國立臺灣師範大學圖書資訊學研究所

碩士學位論文

指導教授：謝　建　成　教授

書目探勘資料之清理研究—以問卷資料為例

A Study of Data Cleaning in Bibliomining ─The Case Study of Questionnaire

研究生：李 威 毅 撰

中 華 民 國 一 ○ 一 年 六 月

摘要

資料清理是書目探勘中的第一步驟，同時也影響書目探勘的結果，但資料本身常具有雜訊的存在，如此可能導致探勘過程中耗費大量時間在解決去除雜訊的問題；同時雜訊過多也會影響書目探勘的結果。在過去研究之中書目探勘的資料清理大多討論內部性資料為主，少有以外部性資料作為資料來源，而圖書館事業中大量的外部性資料可與圖書館自動化系統各個模組資料做結合提供圖書館管理者更加了解圖書館讀者的使用行為。

本研究利用外部性資料作為資料來源，利用去除雜訊、資料整合、資料轉換、資料刪減、實行概念階層等步驟進行資料清理，並透過書目探勘中的迴歸分析與群集分析評估資料清理前後的探勘結果。結果顯示，進行資料清理後迴歸分析的 R^2 與群集分析的解釋變數機率值皆能較執行資料清理前提昇

研究結果顯示本研究中所使用之資料清理方式與步驟有助於提昇書目探勘的準確度。此外，去除雜訊的步驟能有效提昇書目探勘的結果，其後並加以實行各項分群，如：雙變項分群、多變項分群等，皆能提昇書目探勘的結果。

關鍵字：資料清理、書目探勘

Abstract

Data cleaning is the inception of bibliomining, whose results also depend heavily on it. Yet, in the light of the noises encoded in the data in question, the traditional implementation of bibliomining has to sacrifice efficacy for the elimination of these undesired noises.

In past papers most researchers on data cleaning for bibliomining focused on the processing of internal data, only few took external data as their source materials. However, vast external data available in the field of library science can be synthesized with library integrated system, providing librarians a better understanding of the usage behaviors of library users.

In the methodology of our research, we first take external data as our source materials and apply them to different stages of data cleaning, i.e. data integration, data transformation, data reduction and concept hierarchy. Afterwards, we process both the untouched and the processed data with regression and clustering, on whose results we take extensive inspection with an aim to evince our concepts and methodology of data cleaning do facilitate the accuracy in bibliomining.

Our results indicate that we are capable of extracting a much prospering result of variable probability in both R^2 analysis of recession and clustering if data cleaning is adopted in bibliomining. In addition to noise elimination, we found the possibility to further increase the efficacy of bibliomining through dual-variable clustering, multi-variable clustering, to name just a few.

Keyword: Data cleaning, Bibliomining

目　次

表　次

圖　次

第一章　緒論

第一節　研究背景與動機

在數位化時代下，各項的數據資料被單位和組織視為珍貴的資產，決策者可以透過分析資料庫裡的數據資料來擷取出重要資訊以進行決策。但決策者往往必須面對的是龐大的資料量，這些資料量超出決策者可以直接從中獲取資訊的程度，因此需要一種可以自動分析大量資料以找出型樣(Pattern)的技術。而資料探勘(Data Mining)技術提供決策者從資料倉儲或資料庫中挖掘不易發現，但具有一定參考價值的訊息或資訊，從而建構出一套單位或企業可以依循的經營模式，讓單位或企業更有競爭力。

自從圖書館自動化、電子化之後，不斷累積讀者的相關記錄，例如：借閱記錄、檢索記錄等。卜小蝶（2001）提到圖書館借閱記錄是讀者使用圖書館資源的最佳「證據」，也是讀者積極滿足個人資訊需求的行為結果，這類資訊能反映使用者實際的資訊需求，因此對於掌握讀者興趣，作為加強圖書館資源利用的基礎具有一定的參考價值。圖書館擁有大量的讀者基本資料、借閱記錄、館際合作紀錄等,這些資料可以讓圖書館從中發掘出讀者的借閱習性、興趣等，幫助圖書館提供更加優良的讀者服務。

此外，Glesson & Ottensmann（1993）指出，所有類型的圖書館都開始面對越來越多問題，諸如預算縮減、不斷增加的花費、複雜性工作的增加、使用者的要求不斷增多、技術的更新、資訊產品與資訊服務項目的日新月異以及公眾的需求等。為了要應付這些挑戰，以及做出更有效能的計畫與決策，圖書館管理者需要一個能迅速管理與分析資訊的工具。決策支援系統能釐清以上問題並且是一個符合圖書館管理者需求的工具，能作為圖書館資源分配上的一個依據。

而圖書館的自動化系統中的許多記錄，經過適當的處理過程，包括統計分析、資料探勘等方式，可以讓管理者得到所需要的資源與決策參考。

Scott Nicholson（2003）提到書目探勘（Bibliomining）是解決圖書館很多問題的一個研究歷程，圖書館管理者可以使用資料倉儲（Data Warehouse）客觀地獲取資訊，而非依靠自己主觀的判斷來識別這些資料。統計方法及資料探勘技術可以用來分析資料倉儲以瞭解使用者的行為，並建立一套可信的模組，透過瞭解這些模組，可獲得以下四項的優勢：

1. 提出更貼近圖書館讀者的圖書館館藏政策；
2. 藉由建立資料倉儲之後並加以分析使用者使用記錄，使圖書館管理者更了解民眾需求，可以讓圖書館有更好的服務；
3. 制訂圖書館的相關服務工作規範以因應網際網路的時代；
4. 讓使用者更瞭解如何利用圖書館。

謝建成（2008）指出過去書目探勘資料來源主要以圖書館自動化系統或是電子資料庫廠商為主，其中包含有借閱資料、基本資料、館藏資料與流通資料等;另包含有電子資源、如線上資料庫、電子期刊等使用記錄資料。戴玉旻（2000）以交通大學圖書館借閱記錄為資料探勘資料來源；李明修（2007）以某大學館藏借閱記錄做為主要資料來源；張健彥（2008)以國立彰化師範大學圖書館資料為書目探勘的主要資料來源。以上均可以發現以往書目探勘資料來源大多以圖書館自動化系統資料為主要資料來源，缺乏外部資料，如：讀者問卷資料、館員問卷資料等作為輔助書目探勘的資料，如果圖書館自動化系統中資料出現缺漏、錯誤等情形，則書目探勘結果便淪為「Garbage In, Garbage Out」的結果。

而 Brauer（2001）指出在組織中的資料庫，大約有 15%至 20%的資料是錯誤或是無法被使用來做資料探勘的，而重要的組織資料中又有 1%至 10%的資

料內容是缺漏或錯誤的（Laudon,1986）。因此可以推論由圖書館自動化系統或電子資料庫廠商所提供的資料不能夠完全應用於書目探勘上。此外，陳建傑（2009）分析大學圖書館讀者借閱記錄與其興趣是否相關之後，建議除了利用歷史借閱紀錄作為資料來源，也應該一併分析外部資料，如修課記錄，以提高書目探勘的準確率。過去書目探勘工作主要針對圖書館自動化系統中的記錄做探勘，資料內容單一且受限於圖書館自動化系統廠商的限制。資料內容主要為讀者的基本資料、借閱資料與檢索關鍵字等資料，無法真正了解讀者對於借閱書籍的意圖與興趣，導致探勘內容受限於資料本身，因此必須加入外部性資料；但外部資料的複雜性高且不同於以往書目探勘中來自於圖書館自動化系統的記錄，其類型包含有問卷資料、修課記錄等，其中問卷資料在圖書館外部性資料中佔的比例為大宗，包含：圖書館滿意度問卷、讀者使用行為調查、圖書館事業調查、圖書館活動調查等，由於外部資料的來源多、複雜性又高，對於書目探勘而言，需要更加注意資料清理的方式。

資料清理過程涉及：

1. 讀者資料與問卷資料做整合，使外部資料與讀者做正確對應；
2. 透過有效的資料轉換，將整合後的資料轉換為適合探勘的資料形式並去除雜訊；
3. 實行概念階層，將資料提煉出不同的概念主題，以便於進行探勘。

以上皆是必須在資料探勘時考慮的因素，因此如何使用外部資料，並透過資料清理機制而實行書目探勘，實為書目探勘中一項重要工作。

第二節　研究目的與問題

本研究希望以外部資料做為資料來源，先利用資料清理的機制並配合使用書目探勘的相關技術，歸納出適合書目探勘的外部資料清理方式與清理步驟，並以探勘結果評估資料清理的效果，幫助圖書館管理者往後在進行書目探勘時能夠藉由本研究所提之資料清理方式，加速書目探勘之速度與書目探勘的準確度。

本研究之研究目的為：

1. 了解書目探勘對於外部性資料進行資料清理的流程與步驟。
2. 設計外部性資料清理機制，並利用資料探勘技術評估清理前後資料探勘結果之優劣。

為達到以上研究目的，了解在書目探勘中外部性資料的清理機制，相關的研究問題如下：

1. 在使用問卷資料進行資料清理的過程中，如何清理以達到書目探勘所需的資料內容？
2. 哪些資料清理步驟有助於書目探勘的進行？
3. 哪些資料清理技術應用於清理外部性資料能使書目探勘結果更加準確？

第三節　研究範圍與限制

本研究僅針對書目探勘中，以外部性資料做為資料清理的來源，關於圖書館自動化系統各模組中各項資料之清理方式不在此研究討論範圍。此外，本研究利用資料探勘中資料清理方法進行外部資料清理，僅討論探勘結果優劣與否，

資料探勘的演算法效率與使用的資料探勘軟體不同所產生的差異，不在本研究之探討範圍。

第四節　名詞解釋

一、 書目探勘(Bibliomining)

指結合資料探勘與書目計量學於圖書館資料分析應用之中，以分析讀者的行為、採購政策分析、預測經費使用分配、分析讀者社群、提供讀者個人化服務並作為館藏發展的依據。

二、 外部性資料(External Data)

指圖書館自動化系統以外之資料。圖書館自動化系統範圍大多具有以下模組：管理模組、編目模組、流通模組、館藏查詢模組、期刊模組、採訪模組等六大模組，此外，也包含有館際合作之交易紀錄、電子資源使用記錄等。圖書館外部性資料包含:讀者滿意度問卷、讀者使用行為問卷、圖書館評鑑記錄等。

三、 資料清理(Data Cleaning)

本研究所採用的定義基於資料清理主要在消除原始資料中的問題，因此採用 Famili, Shen, Weber, Simoudis (1997)所定義的資料清理是：「至少消除原始資料中的一個問題，且清理過後的資料相較於原始資料是有價值且有用的，能幫助達成資料分析以挖掘出重要資訊。」

第二章　文獻探討

第一節　書目探勘

書目探勘(Bibliomining)一詞是 Scott 與 Stanton 於 2003 年「The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making.」首先提出，在此之前有關資料探勘應用於圖書館領域的研究已有一些成果，但是研究人員所用的詞彙，均是以圖書館的資料探勘（Data Mining in Library）來呈現，因此 Scott 與 Stanton 於 2003 年首先創建了書目探勘學（Bibliomining）一詞，這個名詞的定義就是利用資料探勘 (Data Mining)及書目計量 (Bibliometric)工具於圖書館服務所產生之資料的應用，開創圖書館管理與服務一新的研究領域。

Banerjee（1998）指出引進資料探勘技術於圖書館之中，對於圖書館而言有兩項好處：

1. 相較於紙本目錄資源取得不易，在引入資料探勘之後，電子化的資料能夠更快速且完善的被取用
2. 能讓管理者或使用者不需透過他人幫助，輕易的使用電腦找到其所需的資源。

書目探勘對於圖書館而言，可以將原有的圖書館自動化系統中的各項記錄加以利用，並且將其以電子化方式快速提供給圖書館管理者以及館員做為決策的有效依據。張健彥（2008）將目前書目探勘的相關研究依照功能導向歸為三類：讀者行為模式、協助決策制定、推薦行為

一、讀者行為模式：

透過書目探勘結果，得到讀者使用圖書館的行為模式，同時利用此模式，吸引讀者並達到提升圖書館使用率的研究。吳安琪（2001）以交通大學圖書館的書目、館藏記錄為基礎，以 Aproiri 的演算法探討讀者的社群關係與吸引讀者到館借閱、提升讀者借閱率、提升讀者忠誠度、促進館藏流通率的原因和協助館藏複本採訪政策；而林湧順（2005）以師大附中圖書館館藏記錄配合學生的各項基本資料以及借閱時間等，分析學生借閱行為與其所屬的居住地區、類組、特殊班級(數理班、音樂班)、成績等關係，針對資料做客觀分析反映出學生在借閱行為上的實際情況，提供圖書館經營、館藏發展與個人化服務方面做為依據。

透過書目探勘方式，找尋圖書與使用者之間的關聯，發掘使用者的行為模式，能夠幫助圖書館針對讀者行為制訂對應的經營決策以及個人化的服務，對於圖書館而言，能夠提升讀者使用圖書館意願也能夠幫助讀者更輕易獲取所需的資源。

二、協助決策制定：

透過書目探勘結果找出讀者主要的使用需求，以協助館方針對讀者需求做出適當的館藏發展政策。陳建銘（2001）以台灣科技大學圖書館為研究對象，以霍普菲爾類神經網絡(Hopfield Neural Network)解決關聯規則的部分問題，並應用於圖書館中，找出書籍借閱偏好以及書籍借閱情況，提供學校圖書館做為建構館藏的決策依據並提升圖書借閱率。黃毓菁（2002）將圖書館館藏分類與讀者使用館藏偏好分群做探勘，並透過資料探勘中的關聯規則演算法，期望找出圖書館讀者的潛藏特徵，並且將探勘結果與實際圖書館借閱記錄加以比對，證明其可信賴度以提供圖書館對於館藏預算的分配與提升館藏使用率做為參

