利用資料探勘技術發掘圖書館個人化之書籍推薦

陳垂呈

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

摘要

本論文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讀者曾借閱過的書籍與其興趣度，並以某一讀者之借閱資料為分析的目標，利用資料探勘(data mining)技術中的分類分析(classification)，探討如何發掘此一讀者個人化的書籍推薦。在探勘過程中，比對此一讀者借閱資料與其他借閱資料的相似度，依據其是否符合所設定的條件，來分別設定其與此一讀者借閱資料的關聯性為「高」或「低」，並視其他非此一讀者曾借閱過的書籍項目為影響屬性，然後對讀者的借閱資料進行分類分析。首先，只考量讀者曾借閱過的書籍項目，然後針對借閱資料進行分類分析，藉由所建立的決策樹(decision trees)，可得知那些屬性與此一讀者之關聯性為高，藉以發掘出此一讀者個人化最適性的書籍推薦。再者，考量讀者對曾借閱過之書籍的興趣度，分別將每一書籍分解成其興趣度之數量的項目屬性，然後視分解後非此一讀者曾借閱過之項目屬性為欲分類的屬性，並進行分類分析，藉由所建立的決策樹，可得知那些分解後之項目屬性與此一讀者的關聯性為高，藉以發掘出包含有興趣度之此一讀者最適性的書籍推薦。此探勘結果，對圖書館在擬訂最適性之讀者個人化書籍推薦時，可以提供非常有用的參考資訊。

關鍵詞：資料探勘，分類分析，借閱資料，書籍推薦

一、簡　介

隨著資訊技術的蓬勃發展，促進了電子化圖書館的日漸成熟，圖書館所提供的服務也愈來愈多樣化，電子圖書、網際網路、儲存光碟、多媒體等科技媒體的出現，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，

利用國內外網路上的豐富資源，配合各類電子媒介的輔助，使讀者能以最少的時間，即可享受到最大的服務效益。但如何將傳統圖書館被動式的服務方式，轉變成以主動積極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利用率，是圖書館管理者必須探討的課題之一。

在網際網路服務應用的發展中，其服務形態由過去強調「入門網站」「垂直網站」到目前的「個人化服務網站」，其強調資訊服務必須因人而異，以滿足使用者個人的資訊需求（註1）。在圖書館服務行銷的理念中，辜曼蓉（1999）曾指出圖書館的核心價值是顧客（讀者）和服務人員之間的互動，而館藏及資訊的傳播則扮演提昇圖書館服務品質的輔助角色（註2）。因此，以個人化資訊為導向的服務形態，不僅是網際網路服務應用的潮流，也反映出圖書館服務的本質。針對每一讀者個人化的資訊需求，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與圖書館的經營績效（註3）。

在國內各大專院校之圖書館個人化資訊服務中，交通大學個人化數位圖書館資訊服務是最具有代表性的系統之一（註4），其提供的服務有「智慧型圖書館查詢」、「個人化檢索」、「新書查詢」、「新書列表」、「個人化推薦」、「個人借閱狀況」、「個人書籤」、「個人化桌面」及「使用者設定」等功能，在「個人化推薦」的功能中，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）做為個人化書籍推薦的方法依據。因此，如何將最貼切的書籍推薦給讀者個人，是圖書館個人化資訊服務重要的功能之一。

圖書館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，例如讀者借閱了一本「C語言程式設計」的書籍，我們會發現其中往往也會借閱「資料結構」的書籍，或一些與「C語言程式設計」相關的書籍。因此，如何從累積數量龐大的借閱資料中，找出對讀者有用的資訊或其他知識，即成為圖書館管理者必須思考的問題之一。在本論文中，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經借閱過的書籍項目與其興趣度，並以某一讀者之借閱資料為探勘的目標，假設此一讀者的借閱資料為X，X為包含有興趣度之一個或以上書籍項目所形成的項目組，利用資料探勘（data mining）技術中的分類分析（classification），分別從以下兩方面來發掘讀者個人化之書籍推薦：

（一）只考量書籍項目是否出現在借閱資料中。在探勘過程中，我們將曾經借閱過X之書籍項目達到某一比例值的其他借閱資料，設定其與X的關聯性為「高」，否則設定其關聯性為「低」，並視其他非X之書籍項目為影響屬性，然後對讀者的借閱資料進行分類分析。我們依據ID3演算法所建立的決策樹，可找出那些的影響屬性與X的關聯性為高，藉此做為發掘此一讀者個人化最適性之書籍推薦的依據。

（二）考量出現在借閱資料中之書籍項目包含有興趣度。在探勘過程中，若一借閱資料包含有X之書籍項目達到某一比例值，且這些書籍項目的滿意度都大於或等

於包含於X的書籍項目，則設定其與X的關聯性為「高」，否則設定其關聯性為「低」，並視其他非X之書籍項目為影響屬性。我們分別將借閱資料中的其他非X的各書籍項目，分解成其興趣度單位之個數的項目屬性，然後視各分解後的項目屬性為欲分類的屬性，並以ID3演算法進行分類分析，藉由所建立的決策樹，可得知那些分解後的項目屬性與X的關聯性為高，藉此做為發掘包含有興趣度之讀者個人化最適性的書籍推薦。

我們根據所提出的方法，設計與建置一個讀者個人化最適性之書籍推薦系統。此探勘結果，對圖書館在擬訂最適性之讀者個人化書籍推薦，進而以主動積極的方式來吸引讀者到館借閱，可以提供非常有用的參考資訊。

本論文的架構如下：下一節介紹資料探勘技術，及其在圖書館服務之應用的相關研究；第三節利用分類技術來分析與某一讀者之關聯性為高的書籍屬性，以發掘此一讀者個人化的書籍推薦；第四節考量借閱資料中之書籍項目包含有興趣度，利用分類技術來發掘包含有興趣度之讀者個人化的書籍推薦；第五節根據所提出的方法，設計與建置一個讀者個人化之書籍推薦系統；最後，第六節做一結論。

二、相關研究

資料探勘是從大量資料中找出潛在有用的資訊與知識，其可完成以下工作或更多：分類、關聯規則、分群（clustering）、次序相關分析（sequential pattern analysis），及預測（forecasting）等（註5）。其探勘結果對企業在從事行銷決策及市場預測等活動時，可以提供非常有價值的參考資訊（註6）。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，找尋有興趣或想要借閱的書籍資料，而圖書館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，也造成書籍的借閱率不佳。

