圖書館使用者行爲，亦能藉經探勘得知的新關係，重新規劃館藏發展方向、設計圖書館活動計畫。但亦有學者質疑資料探勘應用於圖書館環境之適用性，主張因圖書館服務內容及館藏特性，貿然引進資料探勘技術將面臨窘境，且又未能達成預期目標。

綜上所述，「資料探勘應用於圖書館」議題實有深入釐清、剖析之必要。因此，筆者藉由文獻分析方式，探討資料探勘之意義、功用，以及一般性的實施過程與應用技術，再針對圖書館服務內容及館藏特質，進一步剖析圖書館應用資料探勘技術之適用範圍、用途，及其困難與限制等亟待研究突破之處，以供圖書館決策者參考。

貳、資料探勘之意義及功用

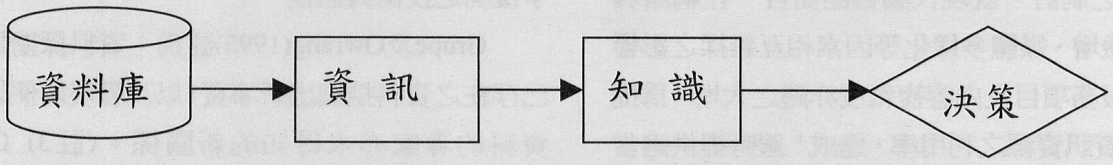
「資料探勘（Data mining）」係針對大量資料進行分類、計算、排序，甚至找出存在資料之間「隱而未覺」之關聯性的一種技術。(註2) 若進一步以企業實務角度來說，資料探勘可謂由資料庫已存在之「資料」中，探勘出「前所未知」、「隱而未覺」、「不明顯」但卻「有意義且可據以行動」之「新關係或事實（有效資訊）」，俾利於企業決策、發展競爭優勢之技術與過程。

Grupe及Owrang(1995)認爲：資料探勘是使用已存在之資料挖掘出新事實，以及發現即便是熟悉資料的專家亦未得知的新關係。(註3) Cabena（1997）指出：資料探勘係「由大型資料庫中，萃取先前未知、有效、可據以行動之資訊，並使用此資訊執行重要企業決策。(註4) Berry及Linoff（1997）說明：資料探勘就是使用自動或半自動的方式對大量資料作分析，以尋找事前未知、有趣且可據以行動的規則或知識。(註5) Peacock（1998）則歸納出三種層次－狹義、廣義、最廣義來解釋資料探勘。狹義意指自動「發現」隱藏在資料庫中「有用」但不明顯的模式，所謂「有用」係指新發現的關係可能對企業策略甚至組織目標造成根本上的影響。廣義重點在於應用傳統方法(如統計)來「證實」發現的新關係。至於最廣義則牽涉到另一重要名詞--「資料庫知識發現（KDD；Knowledge Discovery in Databases）」。(註6)

「資料探勘」、「資料庫知識發現」兩者常被相提並論甚至混爲一談。一般而言，「資料庫知識發現」意即由資料庫中發現知識的過程。再以Frawley及Piatetsky-Shapiro(1991)的解釋爲例，「知識發現（Knowledge Discovery）是從資料中萃取出不明確、前所未知及潛在可用性資訊的過程。」(註7)由上可知，兩者意義似乎難以明確區辨。

究竟資料探勘只是資料庫知識發現（KDD）其中的一個步驟而已？還是兩者同義－資料探勘就是從資料庫中發掘知識的過程？前者以Fayyad(1996)研究爲代表，細述「資料庫知識發現」流程，包括：瞭解資料與應用領域、熟悉相關知識與技術，接著融合與查核資料，並去除錯誤或不一致的資料、發展模式與假設，再應用「資料探勘技術」建立樣式，最後測試與檢核所挖掘的資料，才得到一些有用的知識，進而解釋與使用。(註8)因此，Fayyad認爲資料探勘僅爲資料庫知識發現數項步驟之一。另一方面，就主張「資料探勘即爲資料庫知識發現」的學者而言，Berry及Linoff（1997）說明下列工作內容更可適切解釋資料探勘之意義，包括：分類、推估、預測、關聯規則、群集、描述，此亦「知識發現」之重要工作，因此，Berry及Linoff表示：資料探勘就是從資料庫中發掘知識的過程。(註9)在Peacock（1998）亦明確指出：最廣義的資料探勘即爲「資料庫知識發現(KDD)」。(註10)學者意見分歧、答案見人見智，筆者則較認同後者看法，「資料探勘」應以較宏觀的角度視之，應包含由最初決定目標至最後由資料庫中發掘知識之完整過程，如此資料探勘結果對決策者方有意義。

「資料探勘（data mining）」爲何利於決策？簡言之，資料探勘是一種資料轉換的過程，先由沒有組織的數字與文字集合的資料，轉換爲資訊，再轉換爲知識，最後運用知識產生決策。(註11)如圖一所示：



圖一：資料探勘資料轉換過程

資料來源：Curt H. (1995). The Devile's in The Detail: Techniques, Tool, and Applications for Data mining and Knowledge Discovery-Part 1. Intelligent Software Strategies, 6(9), 3.