考。謝建成、魏儀禎（2003）利用資料倉儲與多維度分析技術，應用於圖書館之中，針對各系所圖書利用情況、各分類館藏圖書情況做為資料倉儲的主題，探討圖書館中館藏利用情況並支援圖書館相關決策。Wu（2003）說明圖書館的管理人員經常需要面對預算分配的問題，正確且有意義的分配預算，對圖書館運作很有助益，從圖書館流通記錄檔中可以發掘出這樣的資訊，因為這些資訊可以反映出讀者的實質需求。此研究旨在說明利用這樣的一個過程，可以讓圖書館的預算分配上有一個可以依循的參考，並建立一個可以應用的模組（Data Mining Based Model，簡稱 DMBA），作為研究的結論。Bleyberg, M. Z., Zhu, D., Cole, K., Bates, D. ＆ Zhan, W.(1999)等人的研究以資料倉儲技術，應用於圖書館之中，主要的目的就是要將所得的結果用來作為圖書館的決策參考，以提升圖書館書籍、期刊及電子資源的使用率，因為讀者對圖書館的使用模式一直的在變動，所以圖書館的管理者必需要不斷的調整這些資源的採購政策與授權資料。鄧世昌（2009）以多層次關聯規則探勘技術，利用圖書館借閱記錄找出圖書館分類架構與讀者借閱書籍的關聯性，進而提供圖書館做為館藏發展與改良圖書檢索系統之用。

從支援決策的觀點來看，書目探勘提供一客觀的角度，透過各項演算方法以及資料倉儲讓資料本身能夠表現出其代表的潛藏意義，提供圖書館管理者做為制定館藏政策、經費增減、服務項目等各項的決策參考指標。卜小蝶（2002）以分類號第三層 000~999 作關聯規則分析，利用相似性比對方法，推估相似借閱行為所反映出的圖書類號，以做為圖書推薦的依據。然而，研究中也指出兩個限制，第一，多數讀者所借閱圖書的類號並不多，因此要得到關聯規則的類號並不容易。第二，讀者借閱圖書的動機可能來自於修課，因此不易釐清類號之間與主題性質上的關聯性。

三、推薦行為：

透過書目探勘結果，圖書館可以依照讀者的興趣、習性、借閱習慣等原則，做為圖書推薦的機制，而此類研究數量也最多。卜小蝶（2001）探索借閱記錄中隱藏的重要規則，包括圖書與讀者、讀者與讀者、圖書與圖書間的關係。圖書館欲主動推薦相關新書或特定讀者輸入關鍵詞查詢出相關圖書時，可以根據借閱記錄，分析有興趣的類號，並透過分類號群集、相似系所分類號群集，以及重要分類號與系所關聯等，進一步將圖書重新排列加以推薦。洪志淵（2001）從圖書借閱資料庫中挖掘出讀者與圖書間的關聯規則，並交由圖書館專家詮釋規則上的知識，以運用於新書推薦。除此之外，更訂出一套 Interesting Rules 的評量方法，以判斷讀者的興趣趨向，並根據讀者族群特性，應用在圖書館的新書推薦上。曾勇森（2002） 將資料探勘技術，應用在圖書館書籍與讀者之間，主要目的是為書籍找尋適性化之讀者，以及為讀者找尋適性化之書籍。首先是將書籍，或是讀者，透過相似度的計算，將所有的書籍或讀者分成多個群組，群組內的書籍或讀者具有高度相似性，而群組間具有高度的差異性，再利用群組內成員的借閱型態趨勢，找尋該群組最適性之讀者或書籍並加以推薦。呂家賢（2005）透過決策樹分析，將讀者做區隔並利用讀者借閱冊數做為目標變數，投入學期成績、學院別、有無申請助學貸款等因素，藉此協助圖書館對於特定目標做行銷並且利用關聯規則建立不同讀者群的借閱關聯性。羅子文、柯皓仁（2007）以 Web2.0 的精神與個人化推薦系統結合，應用於圖書館推薦系統中，並利用關聯規則探勘的方式找出個別讀者的推薦清單，並將書籍以難易的等級加以區分，針對個別讀者提供適合的書籍以提升圖書館推薦系統的品質。謝賓帆（2008）透過決策樹分析，找出書籍與讀者、書籍與書籍之間的關聯特性，並利用讀者借閱類別比例做為調節加權的依據，歸納出適合的讀者與圖書，提供圖書館做推薦圖書時參考之用。楊詠喬（2010）利用關聯規則技術，

對書籍與讀者進行分析，找出各特性相近的讀者借閱圖書的情形，透過關聯規則結果，可能供圖書館推薦讀者有興趣的書籍，以達到提升借閱率的目的。

過去書目探勘中所使用的探勘技術絕大多數為:關聯規則分析、分類分析、群集分析、次序相關分析，表 2-1-1 為林湧順（2005）與陳建傑（2009）所整理之資料探勘技術可應用於圖書館的範疇。

表 2-1-1 資料探勘應用在圖書館的範疇

資料探勘技術	應用範疇
關聯規則分析	1. 找出讀者個人特性與圖書之間的關聯性 2. 利用讀者特性的相似性推薦圖書 3. 將同質性的圖書，推薦給適性的讀者 4. 探求讀者資訊需求特徵，做為圖書館館藏發展政策之依據 5. 圖書館預算分配
分類分析	1. 透過讀者與圖書之間之分類特性，做為圖書推薦機制 2. 建立讀者導向的服務項目
群集分析	1. 找出圖書與圖書、讀者與讀者之間的關係，以探討使用者的集群特性，並找出其借閱行為的傾向
序列分析	1. 依據讀者圖書借閱的順序，來推薦給其他未借閱之讀者 2. 尋找書籍適性化之讀者 3. 圖書館業務人力資源的安排

資料來源：林湧順（2005）。以資料探勘技術探討高中生使用圖書館之行為模式--以國立台灣師範大學附屬高級中學為例。國立臺灣師範大學社會教育學系碩士論文。未出版，台北。陳建傑（2009）。基於借閱目的之資料清理機制研究—以興趣目的為例。國立台灣師範大學圖書資訊學研究所碩士論文。未出版，台北。

第二節　書目探勘技術

書目探勘所應用的探勘技術如前所述，主要包含有關聯規則分析、分類分析、群集分析、序列分析。

一、關聯規則分析(Association Rules)

關聯規則分析主要被用來尋找資料庫中某些資料項目或屬性之間共同發生的關係。最初應用於超級市場的購物籃分析，從購物籃的交易記錄中，可找出相關產品間的關聯，並據以分析顧客購買行為及較常出現購買模式。這樣的結果，可以產生一個行銷的策略就是將經常被一起購買的商品放近一些，以便進一步刺激這些商品一起銷售。

關聯規則運用的原理為條件機率，例如：購買 A 商品時，有多少機率會同時購買 B 商品。而關聯規則以信心水準(Confidence)及支援度(Support)這兩個指標來評斷所找到的規則是否可用。

二、分類分析(Classification)

分類分析是根據一些變數的數值做計算，再依照結果作分類，計算的結果最後會被分類為幾個少數的離散數值，例如將一組資料分為「可能會回應」或是「可能不會回應」兩類。因此會用一些已經分類的資料來研究它們的特徵，然後再根據這些特徵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，或是將一個完整資料庫做部份取樣，再經由實際的運作來測試；譬如利用一個完整資料庫的部份取樣來建立一個分類模式（Classification Model），以後再利用這個模式來對資料庫的其他資料或是新的資料作預測。

三、序列分析(Sequence Analysis)

序列分析是一組按時間順序發生的事件，研究者根據每一固定時間間距依序紀錄事件結果，而時間序列數據最大特點就是當中每一筆緊接著數據的紀錄時間間距均相同。股票市場固定時段價格變化，每月進出口貿易相關數字，每年人口出生率數字等分別為時間序列數據例子。要分析時間序列數據，研究者首先可以使用一些視覺檢查（Visual Inspection）工具（例如立體圖表），從時

間序列數據紀錄，觀察出某些現象特徵及行為，通常時間序列有四種主要的變化：長期或趨勢變化、迴圈變化、季節性變化、非規則或隨機變化。

四、群集分析(Clustering)

群集分析被廣泛應用於社會科學、生物科學、商業和教育等各領域。群集分析是利用一些特性的組合來對樣本作群體的分類，也就是設定一組由多個屬性描述其特性的物件集合，群集分析根據物件間的相似性，將這些物件分成群集，使得每個群集內的成員具有高度的相似性，而不同群集間之物件具有高度的不相似性。

在分群技術中，階層式群集演算法（Hierachical Clustering Algorithms）和切割式群集演算法（Patition Clustering Algorithms）為最常見的分群技術。階層式群集法可分為：凝聚式與分裂式兩種，兩者分別為由下往上（Bottom Up）與由上往下（Top Down）的方式，逐步將物件分為不同群體，此分群方式可以夠過樹狀圖看出各群間的關係，但執行速度緩慢是一主要缺點。

切割式群集演算法需要使用者先確定要切割的 K 個數目，在以群集重心（Cluster Mean）的方式進行分群。K-means 演算法與 E-M 演算法為著名的切割式群集演算法。

五、迴歸分析(Regression)

迴歸分析就是一種統計分析的方法，主要在了解自變數（Independent Variable）與因變數（Dependent Variable）間之數量關係，其主要用處是尋找兩個或兩個以上的變數之間相互變化的關係。當找到這些關係之後，就可以利用結果進行：

(一) 變數間關係敘述（Description）：例如說明節目製作費用與收視率之關係。

(二) 變數間控制（Control）：例如商品價格與需求量有關係，故控制價格，就可以控制需求量。

(三) 對變數值預測（Prediction）：例如若存在製作費與收視率有關係，則可以用此來預估某節目的收視率。

第三節　資料清理

Famili et al.(1997)將資料清理定義為：「至少消除原始資料中的一個問題，且清理過後的資料相較於原始資料是有價值且有用的，能幫助達成資料分析以挖掘出重要資訊。」。資料清理能解決導致分析錯誤的資料問題，同時了解資料的屬性，藉此以更有意義的分析，從資料之中找出有意義的資訊（Famili et al., 1997）。因此可以了解資料清理對於資料探勘而言有著舉足輕重的意義，若是資料清理不完全，或是資料清理不確實，則容易導致分析結果不良甚至錯誤的情況。

曾憲雄、蔡秀滿、蘇東興、曾秋蓉、王慶堯（2005）提到在整個資料探勘歷程中，資料清理通常是花費最多時間的，同時也對探勘品質影響最大。另外，Han & Kamber(2000)指出資料清理過程涉及資料整合、資料轉換、資料刪減的步驟：

一、 資料整合

在資料探勘前多半需要進行資料整合，資料整合將多個資料來源的資料，合併起來放到同一個資料儲存地(資料倉儲)，這些資料可能會來自不同資料表、不同資料庫甚至是一般的檔案來源。

資料整合需要考慮到哪些在資料庫中的資料是屬於相同的實體，而在一般的檔案來源中，則需考慮如何與資料庫中的資料連結，這時候可以透過詮釋資料(Metadata)作為輔助;一般資料來源則可以利用如問卷的編碼簿或外部資料的資料描述文件等做為資料轉換的依據。

另外，資料資料轉換中會因為從不同資料庫資料導出資料，導致資料屬性不一與資料維度命名不同的情況，此時更需要透過詮釋資料與一般資料來源的資料描述文件作為資料轉換的依據。而資料整合中常見的問題為資料值衝突的處理方式。資料來源來自不相同的資料庫或資料表時，可能會因為描述資料所使用的資料單位或屬性值不相同而導致資料整合時衝突，如：公制單位與英制單位彙整時，須考慮紀錄資料單位不同的情形。

二、 資料轉換

資料轉換乃是將資料轉換成適合探勘的型式。資料轉換內容涉及資料平滑化(Smoothing)、資料聚合(Aggregation)、資料一般化(Generalization)、資料正規化(Nomalization)、屬性建構(Attribute Construct)等方式：

1. 資料平滑化：

資料平滑化的方式可以透過分箱(Bining)與迴歸(Reggration)的方式處理。分箱方式先將資料排序後決定要分為多少箱，決定之後將每箱中的資料求其平均值並取代原本的每個值；利用求其資料的迴歸關係式的方法，接著將每個值帶回其迴歸式，使資料平滑化。

2. 資料聚合：

對資料進行匯總與聚集。例如：可以將每日的銷售記錄匯總為每週或每月的銷售記錄；而圖書館資料也可從每日借閱記錄彙整為每週借月記錄等方式。通常此一步驟可用於資料倉儲中將低粒度資料建構成高粒度的資料方塊。

3. 資料一般化

資料一般化又稱為資料廣義化，目的在將概念層級較低的資料轉換成概念層級較高的資料。例如：當原始資料收集時，國文系、英文系隸屬於文學院中，但在資料探勘中依需要可以將其一般化為文學院、此時資料內容則一般化為文學院，不再紀錄國文系、英文系等低層的概念。

4. 資料正規化

資料正規化乃是將屬性資料按比例縮放，使資料內容落入一特定區間中，如-1 至 1，此方法可以利用最大—最小值正規化、Z-score 正規化兩種。此方法對於計算距離的分類演算法而言較不會受到極端值影響。

