南　華　大　學

資訊管理學系碩士論文

應用資料探勘技術於數位圖書館之個人化服務及管理

Apply Data Mining Techniques to Enable Personalized Services and Management on Digital Library

研 究 生：曹　健　華

指導教授：邱　宏　彬

中　華　民　國　九十二　年　六　月

南　華　大　學

碩士學位論文

資訊管理學系

應用資料探勘技術於數位圖書館之個人化服務及管理

Apply Data Mining Techniques to Enable Personalized Services and Management on Digital Library

研究生：曹健華

經考試合格特此證明

口試委員：

指導教授：邱宏彬

所　長：吳光陽

口試日期：中華民國 92 年 6 月 23 日

IV

用資料探勘技術於數位圖書館之個人化服務及管理

學生：曹健華　　指導教授：邱宏彬

南　華　大　學　資訊管理學系碩士班

摘　　要

在數位科技快速成長的時代，已有許多的演算法及技術被提出來加以探究，並且在每天大量產生的電子資料中去分析以探勘出有意義的資訊，以提升數位圖書館知識分享及服務品質。在本研究中，我們建構一個研究模型，以達到個人化服務及數位圖書館管理的目的。以個人化服務為例，即是利用相似特性，如身份別相同之讀者的歷史借閱館藏記錄以產生關聯法則，做為館藏推薦的基礎。除此之外，我們還運用 OLAP 的技術來分析資料方塊，而這些資料方塊是以具有相關聯性的電子資料所產生而來，以獲得使用數位圖書館資源的最新資訊，並有效率地提升數位圖書館的資源管理。我們針對所提出的研究方法以幾個例子加以分析及討論，而數位圖書館的管理人員便可以我們所建議的研究方法在有限的預算中，採購核心及熱門的館藏，以維持及滿足讀者的需求。

關鍵字：資料探勘、關聯法則、個人化服務、數位圖書館，線上分析處理、資料方塊。

Apply Data Mining Techniques to Enable Personalized Services and Management on Digital Library

Studeunt：Julius Tsao　　Advisors：Dr. Hung Pin Chiu.

Department of Information Management
The M.B.A. Pprogram

Nan-Hua University

ABSTRACT

In the era of digital technology growing rapidly, a lot of algorithms and techniques have been proposed to explore and analyze the large quantities of electronic data produced everyday to discover meaningful information so as to enhance the knowledge sharing and service quality of digital library. In this study, we design a system model to enable personalized services and management on digital library. For personalized services, association rules discovered from the books borrowed by the readers in the same cluster are used as the basis of book recommendation. Besides, we apply the OLAP technology to analyze the data cubes, which are created by the relevant electronic data, to obtain the desired up-to-date information for resource usage to effectively promote the resource management of digital library. In this study, several application cases for the proposed approach were analyzed and discussed. Based on the proposed approach, the library managers are expected to purchase the core and hot books in limited budget to maintain and satisfy the requirements of readers.

Keywords：data mining、personalized service、association rules、 OLAP、data cube、digital library

誌　謝

從來都沒有想過有一天能就讀研究所。

首先感謝洪明輝博士及王昌斌博士在高雄學分班時期，對我的啟蒙及提攜，使我有機會進入本校資管所就讀。

更要感謝指導教授邱宏彬博士不厭其煩、悉心指導，才能使資質愚頓的我，完成此篇論文，給您添了很多麻煩，謝謝您的體諒。

回想過去三年點點滴滴，感謝上述三位老師對我的指教及協助，心中的感激非這短短的隻字片語所能表達，在此僅以此「誌謝」向三位老師致上我最高的敬意，祝三位老師身體健康、事業順利。

本篇論文的完成，還要感謝女友靜榆的協助，在這撰稿期間，我的脾氣不好，讓妳受委屈了，謝謝。

同學武隆、國荷還有好多好多幫助我的好朋友、好同學，在此一併向你們大家致謝，祝大家身體健康、事事順心。

最後感謝我的父母，放縱我的任性，使我這二年沒有後顧之憂，專心撰寫論文，祝我的父母福壽綿長。

曹健華　謹誌

民國 92 年 6 月 28 日

目　錄

表　　目　　錄

圖　　目　　錄

第一章、緒論

第一節　研究動機

在網際網路日益普及發達的今日，知識的獲取已不分時地、知識的取得在質量上日益倍增，而圖書館在知識的分享上，因其有系統的儲存知識及透過專業人員的管理，使得圖書館不因時代的變遷而顯得没落，相反的在這知識爆炸、資訊充斥的時代，扮演著不可獲缺、舉足輕重的角色。

然而，傳統圖書館所提供的功能和其所扮演的角色，也需隨著數位科技及網際網路快速成長的步伐，而有所演進。以往有關運用資料探勘技術於數位圖書館【1】【2】方面的論文，著重在依借閱館藏或讀者分群，找出其中的關聯法則以作為推薦的依據。另一方面有關個人化的服務，也偏重在以分群所探勘出的關聯法則結果為基礎，製作一個個人的借閱館藏之資訊使用網頁。因此，在本論文中，將針對數位圖書館的服務及管理所衍生出的兩個課題來加以討論。

壹、針對個人化服務方面：

一、如何根據相似特性，如身份別相同之讀者的歷史借閱館藏記錄來產生關聯法則，並予以推薦？

二、到館新館藏推薦

(一)、主動以 e-mail 通知讀者到館新館藏清單。

(二)、在線上推薦給上線中的讀者。

貳、針對數位圖書館管理(Digital library)方面：

傳統上，圖書館管理人員常常需要得知諸如以下的最新資訊：

一、什麼身分的讀者較常借閱館藏？

二、什麼身分的讀者會較常借閱什麼種類的館藏？

例如：大一新生較常借閱入門書籍，研一新生較常借閱各類期刊論文。

三、借閱館藏的讀者有那些共同點？

例如：常借閱某一類別的館藏，借閱時間平均一次有三週，一學期借閱次數平均有五次，借閱高峰期通常在期中、期末考等。

四、借閱館藏的讀者有那些身份特徵上的區分？

例如：女性較常借閱館藏，男性一次借閱館藏時間較短，研一、大一新生較常借閱館藏。

五、如何擴大讀者群？

以往只針對已借閱館藏的讀者提供服務，但這樣並未影響到

從未借閱館藏，只借閱過一次就未再借閱或已許久未曾借閱館藏的潛在讀者，因此若要擴大讀者群，便非得從這些潛在讀者著手才行。

例如：根據關聯法則得知，企管系大二男性喜歡借閱企業家傳記，因此透過學校 e-mail 信箱主動通知所有企管系大二男性（未曾有借閱記錄或已許久未曾再借閱的讀者）這項訊息。

上述五點可用統計的方法或資料庫語言如 SQL 來處理，但這通常要花一段時間才能得知，而且如 SQL 語言在查詢時得要將要查詢的條件一一輸入，若要查詢的項目很多時，所下的指令相對也較繁複，而統計的方法可能也要花費相當的人工做事後的資料整理，兩者在時間上都相對較不經濟。

因此，我們以 on line mining 的方式，以 OLAP 觀點使用如 IBM DB Miner 等工具，產生一個除了上述兩個方法之外，可以輔助圖書館管理者的管理介面，可突破傳統統計及 SQL 的耗費及可能產生資訊不完整的缺點，使圖書館管理者可立即得知相關訊息，提升管理效率。

第二節　研究目的

壹、個人化服務

從工業時代大量行銷【3】、顧客就接受的產銷模式，至今，強調個人化、專屬化、個性化的商品，不斷地推陳出新，已成為主要之市場導向【4】的模式。同樣地，傳統圖書館也從大學所屬的公務機構，轉變成以服務業性質為導向的數位圖書館模式，如同服務業專業行銷的角色，數位圖書館必須能掌握顧客的消費習性，即讀者的借閱模式、推出符合讀者個人需求的資訊服務、即個人化服務 Personalized Service【5】【1】，如下所言：

以節省時間並提供有效率的館藏搜尋服務及符合讀者需求，並以過去讀者歷史借閱記錄，依讀者個人借閱館藏習慣、興趣、針對這些借閱記錄，經由資料探勘【6】技術分析出讀者的借閱關聯法則，進而主動推薦相關館藏予讀者，除滿足讀者借閱需求外，更進一步提供讀者潛在興趣的館藏可供借閱資訊，即創造需求【2】的概念，以提升圖書館館藏的使用率，並增加讀者的滿意度【3】。

貳、圖書館管理（讀者借閱館藏資訊管理）

為因應教育資源分配有限化的考量，圖書館的經費也在日漸緊縮之中。因此，如何因應在有限經費下，圖書館得以採購質量兼備

的核心館藏及熱門館藏，以維持及滿足讀者的需求，便顯得緊迫及重要。

在此，我們以 On Line Mining【7】的方式，以 OLAP【8】觀點，提供圖書館管理人員以即時的方式，利用 OLAP 觀點所研製的管理軟體【9】【10】，有效率且更貼近讀者使用圖書館資源的最新資訊，創造圖書館管理人員、讀者及校方三贏的局面。

第三節　論文架構

本論文的架構為第一章緒論，包含研究動機、研究目的。

第二章文獻探討： 包含數位圖書館相關文獻回顧、中國圖書分類法範例、個人化及個人化服務的定義、關聯法則、 Apriori 演算法、資料倉儲、線上分析處理及資料方塊。

第三章研究方法包含：研究模型、MUM 演算法、MUM 演算法範例說明、資料方塊範例說明。

第四章實驗結果：為 MUM 演算法實驗及分析。

第五章結論及未來工作。

第二章、文獻探討

第一節　數位圖書館相關文獻回顧

首先我們必須了解什麼是「數位圖書館」？它的定義如下所示：(資料來源：【4】)。

「數位圖書館乃是擁有相關資源(包含軟硬體設備網路、專業人士.....等）以執行下列任務的機構：

對數位形式的館藏進行挑選、組織、提供使用、解譯、傳播、保持完整性、長期保存等工作，並使這些數位形式的館藏能為特定讀者群快速且經濟地運用。」「其館藏包含電子式（與數位式）以及印刷和其它（例如：影帶和聲音）媒材。」

在此我們藉由幾篇先進的研究文獻，來回顧有關數位圖書館方面，大家所提出的問題。

壹、吳安琪【4】的論文所探討的重心是在：

一、讀者難以決定要借那些館藏？

二、讀者不知道有那些相關館藏？

三、館藏數量漸趨龐大，讀者不知道如何利用這些資源？

貳、陳建銘【5】的論文所探討的重心是在：

一、學生進入圖書館所借閱的書籍中，那幾個種類的書籍是同時借閱的，而出現的頻率較高呢？

二、圖書館管理人員若能依圖書館借閱資料庫中所挖掘的資料，多採購借閱率極高的某些類別書籍，……在尋找關聯規則中可挖掘出借閱率高的書籍。

參、戴玉旻【6】的論文所探討的重心是在：

一、提供讀者借閱館藏的建議：讓讀者藉由分析借閱記錄找出讀者經常一起借閱的館藏，再將關聯性館藏推薦給借閱同樣館藏的讀者。

二、推薦讀者新進館藏：藉由借閱類別關聯性，根據讀者的借閱記錄分析興趣類別，進而推薦關聯類別的新進館藏給讀者。

肆、孫冠華【7】的論文所探討的重心是在：

一、某一個人，他對那一類的書，有長期的借閱興趣？

二、對於某一類的新書進入館藏後，我要 e-mail 給那一群人，他們的特性是什麼，如果我 e-mail 給這些人後，我有百分之多少的把握，這些人對於這一類的書有借閱的興趣？