更進一步來說，傳統企業的資訊系統對於大量的資料視爲過去歷史紀錄來用，並未善加利用這些企業經營所保留下來的智慧，事實上，這些資料記錄了企業內決策者的決策過程或消費者的消費決策結果，因此，若能由這些資料所記錄之決策經驗，找出顯著且有效的決策模式或決策法則，進而正確預測未來的行爲，將能賦予組織更多的經營智慧。(註12)因此，資料探勘具有建立「預測」模型的能力，而利於決策。

參、資料探勘之過程

資料探勘過程良窳實爲資料探勘是否成功之關鍵。在諸多「資料探勘」重量級巨作中，均詳述資料探勘過程，擇要陳述如下：

一、Berry、Linoff(1997, 2000) (註13) (註14)

1997年，Michael J.A. Berry與Gordon S. Linoff在《資料探勘技術：行銷、銷售與顧客服務之應用》

一書中，係利用「資料探勘四階段工作循環」之議題，來說明資料探勘過程；而 2000 年，在其新著《資料探勘理論與實務:顧客關係管理之技巧與科學》中，再次強調：資料探勘需藉助上述四階段過程方能成功，更進一步地闡述四階段工作循環之精義。謹將其論點簡要彙整、分述如下：

(一)辨識企業問題（Identifying the problem）

企業可透過定期的經營診斷，或與專家訪談的過程中，開發企業新的機會點，旨於尋找、確認具探勘價值之資料範圍，以便投入下一階段的資料探勘工作。

在此階段重要的是，當資料探勘技術人員進行下一階段的資料探勘之前，絕對需先瞭解商業上真正需要的是什麼，因此每個資料探勘專案，均需要技術人員和瞭解該類型商業的人溝通，也就是和所謂的專家進行溝通，以回答下列問題：

1.資料探勘對解決問題果真必要嗎？投注在資料探勘上的努力是否值得？
2.是否有某一特殊的部分或集群令人最感興趣？
3.所需分析資料是什麼？有那些資料源已無效？某些資料應從何而來？
4.根據專家的直覺和經驗，那些資料是重要的？

而「資料」通常包括瞭解現有客戶及未來潛在客戶之相關資訊、瞭解產品與市場的關聯、瞭解供應者與合夥人的關係，瞭解企業管理程序--包括所蒐集到之相關資料，會影響到的所有部分。

(二)分析資料（Analyzing the data）--將資料轉換為可用以採取行動的資訊

定義企業機會點並決定研究主題後，下一步係選擇合適資料，挖掘隱藏其中但可能有用的資訊，以供制訂後續行動方案。在此階段，主要目標係建立資料探勘模型，重要步驟依序如下：

1.判定及獲得正確的資料

基本上，正確的資料意指資料是簡單、乾淨、完整、可使用、具有適當欄位，資料量越多越好。更重要的是，需判定這些資料是否可解決上述定義之企業問題。

2.將資料淨化並提升資料有效性

在進行資料探勘之前，「資料淨化（Data Cleaning）」是相當重要的。翁頌舜、梁德馨說明「資料淨化」包含資料整理與處理資料中不符定義的數值(例如缺值或年齡爲負值等)，採用淨化後的資料再依專業知識作合理性的判斷(判斷是否在淨化後樣本會偏離母體的情況)，再以此資料做分析，才能得出更嚴謹及正確的結果。(註15)

除了資料需加以淨化之外，爲提升資料有效性，尚須注意下列問題：

(1)欄位樣本數足夠嗎？遺漏的資料會造成大問題嗎？
(2)這些欄位的值有效嗎？是否介於適當範圍之內？
(3)個別欄位之分布情形是否可解釋？

更需注意的是，資料探勘所採用的資料常來自於不同的形式、格式以及系統，不良的資料格式及令人混淆的變數，常成爲影響資料探勘結果的陷阱。

3.資料轉換

使用資料探勘演算法時，需將描述同一對象之資料放在同一列處理，而這些資料須利用正確的層級進行加總（「層級」是指資料被處理之單位）。

4.加入衍生變數

衍生變數的值係由資料內其他值的結合而成，例如：從某一時期開始到結束之間使用率的成長、新費用爲零的月數等。

亦有另一種衍生變數可提供關於其他欄的相關資料，例如：針對每位客戶，以十分位來計算他們在某一時期之總花費，最後將第一級客戶值設爲1，次一級客戶值設爲2，以此類推。

5.確定實際用來建立資料探勘模型之資料

並非資料倉儲內所有資料均需作爲建立資料探勘模型之資料，可根據不同部分資料建立不同的預測模型。在此，需注意是否有過多極端值的問題。

6.選擇模型建立方法及訓練模型

模型建立方法包括類神經網路及決策樹等，其優缺點因資料特性、使用目的而異。在過去，不管是建立或訓練模型都很困難，但目前利用現成的資料探勘工具已可解決此部分的問題。

7.確認模型的表現並選擇最好的模型

由於不同的資料探勘方法，其評估方式亦不同，因此，在使用同一種資料的情況下，需檢查不同模型的效能、確認不同模型的表現，以選出表現最好的模型。

(三)根據資訊行動（Taking action）

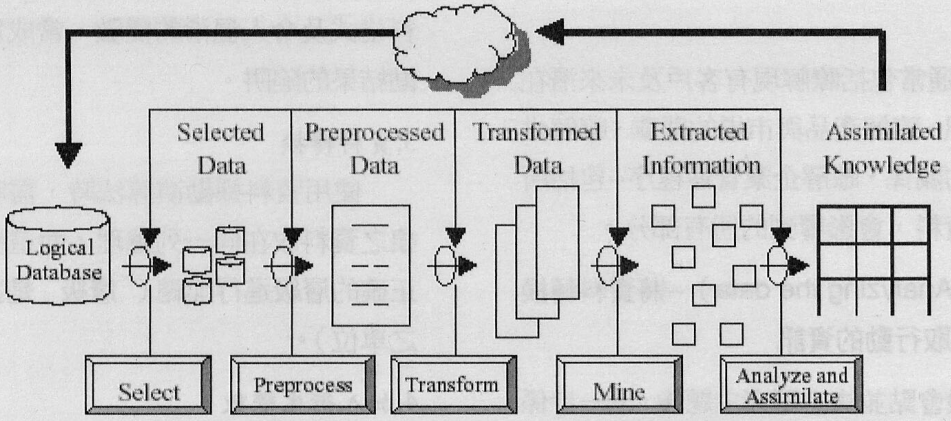
經過資料探勘後，最重要的是，如何將「經過探勘所發現之有用資訊」連結運用於「企業內部流程」上，若未能根據資訊行動，則代表資料探勘並無提供任何益處。

(四)測量、評估探勘結果（Measuring the outcome）

測量結果提供持續改善資料探勘結果的回饋（Feedback）機制，在此所指的測量不僅僅是單純的使用率、成本統計而已，它涵蓋了所有企業價值，這也正是維持良好資料探勘工作循環的重點。