利用資料探勘技術於圖書館經營服務之應用的相關研究有：陳慶瑄（2000）曾利用*k*-means的方法來形成學習社群，以支援電子圖書館之個人化服務（註7）。孫冠華（2003）曾利用關聯規則於數位圖書館的個人化服務及管理（註8）。吳安琪（2001）曾提出利用資料探勘技術來發掘讀者的社群關係，進而達到吸引讀者借閱書籍，以提昇圖書館之借閱率與讀者忠誠度等目的（註9）。洪志淵（2001）曾利用資料探勘技術來找出讀者與圖書之間的一般化關聯規則，做為讀者新書推薦的依據（註10）。張苑菁（2001）曾以模糊理論（fuzzy）與資料探勘技術來分析讀者的借閱資料，進而提供相關的書籍推薦給讀者參考（註11）。余明哲（2003）曾利用關聯規則來找出讀者之間的關係，並考量讀者的興趣，以找出合適的服務推薦給讀者（註12）。因此，如何貼切地提供讀書個人化的書籍推薦，已成為提昇圖書館之經營與服務最重要研究的課題之一，也是資料探勘技術重要的應用主題之一。

分類分析是從已知的物件群中，根據所訂立的屬性條件來進行分類，決策樹（decision trees）與決策法則（decision rules）是分類分析最常用的兩種表示法。例如，

我們可對讀者的借閱資料來進行分類，把借閱資料與某一讀者的關聯性分為「高」與「低」兩種類別型態，再以是否借閱書籍項目來做為影響屬性的分類計算，便可得知影響關聯性之高低的關鍵屬性。

在資料進行分類分析時，一般可以產生出許多分類模式，但其期望得到的分類模式是越精簡越好。以決策樹為例，若決策樹的高度愈小，則表示可用愈少的屬性便能分類出所有物件。因此，一個好的分類技術，應該具有精簡與預測能力佳的特性，目前常被利用的分類技術有ID3（註13）、CN2（註14）、倒傳遞類神經網路（back-propagation）（註15）等。在本論文中，我們以讀者之借閱資料為探勘的資料來源，利用分類分析做為發掘讀者個人化之書籍推薦的方法依據。

三、發掘讀者個人化之書籍推薦

在讀者的借閱資料中，我們往往會發覺在借閱資料中的各書籍項目有其關聯性，因此，如何在大量借閱資料中找出書籍項目之間的關聯性，藉此做為推薦讀者書籍的依據，不只可減少讀者找尋其合適書籍的時間，也可增加圖書館書籍借閱的頻率。在此一節中，我們以讀者之借閱資料為探勘的資料來源，在不考量讀者對書籍之興趣度的情況下，利用分類分析中的ID3演算法，來發掘讀者個人化的書籍推薦。ID3演算法是分類技術中最具代表性的方法之一，在分類的過程中是根據熵值（entropy）和資訊收益（information gain）的大小，來將資料依據屬性加以分類，以找尋最有效益的分類屬性，建構出簡單卻能正確描述資料之間關係的決策樹。此節共分為兩小節：第一小節介紹ID3演算法，並說明如何利用此方法來發掘讀者個人化的書籍推薦；第二小節以一實例做說明。

（一）ID3演算法

ID3演算法是一種決策樹的分類技術，其目的是選擇最佳的屬性來當作節點，以建構出的決策樹為一最簡單狀態，或接近最簡單狀態。最佳節點是依據其節點所產生的熵值所決定，其計算方式如下。

若某一物件集合C，其物件分屬於j個不同類別，則此物件集合之熵值$E(C)$為：

$$E(C)=\sum_{i} p_i \log_2 P_i \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (1)$$

C：物件集合。

i：類別數。

P_i＝（屬於類別i的物件總數）／（C的物件總數）。

設定log20＝0（註16）。

接下來選擇某一屬性X_j為決策樹節點，在此節點下建立m個子節點，並將原本屬於節點的所有物件，分配至具有適當的子節點下。而分配至相同子節點的物件，其屬性X_j值必為相等。故以X_j為節點所產生的子決策樹熵值$E(X_j)$為：

$$E(X_j)=\sum_k (n_k/n)\times E(C_k) \cdots\cdots\cdots\cdots (2)$$

C_k ：物件集合C中其X_j屬性相同的物件子集合k。

$E(C_k)$ ：為物件C_k的熵值。

n ：物件集合C的總物件數。

n_k ：物件子集合C_k的物件數。

資訊收益是原來物件集合的熵值與X_j為決策樹子節點的熵值間得差距，其公式如下：

$$G(X_j)=E(C)-E(X_j) \cdots\cdots\cdots\cdots (3)$$

根據熵值和資訊收益，我們將ID3演算法的執行步驟說明如下（註17）：

1. 首先，設立決策樹的根節點為C，此時所有物件都屬於C的物件集合。

2. 若C中所有的物件都屬於同一類別，則定義C節點為此類別並停止，否則繼續執行步驟3.。

3. 對屬於C的所有物件，分別計算其熵值$E(C)$。

4. 從根節點至目前節點中，若有尚未當過節點的屬性X_j，則以X_j對C物件集合進行分割，並分別計算部分決策樹的熵值$E(X_j)$及資訊收益$G(X_j)$。

5. 選擇具有最大資訊收益的候選屬性，並當做C節點的分類屬性。

6. 在C節點下建立子節點分別為C_1、C_2、…、C_m（假設選擇了m個屬性值做為分類屬性），並將C中的所有物件集合，分配至適合的子節點中。

7. 對每個子節點C_i當做節點C，$1 \le i \le m$，並由2.重覆執行。

我們以讀者之借閱資料為探勘的資料來源，並以某一讀者為探勘的目標，假設此一讀者的借閱資料為X，X、R為一個或以上書籍項目所形成的項目組。在探勘過程中，我們先計算一借閱資料R與此一讀者之借閱資料X間的相似度。其定義如下：