5. 屬性建構

屬性建構乃是透過兩個或兩個以上的屬性加以結合建立新的屬性，例如：將往年單月每人次平均借閱數量與單月到館人次結合而產生出單月借閱總量，透過此方式建構新屬性以符合探勘的資料形式。

三、 資料刪減

Jermyn, Dixon, and Read（1999）指出資料清理過程大約會佔整個探勘計畫的 60~80%的時間。資料探勘中資料清理佔據大多數的時間，而 Jian 與 Jin（2003）與 Jermyn, Dixon, and Read（1999）說明資料清理對於資料本身需要處理的問題有：空缺資料(Missing Data)、錯誤資料(Erroneous Data)、孤立點(Outliers)、

雜訊(Noise)、重複資料(Duplicated Data)、異質性(Heterogeneities)。而其中又以資料異質性(Heterogeneities)最難解決。除了資料本身可能有以上問題需要解決以外，Jermyn, Dixon, and Read（1999）表示資料清理過程中也可能出現以下問題：

1.清理部分資料使資料內容變得無法辨識(One shot cleaning)；

2.對於清理的資料沒有做記錄(No record of cleaning decision)；

3.資料清理程度過低(Cleaning choices made at a low level)；

4.大量使用人工的方式清理資料(Expensive manual methods)；

5.缺乏清楚的清理方法(Lack of clear methodology)。

資料清理過程中不但資料本身容易出現問題，同時清理的方式也必須要審慎評估並記錄詳細步驟，才能避免資料清理時出現錯誤。各項錯誤的資料內容，Jian 與 Jin（2003）指出能透過以下方式處理：忽略、補填資料、分箱(Binning)、群聚(Clustering)、迴歸(Regression)、加入門檻值(Threshold)；Han 與 Kamber（2006）說明缺漏資料也能透過忽略資料、回填原始資料、回填一常數、使用平均數回填資料、回填可能資料；雜訊資料能透過分箱、迴歸、群聚。分箱方式乃是利用分箱方式先將資料排序後決定要分為多少箱，決定之後將每箱中的資料求其平均值並取代原本的每個值；群聚方式乃是設定一組由多個屬性描述其特性的物件集合，群集分析根據物件間的相似性，將這些物件分成群集，而群集的相似性常是以群集中每筆資料的距離決定，距離近代表較相似、距離遠則代表較不相似。表 2-3-1 為各項錯誤資料可行之處理方式：

表 2-3-1 各項錯誤資料之處理方式

資料錯誤類型	可行之處理方式
缺漏資料	忽略資料、回填原始資料、回填一常數、使用平均數回填資料、回填可能資料
孤立點	分箱(Binning)、迴歸(Regression)、群聚(Clustering)、刪

資料錯誤類型	可行之處理方式
	除
雜訊	分箱(Binning)、迴歸(Regression)、群聚(Clustering)、刪除
重複資料	刪除、忽略

資料來源：研究者整理（管志剛、金旭(2003)。數據挖掘中數據預處理的研究與實現。計算機應用研究，7，117-119。 Han J.& Kamber, M. (2000). Data Mining: Concepts and Techniques. San Francisco, CA: Morgan Kaufmann.

四、建立具概念階層的資料集

建立具概念階層的資料集有分箱式(Binning)、直方圖式(Histogram Analysis)、分類分析法(Cluster Analysis)、直觀式離散法(Discretization by Intuitive Partitioning)。分箱式是利用箱中平均值取代箱中的每個值或是利用中位數取代箱中所有的值；直方圖分析法利用設定等深區間的方式將資料分散在各個區間，並使各個區間有相同的樣本數，此時便可以產生多個概念階層的資料集；直觀式離散法乃是基於分箱分式與直方圖式方法所得到的分類區間為 51263.34~60234.13 之間，如此對於直觀法則可以以自然之方式將界限定為 50000~60000 的區間如此一來往後解釋資料結果時可以方便統整為各項的資料。

第四節 資料清理應用於圖書館

資料清理應用於圖書館中之具體研究以陳建傑（2009）以興趣做為借閱目的之資料清理機制研究，有明確探討書目探勘中資料清理機制，其研究透過考量讀者借閱目的中的興趣目的，設計清理機制，並且試圖除去讀者借閱記錄中之非興趣記錄，並以 F-Measure 評估資料清理效果。其餘書目探勘研究或圖書館中之資料探勘研究以針對讀者行為、協助決策制訂、推薦行為等為其主要之研究目標（張健彥，2008）。

陳建傑(2009)表示圖書館中可用來分析的資料可分為三大類:圖書資料、館藏使用記錄與讀者資料，而卜小蝶（2001）提到館藏使用記錄是讀者使用圖書館資源的最佳「證據」；館藏使用記錄包含有歷史借閱記錄、OPAC 查詢記錄、電子資料庫使用記錄等。另外，過去書目探勘研究中，分析讀者借閱記錄的研究,在進行書目探勘前也針對資料的常態性及完整性等進行清理(卜小蝶，2001，2002；王毓菁，2002；余明哲，2003；呂家賢，2004；吳安琪，2000；孫冠華，1999；鄭玉玲，2002；戴玉旻，2002；林湧順，2005；羅子文、柯皓仁，2007；鄧世昌，2009；謝賓帆，2008；楊詠喬，2010）。而針對外在屬性部分研究者會利用系級、年齡、學院等屬性做為分群依據後，再進行分析（余明哲，2003；張苑菁，2001；鄭玉玲，2003）。

資料毀損與雜訊是資料問題中最需要被解決的問題，而雜訊會降低資料的預測能力（famili et al., 1997）。而書目探勘之資料清理也不例外，透過資料清理將可能的雜訊去除能夠使探勘結果更臻完備。以圖書推薦為例，Im and Hars（2007）提出探勘目的會影響讀者對於推薦結果的滿意度，因此針對雜訊需要適當的處理，若是將不需要的資料投入探勘之中，則會引導出錯誤的結果。余明哲（2003）與周黃順加（2006）在進行興趣圖書推薦之前，對於讀者有興趣之項目給予權重，等同於在資料轉換過程中給予權重進行資料清理的步驟。陳垂呈（2005）則是針對讀者首次推薦結果中勾選有興趣項目，並進行加權，也是資料轉換過程中做出特定的資料清理。而陳建傑（2009）透過讀者的借閱目的個別清理資料以提升資料清理的效果，直接針對讀者之興趣書籍作為清理對象。例如：以課業或可視為目的的借閱行為對於圖書館利用興趣圖書推薦時，這些資料便視為是雜訊必須加以清理如此才能提升資料清理的效果。

第五節　圖書館內外部資料

書目探勘乃是利用資料探勘、書目計量學與統計學的方式，在圖書館系統中進行使用者行為基礎(Behavior-Based)研究，從圖書館自動化系統中的記錄、使用者本身的基本資料與圖書館事業的外部資料做連結，使書目探勘能夠運用內部與外部的資料成為探勘的資料來源，而圖書館內部資料內容格式單一，多利用圖書館長期衍生出的記錄格式，因此，在書目探勘之中，對於圖書館的內部資料而言，資料清理的方式相對於外部資料要單純許多；圖書館內部資料可區分為圖書館的記錄資料與圖書館的服務資料，記錄資料包含：

1. 從作品中擷取出的記錄，如：書名、作者等；
2. 從作品中精煉出的記錄，如：主題標目(Subject Headings)、分類；
3. 指出作品型態或位置的記錄，如：URL。

圖書館服務資料則指圖書館提供服務的過程中所紀錄的資料，包含：

1. 讀者的搜尋資料，其中包含有讀者檢索的記錄與檢索的流程；
2. 圖書館流通的記錄，包括讀者的借閱情況與線上公用目錄的資料；
3. 參考服務的記錄，包含參考服務進行的方式、讀者的問題及解答與花費的時間；
4. 推廣與訓練的記錄，其包含活動的類型、活動的主題與參加者的資料；
5. 館際互借與其他服務，包含服務的時間、最終提供服務的單位、花費的成本。

圖書館內部資料以圖書館的記錄資料與圖書館的服務資料為主。圖書館的記錄資料可以發現其註錄的資料類型皆以圖書館中的紀錄方式為基礎，多以

MARC 為主，近期圖書館中也包含有其他類型的註錄資料，如：Dublin Core 的紀錄內容，但其紀錄方式基本上還是以圖書館長期以來的註錄方式為基礎，其資料的複雜性相對較低；而圖書館的服務資料則衍生自圖書館的服務，並且為圖書館記錄資料與外部性資料的連結，藉以評估圖書館服務與瞭解圖書館使用者的基礎。

圖書館外部性資料則是了解圖書館使用者的重要方式之一，透過外部性資料的瞭解與內外部資料的連接，可以更加瞭解到所屬圖書館使用者群體的概況。而相對於圖書館內部資料註錄方式，接以長期以來圖書館界的紀錄方式，外部性的資料記錄內容則相對複雜許多，由於不同圖書館紀錄的方式不同，因此在資料清理上也相對困難許多；此外，內部性資料多屬於常設性質的資料且記錄方式單純，大多紀錄於圖書館之中很長一段時間，或是紀錄之後便不容易加以改變或刪除，所以即使內部性資料量增加，其清理方式也不易變動；而外部性資料對於圖書館而言可能是因為特定因素而產生的，其所紀錄的方式與資料內容大多難以用過往的清理方式完全套用，這也是外部性資料清理不易的重要因素。

圖書館事業中所使用的外部資料項目大多以問卷資料為主，而問卷資料可區分為：圖書館滿意度問卷、讀者使用行為調查、圖書館事業調查、圖書館活動調查。其中圖書館事業調查主要以調查圖書館本身所擁有的資源、設備與人力等為主要調查項目；而圖書館活動調查則針對圖書館個別活動提供參與者回饋意見給館方。其中，圖書館所使用的問卷內容可以分為以下項目：

(一) 基本資料

基本資料部分，瞭解填答讀者的年齡、性別、教育程度、居住地區等，做為區別不同群體讀者的依據。

(二) 滿意度調查

滿意度調查針對圖書館中所提供之設備、人力、資源等各項圖書館服務做調查，瞭解讀者對於圖書館服務的滿意度。

(三) 使用行為

使用行為部分瞭解讀者對於圖書館中服務的使用概況，例如：使用的服務種類、使用時間、使用頻率等作為調查項目。

(四) 使用意願

使用意願部分則調查讀者對於圖書館服務的使用意願，包含現在圖書館服務的使用意願與未來將提供的圖書館服務的使用意願。

圖書館中常見的問卷類型為上述所提，所填的資料類型以名目尺度為主，其他還包含有李克特量表(Likert Scale)與表示時間區段的等距尺度的資料，其相關問卷範例如下：

(一) 名目尺度：

職業：□學生 □軍/公/教人員 □工業/製造/營造相關 □金融/保險/貿易相關 □工商企業投資/自營商 □農/林/漁/牧業 □文化/運動/休閒相關 □專業技術人員(如醫師/律師/資訊/工程/會計/設計等有執照或證照人員) □運輸/倉儲/通信相關 □其他服務業 □家庭管理 □無業/待業 □退休 □其他_____。

(二) 等距尺度

您每次到圖書館平均停留多久時間： □一小時以內 □1-2 小時 □2-3 小時 □3-4 小時 □四小時以上。

(三) 李克特量表

館員服務保持禮貌態度是否重要？ □非常不重要 □不重要 □普通 □重要 □非常重要。

第六節 小結

一、資料清理為資料探勘中最重要環節之一

由上述所提及之文獻可發現到資料清理為資料探勘中重要環節之一，其為資料探勘中最耗時的一項工作，但若是能將資料清理做好，資料探勘工作便也有良好的開始，同時可避免錯誤的資料影響資料探勘的結果。

二、圖書館缺乏以外部資料為主之資料清理研究

過去圖書館中資料探勘的研究主要針對讀者行為、協助制訂決策、推薦行為做討論（張健彥，2008），鮮少以資料清理為其主要研究標的，陳建傑（2009）研究以興趣為借閱目的的資料清理機制，為少數以資料清理為主要研究標的之研究，而資料清理議題在圖書館之中尚處於起步階段，而以外部資料為主的資料清理機制抑是在圖書館中尚未有人提及之研究，因此分析圖書館外部資料前沒有固定的清理機制作為背景。

第三章　研究設計與實施

本研究以外部資料做為書目探勘的資料來源，因此不涉及圖書館自動化系統中的各項紀錄資料。考慮到圖書館外部資料來源可能從出版社、書店等相關企業之銷售資料或問卷資料提供，本研究嘗試以某公司 95 與 97 年度的問卷資料為外部資料來源，做為此次資料清理的基礎。

第一節　研究流程

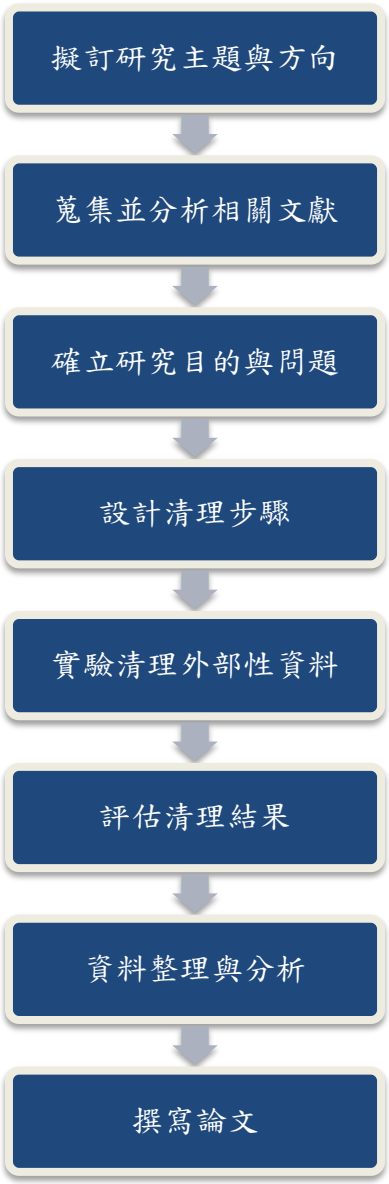

圖 3-1-1 研究流程圖

本研究先確立研究方向與主題，透過蒐集國內外相關文獻了解目前書目探勘的相關研究與資料清理的概況，做為研究設計參考與確立研究題目，之後實際針對外部性資料實驗資料清理方法，評估資料清理方法之優劣，最後撰寫報告。具體研究實施流程如圖 3-1-1。