三、如果一個讀者，在借閱一本書時，我能同時給他一個建議，使

他在借閱的同時，能存更多同類書籍的資訊？

針對上述所提出的問題，我們整理出幾個方向：

一、一次借閱兩本以上的館藏，是否有關聯性存在，若存在關聯性，該用什麼方法找出？

二、讀者與圖書館管理人員如何妥善利用龐大的館藏資源，以達到最大的使用效能？

三、是否能對讀者提供借閱館藏的建議？

四、讀者借閱館藏是否有一些特性存在？

在此，我們利用資料探勘技術中的關聯法則來解決上述的幾個問題，並且加入個人化服務的觀念，並提供圖書館管理人員 On Line Mining 的機制，以強化數位圖書館的館藏資源運用。

第二節　中國圖書分類法範例

為了讓讀者能便於了解中國圖書分類法的分類方式，以下我們將以「中國圖書分類法」這本書做為範例，以圖表的方式來加以說明。

首先讀者在南華大學圖書館的館藏查詢系統，可以搜尋的方式，找到讀者要找的「中國圖書分類法」這本書。

表 2-1 南華大學圖書館館藏查詢資料

#	書刊名 TITLE	作者/ 出版商 AUTHOR/ PUBLISHER	出版年 PUBLISH YEAR	書目索書號 CALL ON	書籍類型 MARC TYPE
□	中國圖書分類法	賴永祥/文華	200	R 023.31 5733 90	中文圖書

資料來源：南華大學圖書館館藏查詢系統網頁

http://libserver2.nhu.edu.tw/11.htm

表 2-2 中國圖書分類綱目表

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 史地總綸	800 語言文字學
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 家專	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 向教	450 礦治	650 史料	850 特種文學
060 普通會社	260 猶太教	460 化學工廠	660 地理	860 東洋文學
070 普通論叢	270 其他宗教	470 製造	670 方志	870 西洋文學
080 普通叢書	280 神話	480 各種營業	680 類志	880 西方諸小國文學
090 群經	290 迷信	490 經營學	690 遊記	890 新聞學
哲學類	自然科學	社會科學類	史地類世界	美術類
100 總論	300 總類	500 總論	700 世界	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 澳洲及其他各洲	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 遊藝：娛樂；休閒

資料來源：【8】

從表 2-1 南華大學圖書館館藏查詢資料內的書目索書號： 023.31 5733 90 為例，對照表 2-2 中國圖書分類綱目表可得到如下分析。

表 2-3 書目索書號分析

書目索書號： 023.31　5733　90
020 代表圖書館總論
023 代表管理
.31 代表中文
5733 代表著者四角號碼
90 代表出版年份

第三節　個人化及個人化服務的定義

壹、個人化（Personalize）的定義：

根據 The American Heritage dictionary 的定義如下：

『To take（a general remark or characterization）in a personal manner』使個人帶有一般的標誌或特徵的行為。由此可推知個人化是具有與眾不同的特色而稱之。

貳、個人化服務（Personalized Service）的定義：

對讀者除了提供一般性的服務之外，另外針對讀者個人的需求量身打造專屬的服務內容，以彰顯其與眾不同的特色，即可稱為個人化服務，如 Amazon.com。

第四節　關聯法則

壹、關聯法則的定義：

令 $I=\{i_1, i_2 ...im\}$ 成為所有項目所成的集合，D 是所有交易記錄 T（Transaction T）的集合，T 是項目集 I 的子集合，是故 $T \subseteq I$，每筆交易記錄給予一個識別碼，稱為 TID。

令 A 是數個項目所成的集合。假設 A 與 B 都是項目的集合，並且都包含在交易記錄 T 之內，則一個關聯法則是：

$A \rightarrow B$

$A \subset I$、$B \subset I$ 及 $A \cap B = \Phi$

一、關聯法則 $A \rightarrow B$ 在交易記錄集 D 中，有著 Support（支持度）S，S 表示在交易記錄集 D 中，$A \cup B$ 的機率，寫成 $P(A \cup B)$。

二、關聯法則 $A \rightarrow B$ 在交易記錄集 D 中，有著 Confidence（信心度）C，C 表示在交易記錄集 D 中，$A \cup B$ 的機率，寫成 $P(A \mid B)$。

也就是說，可得到下列的關聯法則：

Support（A→B）＝P（A∪B）

Confidence（A→B）＝P（A | B）

P 表示 Probability（機率）

舉例來說：

假設有 60% 的讀者借閱 BASCE 入門，有 75% 的讀者借閱 Java 進階，有 40% 的讀者同時借閱 BASCE 入門及 Java 進階，則依上述關聯法則的定義可得出以下的規則。

BASCE 入門→Java 進階〔40% ，60% 〕

貳、相關課題：

在關聯法則研究的領域中，如何在龐大的原始資料庫中以有效率的方式挖掘出隱含且有意義的資訊及規則，一直都是一項重要的課題。

而 Apriori 演算法便是關聯法則研究中的先驅者，而本文所要探討的問題，是建立在以 On Line Mining 為基礎的系統上。因此，除了要能正確無誤的挖掘出關聯法則外，Data Mining 過程的效率便是非常重要的課題。畢竟，在數位圖書館的架構中，透過網路資訊傳

輸速度的快慢，將會對在線上借閱館藏的讀者及圖書館管理者在使用意願上，產生很大的影響。

在效率方面，有二個問題是值得我們來加以探討的。一是演算法本身在 Data Mining 過程的效率問題。另一是如何避免重複的 Mining 原始的資料庫。畢竟演算法本身效率再快，如果在 Data Mining 的過程中必須不斷再重複搜尋原始資料庫，在時間上將是一大浪費。

第五節　Apriori 演算法

現在，讓我們先簡單的回顧一下 Apriori 演算法的主要步驟：

步驟一、反覆的產生候選項目組（Candidate itemset）並搜尋整個資料庫，直到找出所有的大項目組（Large itemset）。

步驟二、利用步驟（1）所找出的大項目組（Large itemset），推導出所有的關聯法則。

演算過程如以下範例所示：

假設原始交易資料庫如下圖

原始交易資料庫 D

TID	List of Item_IDS
100	1,2,5
200	2,4
300	2,3
400	1,2,4
500	1,3
600	2,3
700	1,3
800	1,2,3,5
900	1,2,3

假設最小支持度
（Minimum Support）為 2

C1

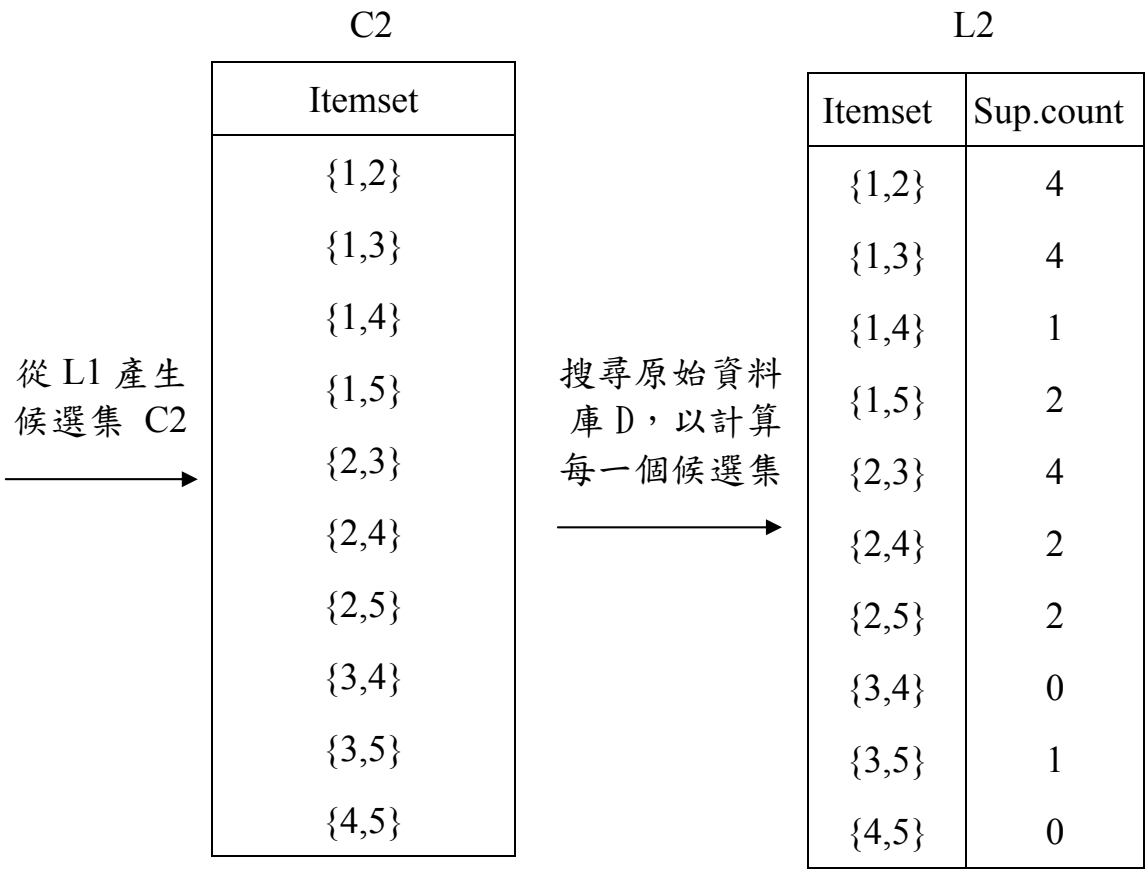
Itemset	Sup.count
{1}	6
{2}	7
{3}	6
{4}	2
{5}	2

搜尋原始資料庫 D 以計算每一個候選集 →

比較候選集支持度與最小支持度 →

L1

Itemset	Sup.count
{1}	6
{2}	7
{3}	6
{4}	2
{5}	2


C2
L2
Itemset
{1,2}
{1,3}
{1,4}
{1,5}
{2,3}
{2,4}
{2,5}
{3,4}
{3,5}
{4,5}
從 L1 產生
候選集 C2
搜尋原始資料
庫 D，以計算
每一個候選集
Itemset
Sup.count
{1,2} 4
{1,3} 4
{1,4} 1
{1,5} 2
{2,3} 4
{2,4} 2
{2,5} 2
{3,4} 0
{3,5} 1
{4,5} 0

L2

Itemset	Sup.count
{1,2}	4
{1,3}	4
{1,5}	2
{2,3}	4
{2,4}	2
{2,5}	2

比較候選集支持度與最小支持度 →

從 L2 產生候選集 C3 →

C3

Itemset
{1,2,3}
{1,2,5}

C3

Itemset	Sup.count
{1,2,3}	2
{1,2,5}	2

搜尋原始資料 D，以計算每一個候選集 →

比較候選集支持度與最小支持度 →

L3

Itemset	Sup.count
{1,2,3}	2
{1,2,5}	2

在此，我們提出幾個想法，試著加以討論：

壹、Apriori 演算法在演算過程中必須多次搜尋原始資料庫，來產生候選項目集（Candidate itemset），並且由於本身演算法的特性，如果原始資料庫交易記錄資料量大時，將會產生數量龐大的候選項目集（Candidate itemset）。