借閱資料R與X之相似度＝（$X \cap R$）的項目個數／X的項目個數。

其表示若相似度愈大，則借閱資料R中愈包含有X中的項目，也就是說，在借閱資料R中包含有非X的書籍項目會與此一讀者具有較高的關聯性。我們探勘的目的就是希望找出那些非X的書籍項目會與此一讀者具有關聯性高的特徵，藉此達到書籍推薦的目的。我們設定一個相似度p，若一借閱資料與X之間的相似度達到大於或等於p，則設定此一借閱資料與X的關聯性為「高」，否則設定關聯性為「低」，並視其他非X之書籍項目（即未曾借閱過的書籍項目）為影響屬性，然後對讀者的借閱資料進行分類分析。其目的就是從借閱資料中，找出與此一讀者具有關聯性高之未曾借閱過的書籍項目。經由上述ID3演算法的計算，我們可以從所建立之決策樹的根節點（root）到類別為關聯性高的路徑（paths）中，找出那些影響屬性會與此一讀者的關聯性高，我們即定義這些影響屬性之項目為此一讀者個人化最適性的書籍推薦。

在探勘過程中，假設借閱資料共有n筆，類別共分為2類（即關聯性高與關聯性低），非X之書籍項目的屬性共有d個，每一屬性之屬性值共有2個（即曾經借閱與未曾借閱），我們考量在最壞的情況下所建立的決策樹其高度為d，其高度從level 1到level d，其所需要的計算時間為：

level 1：計算公式（1）+公式（3）+公式（3）的時間=$d\times(2\times n+2\times 2\times n)+1=6dn+1$
level 2：計算公式（1）+公式（3）+公式（3）的時間=$(d-1)\times(2\times n+2\times 2\times n)+1=6(d-1)n+1$
level 3：計算公式（1）+公式（3）+公式（3）的時間=$(d-2)\times(2\times n+2\times 2\times n))+1=6(d-2)n+1$
⋮
level d：計算公式（1）+公式（3）+公式（3）的時間=$1\times(2\times n+2\times 2\times n)+1=6n+1$

計算其總和為$6nd(d+1)/2+d$，即其時間複雜度（time complexity）為$O(nd2)$（註18）。同樣地，我們考量在最壞情況下所建立的決策樹所需要的記憶體空間為：假設記錄一個類別狀態需要c bytes，記錄一個屬性值需要s bytes，在計算過程中，我們須將借閱資料載入記憶體中，因此所需記憶體為$n(sd+c)$ bytes。

（二）實例說明

我們以一實例來說明發掘某一讀者個人化最適性之書籍推薦的探勘過程。假設{A, B, C, D, E, F, G, X, Y, Z}為書籍項目的集合，共有10個書籍項目，某一讀者的借閱資料為「XYZ」，在不失一般性的條件下，表1為借閱資料庫中部分讀者的借閱資料，共有16筆，相似度p為60%，以下我們說明發掘此一讀者個人化最適性之書籍推薦的探勘過程。

表1 讀者之借閱資料

讀者編號	借閱資料
1	ABDEGXY
2	FX
3	BDXZ
4	ABCDEFXYZ
5	DFG
6	ABCDEYZ
7	ABCDFGYZ
8	AFGZ
9	BX
10	AC
11	ADFG
12	BZ
13	AFY
14	ACE
15	BDFGXY
16	CZ

首先，我們計算各借閱資料與此一讀者之借閱資料「XYZ」間的相似度，若相似度大於或等於60%，則設定關聯性為「高」，否則設定關聯性為「低」，其結果如表2。

表2 借閱資料之關聯性

讀者編號	借閱資料	相似度%	關聯性
1	ABDEGXY	2/3= 67	高
2	FX	1/3= 33	低
3	BDXZ	2/3= 67	高
4	ABCDEFXYZ	3/3= 100	高
5	DFG	0/3= 0	低
6	ABCDEYZ	2/3= 67	高
7	ABCDFGXYZ	3/3= 100	高
8	AFGZ	1/3= 33	低
9	BX	1/3= 33	低
10	AC	0/3= 0	低
11	ADFG	0/3= 0	低
12	BZ	1/3= 33	低
13	AFY	1/3= 33	低
14	ACE	0/3= 0	低
15	BDFGXY	2/3= 67	高
16	CZ	1/3= 33	低

其中A、B、C、D、E、F與G等7個書籍項目，為此一讀者未曾借閱過的書籍項目，因此，我們將表2的資料轉換成表3，其中標示為「1」者表示「曾借閱過」該欄的書籍項目，標示為「0」者表示「未曾借閱過」該欄的書籍項目。

表3 包含未曾借閱過之書籍項目的借閱資料

讀者編號	書籍項目							關聯性
	A	B	C	D	E	F	G	
1	1	1	0	1	1	0	1	高
2	0	0	0	0	0	1	0	低
3	0	1	0	1	0	0	0	高
4	1	1	1	1	1	1	0	高
5	0	0	0	1	0	1	1	低
6	1	1	1	1	1	0	0	高
7	1	1	1	1	0	1	1	高
8	1	0	0	0	0	1	1	低
9	0	1	0	0	0	0	0	低
10	1	0	1	0	0	0	0	低
11	1	0	0	1	0	1	1	低
12	0	1	0	0	0	0	0	低
13	1	0	0	0	0	1	0	低
14	1	0	1	0	1	0	0	低
15	0	1	0	1	0	1	1	高
16	0	0	1	0	0	0	0	低

1：曾借閱過，0：未曾借閱過

首先，我們只考量書籍項目是否出現在借閱資料中：利用公式（1），計算物件集合的熵值，在表3，共有6位與此一讀者借閱資料的關聯性為高，及10位與此一讀者借閱資料的關聯性為低，計算所有讀者的熵值，其計算如下：

E（所有讀者）= –(6/16)$\log_2$(6/16) – (10/16)$\log_2$(10/16)=0.954

接著利用公式（2），計算所有書籍屬性的熵值，我們以A-書籍屬性為例，其熵值計算如下：

E（曾經借閱過A-書籍）= –(4/9)$\log_2$(4/9) – (5/9)$\log_2$(5/9)=0.991
E（未曾借閱過A-書籍）= –(2/7)$\log_2$(2/7) – (5/7)$\log_2$(5/7)=0.863
E（A-書籍）=(9/16)×0.991+(7/16)×0.863=0.577+0.378=0.935

再利用公式（3）來計算A-書籍的資訊收益：

G（A-書籍）=0.954 – 0.935=0.019

依此類推，可計算出其餘各書籍項目屬性的資訊收益如下：

G（B-書籍）=0.548
G（C-書籍）=0.029
G（D-書籍）=0.548
G（E-書籍）=0.143
G（F-書籍）=0.001
G（G-書籍）=0.028