二、Cabena et al. (1998)

1998 年，Cabena 等五人合著之《Discovering Data Mining：From Concept to Implementation》一書中，亦說明在一般性的資料探勘專案應具備之重要過程，及各過程所需付出努力之比重。值得注意的是，作者強調各個過程均需專案團隊投入大量心力，儘管資料探勘的技術不斷發展進步，但它仍是非常勞力密集的活動。專案團隊成員則包括企業分析師(Business analyst)、資料分析師(Data analyst)、資料管理專家(Data management specialist)，通力合作完成專案。資料探勘流程可參見圖二及以下說明：(註 16)



圖二：資料探勘流程圖

資料來源：Cabena, P., Hadjinian, R., Stadler, J. V., & Zanasi, A. (1998). Discovering Data Mining：From Concept to Implementation. N. J.：Prentice-Hall.

(一)決定企業目標(Business Objectives Determination)－20%：

清楚定義企業問題或挑戰，是資料探勘中不可缺少的一部份。雖然這項工作一聽之下感到很直覺、簡單，實則不然，它主導整個資料探勘過程，爲一重要基礎。

(二)資料準備（Data Preparation）－60%：

此爲資料探勘過程最耗神的步驟，約需投入60%的時間與精力。此一步驟又分爲三部份：

1.資料選擇：

識別所有內外部資訊來源，並選擇資料探勘所需資料。

2.資料前置處理：

確保選取資料的品質，爲未來分析作準備，並以決定探勘演算法。

3.資料轉換：

將資料轉換成資料探勘演算法所適合的模式。

(三)資料探勘（Data Mining）－10%：

選擇適當的資料探勘演算法來探勘上一個步驟所轉換出來的資料，此步驟快且爲自動化。

(四)分析結果及吸取知識(Analysis of Results and Assimilation of Knowledge)－10%：

先解釋與評估步驟(三)所探勘出來的資料，再將分析出來的資料加到企業化的組織或資訊系統中。

三、小結

在上述數篇探究資料探勘理論與實務議題之著作中，所闡述之資料探勘過程步驟大致相同。重要的是，皆強調：資料探勘若要成功，其過程必先瞭解企業問題，進行事前規劃與資料準備，再利用資料探勘技術來連結企業需求與資料，以探勘出有用資訊，再進而應用此知識，以解決企業問題(最常見應用於行銷問題上)。換言之，資料探勘若要徹底發揮效能，必須與企業流程緊密結合；而每一資料探勘過程之間，亦需環環相扣，由專案團隊積極溝通、合作，方能成功。

肆、資料探勘之技術

資料探勘(Data Mining)係利用資料來建立模式，並利用這些模式來描述資料特徵(Patterns)以及關係(Relations)，藉此提供決策所需資訊，同時協助進行預測。(註17)依據分析方式與產生的知識型態，最常見之資料探勘技術可分爲下列四類：

一、分類分析(Classification Analysis)

分類分析是從已知類別的物件集合中，依據其屬性(可能影響物件類別的變數)建立一個分類模式(如決策樹或決策法則)來描述物件屬性與類別之關係，然後再根據這個分類模式對其他未經分類或是新的資料做預測。這些用來尋找特徵的已分類資料可能是來自現有的歷史性資料，或是企業蒐集保存的客戶基本資料(Profile)。(註18)

換言之，分類分析是一種依分析對象(資料)屬性分門別類並加以定義，以建立類組（Class）的過程。例如：將信用貸款申請者，區分爲高度風險、中度風險、低度風險的申請者。此技術主要在於利用一些已經分類的資料來研究其特徵，再依據這些特徵對其他未經分類或是新的資料作預測。

二、集群分析(Clustering Analysis)

集群分析係指將所有的資料分成若干集群的過程，也就是根據物件間的相似性(或不相似性)，將所有的物件分成若干個集群（Cluster），使得每個集群內的物件具有高度的相似性（群內同質），而不同集群間具有高度的不相似性（群間異質）。(註19)集群分析的目的是要找出集群之間資料的

差異性，以及集群內資料之相似性。

集群分析與分類分析最大差別在於－集群並沒有事先定義好類別，也沒有訓練樣本（Training Set）。所有紀錄根據彼此的相似程度歸類。(註20)

三、關聯法則分析(Association Rule Analysis)

關聯法則分析係利用「支持度(Support)」與「信度(Confidence)」兩個參數，計算資料庫中所有資料項目間的相關強度，並判斷此一關聯法則是否具有意義，以找出某些資料項目間彼此的關聯性，進而將此分析作爲預測之依據。

「支持度」是資料庫中包含X與Y聯集的項目所佔次數，記爲Support(X∪Y)；「信度」則是定義此關聯法則可以信賴的程度，也就是X出現的條件下，Y也跟著出現的條件機率，記爲Support(X∪Y)/Support(X)。有意義的關聯法則，其支持度與信度必須大於等於使用者所定之最小限制，且信度大於60%。例如在一交易資料庫中，一筆交易記錄中同時購買產品X與產品Y，在此一資料庫中總共出現的次數，便記爲Support (X∪Y)，代表在交易中產品X與Y同時出現的次數。而信度便是在交易資料庫中，單獨購買產品X交易記錄次數的條件下，同時購買產品X與Y的交易記錄次數的百分比。(註21)

就零售業超級市場而言，係針對每一交易項目，分析會一起交易的項目組合有哪些?最著名的使用技巧及範例爲「購物籃分析(Market Basket Analysis)」，此係分析超級市場交易中，很可能會與某一商品一起購買的其他商品(如牛奶與麵包、尿布與啤酒)。藉由這些消費者購物行爲分析，業者確認交叉銷售（Cross-Selling）的機會，以調整貨架上的擺設位置，或調整存貨與訂單數量，進而設計促銷活動，希望更有效銷售貨品。(註22)

四、次序相關分析(Sequential Pattern Analysis)