貳、隨著時間的增加，原始資料庫中的交易記錄將會隨之有所改變，此時 Apriori 演算法在每次有新的交易記錄資料進入原始資料庫時，便必須重新搜尋更動後的整個原始資料庫裡的所有交易記錄資料。

參、0n Line Mining 重要的特性之一，便是能讓數位圖書館管理人員能即時在線上（On line）更動相關參數的設定，即關聯法則中的門檻值（threshold），因此，所使用的演算法必須具有可供機動變更門檻值（threshold）的特性才行。

所以，我們發現在數位圖書館的關聯法則挖掘的應用上，將會遇到以下的問題：

壹、Apriori 演算法運算相當耗時，是否適合用於著重效率的 On Line mining 上？

貳、Apriori 演算法當有一筆新交易記錄進入原始資料庫時，能否不須再重複搜尋整個原始資料庫？只要針對新的交易記錄資料加以處理即可。

參、Apriori 演算法能否動態變更門檻值（threshold），以支援 On Line Mining ？

綜合上述的想法和所提出的問題，我們發現以 Apriori 演算法的特性，能難達成上述問題所提出的要求。因此我們將會使用一種演算法，來償試達成上述我們所提出問題的要求。

第六節　資料倉儲

學者 W.H.Inmon 於 1992 對資料倉儲（Data warehouse）的定義：「資料倉儲其備目標導向、整合性、時間變動性、不擇性等四項特質之摘要性安細部性之資料儲存庫（Data Repository），主要用來支援企業決策過程」【11】。

在原本資料庫的基礎上，資料倉儲對於來自內部及外部的資料進行粹取、轉換、載入、更新等處理程序。

資料倉儲的主要目的在於提供具整合性及集中性的資料，並著重於如何建立及存取資料的技術，所支援的範圍為資料倉儲內的決策資訊。

資料倉儲在建置的過程中，所運用到的關鍵技術緊接著在下一節加以說明。

第七節　線上分析處理及資料方塊

資料倉儲在建置的過程中，運用到兩項關鍵技術，分別是：

壹、線上分析處理 OLAP（On Line Analytical Processing）：

OLAP 可視決策者需要，在多維、多使用者、主從式伺服器計算環境、分析企業資料，處理多維要求，以計算、合併及擷取多維及關聯式資料庫中的資料，來提供資料快速、直覺式線上分析程序，使決策者能快速存取資訊。

OLAP 具有縮放、資料旋轉、複雜計算、趨勢分析及製作模型等特性【12】。因此可針對資料倉儲進行的整理與分析，在資料倉儲作分析與決策時，OLAP 扮演決策分析工具的角色。

OLAP 資料的呈現是以多維的形式來進行資料的處理，所得到的分析結果，又稱多維資料（Multidimensional Data）而以資料方塊（Data Cube）呈現，這也是後續我們在本文中作為數位圖書館管理時所使用到的技術。

貳、多維度模式：Multidimensional Model

多維度模式是一種圖形化的資訊模式，包含以下的組成文件【13】【14】：

一、維度 Dimension：

維度是資料庫大綱中之資料的最基本種類定義，是特定資料的透視或檢視，代表 Data Cube 中表格內的一個單一值的欄位，至少需要二個維度以上才能作有意義的資料參照，維度通常為種類。如讀者維度、館藏維度、時間維度，見圖 2-1 所示。

二、成員 Member：

維度成員為維度內個別元件的元素名稱，即每個維度內所代表的值，維度可以包含無限量的成員。如讀者姓名、館藏名稱、年度月份。

三、量度 Measure：

所要統計的量，如在 2002 年 6 月的館藏借閱數量人數。

四、事實表格 Fact Table：

在關聯式資料庫中所建立的一種表格，保存資料庫的實際資料值、事實表格包含多維資料，綜合各維度表格內的相關資料，加以儲存而成主要表格，如館藏借閱記錄事實資料表，見圖 2-2 所示。

五、維度表格 Dimension Table：

保存關於資料庫成員（Member）及各成員間關係的資料，每一個維度必須用一個表格來定義，而維度表格的內容，如身份別維度表格、部門別維度表格、借閱日期維度表格，也就是 Data Cube 中各個維度所欲統計的資料來源。

六、星狀網目 Star Schema：

星狀網目為關聯式資料庫網目類型，在建立多維資料庫時，OLAP Server 會建立一個主要事實表格（Fact Table）及一組維度表格（Dimension Table）。通常用來連結多個維度表格，因與維度表格（Dimension Table）呈星狀排列，故稱星狀網目（Star Schema），這也是多數資料倉儲用來表示多維度資料模型（multidimensional data model）的邏輯模型，見圖 2-2 所示。

七、星狀表格 Star Table：

星狀網目（Star Schema）所構成的資料表格又稱星狀表格（Star Table），是將事實表格（Fact Table）與星狀網目（Star Schema）的每一個維度表格（Dimension Table）結合，此資料表提供了對多維資料的簡式 SQL 存取方式，可適用於臨時需要的查詢。

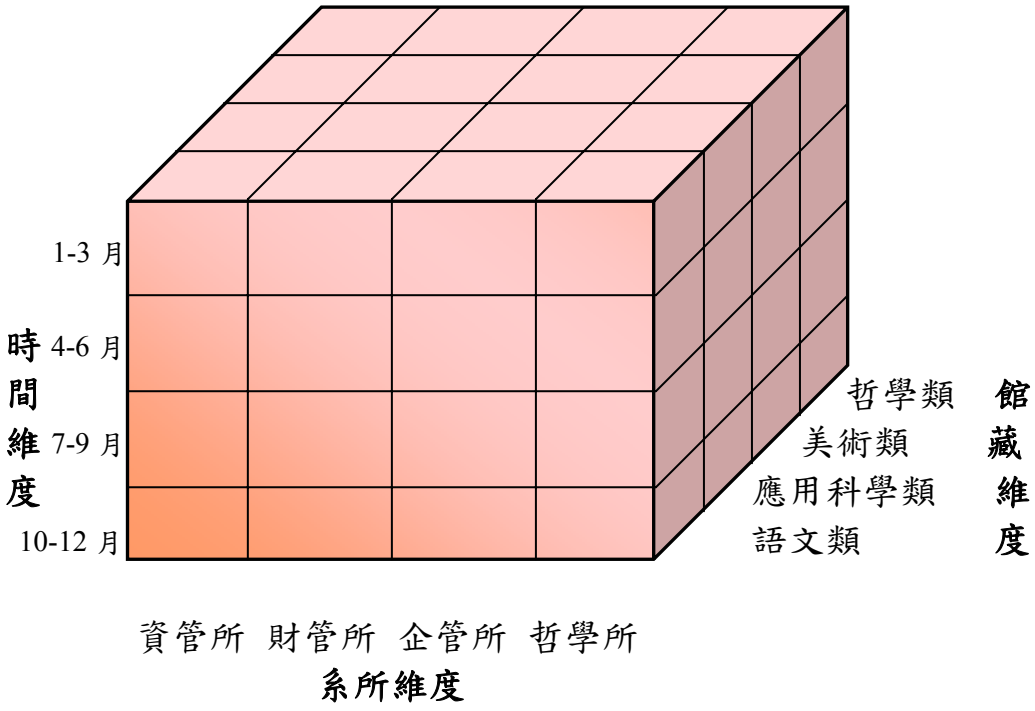


圖 2-1　　資料方塊 多維度資料庫範例圖

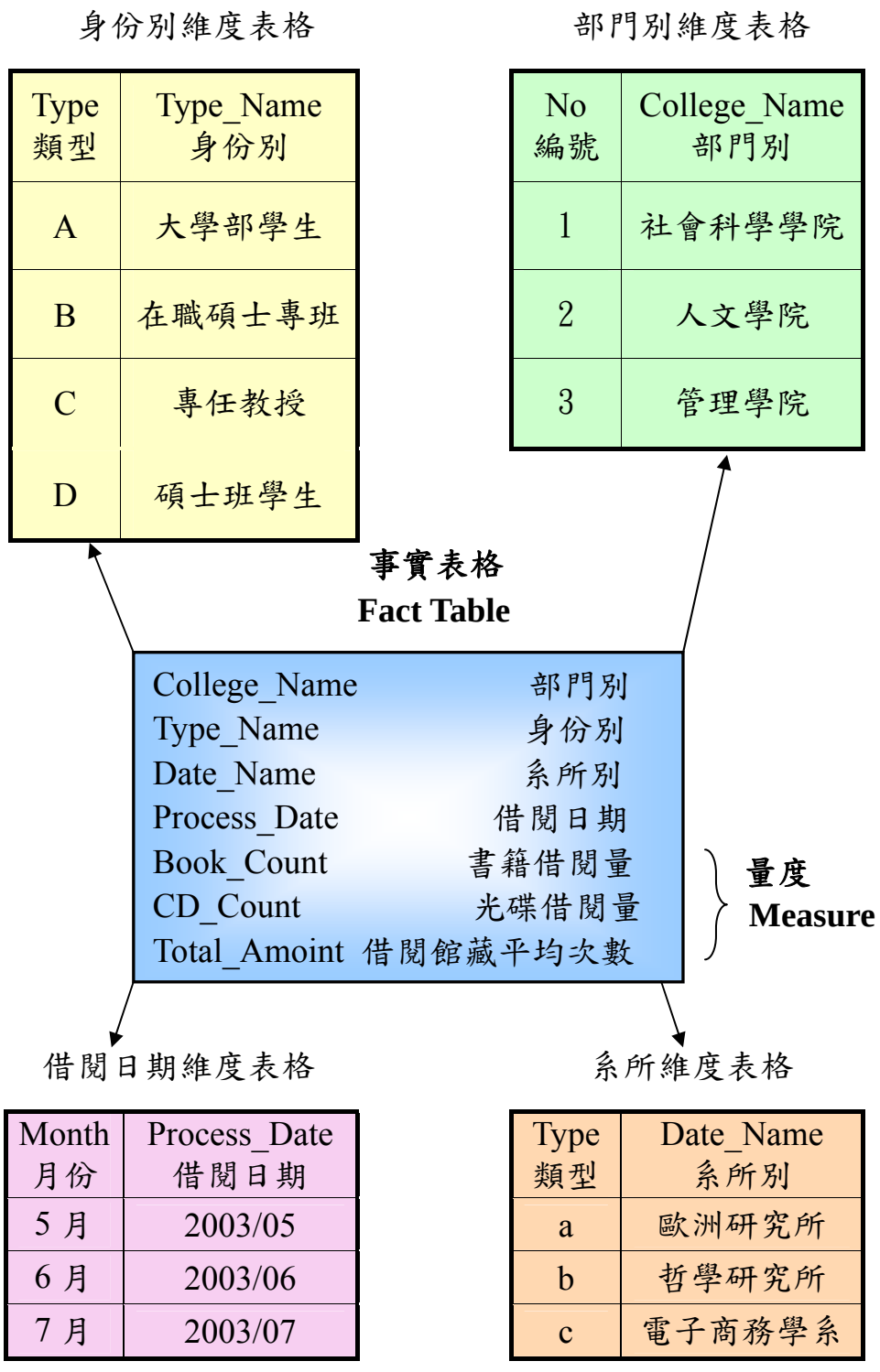

身份別維度表格
Type 類型
Type_Name 身份別
A 大學部學生
B 在職碩士專班
C 專任教授
D 碩士班學生
部門別維度表格
No 編號
College_Name 部門別
1 社會科學學院
2 人文學院
3 管理學院
事實表格
Fact Table
College_Name 部門別
Type_Name 身份別
Date_Name 系所別
Process_Date 借閱日期
Book_Count 書籍借閱量
CD_Count 光碟借閱量
Total_Amoint 借閱館藏平均次數
量度
Measure
借閱日期維度表格
Month 月份
Process_Date 借閱日期
5 月 2003/05
6 月 2003/06
7 月 2003/07
系所維度表格
Type 類型
Date_Name 系所別
a 歐洲研究所
b 哲學研究所
c 電子商務學系