我們選定具有資訊收益最大者B-書籍做為該決策樹的根節點，再把其餘的6個書籍項目加入B-書籍之後，再次計算其熵值。借閱資料經由選定B-書籍之後的分類情況如表4。

由於在B-書籍之屬性值為「0」的情況下，可得到關聯性低的同一類別，因此關於B-書籍之屬性值為「0」的計算就結束。而在B-書籍之屬性值為「1」的情況下，仍可得到不同的類別值，因此必須繼續進行分類的計算，如表5。

我們以表5為分類分析的資料來源，首先再利用公式（1），計算物件集合的熵值，其計算如下：

E（所有讀者）= –(6/8)$\log_2$(6/8) – (2/8)$\log_2$(2/8)=0.811278

接著利用公式（2），計算所有書籍屬性的熵值，我們以A-書籍屬性為例，其熵值計算如下：

E（曾經借閱過A-書籍）= –(4/4)$\log_2$(4/4) – (0/4)$\log_2$(0/4)= 0
E（未曾借閱過A-書籍）= –(2/4)$\log_2$(2/4) – (2/4)$\log_2$(2/4)= 1
E（A-書籍）=(4/8)×0+(4/8)×1 = 0.5

表4　包含未曾借閱過之書籍項目的借閱資料

讀者編號	書籍項目							關聯性
	A	B	C	D	E	F	G	
1	1	1	0	1	1	0	1	高
3	0	1	0	1	0	0	0	高
4	1	1	1	1	1	1	0	高
6	1	1	1	1	1	0	0	高
7	1	1	1	1	0	1	1	高
9	0	1	0	0	0	0	0	低
12	0	1	0	0	0	0	0	低
15	0	1	0	1	0	1	1	高
2	0	0	0	0	0	1	0	低
5	0	0	0	1	0	1	1	低
8	1	0	0	0	0	1	1	低
10	1	0	1	0	0	0	0	低
11	1	0	0	1	0	1	1	低
13	1	0	0	0	0	1	0	低
14	1	0	1	0	1	0	0	低
16	0	0	1	0	0	0	0	低

1：曾借閱過，0：未曾借閱過

表5　包含未曾借閱過之書籍項目的借閱資料

讀者編號	書籍項目						關聯性
	A	C	D	E	F	G	
1	1	0	1	1	0	1	高
3	0	0	1	0	0	0	高
4	1	1	1	1	1	0	高
6	1	1	1	1	0	0	高
7	1	1	1	0	1	1	高
9	0	0	0	0	0	0	低
12	0	0	0	0	0	0	低
15	0	0	1	0	1	1	高

1：曾借閱過，0：未曾借閱過

再利用公式（3）來計算A-書籍的資訊收益：

G（A-書籍）=0.811 – 0.5=0.311

依此類推，可計算出其餘各書籍項目屬性的資訊收益如下：

G（C-書籍）=0.204
G（D-書籍）=0.811
G（E-書籍）=0.204
G（F-書籍）=0.204
G（G-書籍）=0.204

由於D-書籍的資訊收益最大，因此將D-書籍做為連接於B-書籍之屬性值為「1」之路徑的節點，再把其餘5個書籍項目加入D-書籍之後，再次計算其熵值。借閱資料經由選定D-書籍之後的分類情況如表6。

表6 包含未曾借閱過之書籍項目的借閱資料

讀者編號	書籍項目						關聯性
	A	C	D	E	F	G	
1	1	0	1	1	0	1	高
3	0	0	1	0	0	0	高
4	1	1	1	1	1	0	高
6	1	1	1	1	0	0	高
7	1	1	1	0	1	1	高
15	0	0	1	0	1	1	高
9	0	0	0	0	0	0	低
12	0	0	0	0	0	0	低

1：曾借閱過，0：未曾借閱過

由於在D-書籍之屬性值為「0」的情況下，可得到關聯性低的同一類別，因此關於D-書籍之屬性值為「0」的計算就結束。而在D-書籍之屬性值為「1」的情況下，也可得到關聯性高的同一類別，因此結束關於決策樹的分類計算。然後，我們將所有分類屬性依照先後順序連接起來，就形成如圖1的決策樹。

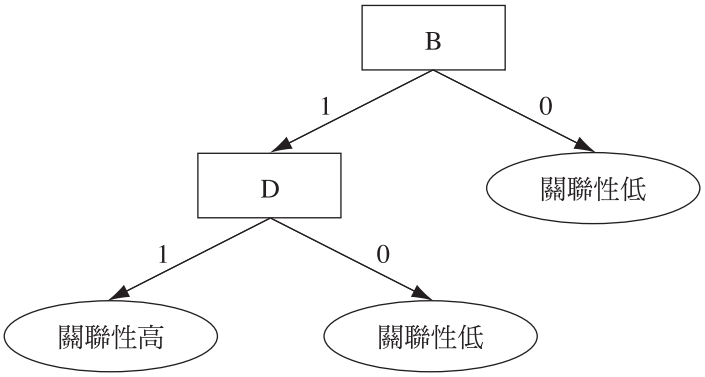

圖1 分類決策樹

從圖1之分類決策樹，我們可從根節點到類別為關聯性高的路徑中，發掘出此一讀者個人化最適性之書籍推薦為：B-書籍或D-書籍。

四、發掘含有興趣度之讀者個人化書籍推薦

在此節中，我們考量讀者對於曾借閱過之書籍項目的興趣度，往往也是會影響其曾借閱過之書籍項目間關聯強度的因素之一。因此，若能在讀者歸還借閱書籍時

或在具有個人化圖書館的功能中，藉由以詢問、填寫或問卷等方式來記錄其對借閱書籍的興趣度值，將曾借閱過之書籍的興趣度也列入考量，則可探勘出更精確之書籍項目彼此的關聯性。在目前提供有個人化資訊服務的圖書館系統中，例如交通大學個人化數位圖書資訊服務（PIE@NCTU），都將讀者個人借閱狀況列為基本的服務功能之一。此功能可列出讀者曾借閱過的書籍項目，因此，只要對此服務功能稍做修改，讓讀者可以填寫其已借閱且歸還之書籍的興趣度值，即可蒐集、記錄到讀者對其曾借閱過之書籍的興趣度值。例如，我們將興趣度值共分為1到5等五個等級，若一讀者R1之借閱資料為A-書籍且興趣度值為4，而有另兩位讀者R2及R3，其借閱資料也都為A-書籍且興趣度值分別為1及3，因此，雖然讀者R1、R2及R3都曾經借閱過A-書籍，但其對書籍的興趣度值不同，相較計算結果，顯示讀者R1與R3之間的興趣度是較R1與R2之間來得相似。