次序相關分析的目的即是由一群有次序性的交易中，找出經常循序出現的交易項目(如大多數的客戶購買洗衣機後會購買烘衣機；購買錄影機，在一段時間後將會購買錄影帶；或當買新房子後一個月內買新爐具的比率爲45%，而二週內買新電冰箱的比率爲60%)，進而瞭解顧客的長期購買行爲。換言之，即從「一段時間」的資料中，找出物品交易的相關性、相依性。經由長期分析購買者行爲，業者可據以調整存貨與訂單數量，進而向已購買第一種產品的顧客促銷與其具有相依性的產品。(註23)

五、小結

欲解決某一種問題時，資料本身特性將影響探勘者所選用的技術，並沒有一種資料探勘技術可應付所有解決問題之需求，故需要用到許多不同技術從資料中找到最佳模式。更重要的是，某種技術或許僅適用於某些領域，在技術與資料之間的配適需格外注意，以達到建構模式之效能。

伍、圖書館應用之資料探勘技術及過程

一、圖書館應用之資料探勘技術及其用途

資料探勘(Data Mining)係藉由分析大量資料以建立模式，並利用這些模式來描述資料特徵以及關係，藉此提供決策所需資訊，同時協助進行預測。(註24)而依據分析方式與產生的知識型態，最常應用之資料探勘技術包括:「關聯法則分析」、「分類分析」、「集群分析」、「次序相關分析」四類。以下謹說明圖書館應用之資料探勘技術及其用途：

(一)關聯法則分析(Association Rule Analysis)

關聯法則分析可謂在現有實證研究中最常應用之資料探勘技術，可利用它針對圖書館個別使用者，提供「顧客化圖書推薦服務」，加強圖書資源之行銷、推廣服務，以提升圖書資源利用率。

目前許多圖書館雖有提供新書到館通知單之

服務，但成效並不顯著，若能針對不同讀者提供符合其興趣之新書，將可有效提升圖書資源利用率。(註25)或是利用它發掘讀者社群關係，找出館藏借閱共同性，運用這些社群關係適時建議讀者借閱其他館藏，以增加讀者繼續借閱館藏之機會，如此一來便可以提昇讀者的忠誠度，使讀者不會只借一、兩次就不再來借閱。(註26)

值得考量的是，在應用關聯法則分析時，「支持度(Support)」與「信度(Confidence)」參數值之高低，大大影響分析結果，過與不及均不能獲得理想、具參考價值之關聯法則，故設定時需格外審慎。

(二)集群分析(Clustering Analysis)

圖書館資料可利用集群分析，找出圖書與圖書之間、讀者與讀者之間的關係，探討使用者群體特性，找出其借閱行為傾向。國內所有實證研究均探討此一技術應用。應注意的是，因事前較不易預設其屬性特性，利用此技術較易尋找隱藏其中的關係集群分析可找出屬性相近的集群，此種分析會因集群數目及借閱總數門檻影響分析結果。(註27)

(三)分類分析(Classification Analysis)

分類分析是從已知類別的物件集合中，依其屬性建立分類模式來描述物件屬性與類別之關係，然後再根據這個分類模式對其他未經分類或是新的資料做預測。(註 28)

圖書館可利用分類分析進行使用者分析，以瞭解哪些族群對圖書館使用率較高。(註 29)

(四)序相關分析(Sequential Pattern Analysis)

圖書館可應用次序相關分析技術，找出讀者館藏借閱順序，進行圖書推薦服務。讀者借閱館藏可能會先借入門的，再借深入的，如果把讀者借閱館藏的順序特性找出來，則下次有某位讀者借入門的館藏時，即可推薦他借閱進階的館藏，讓讀者很容易地知道這本館藏的進階書籍有哪些。(註30)顏嘉惠（2002）曾舉例說明此一概念。(註31)

二、圖書館資料探勘過程

圖書館在引進資料探勘技術之前，除了需遵循資料探勘過程之一般性原則逐步施行，同時為使資料探勘結果可切實改善服務績效並協助決策，應特別注意圖書館服務特性、館藏特質對資料探勘效能所可能造成之影響。以下彙整國內外相關文獻，說明圖書館資料探勘過程：

(一)決定圖書館目標

1.決定圖書館可應用資料探勘技術協助決策之目標

早在1993年，Gleeson及Ottensmann曾說明如何應用流通及編目電腦化系統之資料，以制訂公共圖書館之管理決策。(註32)1996年，《Library Administration & Management》有一特輯－Mining your Automated System，Atkins、Larsen、Mancini和Peters均撰文闡述圖書館應探勘自動化系統內的資料，以利於獲取管理資訊、提升管理績效、制訂管理決策。(註33) (註34) (註35) (註36)

圖書館管理決策最終目標究竟為何？無非是適時提供適當資訊給適當的使用者，以滿足其資訊需求。而在現今使用者需求多元化、資訊類型更多樣化的網路時代，圖書館如何切實瞭解、掌握無論到館或透過網路利用圖書館資源之使用者需求，是提高圖書館資訊資源之利用率、規劃適當服務內容之基本要件。此項工作或可應用資料探勘達成。

1998年，Schulman表示圖書館可應用資料探勘支援決策，作用在於「更加瞭解圖書館使用者行為」。同時，資料探勘亦能讓圖書館重新規劃館藏發展方向，並能根據所瞭解的使用者行為及經探勘得知的新關係，來設計圖書館活動計畫。(註37)

因此，理論上圖書館可利用資料探勘技術協助達成決策目標。

2.資料探勘專案團隊之溝通

Berry和Linoff(2001)在此階段指出：當資料探勘技術人員進行探勘之前，絕對需先瞭解企業所需，因此每個專案均需技術人員和瞭解該類型企業的專家進行溝通。(註38) Cabena(1998)亦同樣強調：儘管資料探勘的技術不斷發展進步，但它仍是非常勞力密集的活動。過程中需要專案團隊投入大量心力、通力合作，而團隊成員包括企業分析師(Business analyst)、資料分析師(Data analyst)、資料管理專家(Data management specialist)。(註39)