圖 2-2　　星狀網目範例 1

第三章、研究方法

第一節　研究模型

我們根據第一章緒論第一節的研究動機，嘗試提出一個研究模型，以便在數位圖書館之運作上提供下列二大功能。

壹、個人化服務：

一、如何根據相似特性，如身份別相同之讀者的歷史借閱記錄來產生關聯法則，並予以推薦？

二、到新館藏推薦

(一)、主動以 e-mail 通知讀者到館新館藏清單。

(二)、在線上推薦給上線中的讀者。

貳、數位圖書館管理：

一、什麼身份的讀者較常借閱館藏？

二、什麼身分的讀者會較常借閱什麼種類的館藏？

三、借閱館藏的讀者有那些共同點？

四、借閱館藏的讀者有那些身份特徵上的區分？

五、如何擴大讀者群？

我們運用資料探勘的技術，以擷取隱藏的知識，並且建立資料方塊以便利用 OLAP 分析讀者使用數位圖書館資源的最新情況，為達成上述的二個目標。本研究提出如圖 3-1 的研究模型。

我們建立的研究模型有以下兩個步驟：

步驟一、資料探勘的階段包含五個階段。

（一）資料來源（Data Source）：

我們以相似特性，如身份別相同之讀者的歷史借閱記錄來產生數位圖書館個人化服務的關聯法則。因此使用五個主要的資料來源，包含：館藏資料檔、書目資料檔、讀者資料檔、館藏借閱記錄資料檔、時間資料檔等，並且依 MUM 演算法在數位圖書館的應用特性上，依時間別將館藏借閱記錄資料檔分割成原始交易記錄及新交易記錄。

（二）資料選擇（Select Data）：

在上述的五個資料檔內，包含了許多屬性項目，但並非每一個屬性項目是我們所需要的，所以我們在此將會做資料的篩選動作。

（三）資料前處理（Preprocess Data）：

由於上述的檔案資料可能都來自不同的資料庫，其中有些在建檔初始可能格式不正確、資料不正確或遺失等情況。 因此針對這些資料必須經過資料前處理才行。

（四）資料轉換（Trans form Data）：

在這個階段，將會視實際的需要，做資料屬性項目的增加，以便於聚集（Aggregation）的運算，可將資料轉換或合併成適當的形式。

（五）資料儲存（Store Data）：

經過上述的幾個階段之後，資料將會被儲存在資料倉儲（Data Warehouse），資料倉儲對於來自內部及外部的資料進行粹取、轉換、載入、更新等處理程序，以利於後續 OLAP 的運作及資料探勘（Data Mining）。

步驟二、在此步驟將分成兩個部份來進行。

第一部份：

（一）建立資料方塊（Data Cubes）：

為了讓數位圖書館管理人員便於檢視計算處理後的結果，我們以視覺化圖形的

方式來呈現。

（二）OLAP 分析：

數位圖書館管理人員可視作業的需要，以計算、合併及擷取多維關聯式資料、分析如讀者借閱館藏資料，快速以了解讀者使用數位圖書館館藏資料的最新情況。

（三）線上分析的結果：

利用 OLAP 分析後的結果，可讓數位圖書館的決策人員了解讀者使用數位圖書館資源的最新情況，以利數位圖書館管理人員做成相關決策。

（四）使用者介面：

建構一個數位圖書館管理系統，可以讓管理人員得以此系統了解讀者使用數位圖書館資源的情形，並且可在線上隨時調整及改變門檻值，以了解不同等級層次的讀者使用數位圖書館資源的情形。

第二部份：

（一）資料探勘（Data Mining）：

將儲在資料倉儲內龐大的資料進資料探勘的工作，以找出有用的資訊。

（二）擷取隱藏的知識：

在本研究中，我們使用關聯法則（Association Rules）來找出隱藏在龐大資料量內有意義的資訊，我們所使用多層更新探勘法（Multilayer Update Miner → MUM）來達成這個目標。

（三）使用者介面：

建構一個個人化服務系統，以相似特性，如身份別相同之讀者的歷史借閱館藏記錄所產生的關聯法則為基礎，進行館藏推薦的功能，並且主動以 e-mail 通知讀者到館新館藏清單及在線上推薦給上線中的讀者。

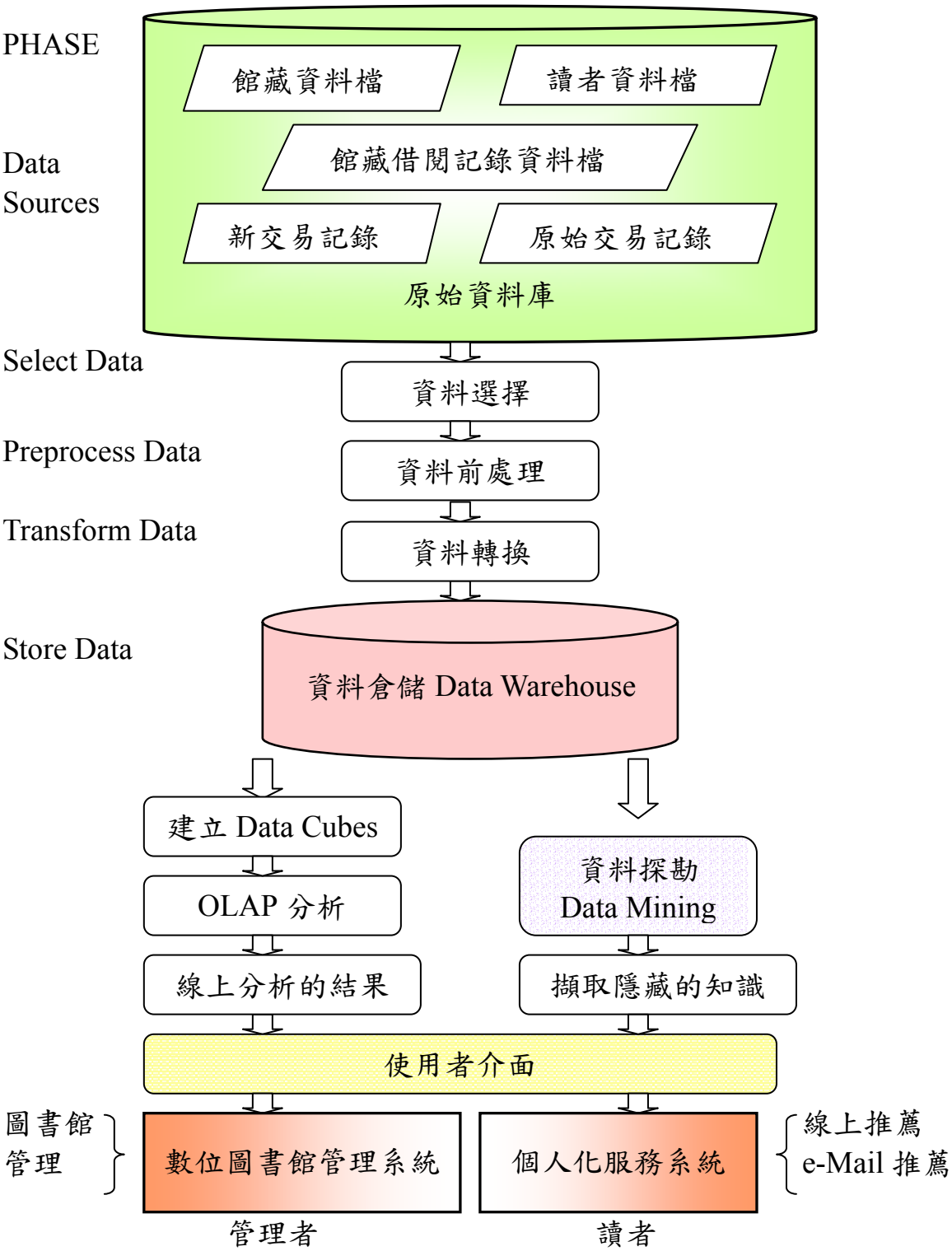

PHASE
Data Sources
館藏資料檔
讀者資料檔
館藏借閱記錄資料檔
新交易記錄
原始交易記錄
原始資料庫
Select Data
資料選擇
Preprocess Data
資料前處理
Transform Data
資料轉換
Store Data
資料倉儲 Data Warehouse
建立 Data Cubes
OLAP 分析
線上分析的結果
資料探勘
Data Mining
擷取隱藏的知識
使用者介面
圖書館管理
數位圖書館管理系統
個人化服務系統
線上推薦
e-Mail 推薦
管理者
讀者

圖 3-1 研究模型圖

第二節　個人化服務

本文以本校南華大學圖書館歷史借閱記錄做為找出讀者借閱關聯法則的資料來源，以此來達到個人化服務的目的。由於圖書館自開館以來，原始資料庫資料高達 554,877 筆，即便以去年 2002 年也有 226,863 筆之多，因為要增加 Data Mining 的效率，因此必須有一個能滿足以下三點的演算法，才能有效率的找出關聯法則：

一、不需重複搜尋原始資料庫。

二、滿足漸進式探勘(Incremental Mining) **【8】【9】【15】**的需求。

三、可支援線上採勘(On Line Mining)的需求。

因此，我們使用多層更新探勘法 (Multilayer Update Miner→MUM) **【10】**，利用其可滿足上列三點需求的特性，以達成個人化服務的功能需求。

壹、MUM 演算法

我們在資料庫或主記憶體（僅可處理較小的資料量）中建立以下的兩個表格分別是：

一、主資料表 Base Table

是一個多層陣列表格（Multilayer Array Table），負責記錄每一筆新增資料所動態產生的相對應的項目集（Itemset），並記錄

這些項目集（Itemset），出現次數的計數值（Count）、支持度（Support）、可信度（Confidence）及所有交易總數等資訊，將符合最小支持度（Minimum Support）的項目集（Itemset）留在本表格中，成為大項目集（Large Itemset）。見表 3-2　MUM 演算法主資料表（Base Table）。