我們仍以讀者之借閱資料為探勘的資料來源，且考量書籍項目包含有興趣的程度，利用ID3演算法來進行分類分析，藉由所建立的決策樹來發掘具有興趣度之讀者個人化的書籍推薦。此節分為兩小節：第一小節說明如何利用ID3演算法來發掘包含有興趣度之讀者個人化的書籍推薦；第二小節以一實例做說明。

（一）分解借閱資料之書籍項目

我們仍以某一讀者為探勘的目標，假設此一讀者的借閱資料為$X=\{u_1x_1, u_2x_2, \cdots, u_jx_j, \cdots, u_kx_k\}$，$k\geq 1$，$k\geq j\geq 1$表示此一讀者曾經借閱過$x_1, x_2, \cdots, x_k$等書籍，且其興趣度分別為$u_1, u_2, \cdots, u_k$，例如一借閱資料為「1A2B3C」，其表示讀者曾經借閱過A-書籍、B-書籍及C-書籍，且興趣度分別為1、2及3，在此設定數字愈大表示興趣度也愈強。在探勘過程中，我們先計算一借閱資料R與此一讀者之借閱資料X間的相似度，其定義如下：

借閱資料R與X之相似度=（R包含有x_j且其興趣度也大於或等於u_i的項目個數）／X之項目個數

其表示若相似度愈大，則借閱資料R愈包含有X的項目，如同前一節所描述，即借閱資料R會與此一讀者具有較高的關聯性。我們設定一個相似度p，若一借閱資料與X之間的相似度達到大於或等於p，則設定此一借閱資料與X的關聯性為「高」，否則設定關聯性為「低」。例如，假設X為「1A2B3C」，另一借閱資料Y為「2A1B3C2D」，則其相似度=2/3。

在進行分類之前，我們必須分別將借閱資料中非X的各書籍項目，分解成其興趣度單位之個數的項目屬性，例如u_A，表示曾借閱過A-書籍並興趣度為u個單位，我們將其分解成「A」,「2A」,「3A」, …,「uA」的u個項目屬性，然後視各分解後的項目屬性為欲分類的影響屬性。我們說明分解各借閱資料如下：

假設$I=\{i_1, i_2, i_3, \cdots, i_b\}$表示共有$b$個書籍項目，對借閱資料中某一書籍項目$i_a$且

興趣度為u個單位而言，$1 \le a \le b$，必須分解成$1i_a, 2i_a, \cdots, ui_a$，共u個單一的書籍項目，其分別表示興趣度為1個單位的i_a-書籍、興趣度為2個單位的i_a-書籍、…，或興趣度為u個單位的i_a-書籍。例如某一借閱資料包含有3A，A∈I，則可分解成1A, 2A, 3A等3個單一書籍項目，並表示都曾經借閱過。同樣地，我們以相同方式處理借閱資料中其他的書籍項目。

我們視各分解後的書籍項目為欲分類的影響屬性，經由ID3演算法的計算，可以從所建立之決策樹的根節點到類別為關聯性高的路徑，得知那些分解後之書籍項目會與此一讀者的關聯性高。在書籍推薦的過程中，我們可以加入興趣度的考量，即對那些在決策樹之根節點到類別為關聯性高的路徑中的影響屬性，若其興趣度大於或等於所設定的條件，我們即定義這些影響屬性為包含有興趣度之讀者個人化最適性的書籍推薦。

在探勘過程中，假設借閱資料共有n筆，類別共分為2類（即關聯性高與關聯性低），非X之書籍項目的屬性共有d個，每一屬性之屬性值共有2個（即曾借閱與未曾借閱），每一屬性的興趣度共有u個級數，經由分解後共有ud個屬性。我們考量在最壞的情況下所建立的決策樹其高度為ud，其高度從level 1到level ud，所需要的時間為：

level 1：計算公式（1）+公式（3）+公式（3）的時間=$ud\times(2\times n+2\times 2\times n)+1= 6udn+1$
level 2：計算公式（1）+公式（3）+公式（3）的時間=$(ud-1)\times(2\times n+2\times 2\times n)+1= 6(ud-1)n+1$
level 3：計算公式（1）+公式（3）+公式（3）的時間=$(ud-2)\times(2\times n+2\times 2\times n))+1= 6(ud-2)n+1$
⋮
level ud：計算公式（1）+公式（3）+公式（3）的時間= $1\times(2\times n+2\times 2\times n)+1= 6n+1$

計算其總和為$6nud(ud+1)/2+ud$，即其時間複雜度（time complexity）為$O(nu^2d^2)$。同樣地，我們考量在最壞的情況下所建立的決策樹所需要的記憶體空間為：假設記錄一個類別狀態需要c bytes，記錄一個屬性值需要s bytes，在計算的過程中，我們須將借閱資料載入記憶體中，因此所需記憶體為$n(sud+c)$ bytes。

（二）實例說明

茲以一實例來說明發掘包含有興趣度之某一讀者個人化最適性的書籍推薦。假設{B, C, D, X, Y, Z}為書籍項目的集合，共有6個書籍項目，各書籍項目之興趣度值共分為1到3等三個等級，某一讀者的借閱資料為「1X2Y3Z」，其表示曾借閱過X-書籍且興趣度為1個單位、Y-書籍且興趣度為2個單位，及Z-書籍且興趣度為3個單位。在不失一般性的條件下，表7為借閱資料庫中部分讀者的借閱資料，共有16筆，相似度p為60%，以下我們說明發掘包含有興趣度之此一讀者個人化最適性書籍推薦的探勘過程。

表7　讀者之借閱資料

讀者編號	借閱資料
1	3B2C2X2Y
2	1D1Y
3	1B1C1X3Z
4	2B2C1D2X1Y3Z
5	1C1Y
6	2B2C1D1X3Y
7	3B1C2D1X2Y1Z
8	1Y
9	1C1X1Z
10	1D1Z
11	3B2Y
12	1C1Y
13	1X1Z
14	1D1X1Y
15	1B1C2X2Y
16	1D2Y