對圖書館資料探勘專案而言，瞭解該類型企業的專家、企業分析師即爲圖書館館方人員，必需回答下列問題：(註40)

(1)資料探勘對解決問題果真必要嗎？投注在資料探勘上的努力是否值得？

(2)是否有某一特殊的部分或集群令人最感興趣？

(3)所需分析資料是什麼？有那些資料源已無效？某些資料應從何而來？

(4)根據直覺和經驗，那些資料是重要的？

上述(2)、(3)、(4)項有關「資料」問題，在下一部分「(二)資料準備－(1)資料選擇」將詳爲探討，此不贅述。而第1項問題：資料探勘對解決問題果真必要嗎？投注在資料探勘上的努力是否值得？值得圖書館決策者審思。

1998年，Banerjee在《Is Data Mining Right for Your Library》一文中，提出：圖書館雖已有適用於各圖書館資訊分享之資料處理標準，但並未建立適用於資料探勘之儲存與檢索標準。由於前者著重各館使用相同標準以利分享，但後者希望的是能表現各館個別特色之標準，兩者之間產生決策上的衝突。若某一圖書館欲自行發展較適用於資料探勘之儲存與檢索標準，可能需面對成本增加及既有資料轉換成功率的風險。(註41)

但同年，Schulman則陳述不同的看法，表示或許應用資料探勘技術並不在圖書館經營需求之內，同時圖書館也不欲投資；不過倘若圖書館已有資料量極大或具特定用途的資料庫在運作中，經營者應會考慮建置實務決策支援系統。且當資料庫持續成長及改變，則經營者幾乎不可能以人工方式來掌握不斷變動的使用者行爲模式及趨勢，亦無法迅速調整出最佳的館藏發展方向、提供最新資訊服務。(註42)而資料探勘技術正可彌補上述不足之處。

因此，圖書館在應用資料探勘技術之前，首要之務在於審慎評估資料探勘技術是否適用於圖書館現有條件，以及可否協助達成組織目標。評估結果若認爲資料探勘對解決問題確有其必要性，則可進一步執行下列步驟。

(二)資料準備

1. 資料選擇--尋找、確認具探勘價值之資料範圍

在決定圖書館目標後，進入「資料準備」階段首先要做的是--識別所有內外部資訊來源，選擇資料探勘所需應用之資料。此階段對於最終之資料探勘結果是否具決策應用之價值，影響殊深。析要如下：

(1)產生圖書館服務交易資料之服務項目

目前資料探勘技術多成功應用於商業或科技業領域，分析銷售紀錄之類已結構化之交易資料；但就圖書館而言，若欲完整分析圖書館服務交易情況，則其中必包含大量、非結構化、資料來源多樣化的交易內容(如紙本圖書、紙本期刊、線上資料庫等館藏資源)及交易資料(如圖書借閱紀錄、館藏資源使用統計、資料庫檢索紀錄等)。

Guenther(2000)說明大多數圖書館之交易資料來自於下列服務項目：「館藏發展」、「流通

服務」、「參考服務」，後兩者是使用者與圖書館之間產生交易的服務項目。例如在進行「流通服務」時，圖書館需處理文獻傳遞或館際互借等任何屬於流通服務的交易需求；在進行「參考服務」時，圖書館須整合蒐集各類符合使用者需求之資源；「館藏發展」是圖書館爲維護及發展使用者所需資訊所進行之服務，故其資料可謂「流通服務」及「參考服務」交易資料的整合。(註43)

惟目前圖書館資料探勘之實證研究中，探勘範圍多侷限於圖書借閱資料。誠如Schulman所言，圖書館使用者行爲模式並不僅指借閱，若圖書館系統可識別遠端網路連線使用者身份，並進行資料庫使用統計，則圖書館方有可能勾勒出圖書館資源及服務如何被使用之完整藍圖。(註44)因此，欲進行圖書館服務交易資料探勘時，不可忽視圖書借閱服務之外可產生圖書館交易資料之服務項目。

(2)具探勘價值之圖書館交易資料類型

基本上，並非所有圖書館服務交易資料均需利用資料探勘技術處理，端視圖書館管理目標及使用者需求而定。惟目前依圖書館交易資料可得性、交易資料重要性、相關技術發展性等因素，可概分爲兩方向說明較具探勘價值之圖書館交易資料類型：

A.最常應用資料探勘技術之交易資料--圖書館自動化系統之「圖書借閱紀錄」

如前述，1996年，Atkins、Larsen、Mancini及Peters均撰文闡述圖書館應探勘自動化系統內的資料，以利於獲取管理資訊、提升管理績效、制訂管理決策。

而在已發展多年之圖書館自動化多種系統模組中，功能最齊全、完善者首推「圖書流通模組」，累積許多值得探勘之圖書借閱紀錄。圖書借閱紀錄向來是讀者實際使用圖書館資源之「證據」，也是讀者積極滿足個人資訊需求之行爲結果，其中潛藏大量圖書與讀者互動之歷史紀錄、有意義之關係或規則。因此，圖書借閱紀錄能反映使用者實際之資訊需求，對於掌握讀者興趣，進而作爲加強圖書資源利用之基礎，具有一定之參考價值。(註45)

就國內實證研究而言，均僅限於使用該館圖書館自動化系統之讀者圖書借閱紀錄進行資料探勘。以卜小蝶教授進行之研究爲例，以世新大學圖書館所提供1996年9月23日至1998年9月22日兩年之借閱歷史檔爲基礎，共計171,2841筆紀錄(交易資料)。由於一般圖書館自動化系統多未考量到資料探勘所需要之檔案內容，因此首先必需針對現有檔案進行篩選，主要檔案除圖書借閱紀錄檔外，尚包括讀者檔(11,616筆)、館藏檔(135,981筆)、書目檔(94,000筆)，及其他相關代碼檔，如系所代碼對照檔、讀者身份代碼對照檔等。均僅限於使用該館圖書館自動化系統之讀者圖書借閱紀錄資料進行資料探勘。(註46)