二、備用資料表（Temp Table）

負責收留在主資料表（Bust Table）中，不符合最小支持度（Minimum Support）的項目集（Itemsets）。此表格最大的貢獻是：

（一）可減少主資料表（Base Table）的資料量（Data Size），縮短搜尋原始資料庫的時間。

（二）將每一筆交易記錄項目集（Itemset）拆解成各個項目（Item）時，在第一次被判定不符合最小支持度（Minimum Support）而被歸類到備用資料表（Temp Table）的項目（Item），它們在交易記錄中出現的次數之計數值（Count）仍舊會被繼續累計，而不玫被遺忘。當持續累積到符合最小支持度（Minimum Support）時，便可回到主資料表

（Base Table）中，成為大項目集（Large Itemset）。

（三）扮演輔助主資料表（Base Table）的角色，可避免刪除一些在交易記錄中，交易出現次數緩步成長的項目（Item）。見表 3-6　MUM 演算法備用資料表（Temp Table）。

貳、MUM 演算法處理作法如下：

MUM 演算法將原始資料庫內的交易記錄逐筆分解成 1-Itemset、2-Itemset…、n-Itemset 的各個項目集（Itemset），並儲存至對應的資料表中。其演算法處理作法如下。

一、讀入一筆借閱記錄。

二、動態將借閱記錄分解成 1-Itemset，2-Itemset，、、、n-Itemset。

三、建立主資料表格及備用資料表格。

四、掃描對應的資料表中是否已有該項目組存在。

五、已存在：

將對應項目組 count 數加 1。

未存在：

新增項目組，count 數設為 1。

六、計算此項目組的 support 是否達到目前的 Minimum Support？

若是：

則留在至主資料表中。

若否：

則留在備用資料表中。

參、MUM 演算法範例說明

首先我們假設圖書館原始資料庫借閱記錄如表 3-1 所示：

表 3–1 圖書館原始資料庫借閱記錄

借閱記錄 TID	項目集 （Itemsets）
1	鹿鼎記、一手車訊、Here、哈利波特 VCD
2	Here、Java 入門、Java Script
3	C++、Here、一手車訊
4	Java 入門、活用 Windows XP、Mook
5	活用 Windows XP、Here、天龍八部、Java 入門
6	Here、哈利波特 VCD、Java Script、Java 入門
7	Java 入門、Java Script、鹿鼎記
8	Java 入門、C++、活用 Windows XP、Here
9	Mook、一手車訊
10	類神經網路、西洋美術史、哈利波特 VCD

假設最小支持度（Minimum Support）設定分別為 1-Itemset 等於 20% ，2-Itemset 等於 15% ，3-Itemset 等於 10% 。

表 3-2　MUM 演算法主資料表 Base Table

1-Itemset	Count
一手車訊	3
Here	5
哈利波特 VCD	2
Java 入門	6
Java Script	3
活用 Windows XP	3

1-Itemset Table

2-Itemset	Count
一手車訊、Here	2
哈利波特 VCD、Here	2
Java 入門、Java Script	3
Here、Java Script	3
Here、Java 入門	3
Java 入門、活用 Windows XP	2

2-Itemset Table

3-Itemset	Count
Java 入門、Here、Java Script	2
Here、Java 入門 、活用 Windows XP	2

3-Itemset Table

表 3-3　1-Itemset 對應的資料表 Minimum Support＝20%

借閱記錄 TID		Java Script	天龍八部	鹿鼎記	一手車訊	Here	Java 入門	C++	活用 Windows XP	Mook	哈利波特 VCD	類神經網路	西洋美術史
1	鹿鼎記、一手車訊 Here、哈利波特 VCD	/	/	1/100	1/100	1/100	/	/	/	/	1/100	/	/
2	Here、Java 入門 JavaScript	1/50	/	1/50	1/50	1/50	1/50	/	/	/	1/50	/	/
3	C++、Here、一手車訊	1/33.3	/	1/33.3	2/66.6	2/66.6	1/33.3	1/33.3	/	/	1/33.3	/	/
4	Java 入門、Mook 活用 Windows XP	1/25	/	1/25	2/50	2/50	2/50	1/25	1/25	1/25	1/25	/	/
5	活用 Windows XP、Here 天龍八部、Java 入門	1/20	1/20	1/20	2/40	3/60	3/60	1/20	2/40	1/20	1/20	/	/
6	Java Script、Java 入門、 哈利波特 VCD、Here	2/33.3	1/16	1/16	2/33.3	4/66.6	4/66.6	1/16	2/33.3	1/16	2/33.3	/	/
7	Java 入門、鹿鼎記 Java Script	3/42.8			2/28.5	4/57.1	5/71.4		2/28.5		2/28.5	/	/
8	Java 入門、C++、Here 活用 Windows XP	3/37.5			2/25	5/62.5	6/75		3/37.5		2/25	/	/
9	Mook、一手車訊	3/33.3			3/33.3	5/55.5	6/66.6		3/33.3		2/22.2	/	/
10	類神經網路、西洋美術史、哈利波特 VCD	3/30			3/30	5/50	6/60		3/30		2/20	1/10	1/10

表 3-4　2-Itemset 對應的資料表 Minimum Support＝15%

借閱記錄 TID		一手車訊、Here	哈利波特 VCD Here	C++、Here	Java 入門 JavaScript	Here JavaScript	Here 活用 Windows XP	Here Java 入門	Java 入門、活用 Windows XP
1	鹿鼎記、一手車訊 Here、哈利波特 VCD	1/100	1/100	/	/	/	/	/	/
2	Here、Java 入門 JavaScript	1/50	1/50	/	1/50	1/50	/	1/50	/
3	C++、Here、一手車訊	2/66.6	1/33.3	1/33.3	1/33.3	1/33.3	/	1/33.3	/
4	Java 入門、Mook 活用 Windows XP	2/50	1/25	1/25	1/25	1/25	/	1/25	1/25
5	活用 Windows XP、Here 天龍八部、Java 入門	2/40	1/20	1/20	1/20	1/20	1/20	2/40	1/40
6	Java Script、Java 入門、哈利波特 VCD、Here	2/33.3	2/33.3	1/16	2/33.3	2/33.3	1/16	3/50	2/33.3
7	Java 入門、鹿鼎記 Java Script	2/28.5	2/28.5	1/14.3	3/42.8	2/28.5	1/14.3	3/42.8	2/28.5
8	Java 入門、C++、Here 活用 Windows XP	2/25	2/25		3/37.5	3/37.5		3/37.5	2/25
9	Mook、一手車訊	2/22.2	2/22.2		3/33.3	3/33.3		3/33.3	2/22.2
10	類神經網路、西洋美術史、哈利波特 VCD	2/20	2/20		3/30	3/30		3/30	2/20

表 3-5　3-Itemset 對應的資料表 Minimum Support＝10%

借閱記錄 TID		Java 入門、Here JavaScript	活用 Windows XP Here、Java 入門
1	鹿鼎記、一手車訊 Here、哈利波特 VCD	/	/
2	Here、Java 入門、JavaScript	1/50	/
3	C++、Here、一手車訊	1/33.3	/
4	Java 入門、Mook、活用 Windows XP	1/25	/
5	活用 Windows XP、Here、天龍八部、Java 入門	1/20	1/20
6	Java Script、Java 入門、哈利波特 VCD、Here	2/33.3	1/16
7	Java 入門、鹿鼎記、Java Script	2/28.5	1/14.2
8	Java 入門、C++、Here、活用 Windows XP	2/25	2/25
9	Mook、一手車訊	2/22.2	2/22.2
10	類神經網路、西洋美術史、哈利波特 VCD	2/20	2/20

表 3-6 MUM 演算法備用資料表 Temp Table

		2-Itemset	Count
1-Itemset	Count	C++、Here	2
天龍八部	1	Here、活用 Windows XP	2
鹿鼎記	2		
C++	2		
Mook	2		
類神經網路	1		
西洋美術史	1		

在表 3-2 1-Itemset 主資料表（Base Table）資訊在處理過程的變化，左方的部份為圖書館原始資料庫借閱記錄，上方為 1-Itemset 的項目，而表格中的資訊則是其出現次數及支持度。在表 3-3 1-Itemset 對應的資料表所設定的 Minimum Support＝20% ，可以看出其中如天龍八部在這總數十筆的借閱交易記錄中，只出現一次借閱記錄。其 Support 值只有 16% ，因此被移入備用資料表（Tem table），並在 Count 值記錄為一筆，其餘的館藏也依此類推，詳見表 3-6 MUM 演算法備用資料表。

而 2-Itemset 對應的資料表所設定的 Minimum Support＝15% 的處理結果詳見表 3-4 所示，及 3-Itemset 對應的資料表所設定的 Minimum Support＝10% 的處理結果詳見表 3-5 所示。

由本節 MUM 演算法的範例說明可以得知 MUM 演算法，具有以下三個特性：

一、當 Minimum Support 值變動時，不需重複搜尋原始資料庫。

二、採用多層更新（Multilayer Update）的方法。即建置主資料表（Base Table）和備用資料表（Temp Table）與對應的資料表，可加快搜尋原始資料庫的速度，滿足漸進式探勘（Incremental Mining）的需求。

三、由於具有快速處理新增資料的能力，因此可支援線上探勘（On Line Mining）的需求。

第三節　數位圖書館管理

在此，我們以第一章緒論第一節「針對圖書館管理方面」的研究動機內的問題，以資料方塊的方式來呈現其解決方式。

壹、『什麼身份的讀者較常借閱館藏？』

首先要知道「較常」的數據其定義為何？我們的步驟是：

一、先統計出各種身份別讀者某月的借閱館藏次數總和。

二、統計實際借閱館藏人數（即扣除借閱館藏次數內之重複個數）。

三、計算出各種身份別讀者借閱館藏平均次數。

四、比對出超過此借閱館藏平均次數的各種身份別讀者。

貳、『什麼身分的讀者會借閱什麼種類的館藏？』

如碩士班學生的讀者會借閱語文類的館藏。

經由預作統計的處理而得到表 3-7 事實表格，接著並產生圖 3-2 星狀網目範例 2 及藉由 OLAP 分析可以快速得到上述二個問題的最新資訊如圖 3-3 的資料方塊圖 1。

表 3-7 事實表格範例 1

Type_Name 身份別	Book_Type 館藏種類	Process_Date 借閱日期	B_Count 借閱館藏次數	P_Count 借閱館藏人數	B_Avg 借閱館藏平均次數
1	A	05	518	210	2.47
		06	427	180	2.37
		07	482	192	2.51
	B	05	223	55	4.05
		06	152	61	2.49
		07	245	101	2.42
	C	05	784	218	3.59
		06	812	323	2.51
		07	756	456	1.65
	D	05	366	118	3.10
		06	386	92	4.19
		07	410	182	2.25
2	A	05	120	53	2.26
		06	142	61	2.32
		07	95	36	2.63
	B	05	26	12	2.16
		06	38	9	4.22
		07	21	11	1.90
	C	05	181	61	2.96
		06	201	83	2.42
		07	153	45	3.40
	D	05	138	39	3.53
		06	120	58	2.06
		07	72	27	2.66

表 3–7 事實表格範例 1 （續）

3	A	05	220	20	11.00
		06	172	17	10.11
		07	162	29	5.58
	B	05	68	25	2.72
		06	31	12	2.58
		07	76	31	2.45
	C	05	280	55	5.09
		06	198	51	3.88
		07	265	66	4.01
	D	05	168	32	5.25
		06	145	28	5.17
		07	175	28	6.25
4	A	05	855	301	2.84
		06	910	311	2.92
		07	720	295	3.51
	B	05	430	127	3.55
		06	411	105	3.91
		07	313	108	2.89
	C	05	978	313	2.62
		06	918	386	2.37
		07	822	291	2.82
	D	05	588	210	2.8
		06	663	227	2.92
		07	346	102	3.39