首先，我們計算各借閱資料與此一讀者之借閱資料「1X2Y3Z」間的相似度，若相似度大於或等於60%，則設定關聯性為「高」，否則設定關聯性為「低」，其結果如表8。

表8　借閱資料之關聯性

讀者編號	借閱資料	相似度%	關聯性
1	3B2C2X2Y	2/3= 67	高
2	1D1Y	0/3= 0	低
3	1B1C1X3Z	2/3= 67	高
4	2B2C1D2X1Y3Z	2/3= 67	高
5	1C1Y	0/3= 0	低
6	2B2C1D1X3Y	2/3= 67	高
7	3B1C2D1X2Y1Z	2/3= 67	高
8	1Y	0/3= 0	低
9	1C1X1Z	1/3= 33	低
10	1D1Z	0/3= 0	低
11	3B2Y	1/3= 0	低
12	1C1Y	0/3= 0	低
13	1X1Z	1/3= 33	低
14	1D1X1Y	1/3= 33	低
15	1B1C2X2Y	2/3= 67	高
16	1D2Y	1/3= 33	低

其中B、C與D等3個書籍項目，為此一讀者未曾借閱過的書籍項目，依據興趣度值來分別分解各書籍項目，因此將表8的資料轉換成表9，其中第2欄到第8欄分別表示1B、2B、3B、1C、2C、1D及2D等7個單一書籍項目，其中打「1」者表示曾借閱過該欄的書籍項目，打「0」者表示未曾借閱過。

表9 包含未曾借閱過之書籍項目的借閱資料

讀者編號	書籍項目							關聯性
	1B	2B	3B	1C	2C	1D	2D	
1	1	1	1	1	1	0	0	高
2	0	0	0	0	0	1	0	低
3	1	0	0	1	0	0	0	高
4	1	1	0	1	1	1	1	高
5	0	0	0	1	0	0	0	低
6	1	1	0	1	1	1	0	高
7	1	1	1	1	0	1	1	高
8	0	0	0	0	0	0	0	低
9	0	0	0	1	0	0	0	低
10	0	0	0	0	0	1	0	低
11	1	1	1	0	0	0	0	低
12	0	0	0	1	0	0	0	低
13	0	0	0	0	0	0	0	低
14	0	0	0	0	0	1	0	低
15	1	0	0	1	0	0	0	高
16	0	0	0	0	0	1	0	低

1：曾借閱過，0：未曾借閱過

首先利用前文第三節公式（1），計算物件集合的熵值，在表2中共有6位與此一讀者之關聯性為高，及10位與此一讀者之關聯性為低，計算所有讀者的熵值，其計算如下：

$$E(\text{所有讀者}) = -(6/16)\log_2(6/16) - (10/16)\log_2(10/16) = 0.954$$

接著利用公式（2）計算出各個屬性下其子節點的熵值，以單一書籍項目「1B」為例，其屬性值的熵值計算如下：

$$E(\text{曾借閱過1B-書籍}) = -(6/7)\log_2(6/7) - (1/7)\log_2(1/7) = 0.592$$

$$E(\text{未曾借閱過1B-書籍}) = -(0/9)\log_2(0/9) - (9/9)\log_2(9/9) = 0$$

$$E(2B) = (7/16) \times E(\text{曾借閱過1B-書籍}) + (9/16) \times E(\text{未曾借閱過1B-書籍}) = 0.259$$

再者，利用公式（3）計算資訊收益：

$$G(1B) = 0.954 - 0.259 = 0.659$$

依此類推，可計算出其餘各屬性項目的資訊收益如下：

$G(2B) = 0.566$

$G(3B) = 0.059$

$G(1C) = 0.437$

$G(2C) = 0.321$

$G(1D) = 0.006$

$G(2D) = 0.199$

我們選定具有資訊收益最大者1B-書籍做為該決策樹的根節點，再把其餘6個屬性項目加入1B-書籍之後，再次計算其熵值。借閱資料經由選定1B-書籍之後的分類情況如表10。

表10　包含未曾借閱過之書籍項目的借閱資料

讀者編號	書籍項目							關聯性
	1B	2B	3B	1C	2C	1D	2D	
1	1	1	1	1	1	0	0	高
3	1	0	0	1	0	0	0	高
4	1	1	0	1	1	1	1	高
6	1	1	0	1	1	1	0	高
7	1	1	1	1	0	1	1	高
11	1	1	1	0	0	0	0	低
15	1	0	0	1	0	0	0	高
2	0	0	0	0	0	1	0	低
5	0	0	0	1	0	0	0	低
8	0	0	0	0	0	0	0	低
9	0	0	0	1	0	0	0	低
10	0	0	0	0	0	1	0	低
12	0	0	0	1	0	0	0	低
13	0	0	0	0	0	0	0	低
14	0	0	0	0	0	1	0	低
16	0	0	0	0	0	1	0	低

1：曾借閱過，0：未曾借閱過

由於在1B-書籍之屬性值為「0」的情況下，可得到關聯性低的同一類別，因此關於1B-書籍之屬性值為「0」的計算就結束。而在1B-書籍之屬性值為「1」的情況下，仍處於不同的類別值，因此必須繼續進行分類的計算，如表11。

我們以表11做為分類分析的資料來源，首先再利用公式（1），計算物件集合的熵值，其計算如下：

$$E(\text{所有讀者}) = -(6/7)\log_2(6/7) - (1/7)\log_2(1/7) = 0.592$$

接著利用公式（2），計算所有書籍屬性的熵值，我們單一書籍項目「1C」為例，其熵值計算如下：

表 11 包含未曾借閱過之書籍項目的借閱資料

讀者編號	書籍項目						關聯性
	2B	3B	1C	2C	1D	2D	
1	1	1	1	1	0	0	高
3	0	0	1	0	0	0	高
4	1	0	1	1	1	1	高
6	1	0	1	1	1	0	高
7	1	1	1	0	1	1	高
11	1	1	0	0	0	0	低
15	0	0	1	0	0	0	高

1：曾借閱過 0：未曾借閱過

E（曾經借閱過 1C 書籍）$= -(6/6)\log_2(6/6) - (0/6)\log_2(0/6)=0$

E（未曾借閱過 1C 書籍）$= -(0/1)\log_2(0/1) - (1/1)\log_2(1/1)=0$

E（1C 書籍）$=(6/7)\times 0+(1/7)\times 0=0$

再利用公式（3）來計算 1C- 書籍的資訊收益：

G（1C 書籍）$=0.592 - 0=0.592$

依此類推，可計算出其餘各書籍項目屬性的資訊收益如下：

G（2B 書籍）$= 0.076$

G（3B 書籍）$= 0.199$

G（2C 書籍）$= 0.129$

G（1D 書籍）$= 0.129$

G（2D 書籍）$= 0.076$

由於 1C- 書籍的資訊收益最大，因此將 1C- 書籍做為連接於 1B- 書籍之屬性值為「1」之路徑的節點，再把其餘的 5 個書籍項目加入 1C- 書籍之後，再次計算其熵值。借閱資料經由選定 1C- 書籍之後的分類情況如表 12。