第二節　研究工具

本研究所使用之輔助軟體包含有 Excel 與資料庫管理與資料探勘軟體 Microsoft SQL Server 2008 Enterprise 版。Excel 軟體主要做為資料清理的工具，透過 Eecel 將資料做整合、清理、轉換等工作，並匯出為 CSV 檔，做為匯入 Microsoft SQL Server 之用。

Microsoft SQL Server 原是資料庫管理系統，自從 7.0 版提供資料轉換服務與線上分析處理後，便能夠進行資探勘的操作。其後 2000 版提供決策樹與群集分析兩種演算法，2005 版提供有九種演算法，包含：決策樹、分群法、時間序列、時間群集、關聯規則、單純貝式機率分類、類神經網絡、線性迴歸、羅吉斯迴歸等九種，其透過整合的資料分析服務將商業理解、資料理解、資料預備、塑模、評估、部署等六項標準的資料探勘作業整合其中，其後的 2008 Enterprise 版亦是一套完整的資料庫管理系統與商業情報管理系統，承襲 2005 版能夠進行資料探勘中的九項演算法，並提供倉儲系統透過 OLAP 能夠提供決策的依據。

第三節　研究方法與設計

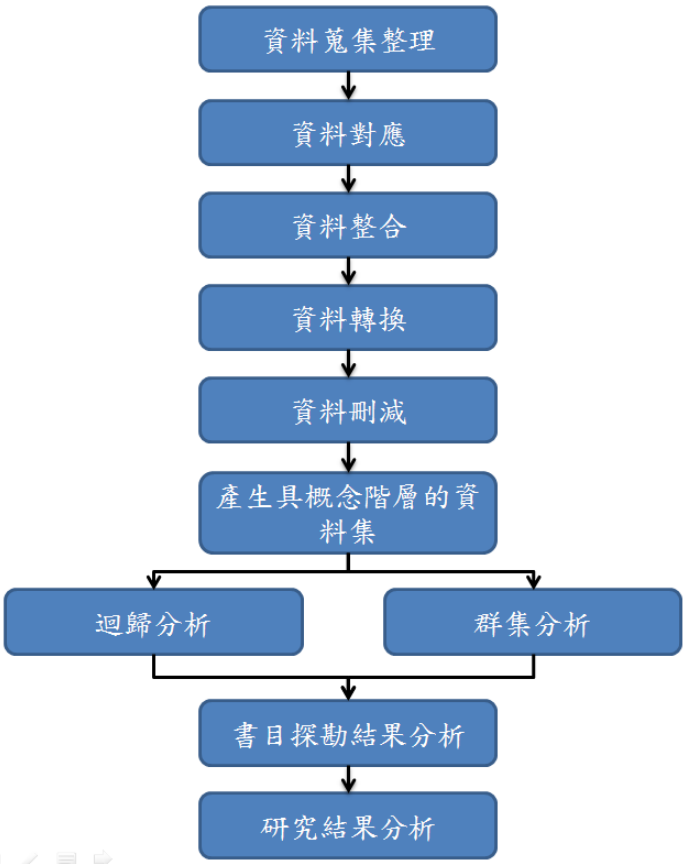

圖 3-3-1 研究方法流程圖

本研究的研究方法透過書目探勘中群集分析與迴歸分析的結果，做為評估資料清理機制好壞的方式，由圖 3-3-1 得知本研究從資料蒐集與整理起，經過資料整合、資料轉換、資料刪減、產生具概念階層的資料集最後透過書目探勘的技術評估資料清理的結果，最後綜合評估本研究結果。

一、資料對應

本研究將使用的問卷資料與圖書館事業中常見的問卷資料加以整理，透過研究者整理出圖書館事業中常使用的問卷類型，與本研究問卷資料加以對應，以利用本研究所使用之資料與圖書館事業中問卷資料能夠成功對應，以便於之後對於本研究問卷進行資料清理、資料探勘評估等步驟。

二、資料整合

本研究中資料清理步驟，主要透過 Excel 做為資料清理的工具。圖書館自動化系統資料與圖書館外部資料之整合可以透過讀者編號的對應方式針對每一筆外部資料與其圖書館自動化系統中的讀者資料加以整合，可彙整成讀者借閱紀錄與外部資料的整合型式，或是讀者於圖書館自動化系統中填寫興趣圖書的資料與外部資料整合。可利用讀者姓名與讀者年級等資料加以整合，若是所獲得之外部資料僅填答至系所、年級等資料粒度較大之資料，則必須將圖書館自動化系統中的所屬類別之資料加以彙整，以便於進行資料探勘。

本研究在資料整合步驟透過研究者自行整理某公司 95、97 年度的問卷資料編碼簿為基礎，將兩年度的同質性資料做直接整合的步驟，並再針對兩年度異質性的資料做處理，異質性資料處理可以透過兩年度的編碼簿作為整合基礎，分為異質性紀錄值的處理、異質性意義的處理。異質性紀錄值需要透過了解兩年度編碼簿各自不同的紀錄方式加以將記錄值彙整為單一種紀錄方式；異質性意義的處理需要利用人工方式查閱兩年度每項欄位是否有意義相類似的欄位才得以彙整為單一記錄，如圖 3-3-2 所示。

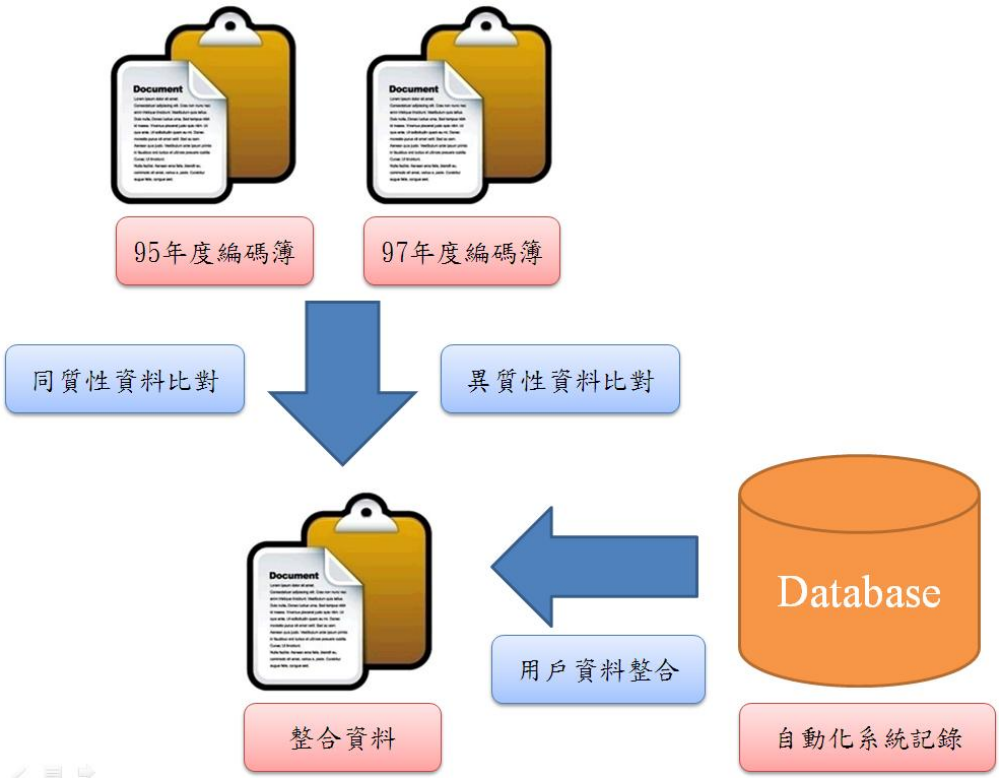

圖 3-3-2 資料整合示意圖

三、資料轉換

資料轉換方法包含有：資料平滑化(Smoothing)、資料聚合(Aggregation)、資料一般化(Generalization)、資料正規化(Normaliztion)、屬性建構(Attribute Construct)。

(一) 資料平滑化(Smoothing)

資料平滑化透過分箱(Binning)、迴歸(Reggration)等方式，消除資料中的雜訊(Noise)，讓書目探勘過程中，不會因為資料內容雜訊過多而減低書目探勘的效果。例如：分箱方式將參考相鄰的值作為平滑的方式如資料內容為 Bin1：4, 8, 15、Bin2：21, 21, 24 利用平均數平滑的方式可以將其平滑為 Bin1：9, 9, 9、Bin2：22, 22, 22；而迴歸的方式則

是讓資料調適而符合一個函數(迴歸函數)，如此當知道一個變數之後即可預測另一個變數的值。

(二) 資料聚合(Aggregation)

資料聚合係利用資料粒度(Granularity)較低的資料做為聚合的目標，例如：將每日資料聚合成每週、每月、每季的資料，透過這樣的方式可將資料粒度過低的資料提升成我們所需要的資料層級，達到書目探勘結果符合我們的需求。

(三) 資料一般化(Generalization)

資料一般化係將較低層級(Low-level)的資料提升其資料層級，將資料轉化成我們一般常用的資料概念，例如：街道資料轉化成縣市資料、縣市資料轉化成北中南三區資料等。幫助研究者能獲取所需資料。

(四) 資料正規化(Normalization)

為避免資料內容的數值變化過大或各資料欄為之間編碼方式差異過大，透過正規化的方式能幫助減少因為數值差異太大所造成的資料探勘結果偏誤，例如：。

(五) 屬性建構(Attribute Construct)

透過兩個或兩個以上的屬性結合，得到新的屬性，例如：透過長度與寬度相乘能得到面積資料、透過平均使用時數與數量相乘得到總使用時數等方式。透過這個方式幫助建立起隱藏在資料之中的屬性。

四、資料刪減

資料刪減方法包含有：從資料方塊聚合(Data Cube Aggregation)、選擇屬性的子集合(Attribute Subset Selection)等。利用從資料倉儲中的資料方塊將資料聚合成粒度較高的資料，刪去過於細節的資料避免書目探勘過程，因為資料粒度太低而導致無法拆解的情況產生；刪除與主題較不相關的資料部分，透過資料子集合的資料，能夠更加貼近研究主題所需的資料。

五、產生具概念階層資料集

建立概念階層方式有許多種，包含有：分箱式、直方圖式、分類分析法、直觀式離散法。本研究將以直觀式離散法為主要概念階層的建立方式。

六、評估資料清理結果

本研究以群集分析與迴歸分析結果為評估資料清理機制方式：

(一) 群集分析

群集分析結果將以結果中樣本數量最大的群集與樣本數量次大之解釋變數的機率值為評估結果，解釋變數的機率值利用群集分析中各項變數的差異而顯示各項變數能夠解釋群集差異的機率，透過此方法作為評估群集分析結果之有效程度，如圖 3-3-3。

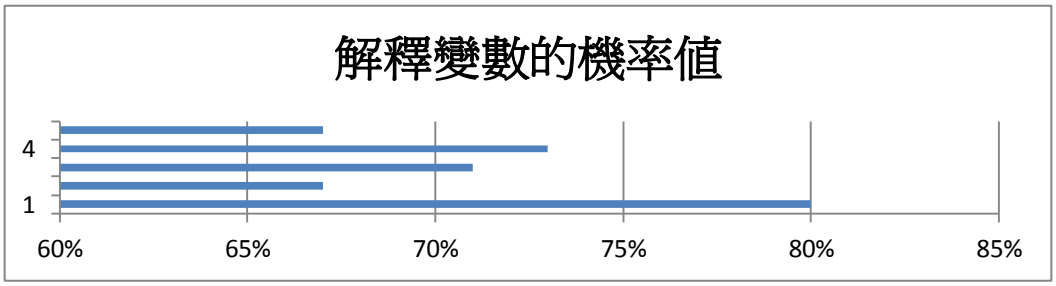

圖 3-3-3 群集分析結果評估圖

群集分析結果以兩個群體間的辨識率做為評估結果的方式，若是兩個結群體間的辨識率顯著提高，則表示透過清理結果能有效提升書目探勘結果的良率。

(二) 迴歸分析

迴歸分析透過投入的自變數與應變數了解其數量關係，能夠找出兩個或兩個以上的變數間變化情況，例如：電器數量、電器使用時間與用電量之間的關係。本研究評估結果利用迴歸分析中的 R^2 做為評估資料清理結果的方式

第四章　研究結果與分析

本章將依照研究步驟，分為四個小節，分別敘述本研究之研究過程與結果。依序分為：問卷資料對應與描述、清理方式、資料清理結果評估與小結。

問卷資料清理與評估描述本研究中所使用問卷對應至圖書館事業常見問卷的題項；清理方式描述本研究中如何清理外部性資料，其步驟與方式以及如何實行各種分群方式；資料清理結果評估描述透過本研究中所提之清理方式，利用群集分析與迴歸分析兩種方式各項分群的結果能提高書目探勘的有效性；小節部分則統整本章中各項研究過程之記錄提供一個概觀的回顧。