其餘研究所選取之資料檔案，與上述大致相同；至於資料類型係僅限圖書或包括期刊、論文，或是否扣除複本館藏檔紀錄，則依研究者或該校圖書館自動化系統而定。

B.不可忽視之圖書館交易資料類型--電子資源使用統計

在現今資訊網路發達、資訊量激增、資訊媒體多樣化等因素相互雜揉之影響下，圖書館必需體認電子圖書館、虛擬圖書館時代已經到來，絕對無法僅將服務焦點擺在以實體呈現之有形圖書。對此，Banerjee(1998)亦指出：即使紙本資料未來仍是圖書館重要的館藏型式，但未來只以電子型式出版之資訊資源將大量增

加。(註47)

同時，許多使用者習慣在自家電腦前完成知識之查找、擷取、彙整與創新，一氣呵成、迅速俐落！因此，透過網路使用圖書館資料庫檢索、電子期刊等電子資源之使用者，較之到館借閱圖書之使用者，在未來或甚至是現在，數量只會有增無減。Peters亦言，在許多圖書館中，大部分使用者已經成爲主要或僅使用遠端服務之使用者。(註48)這也表示此類使用者更需圖書館投入心力瞭解其資訊需求。就圖書館經營成本效益觀點而言，每年採購線上資料庫及電子期刊之費用，遠勝紙本圖書不止數倍，更需要圖書館審慎採購最能滿足使用者需求之電子資訊。

因此，圖書館經營者需正視電子資源使用紀錄統計之重要性，若應用資料探勘技術處理之，可更加瞭解圖書館使用者資訊尋求行爲。此外，Peters指出：電子資源使用紀錄統計可讓圖書館進行電子資源採購效益評估，作爲圖書館決策者管理資訊系統之基礎，而以往圖書館員憑藉直覺與經驗執行館藏發展工作，現可利用此種科學方法協助進行。(註49)

惟電子資源使用統計在近年突破技術限制後，在相關標準、解釋及實務應用上，雖較以往有良好的發展，但尚有一些瓶頸亟待克服，對圖書館來說，最主要的問題來自於資料可得性。電子資源使用紀錄不若圖書借閱紀錄儲存於圖書館本身自動化系統內，圖書館或多或少需仰賴不同資料庫出版者或代理商提供使用紀錄統計資料，但兩者立場可能產生衝突，一則來自於資料庫廠商不願讓圖書館掌握有可能爲低使用率之事實，二則不願免費提供這項可能成爲附加、可收費之新產品服務項目。(註50)

就實務來看，圖書借閱紀錄雖是圖書館最易取得、資料量最多、最利於應用資料探勘之自動化資料，惟它並未完整反映圖書館使用者之資訊需求及其使用狀況。雖然圖書館可依本身管理及服務需求，決定所欲探勘之資料範圍，但倘若探勘目標在於欲以結果來改善圖書館全面性服務效能，則不可忽略電子資源使用紀錄統計資料之重要性，以免失之偏頗。

(3)「整合」再「整合」之困難--交易內容及交易資料

目前資料探勘技術多成功應用於商業或科技業領域，分析一些銷售紀錄之類已結構化之交易資料；就圖書館而言，上述已說明「完整」分析圖書館服務交易資料之重要性，但實務上圖書館服務包含大量非結構化、資料來源多樣化的交易內容及交易資料，難題即來自於建構整合性平台，包括：

A.整合大量且不同類型服務交易內容的平台--如紙本圖書、電子期刊、線上資料庫文獻等，包含結構化或非結構化的資訊。

B.整合大量且不同型式服務交易資料的平台--如自動化系統之圖書借閱紀錄、文獻傳遞服務申請單、館際互借服務申請單等。

就交易內容而言，各種資料類型的館藏就如同圖書館銷售的產品，像大型企業一樣，圖書館也希望將館藏資料基於管理決策及使用者需求進行處理，最完美的方式即是圖書館及使用者均可利用整合性系統同時處理及使用不同資料類型之資源，但在現實世界裡，我們所面臨的卻是需要周旋在各不相容的系統中，方能達成管理及使用目的。即便解決了整合技術的問題，但線上資料庫、電子期刊及電子書銷售廠商亦未必願意將其提供的資訊資源，供圖書館進行探勘所需的處理。因此，整合大量且不同類型交易內容的平台，實有其困難。(註51)以交易資料而論，圖書

館面臨如何整合發生在不同環境(紙本或電子)、服務項目下的交易資料處理問題。簡言之，來自交易內容及交易資料整合之複雜度，導致資料處理困難度激增。

(4)小結

在此情況下，誠如Berry及Linoff(1997)所提醒：資料探勘所採用的資料常來自於不同的形式、格式以及系統，常成爲影響資料探勘結果的陷阱。圖書館需權衡輕重，在考量管理目標、整合技術限制之下，也許短期僅能選擇雖爲片面但較適用於資料探勘技術之資料(如經自動化系統處理之圖書借閱紀錄)，亟待未來技術突破後，再尋求問題解決之道。

2. 資料前置處理

此階段旨在確保選取資料的品質，爲未來分析作準備，並以決定探勘演算法。

以「資料淨化」及「加上衍生變數」之步驟來說，卜小蝶（2001）表示：每個資料檔案均需審慎考量其用途，除了需要作適當的清理及修改外（如錯誤資料清理、將各項代碼一致化等），對於資料層次的分析也很重要，如加入多層次(Multi-level)的欄位(如讀者系所可加上院別或年級等層次)。(註52)此外，爲提升資料有效性，尚須注意「欄位」選擇問題。

3. 資料轉換

意即將資料轉換成資料探勘演算法所適合的模式。以陳建銘（2002）研究爲例，該研究目的爲瞭解某一人同時借閱了哪些種類的書籍，其資料轉換步驟簡述如下：(註53)