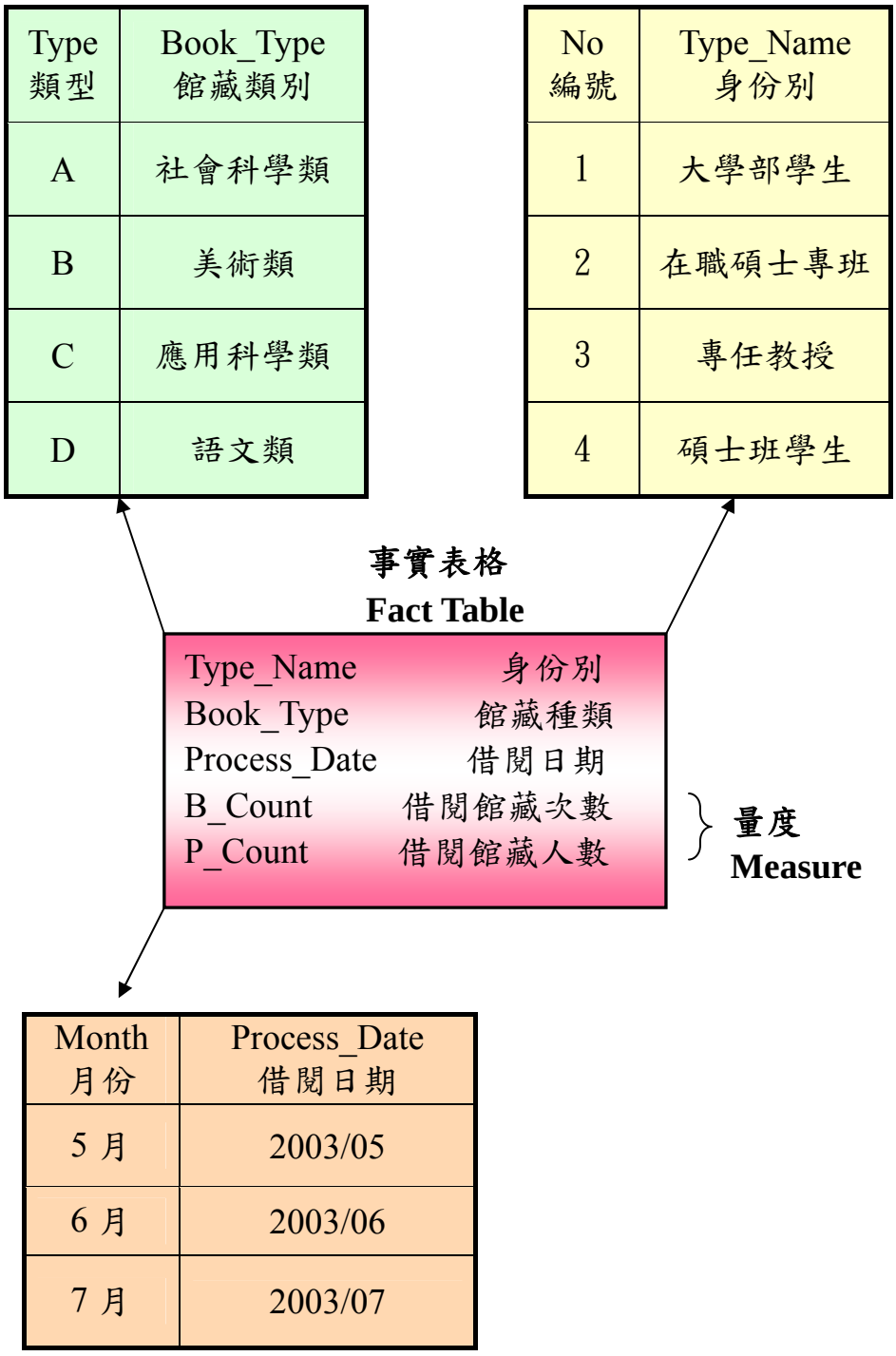

Type
類型
Book_Type
館藏類別
A
社會科學類
B
美術類
C
應用科學類
D
語文類
No
編號
Type_Name
身份別
1
大學部學生
2
在職碩士專班
3
專任教授
4
碩士班學生
事實表格
Fact Table
Type_Name 身份別
Book_Type 館藏種類
Process_Date 借閱日期
B_Count 借閱館藏次數
P_Count 借閱館藏人數
量度
Measure
Month
月份
Process_Date
借閱日期
5 月
2003/05
6 月
2003/06
7 月
2003/07

圖 3-2 星狀網目範例 2

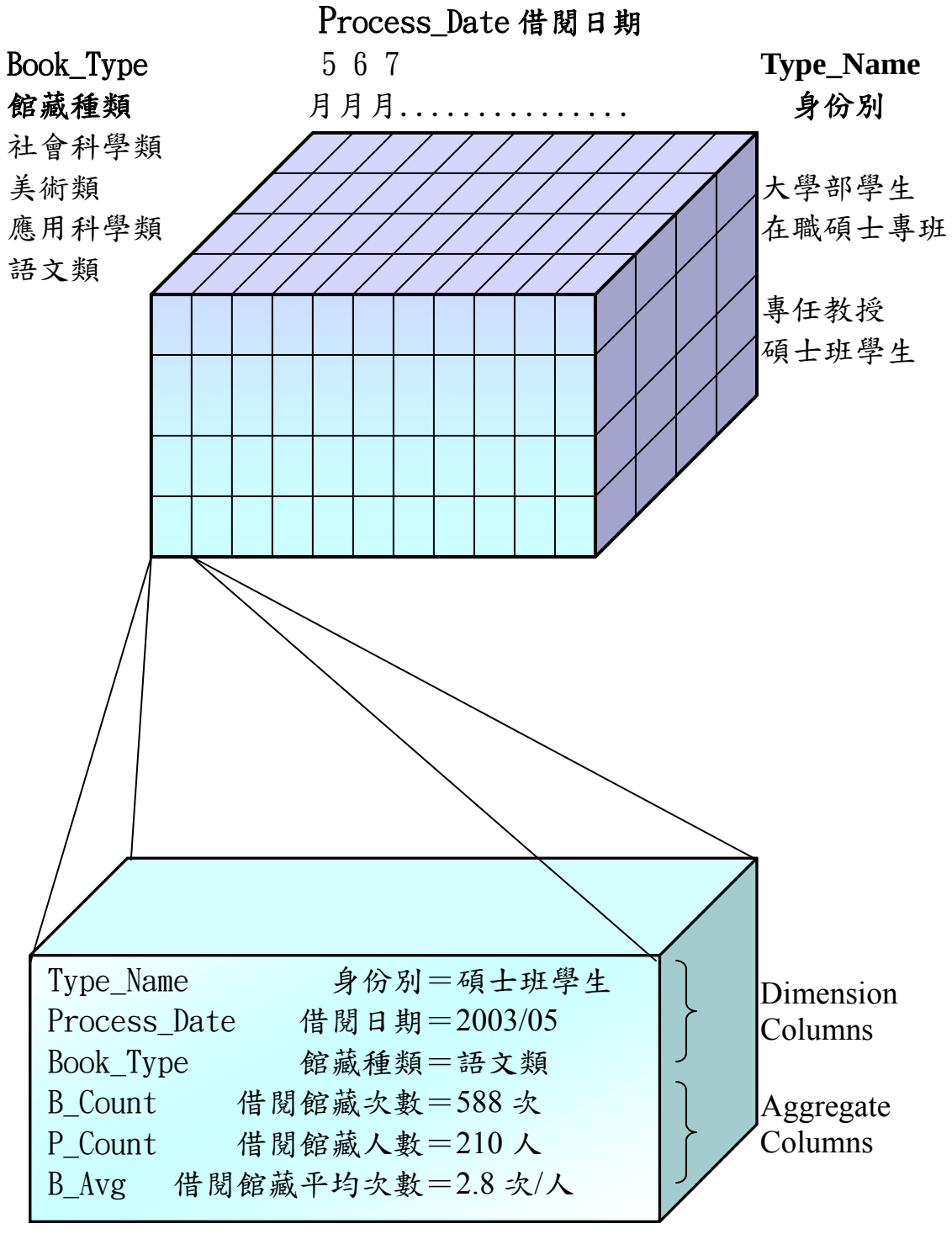

Process_Date 借閱日期
Book_Type
館藏種類
社會科學類
美術類
應用科學類
語文類
5 6 7
月月月.....
Type_Name
身份別
大學部學生
在職碩士專班
專任教授
碩士班學生
Type_Name 身份別＝碩士班學生
Process_Date 借閱日期＝2003/05
Book_Type 館藏種類＝語文類
B_Count 借閱館藏次數＝588 次
P_Count 借閱館藏人數＝210 人
B_Avg 借閱館藏平均次數＝2.8 次/人
Dimension Columns
Aggregate Columns

圖 3-3 資料方塊圖 1

參、『借閱館藏的讀者有那些共同點？』

肆、『借閱館藏的讀者有那些身份特徵上的區分？』

經由預作統計的處理而得到表 3-8 事實表格 2，接著並產生圖 3-4 星狀綱目範例 3 及藉由 OLAP 分析可以快速得到上述二個問題的最新資訊如圖 3-5 的資料方塊圖 2。

表 3-8 事實表格範例 2

College_Name 部門別	Gender 性別	Process_Date 借閱日期	Day_Count 借閱館藏天數	P_Count 借閱館藏人數	Day_Avg 借閱館藏平均天數
1	M	05	12,830	528	24.3
		06	7.833	361	21.7
		07	11,039	415	26.6
	F	05	16,012	556	28.8
		06	16,936	582	29.1
		07	16,104	610	26.4
2	M	05	7,238	312	23.2
		06	8,925	423	21.1
		07	5,278	207	25.5
	F	05	9,537	422	22.6
		06	11,030	383	28.8
		07	15,511	562	27.6
3	M	05	20,576	712	28.9
		06	14,522	685	21.2
		07	16,201	587	27.6
	F	05	20,253	696	29.1
		06	18,418	728	25.3
		07	16,589	619	26.8