第一節　問卷資料對應與描述

本研究使用之資料內容為某公司 95、97 年兩年度之問卷資料，原始資料各包含 1646 與 284 個欄位，其問卷透過 Excel 整理其記錄的資料類型。兩年度的欄位數相差甚大，並且須配合其系統中之使用者年度記錄作為資料整合的依據。其記錄方式包含有名目尺度、等距尺度、李克特量表等紀錄方式，與圖書館事業中常見的外部資料紀錄方式相符。並且資料欄位數與問卷題項較圖書館事業中的問卷題項來得多，因此本次透過此兩年度問卷作為資料來源，表 4-1-1 為圖書館問卷記錄方式與本次資料問卷對應題項。

表 4-1-1 圖書館問卷記錄方式與本次資料問卷對應題項

圖書館記錄方式	此研究資料之問卷題項
名目尺度	Q1,Q2,Q3,Q4,Q5,Q7,Q8,Q9,Q12,Q14,2,5,6,13,14,15,16,21,22
等距尺度	Q4,Q5,Q10,Q10-1,Q11-1,Q12,11,13,22
李克特量表	Q10-4,19,23,24,

在本研究中名目尺度題項為：請問您家中是屬於哪一種住戶類型？

□(1) 空戶（無人居住，例如尚未遷入之新屋或久無人活動之空屋）□(2) 住戶（純住宅） □(3) 住商合一（店面與住家在一起，例如：前面雜貨店、後面為住家等） □(4) 非住戶（樓梯間、大樓管理間、養豬間、魚池等，不可住人但設有電表的空間）

等距尺度題項為：請問您家中(或店中)多久有打算添購或換購新的冷氣機？

□(1) 半年內 □(2)一年內 □(3)兩年內 □(4)不一定

李克特量表題項為：請問貴用戶未來二年有意願購買變頻電冰箱嗎？

□非常有意願 □有意願 □普通 □沒意願 □非常沒意願

此研究所使用的資料中，包含有圖書館外部資料較少使用的等比尺度，其題項分別為：7, 8, 13, 14, 16, 22，等比尺度在本研究中題項為：夏季平均每次（每 2 個月）電費：_____元/2 月。

並且將其依照年度與使用者類型加以區分如圖 4-1-1。

圖 4-1-1 兩年度與使用者類型示意圖

依照所分類的兩個年度資料與 A、B 兩型使用者做為區隔，分別對於其各自做統計描述如表 4-1-2、表 4-1-3、表 4-1-4 與表 4-1-5：

表 4-1-2　95 年度 A 型使用者使用季節概述

類型		1 使用季節	2 使用季節	3 使用季節	4 使用季節
1	平均數	1027.53	1671.00	1007.42	1101.02
	個數	748	748	748	748
	標準差	2395.852	4416.151	3092.803	2666.357
	變異數	5740105.746	19502392.653	9565429.591	7109461.27
2	平均數	854.74	1376.65	847.71	923.86
	個數	2764	2764	2764	2764
	標準差	1116.375	2477.906	2583.454	2157.458
	變異數	1246292.110	6140016.324	6674232.055	4654624.22
3	平均數	942.18	1707.09	1059.25	1101.12
	個數	1054	1054	1054	1054
	標準差	1328.412	2558.554	1302.747	1641.948
	變異數	1764677.825	6546197.054	1697151.042	2695991.89
4	平均數	1140.02	1889.34	1154.65	1243.76
	個數	279	279	279	279
	標準差	2725.270	3591.244	1810.672	2667.852
	變異數	7427094.399	12897030.757	3278534.178	7117436.84
5	平均數	3977.50	5982.09	3407.93	4241.70
	個數	149	149	149	149
	標準差	15629.471	23685.372	12802.511	17377.283
	變異數	244280363.914	560996845.891	163904288.685	30196995.66
總和	平均數	1008.63	1656.96	1010.23	1104.94
	個數	4997	4997	4997	4997
	標準差	3138.086	5062.923	3271.861	3722.912
	變異數	9847581.525	25633184.635	10705071.515	1386007.5

在 95 年 A 型使用者的使用情況部分，皆以第二個使用季節的使用平均數最高，其五類型使用者各自平均為 1670、1376、1707、1889 與 5982，而此類型之平均為1656，而標準差在四個季節之中，皆高於其對應使用季節的平均數，

顯示其資料內容分部極為分散，此外，在此分類中五類使用者之中以第 5 類使用者其平均數最高。

表 4-1-3　95 年度 B 型使用者使用季節概述

類型		1 使用季節	2 使用季節	3 使用季節	4 使用季節
1	平均數	6353.63	10553.11	5968.94	7449.73
	個數	146	146	146	146
	標準差	16719.048	26765.078	15431.861	18541.627
	變異數	279526552.06	716369405.629	238142326.32	343791945.59
2	平均數	3471.65	5730.07	3190.83	3772.43
	個數	463	463	463	463
	標準差	12287.282	18128.621	9901.977	11982.448
	變異數	150977298.34	328646900.00	98049143.720	143579050.77
3	平均數	5302.25	9236.71	5178.92	6420.01
	個數	157	157	157	157
	標準差	22729.741	36138.380	19531.798	24961.214
	變異數	516641121.40	1305982476.11	381491137.512	623062204.32
4	平均數	4305.73	7833.05	4542.56	5072.13
	個數	131	131	131	131
	標準差	14101.581	26975.064	15832.647	16838.968
	變異數	198854581.35	727654096.4	250672700.9	283550831.99
5	平均數	2473.29	3821.99	2252.07	2659.16
	個數	150	150	150	150
	標準差	6329.982	8481.314	4910.518	6345.571
	變異數	40068672.155	71932693.664	24113186.995	40266270.028
總和	平均數	4105.78	6911.99	3907.44	4681.03
	個數	1048	1048	1048	1048
	標準差	14635.591	23327.970	13006.204	15724.229
	變異數	214200524.0	544194202.010	169161353.57	247251362.63

在 95 年 B 型使用者的使用情況部分，也是以第二個使用季節的平均數最高，其五類型使用者中其各自的平均數為 10553、5730、9236、7833 與 3821，此分類的平均為 6911，而在這分類中五類使用者的使用概況則是以第 1 類使用

者的平均數最高，與 95 年 A 型使用者相同的是其標準差相對於其平均數而言相當大，顯示 95 年 B 型使用者之資料內容極為分散。

表 4-1-4　97 年度 A 型使用者使用季節概述

類型		1 使用季節	2 使用季節	3 使用季節	4 使用季節
1	平均數	1101.40	1788.79	711.18	910.81
	個數	734	734	734	734
	標準差	4188.344	5706.736	2451.345	2911.176
	變異數	17542227.402	32566837.034	6009093.371	8474944.727
2	平均數	1262.68	2105.69	789.84	1104.69
	個數	2569	2569	2569	2569
	標準差	6331.352	6668.944	2164.774	5054.483
	變異數	40086013.780	44474812.478	4686244.884	25547796.747
3	平均數	1351.90	2296.15	881.01	1301.57
	個數	1230	1230	1230	1230
	標準差	7622.201	7622.084	3253.310	7182.154
	變異數	58097953.924	58096158.155	10584024.456	51583341.765
4	平均數	1118.92	2034.96	787.92	1039.39
	個數	486	486	486	486
	標準差	2325.354	3637.110	1681.174	2553.676
	變異數	5407269.408	13228572.200	2826345.601	6521262.387
5	平均數	1532.87	2565.08	953.67	1374.21
	個數	526	526	526	526
	標準差	5927.953	7828.545	3266.986	4782.971
	變異數	35140627.364	61286116.392	10673196.260	22876812.950
總和	平均數	1272.89	2141.16	814.18	1141.50
	個數	5555	5555	5555	5555
	標準差	6124.910	6686.517	2561.767	5205.541
	變異數	37514528.224	44709506.202	6562651.288	27097658.470

在 97 年 A 型使用者的使用情況部分，如同 95 年 A 型使用者在使用季節的部分，也是以第 2 季較高，而標準差也遠大於平均數，而在此分類中的五類使用者也是以第 5 類之使用者的平均數最高。

表 4-1-5 97 年度 B 型使用者使用季節概述

類型		1 使用季節	2 使用季節	3 使用季節	4 使用季節
1	平均數	2681.61	3870.32	1593.63	2280.79
	個數	87	87	87	87
	標準差	7179.620	7394.289	3816.507	6039.395
	變異數	51546941.101	54675509.802	14565724.677	36474293.608
2	平均數	1947.72	3560.49	1448.73	2126.61
	個數	235	235	235	235
	標準差	4967.944	7512.069	3119.973	5334.394
	變異數	24680470.329	56431179.610	9734228.428	28455759.708
3	平均數	1771.28	3359.11	1322.81	1817.17
	個數	114	114	114	114
	標準差	4390.602	6286.662	2713.462	4543.842
	變異數	19277381.584	39522117.766	7362877.874	20646502.600
4	平均數	1227.95	2307.05	857.63	1045.15
	個數	40	40	40	40
	標準差	1543.750	2537.355	939.692	1254.684
	變異數	2383163.279	6438169.023	883021.984	1574232.797
5	平均數	2010.68	3838.24	1353.16	2011.82
	個數	38	38	38	38
	標準差	3925.333	6418.667	2382.721	3428.139
	變異數	15408241.789	41199282.348	5677357.326	11752139.506
總和	平均數	1977.60	3486.94	1392.69	1987.56
	個數	515	515	515	515
	標準差	5053.079	6865.815	2999.136	4966.728
	變異數	25533605.291	47139421.026	8994814.047	24668382.822

在 97 年 B 型使用者的使用情況部分，與 95 年 B 型使用者相同，是以第 2 季節的使用者之平均數最高，而其分類中也以第 5 類使用者之使用平均數為最高。

由表 4-1-2、表 4-1-3、表 4-1-4 與表 4-1-5 發現兩年度 B 型使用者平均數皆比 A 型使用者高；而 95 年度的 A、B 兩型使用者間的使用差距大於 97 年度 A、B 兩型使用者之間的差距；95 年度的 A 型使用者平均數小於 97 年度的 A 行使

用者；95 年度的 B 型使用者則大於 97 年度的 B 型使用者。另外，能歸納出兩年度在 A 型使用者中皆是以第 5 分類使用者的使用率最高;而兩年度在 B 型使用者之中使用的情形則是分別為第 1 及第 2 兩個類型使用者的使用率為最高。另外，從標準差能夠發現，兩年度的資料內容相當分散，標準差幾乎都大於平均數，而 A、B 類型使用者皆在第 2 季節的時候使用率會遠大於其他季節，如表 4-1-6。

表 4-1-6 各年度、各類型使用率最高者之整理。

	95 年度 A 型使用者	95 年度 B 型使用者	97 年度 A 型使用者	97 年度 B 型使用者
使用率 最高之季節	第 2 季	第 2 季	第 2 季	第 2 季
使用率 最高之類型	第 5 類	第 1 類	第 5 類	第 2 類

本研究將圖書館事業外部資料與本研究中的資料集做問卷題項的對應，而本研究中原始資料能完全對應至圖書館事業中的外部資料紀錄方式。因此，能夠利用本研究之資料集進行測試，實行本研究資料清理方式，而圖書館事業中的常見紀錄方式為名目尺度、等距尺度與李克特量表三種，在本研究資料集皆具有此三種資料記錄。

第二節　清理方式

本節描述在資料清理時所進行的詳細內容，包含將 95、97 兩年度問卷資料整合、資料轉換、資料刪除與建立具概念階層資料集。

一、 兩年度問卷資料整合

原始資料是以 95 年度及 97 年度問卷調查所蒐集的問卷內容為基礎，輔以 95 年度與 97 年度樣本用戶對應的資料，為此次資料探勘之原始資料。95 年度有效樣本共有 6045 個；97 年度有效樣本共有 6070 個，兩年度在問卷收集的資料內容相差甚大，編碼方式也不相同。因此，在原始資料整理部分便先由相同的欄位做為初步的清理方向。

資料整合部分將兩年度資料欄位相同部分加以整合，並針對資料欄位相似的部分加以整合，例如：97 年的行政區與 95 年的行政區域。此項資料整合的目的在於將兩年度相似的資料整合為相同的資料格式以及編碼方式，以便於往後再進行資料探勘之時可以做為兩年度資料的彙整探勘。此外，資料整合部分仍需考慮將樣本戶的問卷資料與其該年度的使用者資料做整合，以利後續在進行資料轉換時可以將使用者資料的因素考慮入內，避免資料在資料轉換時無法將原有的使用者資料與對應的樣本問卷接合。透過相同的資料欄位可以將資料先做初步的統整，表 4-2-1 為欄位相同中可直接合併資料：

表 4-2-1 兩年度問卷資料欄位對應表

95 年度資料欄位	97 年度資料欄位
Q1	VAR00004
Q301	VAR00006
Q5191	VAR00126
Q5091	VAR00133
Q5101	VAR00140

95 年度資料欄位	97 年度資料欄位
Q5211	VAR00147
Q5081	VAR00154
Q5251	VAR00161
Q5201	VAR00168
Q5261	VAR00175
Q4051	VAR00182
Q4071	VAR00187
Q5171	VAR00192
Q5161	VAR00197
Q4061	VAR00202
Q5151	VAR00207
Q5141	VAR00212
Q4041	VAR00227
Q5121	VAR00232
Q5131	VAR00237
Q9	VAR00396