(1)先將文字檔形式的資料庫轉換成EXECL檔案格式。

(2)以身份別將借閱資料庫拆解成數個較小的資料庫，包括：本校教職員工、圖書館員、畢業生、研究生與大學生等10個資料庫。

(3)剔除所有借閱之西文圖書的紀錄，因爲所借閱的讀者極少，且其分類法與中國圖書分類法不同。

(4)按照中國圖書分類法將借閱紀錄分成10個種類。

(5)由於找關聯法則需要兩種或兩種以上借閱不同書籍的資料，所以必須剔除僅借閱一個種類書籍的紀錄，留下借閱兩種或兩種以上的讀者借閱紀錄，供 Apriori 關聯規則的輸入資料(Input Data)使用。

(6)另依據中國圖書分類法書籍種類分成十大類，將資料分成10個欄位(column)，對應10大圖書種類，供霍普菲爾－坦克類神經網路使用的輸入資料。

(三)資料探勘

選擇適當的資料探勘演算法來探勘上一個步驟所轉換出來的資料，此步驟完成速度快且爲自動化。

至於後續之「根據資訊行動（Taking action）」、「測量、評估探勘結果（Measuring the outcome）」之步驟，在圖書館實際引進資料探勘技術後，可進行館藏資源使用統計或使用者服務滿意度調查，藉此評估資料探勘成效。

三、國內相關實證研究

國外研究探討「資料探勘應用於圖書館環境」議題時，焦點多集中在其適用性，國內研究則以實際進行資料探勘之實證研究爲主，尤以學位論文居多。謹將國內實際應用資料探勘技術於圖書館服務(交易)資料分析之實證研究，擇要列表整理如下：

表一：國內應用資料探勘技術於圖書館之實證研究一覽表

研究者	研究主題	研究目的	資料探勘技術及演算法
卜小蝶（2001）(註54)	以圖書借閱記錄探勘加強圖書資源利用之探討	1.改進現行圖書檢索系統 2.有助於推展圖書資源服務 3.有助於規劃圖書館館藏	利用關聯法則之 Apriori 演算法，以及 K-Means 法進行資料分群
張菀菁（2000）(註 55)	以模糊理論建構之圖書推薦系統	1.分析讀者屬性及其借書習性 2.針對讀者屬性，找出與他相似的讀者集群及其借閱書籍，以提供推薦書單供讀者參考	結合模糊理論，利用 Semi-Supervised c-Means 分群法，以及關聯規則技術
陳建銘（2000）(註 56)	類神經網路於 Web Mining 之應用（以圖書館學生圖書借閱紀錄爲研究對象）	1.供圖書館員掌握該校學生借閱書籍偏好與借閱情況 2.提高圖書館借閱率 3.結果提供圖書館作爲其他決策之參考。	以霍普菲爾類神經網路(Hopfield Neural Network)解決關聯規則挖掘找出高頻項目演算法
吳安琪（2000）(註 57)	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	1.探索讀者社群特性 2.運用社群關係達到下列目的：吸引讀者到館借閱、提昇館藏借閱率、提昇讀者忠誠度、協助館藏複本採訪政策及促進館藏流通率	採用Apriori演算法找出讀者借閱館藏的共同性，並改進部份GSP的方法，找出讀者借閱館藏的順序
王毓菁（2002）(註 58)	圖書館閱覽者群組潛在特徵探勘資訊系統	根據不同館藏資料的分類，閱覽者的基本資料、及閱覽者借閱館藏資源的借閱記錄，探勘圖書館閱覽者群組潛在特徵	使用天真貝氏分類法(Naïve Bayes Classification)將未經分類的資料加以自動分類。再使用關聯規則探勘圖書館閱覽者群組的潛在特徵

資料來源：本文作者整理

由上表分析，資料探勘協助圖書館決策之預期目標，包括：1.探討使用者個別資訊需求與群體特性，找出使用者借閱行爲傾向；2.提昇圖書資源之使用率，加強圖書資源行銷、推廣服務；3.針對圖書館個別使用者，發展進行顧客化書目推薦服務。惟上述研究侷限於分析圖書借閱使用者之資料，而未包括線上資料庫、電子期刊或其他服務項目之使用者資料。另外，由於國內實證研究多爲學位論文性質，並非由圖書館主導之專案計畫，因此未能得知圖書館服務交易資料經資料探勘後所得之資訊，是否確能達成上述協助圖書館決策之目的。

陸、結論

爲達成滿足使用者資訊需求之目標，圖書館實

有切實掌握使用者需求，並主動行銷及推廣服務之必要性。而近年極熱門應用於企業行銷及顧客關係管理決策之資料探勘技術，是否也能應用於圖書館環境？國內外相關研究文獻爲數不多，適用與否之正反意見均存在，故此議題實有深入釐清、剖析之必要。

「資料探勘（Data Mining）」係指由資料庫已存在之「資料」中，探勘出「前所未知」、「隱而未覺」、「不明顯」但卻「有意義」且「可付諸行動」之「新關係或事實（有用資訊）」，以利於決策之過程。

資料探勘過程之良窳實爲它能否成功達成目標之關鍵。就圖書館之資料探勘過程而言，亦如一般性程序，需先瞭解組織問題，進行事前規劃與資料準備，方運用資料探勘技術將資料轉換爲有用資訊，最後應用所獲得之知識以解決問題、達成目標。理論上，圖書館應可運用資料探勘提升管理績效、支援管理決策，亦能更加瞭解圖書館使用者行爲，以重新規劃館藏發展方向及資訊服務項目。就探勘技術類型來說，以應用關聯法則分析及集群分析技術爲主。

但實務上，圖書館若欲應用資料探勘技術以達成目標，尚有些許困難與限制亟待突破。目前探勘資料範圍多限於使用圖書館自動化系統之讀者圖書借閱紀錄，但可提供探勘資料之服務項目並不止於此，故其探勘結果並未「完整」反映圖書館使用者之資訊需求及其使用狀況，「遺漏資料」包括資料庫及電子期刊等文獻檢索使用紀錄，以及其它亦能產生圖書館交易資料之服務項目(例如館藏發展、參考服務)之紀錄。此外，由於圖書館具有不同資料類型之交易內容，及不同服務項目之交易資料，導致資料處理之複雜度激增，而需突破各自建構整合不同交易內容、交易資料之平台，再予以整合所有資料之平台的技術，如想提升圖書館資料探勘分析結果的品質，此亦是一項挑戰。

因此，就現階段資料探勘技術發展狀況及圖書館本身環境條件來說，尚無法全面性達成上述協助圖書館決策之目的，未來研究可朝向探勘電子文獻檢索紀錄資料之技術發展，以瞭解其使用者需求進而改善服務。更重要的是，發展各自建構整合不同交易內容、交易資料之平台技術，以及整合所有資料之平台技術，再應用資料探勘技術分析資料，以獲取能完整反映圖書館使用者之資訊需求及其使用狀況之資訊，協助圖書館決策之用。

(收稿日期：2003 年 9 月 17 日)

註　釋：

註 1： Cabena, P., et al., Discovering Data Mining: From Concept to Implementation (N. J.: Prentice-Hall, 1998).