表 12 包含未曾借閱過之書籍項目的借閱資料

讀者編號	書籍項目						關聯性
	2B	3B	1C	2C	1D	2D	
1	1	1	1	1	0	0	高
3	0	0	1	0	0	0	高
4	1	0	1	1	1	1	高
6	1	0	1	1	1	0	高
7	1	1	1	0	1	1	高
15	0	0	1	0	0	0	高
11	1	1	0	0	0	0	低

1：曾借閱過，0：未曾借閱過

由於在1C-書籍之屬性值為「0」的情況下，可得到關聯性低的同一類別，因此關於1C-書籍之屬性值為「0」的計算就結束。而在1C-書籍之屬性值為「1」的情況下，也可得到關聯性高的同一類別，因此結束建構決策樹的分類計算。然後，將所有分類屬性依照先後順序連接起來，就形成如圖2的決策樹。

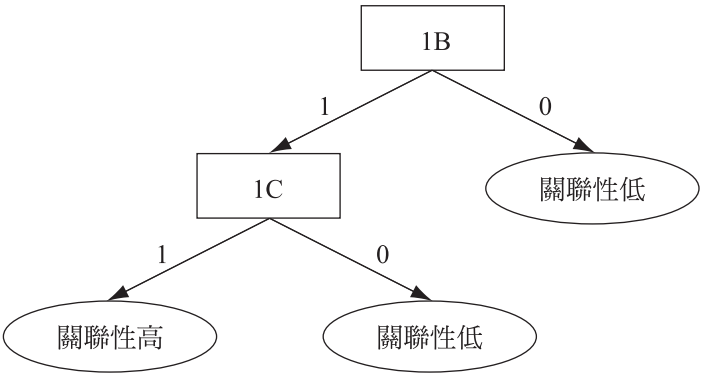

圖2　分類決策樹

從圖2分類決策樹，若設定推薦的興趣度必須大於或等於1，則可發掘包含有興趣度之此一讀者個人化最適性的書籍推薦為：B-書籍或C-書籍。

五、讀者個人化最適性之書籍推薦系統建置

我們利用前文各節所描述的方法，應用到探勘讀者個人化最適性之書籍推薦的系統實作上。我們以C#為撰寫的程式語言，在不失一般性的條件下，假設書籍項目全部有26項，分別以A, B, C, …, Z表示之，以亂數隨機產生每一讀者的借閱資料，借閱資料中的各書籍其興趣度值最多為5個單位，共產生100筆借閱資料。以下為此一系統探勘過程的執行畫面。

圖3為系統的借閱資料，包含有「讀者編號」與「曾經借閱過的書籍項目」等欄位資料。


借閱資料：資料表

讀者編號	曾經借閱過的書籍項目
1	1A,1B,3C,2E,1F,2H,4K,2P,4Q,5W,
2	2G,1I,2K,3M,2U,1Z
3	1C,2E,1J,2Q,4U
4	1B,2C,3D,2E,1F,2G,4H,5M,3N,1V
5	1E,2G,3H
6	2B,1C,3D,1E,2F,2H,1I,1K
7	2B,3C,1D,2E,4G,2H,1L,2N,3P,4R,1
8	2B,4G,2H,1J,2N,1Q,2V
9	1C,2D,4E
10	2B,3D,1J

記錄: 1 之 100

圖3　借閱資料

圖4為探勘畫面，其中包含有兩項功能選項：「探勘個人化最適性之書籍推薦」及「探勘包含有興趣度之個人化最適性的書籍推薦」。假設目前點選「探勘個人化最適性之書籍推薦」功能，因此，在探勘過程中將忽略書籍項目的興趣度，在「請輸入欲探勘之讀書編號」欄位中填入「9」，經由第三節演算法的計算過程，可在「推薦書籍項目」欄位中顯示出探勘結果，如圖4，在建構出的決策樹中，我們簡化書籍推薦的陳述，只列印出與欲探勘之讀者具有關聯性高之路徑中的書籍項目。

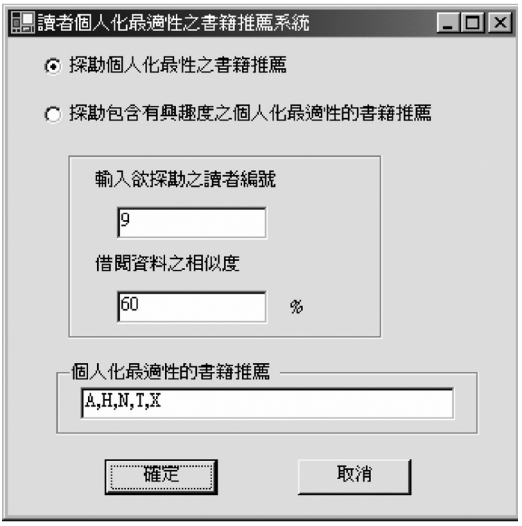

圖4　探勘個人化最適性書籍推薦執行畫面

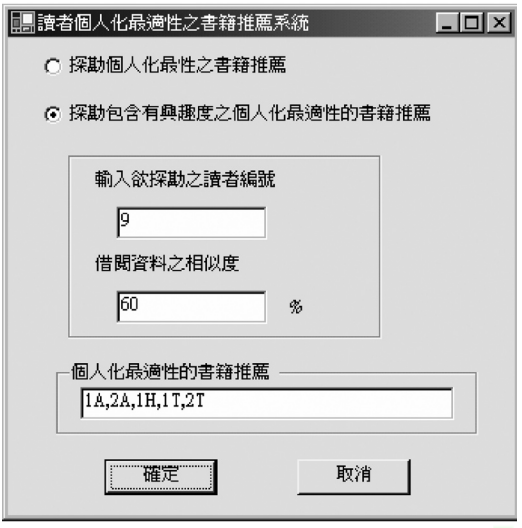

圖5　探勘含有興趣度之個人化最適性書籍推薦執行畫面

圖5為點選「探勘包含有興趣度之個人化最適性的書籍推薦」功能，在「請輸入欲探勘之讀書編號」欄位中填入「9」，經由第四節演算法的計算過程，可在「推薦書籍項目」欄位中顯示出包含有興趣度大於或等於1的探勘結果，如圖5，同樣地，我們只列印出與欲探勘之讀者具有關聯性高之路徑中的書籍項目。