兩年度問卷資料共有 21 項記錄項目相同，但其問卷編碼方式則不一定完全一致，如：95 年度問卷資料，區域是分成北中南東之後，在加上其每個區域的流水號最為其區域的編碼方式。

二、兩年度問卷資料轉換

而 97 年度問卷資料的區域是以 25 個縣市分別給與流水號的方法編碼，因此兩年度在編碼部分不盡相同，在問卷題項相同的情況之下，仍需要將兩年度的編碼修正為相同的編碼方式才能夠使用。其編碼修改主要針對區域做修正，編碼修改為求統一以方便資料探勘結果分析，以 97 年度問卷資料的編碼為主，表 4-2-2 為區域修改項目與修改後的編碼。

表 4-2-2 區域編碼修改表

區域	95 年度編碼	97 年度編碼	合併後編碼
臺北縣	11	1	1

區域	95 年度編碼	97 年度編碼	合併後編碼
宜蘭縣	17	2	2
桃園縣	14	3	3
新竹縣	16	4	4
苗栗縣	21	5	5
臺中縣	23	6	6
彰化縣	24	7	7
南投縣	26	8	8
雲林縣	25	9	9
嘉義縣	32	10	10
臺南縣	34	11	11
高雄縣	36	12	12
屏東縣	37	13	13
臺東縣	38	14	14
花蓮縣	27	15	15
澎湖縣	40	16	16
基隆市	13	17	17
新竹市	15	18	18
臺中市	22	19	19
嘉義市	31	20	20
臺南市	33	21	21
臺北市	12	22	22
高雄市	35	23	23
金門縣	41	24	24
連江縣	39	25	25

兩年度資料整合部分主要針對兩年度問卷資料題項中，紀錄項目相同的部分做為整合的目標，若能直接整併於同一項目之下的即直接整合，如：問卷題

像中,所有以數量為問項的問題,而其他需要整合的部分主要以區域問題為主，由於兩年度的問卷資料在進行區域編碼時的不同而需要整合，以方便後續在彙整探勘時能夠共同使用兩年度資料。

三、資料刪減

此部分主要去除在資料內容中的雜訊，以偏誤值、缺漏值與重複值為主要清理的內容，以下將分述其清理方式與情況：

(一) 去除偏誤值

在探勘中去除偏誤值是一項重要的工作，偏誤值會造成其資料內容的複雜性同時無法找出資料的共同性，因此在去除偏誤值除了錯誤質的問題以外，偏誤值過多或過大也是必須注意的。本研究問卷中偏誤值如：季節使用值上限部分 10000 以上，對於本研究資料而言，偏離平均值過多；而季節使用值 50 以下，則是由於在平均數兩個標準差後即為負數，但從資料內容而言，季節使用值小於 50 屬於非自然現象，因此將偏誤值加以刪減。而 95 年度資料所刪減筆數為 189 筆資料，而 97 年度資料刪減筆數為 522 筆資料。

(二) 去除缺漏值

缺漏值的處理以兩種方式為主：刪除該筆資料、補回缺漏值。

刪除資料的條件為當無法判斷其資料之正確記錄內容或填寫資料有相互矛盾之情形，如：勾選無使用過圖書館數位學習服務，但其後仍選擇其使用數位學習服務之頻率等情況視為填寫資料有相互矛盾之情形，因此予以將此筆資料刪除；而補回缺漏值的方式包含有回填原始資料內容與回填平均數，此研究所使用的資料來源是以

問卷資料為主，而不涉及其問卷轉錄為資料的過程，並無法回填原始資料，因此在缺漏值的部分僅以回填平均數為主。

(三) 去除重複值

此研究中去除重複值的方式以用戶的識別編號為標準，若兩筆資料其用戶的識別編號相同，即認為其兩筆資料為相同使用者所填寫，應予以將其資料刪除避免資料重複。

四、產生具概念階層資料集

圖書館事業中外部資料以問卷為主，而問卷資料內容則大多包含有使用者個人資料，如：居住地區、年齡、性別；另外也包含使用圖書館服務之經驗、頻率等資料內容；此部分透過多變項使用者類型、區域使用者類型、雙變項使用者類型作為分群的基準，以符合圖書館外部資料中多數能做為分群之標準：

(一) 多變項使用者分群

多變項使用者依照其擁有特性作為分群標準，瞭解擁有不同特性的使用者對於其使用資源是否有特定的影響，多變項使用者類型包含以從事職業、居住類型、居住環境等特性為分類標準，其從事職業詳細內容包括使用者其所從事的職業是製造相關、專業技術人員、學生、教師、軍公教人員等；所居住類型包含其居住地為平房、二樓以上透天厝、六樓以下公寓、七至十一樓大廈、十二樓以上大廈等。在本研究中多變項使用者以五個變項為分群基準，此五類型使用者，95 年度與 97 年度資料皆以第 2 類型居多，而資料量最少的皆是第 5 類型。其樣本詳細數量如表 4-2-3：

表 4-2-3 五類型使用者，兩年度問卷人數

類型	95 年度樣本人數	97 年度樣本人數
第 1 類型	894	821
第 2 類型	1317	2804
第 3 類型	1099	1344
第 4 類型	1271	526
第 5 類型	299	564

(二) 以區域做為分群標準，主要分為北中南三區：

此分類主要透過不同區域的分群方式了解不同區域的概況。北區包含宜蘭縣、基隆市、台北縣、台北市、桃園縣；中區包含：新竹縣市、苗栗縣市、台中縣市、彰化縣市、南投縣、花蓮縣、台東縣；南區包含：雲林縣、嘉義線、台南縣市、高雄縣市、屏東縣市、金門縣、澎湖縣、馬祖縣，其中外島地區由於使用概況與南部地區相似，故分群進入南區。分區資料兩年度皆以北區的人數最多，而以中區的人數最少顯示並且在百分比的分布差距上相去不遠，約在1~2%之間。此分群方式主要瞭解在不同區域之中會不會各自擁有其特定的屬性，與原本的縣市分群相比能了解是否有北中南三區各自的情況發生。表 4-2-4 為不同分區中不同年度的樣本數量概況：

表 4-2-4 北中南區概況表

	95 年度樣本人數	97 年度樣本人數
北區	2169(35.8%)	2220(36.5%)
中區	1740(28.7%)	1797(29.5%)
南區	2135(35.3%)	2054(33.8%)

(三) 以雙變項類型使用者為分群標準：

此雙變項類型的使用者，兩年度分群樣本數如表 4-2-5，常見雙變項的問卷類型包括男與女、營業與非營業等變項。雙變項的分群方式

在所有的問卷問像中幾乎都會出現，尤其以男和女的區分情形最為常見，其他可能出現的雙變項問題像是以有無為詢問方式的問題，如：請問是否有使用過本館的數位學習專區？此類問項也常使用雙變項的問卷內容，但較少做為分群的依據。此雙變項類型的使用者分群主要瞭解兩個分群其各自所屬的屬性以及概況。

表 4-2-5 A、B 型使用者概況表

	95 年度樣本人數	97 年度樣本人數
A 型使用者	4997	5555
B 型使用者	1049	515

(四) 同類合併

同類合併的情況使用在當類別過多時，而這些類別都屬於需要考慮的因素，但過多的類別對於探勘而言，會增加資料的複雜情形，此時則須要將這些考慮的因素，以類似意義做為基礎加以合併，避免太多的因素投入探勘之中，導致探勘結果無法聚焦。例如：圖書館事業之中，借閱書籍的數量、文獻傳遞的服務、館際互借這些情形都算是借閱圖書資訊的一種方式，在這時候可以將其歸類為一項，如：閱覽服務數量這樣可以簡化三個類別的內容並且也可以不完全失去原先資料的情形，期望從中探勘出現規則。本研究中將五項因素加以合併為一項，期望減低過多的項目導致資料複雜性過高而探勘結果不佳。

而整體的資料清理動作包含資料對應、資料整合、資料轉換，資料刪減，接著將 95、97 兩年度資料加以建立具概念階層的資料級，分別依照雙變項、

多變項、區域變項與同類合併方式產生概念階層的動作。圖 4-2-1 為資料清理後資料分群狀況示意圖。

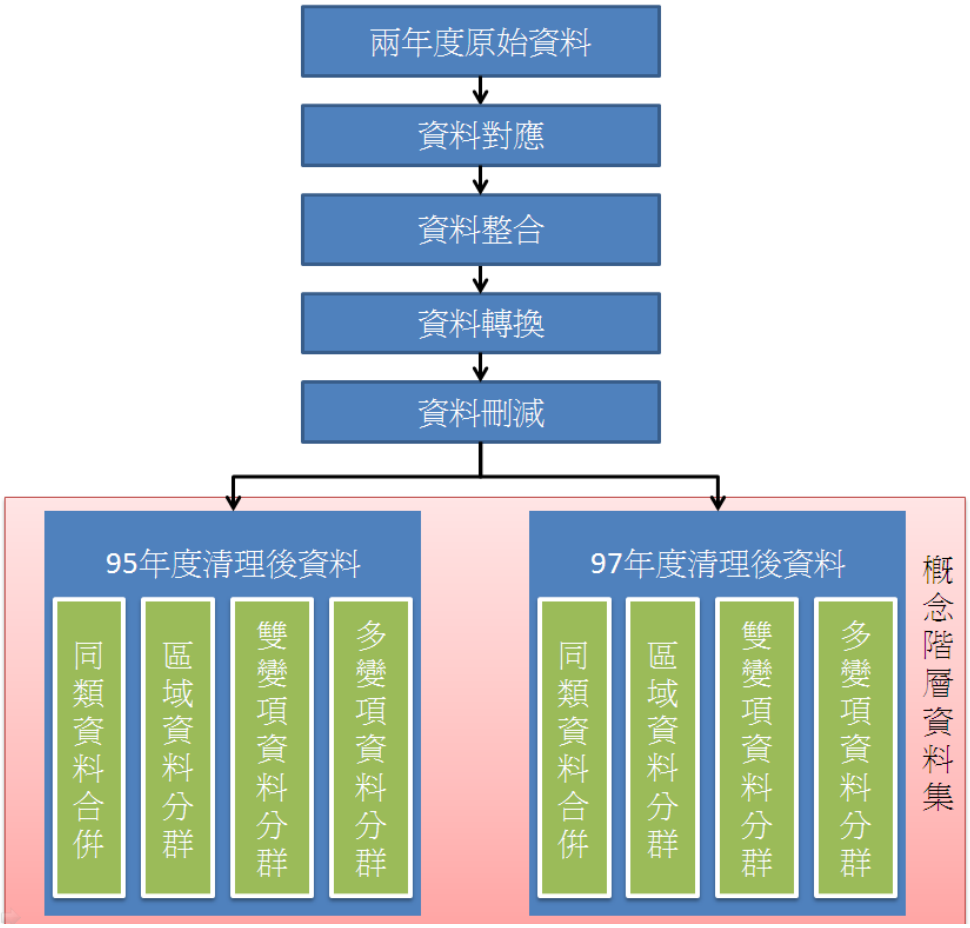

圖 4-2-1 資料清理後資料分群狀況示意圖。

第三節　資料清理結果評估

本研究將清理過後之資料，分別以群集分析和迴歸分析兩種資料探勘方式進行評估此研究中資料清理的方法是否有助於書目探勘的進行，並利用 Microsoft SQL server Enterprise 版中的商業智慧採礦工具進行資料探勘。另外，本研究僅討論探勘結果是否因為本研究資料清理過後而使資料探勘的良率提升，並不討論其軟體之特定參數的因素，以下分就群集分析與迴歸分析兩種方式分別敘述：

一、群集分析

(一) 清理前資料探勘之結果

資料清理前 95、97 年資料試行群集分析，其最大群集與次大群集的其解釋變數的機率值分別為 95 年：41%、38%、40%、33%、31% 與 97 年：39%、39%、42%、36%、35%。

可以發現到兩年度未清理資料的最大群集與次大群集的變數解釋機率值皆不高，其不同群集間的差異度不大，大約在 30~40%左右，

(二) 清理後資料探勘之結果

資料清理後結果分就兩年度資料刪減後的其解釋變數的機率值、實行概念階層解釋變數的機率值。

95 年度在資料刪減之後其解釋變數的機率值提升為 50%、44%、49%、52%、56%。

以區域資料分類之後其解釋變數的機率值提升為 63%、68%、61%、66%、65%。

以雙變項分群之後其解釋變數的機率值分別改變其解釋變數的機率值為 67%、71%、71%、72%、69%與 68%、73%、69%、74%、77%。

而多變項分群之後其解釋變數的機率值分別為68%、73%、69%、74%、77%；74%、71%、77%、73%、72%；75%、74%、71%、69%、73%；70%、73%、72%、71%、72%；72%、73%、68%、71%、79%。

將同類資料合併後解釋變數的機率值為 71%、75%、72%、71%、74%。97 年度在資料刪減之後其解釋變數的機率值提升為 49%、47%、44%、45%、53%。以區域資料分類之後其解釋變數的機率值提升為 61%、62%、68%、69%、71%。以雙變項分群之後其解釋變數的機率值分別改變其解釋變數的機率值為 63%、68%、70%、69%、73%與 69%、78%、79%、80%、77%。

多變項分群之後其解釋變數的機率值分別為 68%、75%、76%、72%、77%；74%、72%、75%、76%、72%；69%、72%、73%、70%、72%；71%、74%、73%、75%、78%；78%、74%、70%、75%、71%。將同類資料合併後解釋變數的機率值為 73%、71%、72%、69%、72%。