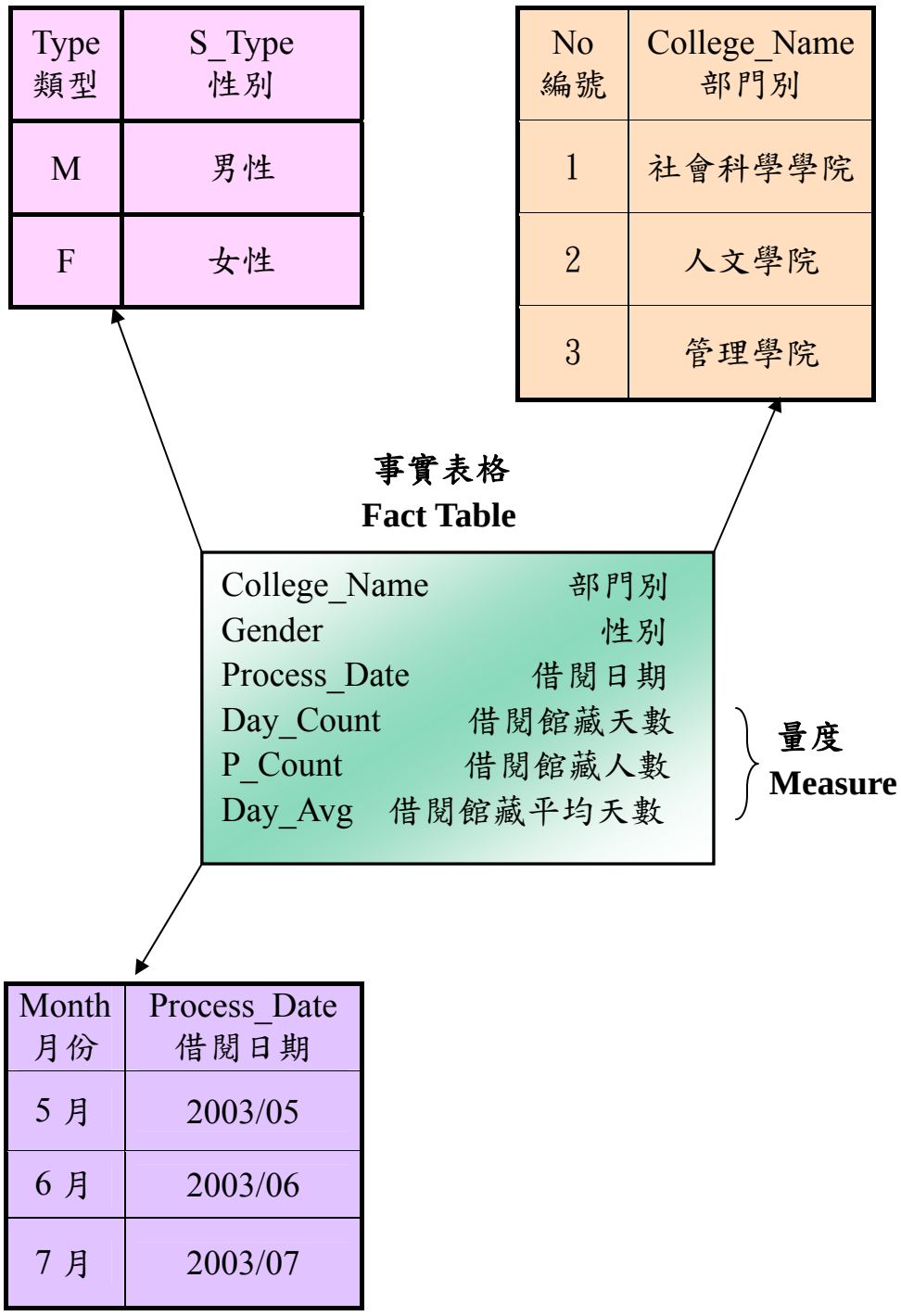

Type 類型
S_Type 性別
M
男性
F
女性
No 編號
College_Name 部門別
1
社會科學學院
2
人文學院
3
管理學院
事實表格
Fact Table
College_Name 部門別
Gender 性別
Process_Date 借閱日期
Day_Count 借閱館藏天數
P_Count 借閱館藏人數
Day_Avg 借閱館藏平均天數
量度
Measure
Month 月份
Process_Date 借閱日期
5 月
2003/05
6 月
2003/06
7 月
2003/07

圖 3-4 星狀綱目範例 3

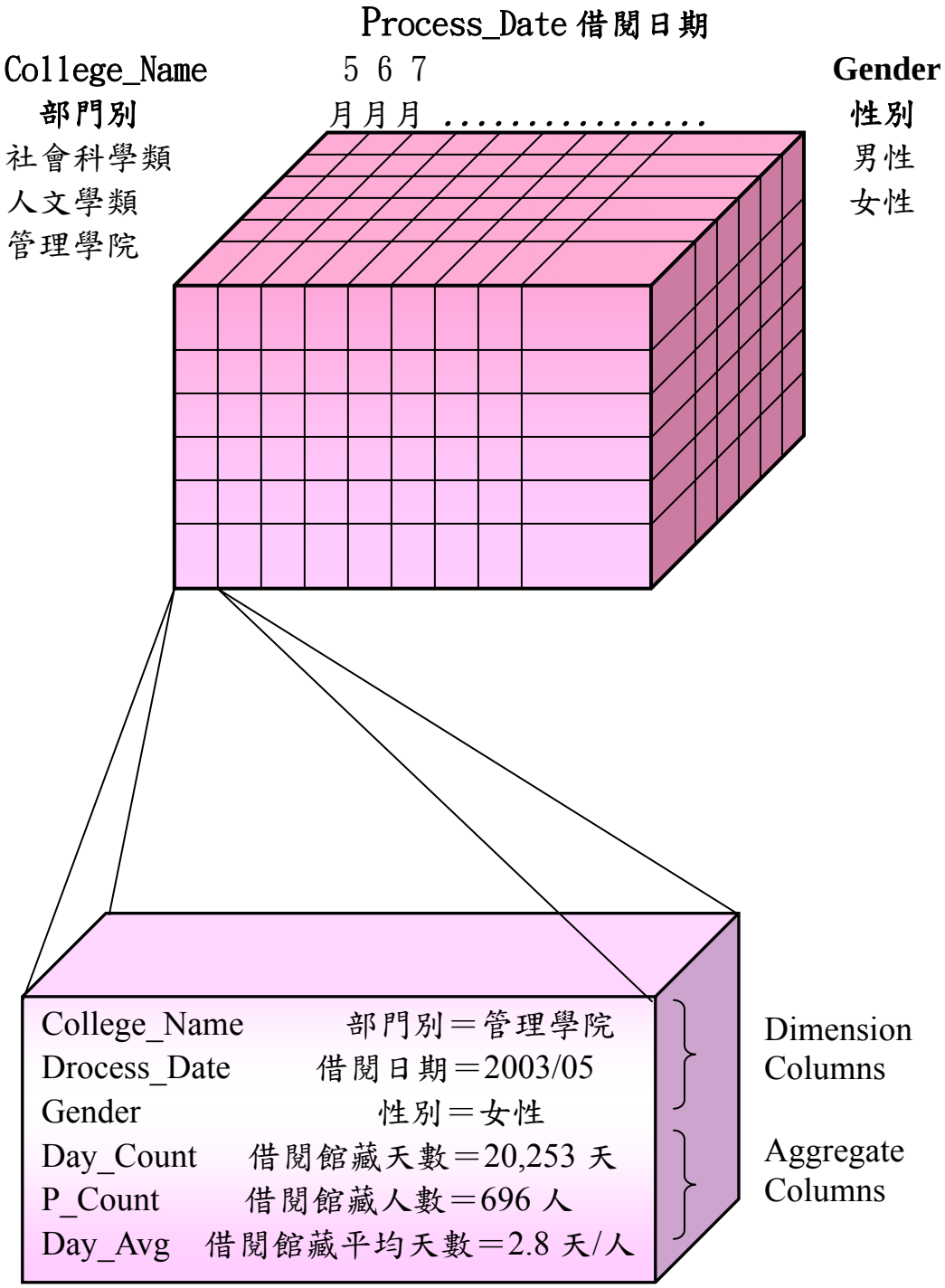

Process_Date 借閱日期
College_Name
部門別
社會科學類
人文學類
管理學院
5 6 7
月月月 .....
Gender
性別
男性
女性
College_Name 部門別＝管理學院
Drocess_Date 借閱日期＝2003/05
Gender 性別＝女性
Day_Count 借閱館藏天數＝20,253 天
P_Count 借閱館藏人數＝696 人
Day_Avg 借閱館藏平均天數＝2.8 天/人
Dimension Columns
Aggregate Columns

圖 3－5 資料方塊圖 2

第四章、實驗結果

本章節將針對 Apriori 演算法及 MUM 演算法在執行效能上的差異，並進一步分析 MUM 演算法的優缺點及探討是否有可能改善的空間及方式，以做未來更進一步的研究。

第一節　實驗環境

壹、硬體設備：

本實驗所使用的硬體設備為一部個人電腦，主要規格如下：

CPU：AMD Athlom 1G

RAM：512MB

HD：1BM 40G

貳、軟體設備：

Windows 2000 Profexional with SP2

SQL 7.0 with SP2

Virtual Basic 6.0

參、資料來源：

以 VB 6.0 程式產生亂數，寫入資料庫，做為本實驗的資料來源，主要目的在於驗証 Apriori 演算法及 MUM 演算法在資料探勘速度上

的比較，來做有分析結果的依據。

產生實驗資料時需設定三個參數，分別是：

一、交易資料的筆數。

二、每筆資料最多的項目數。

三、每個項目變化的範圍。

第二節　實驗結果

分為兩個部份：

壹、漸進式探勘（Incremental Mining）實驗：

一、假設圖書館原始資料庫有 10,000 筆資料。

二、每次動態新增 100 筆記錄，並將 Minimum Support 值設為 100。

表 4-1 漸進式探勘實驗數據比較表（單位：秒）

	10,000 筆	10.100 筆	10,200 筆	10,300 筆
Apriori 演算法	977	1,005	1,036	1,076
MUM 演算法	9,463	132	134	132
MUM／Apriori 多耗時間倍數	9.68 倍	4.84 倍	3.22 倍	2.4 倍

資料來源：【11】及本研究整理

由此實驗結果可以看出，在剛開始第一次搜尋原始資料庫時，Apriori 演算法只花了極少的時間完成搜尋，而 MUM 演算法比 Apriori 演算法多花了 9.68 倍的搜尋時間，這是因為 MUM 演算法的特性所致。MUM 演算法必須拆解圖書館原始資料庫內每一筆交易記錄的項目（Item），以製成對應資料表（Reference Table）所致。

但從表上也可看出，當每一次新增一定量筆數（Minimum Support＝100）時，Apriori 演算法仍須再重新搜尋原始資料，以致所花時間愈來愈多，而 MUM 演算法卻只用了 Apriori 演算法 13.13% （10,100 筆）秒；12.93% （10,200 筆）秒；及 12.26% （10,300 筆）秒。

並可從表 4-1 可推估，隨著原始資料庫交易時間日積月累的資料量，MUM 演算法的運算效率將被突顯出來。

貳、線上探勘（On Line Mining）實驗：

一、假設原始資料庫有 10,000 筆資料。

二、Minimum Support 值分別為 150、125、100 及 75。

表 4-2 線上探勘時實驗數據比較表（單位：秒）

	Min-Sup＝150	Min-Sup＝125	Min-Sup＝100	Min-Sup＝75
Apriori 演算法	617	637	977	2,159
MUM 演算法	2	2	2	2

資料來源：【11】

從表 4-2 可以看出，由於 Minimum Support 值不斷改變，Apriori 便必須隨之不斷的重複搜尋原始資料庫，更發現當 Minimum support 愈小時，所須花費的時間將會更為驚人，而 MUM 演算法由於已在第一次搜尋圖書館原始資料庫時，便已建置了對應的資料表，當 Minimum Support 值改變時，只須搜尋對應的資料表，並將結果分別歸類到主資料表（Base Table）或備用資料表（Temp Table）即可，而不必再重新探勘整個圖書館原始資料庫，因此資料挖掘（Data Mining）的時間不會隨著 Minimum Support 值的改變而有所變化。