六、結論與未來研究

在讀者曾經借閱過的書籍與其興趣度中，可反映出讀者對書籍的偏好特徵，若能從這些資料中找出讀者對不同書籍彼此間的關聯性，對圖書館經營者在擬訂讀者個人化的書籍推薦時，必可提供相當有用的資訊。在本論文中，我們以某一讀者為探勘的目標，利用分類技術分別考量書籍項目是否出現在借閱資料中、及書籍包含有興趣度，來找出此一讀者個人化最適性的書籍推薦。從資料的蒐集、分析、方法設計，及結果推導出的個人化書籍推薦，顯示我們所提出之探勘過程具有實務應用及方法創新的學術價值。

目前已有許多相關的研究，探討如何利用資料探勘技術以提昇圖書館對讀者個人化的服務（註19），但鮮少利用分類技術做分析，及針對讀者對借閱書籍的興趣度做考量。因此，本論文利用分類技術並考量讀者對書籍的興趣度，來探勘借閱資料中書籍項目與讀者的關聯性，勢必能找出對讀者個人更貼切與適性的書籍推薦。

本研究僅就如何利用分類技術在發掘圖書館個人化之書籍推薦的探勘過程作探討，並根據所提出之探勘方法，設計與建置一個讀者個人化最適性之書籍推薦系統的操作雛型，以具體顯示書籍推薦系統最基本的操作功能。對於未來繼續從事之相關研究有：

（一）對借閱資料之格式與形態的有效分析與運用。

（二）考量讀者個人特徵資料與其閱讀領域的興趣（例如文藝類、旅遊類、科幻小說類等閱讀興趣）等因素。

（三）應用其他資料探勘技術在此研究問題的可行性。

（四）利用網頁設計功能改良探勘系統操作介面的友善性，並做實際的應用及分析驗證。

誌　謝

作者感謝兩位匿名審查委員寶貴的意見和指正。

註　釋

註1　K. C. Laudon, & J. P. Laudon, Management information systems: Managing the digital firm, 8th ed. (Prentice Hall, 2002).

註2　辜曼蓉，「讀者資訊尋求行為與以讀者為中心的圖書館行銷」，書府，20（1999）：頁81-111。

註3 J. Ou, S. Lin, & J. Li, “The personalized index service system in digital library,” *Cooperative Database Systems for Advanced Applications*, (2001), pp. 92-99.

卜小蝶，「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(民國87年10月)。

註4 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資訊服務(PIE@NCTU)系統」，國立成功大學圖書館館刊，9(民國91年4月)。

交通大學個人化數位圖書館資訊服務，http://mylibrary.e-lib.nctu.edu.tw/

註5 M. S. Chen, J. Han, J., & P. S. Yu, “Data mining: An overview from a database perspective,” *IEEE Trans. on Knowledge and Data Engineering*, 8：6(1996)：866-883.

魏志平、董和昇，電子商務理論與實務(台北市：華泰書局，2002)，頁167-205。

註6 M. J. A. Berry, & G. Linoff, *Data mining techniques for marketing, sales, and customer support* (New York: John Wiley, 1997).

註7 陳慶瑄，「學習社群對電子圖書館個人化服務之影響」(碩士論文，國立中正大學資訊管理研究所，2000)。

註8 孫冠華，「應用資料探勘技術於數位圖書館之個人化服務及管理」(碩士論文，南華大學資訊管理學研究所，2003)。

註9 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，2001)。

註10 洪志淵，「圖書流通記錄之一般化相關規則找尋之研究」(碩士論文，國立中山大學資訊管理研究所，2001)。

註11 張苑菁，「以模糊理論建構之圖書推薦系統」(碩士論文，淡江大學資訊工程研究所，2001)。

註12 余明哲，「圖書館個人化館藏推薦系統」(碩士論文，國立交通大學資訊科學研究所，2003)。

註13 J. R. Quinlan, “Induction of deciscion trees,” *Machine learning,* 1(1986)：81-106.

註14 P. Clark, & T. Niblett, “The CN2 induction algorithm,” *Machine Learning*, 3 (1989)：261-283.

註15 E. Rich, & K. Knight, Learning in neural network, 2nd ed. (New York：McGraw-Hill, 1991).

註16 同註13。

註17 同註5，魏志平、董和昇；同註13，Quinlan。

註18 P. E. Utgoff, “Incremental induction of deciscion trees,” *Machine learning*, 4(1989)：61-186.

註19 同註4，湯春枝；同註7，陳慶瑄；同註8，孫冠華；同註12，余明哲。

Using Data Mining Techniques to Discover Personalized Book Recommendation for Library

Chui-Cheng Chen

Associate Professor
Department of Information Management
Southern Taiwan University of Technology
Tainan, Taiwan, R.O.C.
E-mail: ccchen@mail.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 with the degree of interest. We let one reader as the target of mining and use classification analysis to discover the personalized book recommendations for the reader. In the mining process, we compute the degree of similarity of borrowing history records between the reader and other. If the degree conform the given condition, we assign the association level between the both readers is "high". Otherwise, it is "low". For books not borrowed by the reader, we treat those books as attributes for classification. First, we only consider readers ever borrowed books, and classify the borrowing history records to construct a decision tree. We can find the association level to be "high" between some attributes and the reader according to the decision tree. It is the basis to discover the most adaptive book recommendations for the reader. Moreover, we consider books with readers interests in the borrowing history records. Each book is divided to u unit items where u is the degree of the interest, u is positive integer, and the degrees of interest of these items are, respectively, from 1 to u. For books not borrowed by the reader, we divide those books to unit items and treat those items as attributes for classification. We can construct a decision tree after classifying the borrowing history records. According to the decision tree, we can find the association level to be "high" between some attributes and the reader. It is the basis to discover the most adaptive book recommendations for considering the reader's interesting. The results of the mining can provide very useful information to recommend the most adaptive books for individual reader.

Keywords: *Data mining; Classification analysis; Borrowing history records; Book recommendations*