表 4-3-1 為 95、97 年度清理後解釋變數機率值的對照表

	95 年度 解釋變數的機率值					97 年度 解釋變數的機率值				
資料刪減	50%	44%	49%	52%	56%	49%	47%	44%	45%	53%
區域資料分類	63%	68%	61%	66%	65%	61%	62%	68%	69%	71%
雙變項分群 1	67%	71%	71%	72%	69%	63%	68%	70%	69%	73%
雙變項分群 2	68%	73%	69%	74%	77%	69%	78%	79%	80%	77%
多變項分群 1	68%	73%	69%	74%	77%	68%	75%	76%	72%	77%

	95 年度 解釋變數的機率值					97 年度 解釋變數的機率值				
多變項分群 2	74%	71%	77%	73%	72%	74%	72%	75%	76%	72%
多變項分群 3	75%	74%	71%	69%	73%	69%	72%	73%	70%	72%
多變項分群 4	70%	73%	72%	71%	72%	71%	74%	73%	75%	78%
多變項分群 5	72%	73%	68%	71%	79%	78%	74%	70%	75%	71%
同類合併	71%	75%	72%	71%	74%	73%	71%	72%	69%	72%

由表 4-3-1 可以發現到在資料刪減之後,兩年度群集中變數的解釋機率值皆以上升至 50%左右；而將區域分群之後則變數的解釋機率值上升至 65%左右；雙變項分群之後變數的解釋機率值也是上升至 68%左右；而多變項分群之後則可以發現變數的解釋機率值上升至 70%；同類合併之後變數的解釋機率值也大約上升至 70%。

(三) 清理前後比較

95 年度清理前後比較，可以發現到在將偏誤值、重複值、缺漏值去除之後，其變數的解釋機率值即有成長，而之後的各項分群對於資料探勘而言,是而在第五個變數方面去除偏誤值之後成長最多，而在各項分類分群部分，因為皆先去除偏誤值部分的因素，所以成長幅度皆高過去除偏誤值這部分，可推測此研究中的的清理為有效步驟，圖 4-3-1 為 95 年度各項清理方式的變數的解釋機率值。

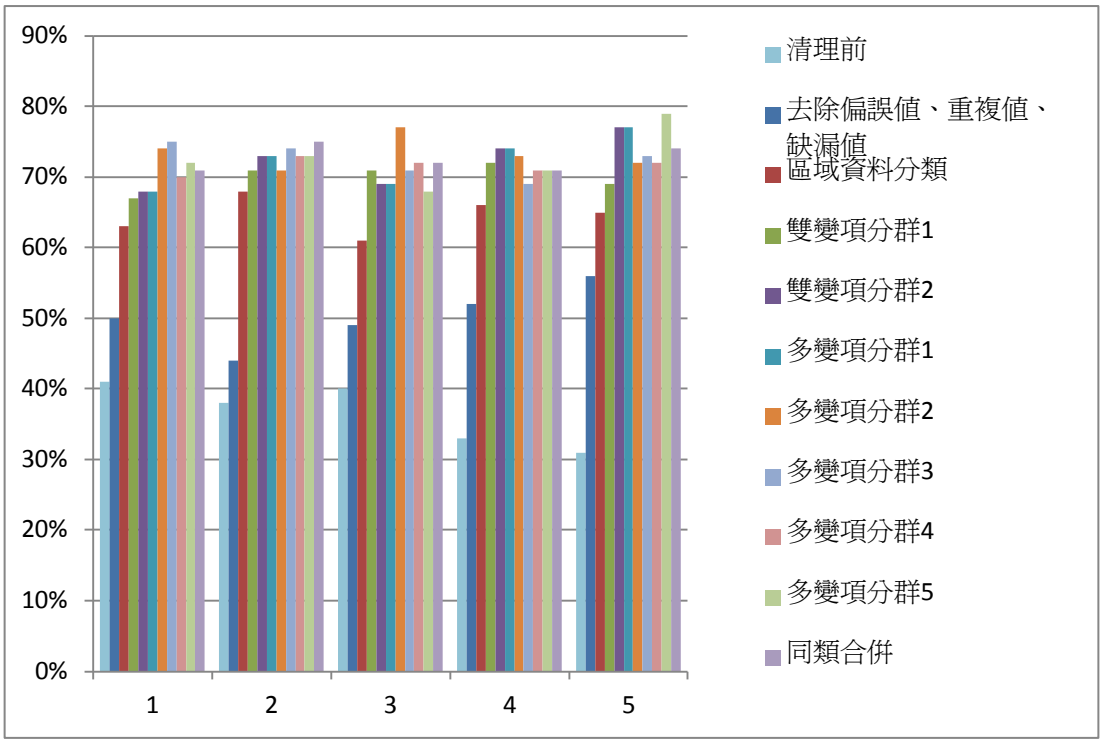

圖 4-3-1 95 年度各項清理方式的變數的解釋機率值

97 年度清理前後比較可以發現在清理前與清理過偏誤值等部分之後，呈現變數的解釋機率值上升的情況，而在各分類分群清理過後，從圖 4-3-2 可以發現其上升幅度不一，但皆是上升的情況，可以推測此次資料清理步驟是有效的。

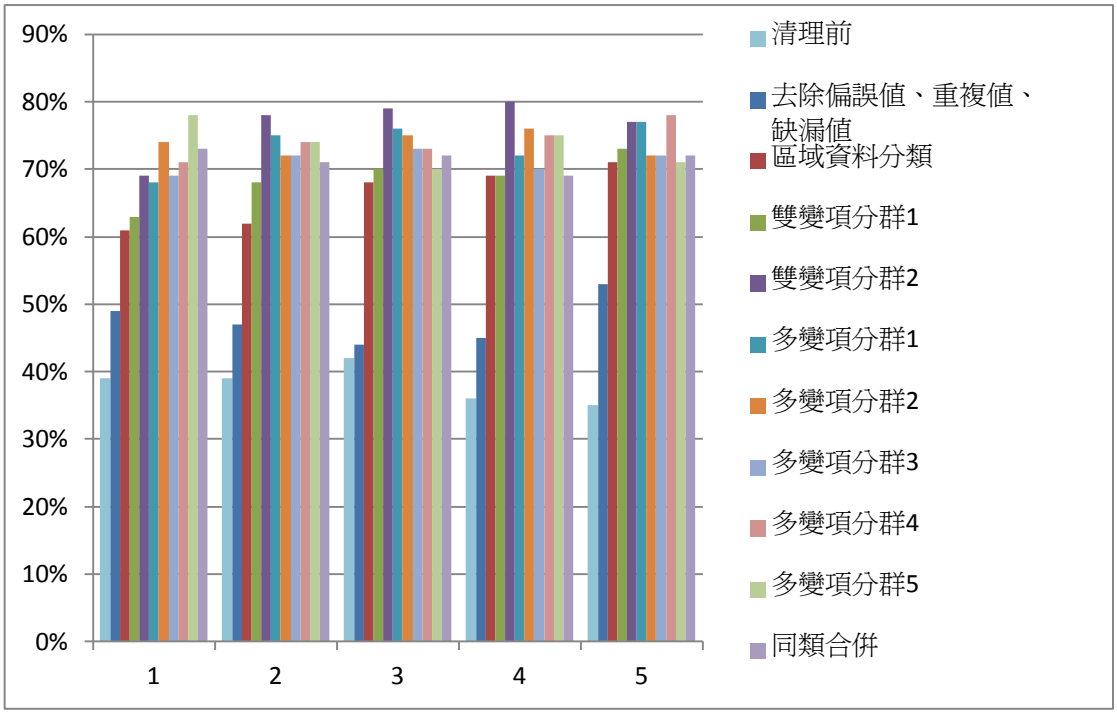

圖 4-3-2 97 年度各項清理方式的變數的解釋機率值

綜觀兩年度的清理前後之比較，可以發現到去除偏誤值、重複值、缺漏值皆有助於提升變數的解釋機率值，另外，在各分類分群的部分，變數的解釋機率值皆高過去除偏誤值的部分，但提升的情況不一。

二、迴歸分析

(一) 清理前資料探勘之結果

資料清理前 95、97 年度資料試行進行迴歸分析其 R^2 分別為 0.26 與 0.31，兩年度資料的 R^2 皆小於 0.6 顯示其在未進行資料清理前，資料內容在進行迴歸分析時可信度過低且無法顯示出其資料探勘結果能夠解釋其迴歸關係式，因此需要進行資料清理的動作。

(二) 清理後資料探勘之結果

資料清理後結果分就兩年度資料刪減後的 R^2、概念階層的 R^2 分做敘述。

95 年度在資料刪減之後其 R^2 為 0.51；以區域資料分類之後其 R^2 提升為 0.74；以雙變項分群之後其 R^2 分別改變為 0.65、0.71；而多變項分群之後其 R^2 分別為 0.70、0.73、0.86、0.69、0.77；而將同類資料合併後 R^2 為 0.77。

97 年度在資料刪減之後其 R^2 為 0.52；以區域資料分類之後其 R^2 提升為 0.77；以雙變項分群之後其 R^2 分別改變為 0.66、0.74；而多變項分群之後其 R^2 分別為 0.71、0.69、0.71、0.82、0.77；而將同類資料合併後 R^2 為 0.72。表 4-3-2 為 95、97 年度資料清理後 R^2 對照表。

表 4-3-2 為 95、97 年度資料清理後 R^2 對照表。

	95 年度 R^2	97 年度 R^2
去除偏誤、重複、缺漏值	0.51	0.52
區域資料分類	0.74	0.77
雙變項分群 1	0.65	0.66
雙變項分群 2	0.71	0.74
多變項分群 1	0.70	0.71
多變項分群 2	0.73	0.69
多變項分群 3	0.86	0.71
多變項分群 4	0.69	0.82
多變項分群 5	0.77	0.77
同類合併	0.77	0.72

由表 4-3-2 可以發現在兩年度資料的清理上，將偏誤值、重複值、缺漏值除去之後 R^2 即有改善，顯示資料清理的第一步已有效果；而

其後的不同分群方式，兩年度的各群資料皆達到 0.65 以上的水準也對資料探勘的結果有幫助，表示若將資料去除雜訊之後並將其有效分群，則對於迴歸分析中的可信度 R^2 有幫助。

(三) 清理前後比較

95 年度資料清理前後之比較，由圖 4-3-3 可以發現資料清理前與資料清理之後其迴歸分析中 R^2 的改變，在一開始將偏誤值、重複值與缺漏質去除之後其 R^2 即已提升，而其後資料清理中，各項分群的結果對於資料探勘而言影響不一，但皆是上升的情形，可以推測在此研究中資料清理的動作有助於資料探勘的進行並且能幫助提升資料探勘的有效率。表 4-3-3 為各項清理結果與位資料清理前之提升情形。

表 4-3-3 各項清理結果與位資料清理前之提升情形

	95 年度 R^2	清理後上升比例	97 年度 R^2	清理後上升比例
資料刪減	0.51	0.25	0.52	0.21
區域資料分類	0.74	0.48	0.77	0.46
雙變項分群 1	0.65	0.39	0.66	0.35
雙變項分群 2	0.71	0.45	0.74	0.43
多變項分群 1	0.70	0.44	0.71	0.4
多變項分群 2	0.73	0.47	0.69	0.38
多變項分群 3	0.86	0.6	0.71	0.4
多變項分群 4	0.69	0.43	0.82	0.51
多變項分群 5	0.77	0.51	0.77	0.46
同類合併	0.77	0.51	0.72	0.41

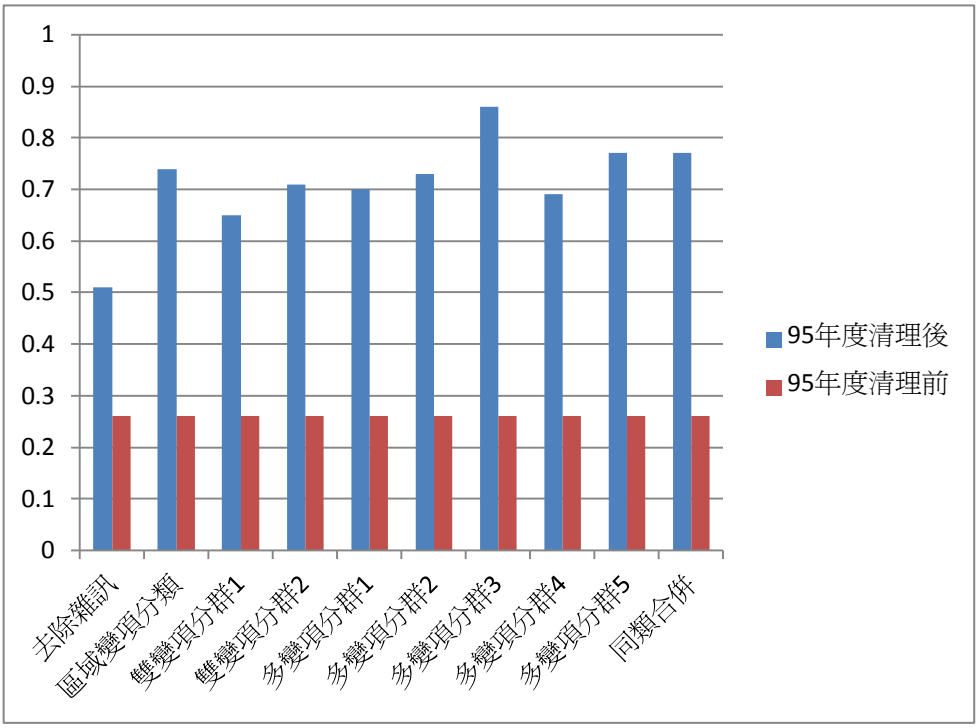

圖 4-3-3 95 年度迴歸分析 R^2 清理前後概況

97 年度資料清理前後之比較，由圖 4-3-4 可以看發現資料清理前與資料清理之後其迴歸分析中 R^2 的改變，在去除雜訊之後其 R^2 即有所改善，並且其後的各項分群方式中，迴歸分析 R^2 的變化與 95 年度資料清理前相比皆是上升的情況，顯示此研究中的資料清理步驟確實有助於迴歸分析中 R^2 的提升。