第三節　MUM 演算法的缺點及改進方式

壹、如果將 Minimum Support 值設定成非常的小，如 0.5、0.01 等，則所產生的非大項目集（Non-Large Itemset）數目將會非常的多，將會產生一張資料量相當大的備用資料表（Temp Table），而佔用大量的儲存空間。

解決之道：

步驟一：

如果 Minimum Support 值設定為 0.05，按照原有做法，符合 Minimum Support 值會被歸類到備用資料表（Temp Table），此時再設定一個比原有的值更低的 Minimum Support 值如 0.025。

也就是說：將備用資料表內的項目集（Itemset）再做一次篩選，符合 Minimum Support 值 0.025 的大項目集（Large Itemset）當做候選集（Candidate Itemset）。

步驟二：

將備用資料表（Temp Table）的候選集（Candidate Itemset）與原有主資料（Base Table）的大項目集（Large Itemset）做比對動作。將二者重複的項目集予以刪除，如

此便有較高的正確性。

步驟三：

將已經篩選的候選集（Candidate Itemset）移入主資料表（Base Table）後，即告完成。

貳、如果拆解出 1-Itemset，2-Itemset、、、n-Itemset，而 n-Itemset 過多時，也會產生一張資料量相當大的對應的資料表，而佔用大量的儲存空間。

解決之道：

首先要談的是圖書館資料庫特性的問題。我們發現，在實際的館藏借閱交易記錄，一次借閱一本，一次借閱三本，一次借閱五本的情形佔大多數。而一次只借閱一本的交易記錄在一開始時，資料探勘出關聯法則，便必須將此類交易記錄予以刪除。因此已大幅降低了圖書館資料庫的交易記錄。

而一次借閱七本以上館藏的情形較為少見。原因是圖書館的借閱規則規定一本館藏原則上一次只能借閱一個月，若屬熱門館藏則只有兩週的借閱期。若一次借閱七本以上，同樣總共只有一個月的閱讀期，並不符合讀者的利益。

因此，一次借閱三本、五本館藏的借閱記錄會佔剩餘資料量的

大宗，因此所產生的 n-Itemset 的對應的資料表便不致到無法控制的地步。

參、若圖書館原始資料庫取樣不足（如只取兩萬筆），或日後新增資料大幅增加（如一次增加 5,000 筆），則也會產生過多的非大項目集（Non- Large Itemset），以致備用資料表（Temp Table）資料量大增。

解決之道：

步驟一：

以南華大學圖書館為例。目前累計館藏借閱記錄為 554,877 筆左右，若扣除一次只借閱一本館藏的借閱記錄，假設仍有二十三萬筆，以民國九十年十二月三十一日做為截算日，館藏借閱記錄日期在截算日之前的資料稱為圖書館原始資料庫記錄，在截算日之後的記錄稱為新增記錄。如此一來，圖書館原始資料庫的記錄量便與新增記錄量呈現極大的差值。

步驟二：

可自行機動設定截算日。如一年一次，或間隔半年、三個月一次皆可，目的都為了減少新增的記錄量。

第五章、結論及未來研究工作

壹、針對個人化服務方面：

一、如何根據相似特性，如身份別相同之讀者的歷史借閱館藏記錄來產生關聯法則，並予以推薦？

我們以多層更新探勘法（Multilayer Update Miner→MUM）來找出借閱館藏記錄中的關聯法則，並以此結果加以推薦。

二、到館新館藏推薦

（一）主動以 e-mail 通知讀者到館新館藏清單。

通常學校建置了全校的 e-mail 帳號，因此圖書館管理人員可將到館新館藏分類後，根據讀者以往的館藏借閱記錄所探勘出的關聯法則，將到館新館藏列出清單後 e-mail 給讀者

（二）在線上推薦給上線中的讀者。

由於讀者可能尚未閱讀由數位圖書館系統所寄發的 e-mail，因此在讀者登入數位圖書館個人化服務系統之後，可將依上述方法所得到的新館藏推薦清單顯示在讀者所瀏覽的網頁上供讀者選擇。

貳、針對圖書館管理方面：

一、什麼身份的讀者較常借閱館藏？

經由事先預作統計計算處理後可得知，在 5 月、6 月、7 月這三個月內借閱館藏平均次數（B_Avg）達到 3.41 次/月身份別的讀者將被定義為『較常』借閱館藏。因此將發現一個現象，就是透過統計的結果,四種身份別的讀者在不同的借閱日期都符合了『較常』借閱館藏的定義，然而透過 OLAP 的檢視可清楚的看出，在大學部學生、在職碩士專班、專任教授及碩士班學生這四種身份別中，專任教授除了借閱館藏次數大多符合『較常』借閱館藏的定義之外，在不同的借閱日期中也維持著較高的借閱率，由此可看出，透過圖形視覺化的檢視，OLAP 更能比傳統的方式更能有效掌握正確的資訊。因此本問題的答案理應選擇專任教授作為較常借閱館藏的讀者比較恰當。

二、什麼身份的讀者會較常借閱什麼種類的館藏？

透過 OLAP 的檢視可看出：

（一）大學部學生較常借閱美術類、應用科學類及語文類的館藏。

（二）在職碩士班學生較常借閱美術類及應用科學類的館藏。

（三）專任教授較常借閱社會科學類、應用科學類及語文類的館藏。

（四）碩士班學生較常借閱社會科學類及美術類的館藏。

三、借閱館藏的讀者有那些共同點？

經由預作統計的處理可得知，借閱館藏平均天數（Day_Avg）為 25.81 天/人。因此透過 OLAP 的檢視可看出：

（一）社會科學學院、人文學院、管理學院三者的共同點是三個學院的女性讀者在 7 月份都有較高的借閱館藏平均天數。

（二）社會科學學院、人文學院、管理學院三者的共同點是三個學院的男性讀者在 6 月份借閱館藏平均天數都沒有達到 25.81 天/人的平均值。

（三）管理學院的男性及女性讀者的共同點是，5 月份及 7 月份的借閱館藏平均天數都高於 25.81 天/人的平均值。

四、借閱館藏的讀者有那些身份特徵上的區分？

透過 OLAP 的檢視可看出：

（一）社會科學學院及人文學院的男性讀者 5 月份及 6 月份借閱館藏平均天數較女性讀者為少。

（二）5 月份借閱館藏平均天數，管理學院男性讀者高於 25.81

天/人的平均值，而社會科學學院及人文學院的男性讀者則皆低於平均值。

五、如何擴大讀者群？

依關聯法則及 OLAP 觀點分析出的借閱館藏族群特徵（依身份別、館藏別、時間別）所分析過的資料；主動透過學校全體人員的 e-mail 信箱，主動推薦令人可能感興趣的館藏。

我們以多層更新探勘法（Multilayer Update Miner ⟶ MUM ）來處理關聯法則的問題，並且指出演算法上的缺失及改善方法，相信對提升此演算法在數位圖書館方面的應用上有相當的幫助。

此外，我們提出一個數位圖書館個人化服務及管理的研究模型，在此研究模型下，可發展使用者介面便於進一步完成：

壹、個人化服務系統：

根據身份別相同讀者的歷史借閱記錄所產生的關聯法則在線上推薦，並主動以 e-mail 通知讀者有新館藏到館，以增加數位圖書館館藏借閱的使用率。

貳、數位圖書館管理系統：

透過 OLAP 的分析，可即時在線上得知讀者借閱館藏後所產生

的各項資訊，針對這些資訊加以分析，擷取其中有效益的訊息，以增進數位圖書館資訊管理的效率，進一步提升服務品質。

未來研究工作將可朝個人化服務系統及數位圖書館管理系統的實作邁進。

文獻參考

中文文獻參考：

【1】吳凱雯，利用資料挖掘技術提供網際網路使用者個人化服務， 台中縣，靜宜大學資訊管理學系，民國 90 年。

【2】方佳琳，花招乎?口碑乎?客服仍需努力，台北市，漢藝整合公關顧問公司，及時公關電子報第 18 期，民國 92 年。

【3】綾野克俊，回歸 TQC 的本質:掌握顧客滿意”， 品質管理，43 (May)，民國 81 年。

【4】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，新竹市，國立交通大學資訊科學研究所，民國 90 年。

【5】陳建銘，類神經網路於 Web Mining 之應用，台北市，國立台北科技大學商業自動化與管理研究所，民國 90 年。

【6】戴玉旻，圖書館借閱記錄探勘系統，新竹市， 國立交通大學資訊科學研究所，民國 90 年。

【7】孫冠華，圖書館新書推薦之個人化服務方法，高雄市，國立中山大學資訊管理研究所，民國 89 年。

【8】賴永祥，中國圖書分類法，文華出版社，台北市，民國 90 年。

【9】留乃俊，大型資料庫中高效率之漸進式關聯規則探勘方法，台南市，國立成功大學資訊工程研究所碩士論文，民國 90 年。

【10】王慶堯，利用準大項目集之漸進式挖掘，高雄縣，義守大學資訊工程研究所碩士論文，民國 89 年。

【11】毛立仁、楊昇樺，關聯法則之多層更新挖掘法及其應用，嘉義縣，南華大學資訊管理研究所碩士論文，民國 90 年。

英文文獻參考：

【1】Digital Library Federation ,A Working Definition of Digital Library ,1998。

【2】Cleveland,G,Digital Libraries:Definitions,Issue and Challenges,UDT Occasional paper #8 ,1998。

【3】Philip Kotler,Marketing Management,Analysis,Planning,Implement and control,7thed,Englewood Cliffs: Pewntice-Hall Inc1991。

【4】Ajay K.Kohli &Bernard J.Jaworski,Market Orientation Implications Journal of Marketing, April 1990。

【5】Christoph M.Jansen, Eva-Maria Wiemann,,Volker Bach, Knowledge Platform for Electronic Customer Care, Proceedings of the INET'99,1999。

【6】Jiawei Han and Micheline Kamber, Data Mining: Concept and Techniques,Morgan Kaufmann,2000。

【7】C.Hidber, Online Association Rule Mining, in Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, pp.145－156, 1999。

【8】Surajit Chaudhuri and Umesh Dayal, An Overview of Data Warehousing and OLAP Technology, ACM SIGMOD Record, vol.26, no.1, Mar.1997。

【9】IBMOLAP Server, URL:http://www3.ibm.com/software/data/db2/db2olap/miner.htm.

【10】IBM DB2 Intelligent Miner,URL:http://www14.software.ibm.com/webapp/download/product.is.

【11】William H.Inmon, Building the Data Warehouse, QED Technical Publishing Group. Wellesley Mass.,1992。

【12】IBM, IBM OLAP Spread Sheet Add-in User Guide,2002。

【13】IBM, IBM DB2 OLAP Server,2002。

【14】IBM, IBM DB2 Starter Kit ch.10. p.131, p.182,p.184,2002。

【15】D.W.Cheung,J.Han,V.T.Ng and C.Y.Wong. Maintenance of Discovered Association Rules in Large Databases : An Incremental Updating. Technique In Proc.of the International Conference on Date Engineering, Pages106-114,1996。