註 2： 張真誠、蔡文輝、林敏惠著。挑戰資料庫管理系統（台北市：旗標，民國 92 年）。

註 3： Grupe, F. H. and Owrang, M.M., "Database Mining Discovering New Knowledge and Cooperative Advantage.," Information Systems Management, 12:4 (1995), pp.26-30.

註 4： 同註 1。

註 5： Berry, M. and Linoff, G., Data Mining Techniques for Marketing, Sales, and Customer Support (John Wiley & Sons, Inc. 1997).

註 6： Peacock P. R., "Data Mining in Marketing.: Part 1". Marketing Management, 6:4(1998), pp.9-18.

註 7： Piatetsky-Shapiro, G. and Frawley, W. J., Knowledge Discovery in Databases (California: AAAI/MIT Press.1991).

註 8： Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P., "From Data Mining to Knowledge Discovery in Databases," AI Magazine 17 (1996), p.37.

註 9： 同註 5。

註 10： 同註 6。

註 11： Curt H., "The Devile's in The Detail: Techniques, Tool, and Applications for Data mining and Knowledge Discovery-Part 1," Intelligent Software Strategies 6:9 (1995), p.3.

註 12： 同註 5。

註 13： 同註 5。

註 14： 裴瑞(Berry, M.)、林諾夫(Linoff, G.)合著，資料採礦理論與實務：顧客關係管理之技巧與科學（Mastering Data Mining, The Art & Science of Customer Relationship Management），吳旭智、賴淑貞譯（台北市：維科，民國 90 年）

註 15： 翁頌舜、梁德馨，「資料採礦資料缺值插補之變異數分析」，輔仁管理評論 9 卷 3 期（民國 91 年），頁 163-180。

註 16： 同註 1。

註 17： 鄭宇庭、蘇志雄，「商業智慧的工具－資料採礦」，輔仁管理評論 9 卷 3 期（民國 91 年），頁 11-34。

註 18： 邱義堂，「通信資料庫之資料探勘：客戶流失預測之研究」（碩士論文，國立中山大學資訊管理學系研究所，民國 90 年）。

註 19： 同前註。

註 20： 許哲瑋，「資料挖掘與統計方法應用於資料庫行銷之實證研究－美妝保養品業爲例」（碩士論文，國立臺北大學企業管理研究所，民國 91 年）。

註 21： 陳建銘，「類神經網路於 Web Mining 之應用」（碩士論文，國立臺北科技大學商業自動化與管理研究所，民國 89 年）。

註 22： 同註 18。

註 23： 同註 18。

註 24： 同註 17。

註 25： 卜小蝶，「以圖書借閱記錄探勘加強圖書資源利用之探討」，中國圖書館學會會報 66 期（民國 90 年），頁 59-72。

註 26： 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」（碩士論文，國立交通大學資訊科學研究所，民國 91 年）。

註 27： 同註 25。

註 28： 同註 18。

註 29： 顏嘉惠，「資料探勘於圖書館行銷及顧客關係管理之應用」，圖書與資訊學刊 42 期（民國 91 年），頁 58-68。

註 30： 同註 26。

註 31： Schulman, S., "Data Mining：Life after report generators," Information Today 15:3 (1998), p52.

註 32： Gleeson, M. E. and Ottensmann J. R., "Using Data from Computerized Circulation and Cataloging Systems for

Management Decision Making in Public Libraries," Journal of The American Society For Information Science 44:2 (1993), pp.94-100..

註 33： Atkins, S., "Mining Automated Systems for Collection Management," Library Administration & Management, 10:1 (1996), pp.16-17.

註 34： Larsen, P., "Mining Your Automated System for Better Management." Library Administration & Management, 10:1 (1996), p.10.

註 35： Mancini, D. D., "Mining Your Automated System for Systemwide Decision Making," Library Administration & Management 10:1 (1996), pp.11-15.

註 36： Peters, T., "Using Transaction Log Analysis for Library Management." Library Administration & Management, 10:1 (1996), pp.20-25.

註 37： 同註 31。

註 38： 同註 14。

註 39： 同註 1。

註 40： 同註 14。

註 41： Banerjee, K., "Is Data Mining Right for Your Library?," Computers In Libraries, 18:12 (1998), pp.28-31.

註 42： 同註 31。

註 43： Guenther, K., "Applying Data Mining Principles to Library Data Collection," Computers In Libraries, 20:4 (2000), pp.60-63.

註 44： 同註 31。

註 45： 同註 25。

註 46： 同註 25。

註 47： 同註 41。

註 48： Peters, T., "What's the Use? The Value of E-resource Usage Statistics. New Library World, 103:1/2 (2002), pp.39-47.

註 49： 同註 48。

註 50： 同註 48。

註 51： 同註 43。

註 52： 同註 25。

註 53： 同註 21。

註 54： 同註 25。

註 55： 張菀菁，「以模糊理論建構之圖書推薦系統」（碩士論文，淡江大學資訊工程學系，民國 89 年）。

註 56： 同註 21。

註 57： 同註 26。

註 58： 王毓菁，「圖書館閱覽者群組潛在特徵探勘資訊系統」（碩士論文，華梵大學工業管理學系，民國 91 年）。