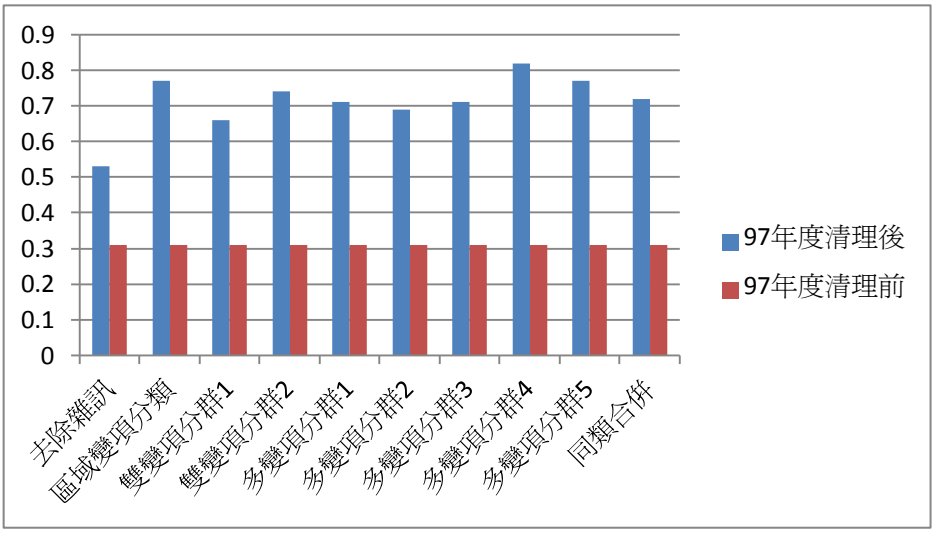

圖 4-3-4 97 年度年度迴歸分析 R^2 清理前後概況

由圖 4-3-3 和圖 4-3-4 可以看出在迴歸分析之中資料清理前與資料清理之後 R^2 的變化，在資料刪減之後，R^2 皆有所提升，而概念階層之分群方法也同時上升，但在上升情況中各概念階層之分群方法則皆高於資料刪減之後，但其上升情況不一，但從比較圖中可以發現此研究所進行的資料清理能夠使資料探勘結果提升 R^2 的數值，達到資料清理的目的。

第四節　小結

本研究利用問卷對應的方式是將此研究所使用之資料對應圖書館事業中可能面臨的問卷資料與問卷類型，透過這樣對應的方式推定這次的資料有助於在書目探勘之中進行資料清理的效果，而圖書館事業中常見的問卷題項類型包含名目尺度、等距尺度、李克特量表。而本研究所使用之資料也同時包含這樣的資料內容並加以對應。

在本研究中利用資料清理的方式清理外部性資料，而清理方法包含：資料對應、資料整合、資料轉換、資料刪減、建立具概念階層的資料集包含以區域變項分群、雙變項分群、多變項分群、同類合併等方法加以實作。

一、 資料對應

資料對應部分，問卷資料為圖書館外部性資料之大宗，透過歸納問卷資料中的題項了解到圖書館紀錄方式以名目尺度、等距尺度、李克特量表，而在本研究中資料集皆具有此三項資料記錄，並透過問卷資料對應表做歸納，表示本資料集得涵蓋圖書館問卷資料的紀錄類型。

二、 資料清理

資料清理部分分為資料刪減、資料整合、資料轉換並建立具概念階層的資料集。資料刪減部分主要處理資料偏誤值、重複值、缺漏值等，偏誤值部分主要刪減為季節使用值 50 以下與 10000 以上，兩年度共 711 筆資料；重複值部分則是刪去重複的筆數；缺漏值透過季節使用值之平均數做為回填的數值。

資料整合部分以編碼簿為基礎，將兩年度資料可直接合併的欄位合併，並且針對，同時找出兩年度資料中相似的資料欄位加以整併，以利往後資料轉換時，能夠相似資料欄位做合併。

資料轉換部分，將兩年度資料編碼不同的紀錄以 97 年度資料為主，將其固定為同一個紀錄格式，並且將避免往後在進行探勘時，因為資料紀錄格式不同導致資料結果的判讀錯誤。

建立具概念階層資料集則是透過區域變項分群、雙變項分群、多變項分群與同類合併四種方式進行，透過此四項分群方式符合圖書館問卷中常見之題項，並進行資料探勘結果的評估，確保此四項概念階層的方式能適當運行書目探勘之資料清理工作之中。

三、 探勘分析評估

將 95、97 年度問卷資料利用前述各項清理方式進行資料清理的步驟後，即可看到資料探勘結果的改善；進行資料刪除的步驟後，在迴歸分析的分析結果中，其 R2 的數值可比未進行資料清理步驟的原始資料提升大約 0.3 左右；而在變數的解釋機率值部分大約可以提升 10%，因此可以確定在資料探勘中清除雜訊是必要的，同時去除雜訊的方式也可以依照本研究中所提及的方法做處理；而去除雜訊之後分群的部分，兩年度資料呈現的結果皆是本研究資料清理方式

有助於改善資料探勘的結果，與去除雜訊之後相比 R2 大約可以提升 0.2~0.3 左右；而變數的解釋機率值部分則可以提高大約 20%左右。

第五章　結論與建議

本研究旨在透過進行外部資料的清理歸納出外部資料資料清理的方式以提升書目探勘的有效度，並透過不同分群方式實驗是否能有效提升資料探勘的結果。本章第一節為研究結論，綜述研究結論並回應研究目的與問題；第二節提出未來研究建議供往後圖書館相關研究做為參考。

第一節　研究結論

一、 去除雜訊有助於幫助書目探勘的進行

在資料清理步驟方面，本研究資料清理過程以去除雜訊為原則，其中包含有去除偏誤值、重複值、缺漏值三項，原因是雜訊最容易在資料探勘之中影響結果，因此將其去除之後，從本研究的數據而言，也有助於提升資料探勘的有效性(本研究中為 R^2 與變數的解釋機率值)，因此清理步驟上以去除雜訊值最為優先，另外，各項分群方式則視資料內容做為調整，分別可進行雙變項分群、多變項分群、區域分類、同類合併等方式，以面對不同需求的資料探勘。

二、 透過分群使書目探勘結果更加準確

清理外部性資料時，以本研究而言為達到書目探勘所需之資料內容，必須確定其所記錄資料之變項為何？若使用是區域變項則需要透過轉換方式將其加以轉換為區域資料；若是以雙變項分群方式則需要將其資料分開另行資料探勘的方式，簡化原先聚合在一起的資料，此一方法有助於減低資料複雜度，進而提升資料探勘結果的有效性；若是以多變項分群的方式也如同雙變項方式相同，但需要注意是否變項內容會過於繁多導致各分群資料量不足的情況；若是以同類合併的方式則須考慮到所合併的項目是否是該當合併的，或是其合併之後不致使資料所隱含的資訊消失。

而從本研究中發現，要使書目探勘的結果更加準確當方式，首先，需要清理資料中的雜訊包含有偏誤值、重複值與缺漏值，雜訊對於書目探勘而言會導致探勘結果不準確同時，無法使探勘結果聚焦，會導致 R^2 過低或是變數的解釋機率值減低，因此去除雜訊為書目探勘中必須之項目，而其後為提升書目探勘之準確性則必須實行不同的資料轉換方式，包含雙變項分群、多變項分類、同類合併、區域變項分群等方法這些方法對於書目探勘而言都有其有效性。

第二節　研究建議

針對上述之結論，本研究提出幾項建議，供未來在資料清理與書目探勘時做為參考。

一、外部資料收集須要有計畫性蒐集

本研究中所使用兩年度資料因為資料內容收集迥異，因此在資料整合與資料轉換上僅能擷取少部分的資料作為探勘之用，而其他記錄之資料，因為所記錄資料項目不同或沒有相對應的資料可以合併，因此資料減低其被使用率，若後續研究者能有計畫蒐集資料，使各年度資料能夠加以整合，則更可以看出資料變化的趨勢。

二、確立探勘目的後再進行資料清理

本研究中資料清理方式為努力達到資料探勘中結果的準確，在不確定探勘目的為何的情況下進行資料清理，因此實行了許多不同方式的分群與資料合併，希望探討每個分群方式結果是否皆能提升資料探勘的準確度，但多種的分群方式也導致探勘主題分散，較不專一，往後研究若可以確定探勘主題之後加以清理或許能有更明確的清理步驟提出。

三、探勘結果各概念階層分群方式不一的情況可再深入探討

本研究中利用圖書館問卷中常見之各項分群方式進行概念階層的實行，並針對每項分群方式逐一評估其結果，但從結果來看，雖對於資料探勘而言，準確性有所提升，但各項分群結果準確度成長幅度不一，建議未來研究可針對各項分群方式分別評估其成長原因，以及如何使特定分群方式的探勘結果準確性成長。

參考文獻

中文參考文獻

卜小蝶（2001）。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，66，59-72。

卜小蝶（2002）。使用者導向之圖書分類關聯分析研究。圖書資訊學刊，17，81-94。

王毓菁（2002）。圖書館閱覽者群組潛在特性探勘資訊系統。華梵大學工業管理系碩士論文。未出版，台北。

余明哲（2002）。圖書館個人化館藏推薦系統。國立交通大學資訊科學系碩士論文。未出版，新竹。

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

呂家賢（2005）。運用資料探勘技術於大學圖書館圖書資源推廣利用之研究。銘傳大學管理研究所碩士在職專班碩士論文。未出版，桃園。

林湧順（2005）。以資料探勘技術探討高中生使用圖書館之行為模式--以國立台灣師範大學附屬高級中學為例。國立臺灣師範大學社會教育學系碩士論文。未出版，台北。

柯皓仁、楊雅雯、吳安琪、戴玉旻（2002）。個人化及群體化圖書館資訊服務初探。國家圖書館館刊，91(1)，161-195。

曹健華（2002）。應用資料探勘技術於數位圖書館之個人化服務及管理。南華大學資訊管理學研究所碩士。未出版，嘉義。

陳建傑（2009）。基於借閱目的之資料清理機制研究—以興趣目的為例。國立台灣師範大學圖書資訊學研究所碩士論文。未出版，台北。

陳建銘（2001）。類神經網路於 Web Mining 之應用。國立台北科技大學商業自動

化與管理研究所碩士學位論文。未出版，台北。

曾勇森（2002）。利用資料探勘技術增進圖書館之服務效益。南台科技大學資訊管理系碩士論文。未出版，台南。

曾憲雄、蔡秀滿、蘇東興、曾秋蓉、王慶堯（2005）資料探勘。台北，旗標。

黃俊榮（2005）。利用分群化技術發掘圖書館書籍借閱之推薦服務。南台科技大學資訊管理系碩士。未出版，台南。

楊詠喬（2010）。應用資料探勘技術於圖書館藏推薦之研究。醒吾技術學院資訊科技研究所。未出版，台北。

鄧世昌（2009）。以多層次關聯規則探勘技術探索圖書館使用者借閱行為模式。樹德科技大學資訊管理研究所碩士論文。未出版，高雄。

賴雨廷（2002）。利用資料探勘技術應用於圖書館新書推薦之研究。國立中山大學資訊管理學系研究所碩士論文。未出版，高雄。

戴玉旻（2001）。圖書館借閱記錄探勘系統。國立交通大學資訊科學研究所碩士論文。未出版，新竹。

謝建成、魏儀禎（2003）。資料倉儲於圖書館管理應用之研究—以分析館藏圖書探討圖書採購決策。教育資料與圖書館學，40: 3，345-356。

謝賓帆（2008）。利用興趣加權分類技術發掘書籍借閱之適性化推薦。南台科技大學資訊管理系碩士論文。未出版，台南。

羅子文、柯皓仁（2007）。Web 2.0 概念的圖書館個人化推薦系統。台北市立圖書館館訊，24(4)，1-30。

英文參考文獻

Banerjee, K.（1998）. Is data mining right for your library? Computers in Libraries, 18（10）, 28-31.

Brauer, B. (2000). Data Quality: Spinning Straw into Gold. Retrieved Mar, 20, 2010,

from http://www2.sas.com/proceedings/sugi26/p117-26.pdf

Famili, A., Shen, W.M., Weber, R.,& Simoudis, E.(1997). Data preprocessing and intelligent data analysis. Intelligent Data Analysis, 1 (1),1–28.

Han J.& Kamber, M. (2000). Data Mining: Concepts and Techniques. San Francisco, CA: Morgan Kaufmann.

Jermyn, P., Dixon, M., Read, B. J.(1999). Preparing clean views of data for data mining. Retrieved Jan, 10, 2010, from http://www.ercim.eu/publication/ws-proceedings/12th-EDRG/EDRG12_JeDiRe.pdf

Laudon, K. C.（1986）. Data quality and due process in large interorganizational record systems. Communications of the ACM, 29（1）, 4-11.

Nicholson, S.（2003）. The Bibliomining Process: Data Warehousing and Data Mining for Library Decision Making. Information Technology & Libraries, 22（4）, 146-151.

Wu. C.H (2003). Data mining applied to material acquisition budget allocation for libraries: design and development, Expert Systems with Applications, 25(3), 401-411.