國立交通大學

理學院網路學習碩士專班

碩 士 論 文

結合資料倉儲與資料探勘的技術
分析中小學數位落差

Applying Data Warehousing and Data Mining Techniques

to Analyze The Digital Divide of K-12

研 究 生：蕭斯聰

指導教授：曾憲雄　博士

中 華 民 國　九 十 三　年 六 月

結合資料倉儲與資料探勘的技術
分析中小學數位落差
Applying Data Warehousing and Data Mining Techniques to Analyze The Digital Divide of K-12

研 究 生：蕭斯聰　　Student：Hsi-Tsung Hsiao

指導教授：曾憲雄　　Advisor：Shian-Shyong Tseng

國 立 交 通 大 學
理學院網路學習碩士專班
碩 士 論 文

A Thesis
Submitted to Degree Program of E-Learning
College of Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in

Degree Program of E-Learning

June 2004

Hsinchu, Taiwan, Republic of China

中華民國九十三年六月

結合資料倉儲與資料探勘的技術
分析中小學數位落差

Applying Data Warehousing and Data Mining Techniques to Analyze the Digital Divide of K-12

研究生：蕭斯聰　　指導教授：曾憲雄博士

國立交通大學

理學院網路學習碩士專班

摘要

對於一個商業或研究機關團體在從事主題研究後，累積了大量的研究資料時，要如何有效的管理及善用這麼龐大的重要資源，成為每一位研究工作者所要面對的課題。本論文主要在提出如何應用「資料倉儲」(Data Warehousing) 及線上分析處理(On-Line Analytic Processing, OLAP)的技術，完成「應用資料倉儲技術之問卷分析」的架構與設計，並利用「資料探勘」(Data Mining)技術來對「中小學數位落差資料」進行分析與歸納的研究。

研究過程可概分為三個階段**(1)資料前處理階段:**將蒐集到的資料進行過濾、整合、轉換等資料前處理程序後，成為可適用於資料倉儲的資料格式。**(2)資料倉儲建置階段:**將資料前處理程序後的資料，建立成具有多維度資料模型結構的資料方塊體（Data Cube）後，存入「中小學數位落差資料倉儲」。**(3)線上分析及資料探勘階段:**在多維度資料倉儲建置完成後，便可進行「線上分析」及「資料探勘」的處理，產生有意義的資訊或特徵。而且，使用「資料探勘」的技術，結合受訪者的相關背景特徵資訊來進行群集(Clustering)分析，利用分群後的群集個體差異，來建立可代表各群集的「數位學習落差形成因素」決策樹(Decision Tree)，再從決策樹歸納出有效的規則，可供學生、教師、學校作進一步的調查或研究之用，也可作為推行中小學資訊教育計畫決策的依據或相關研究工作的使用。

關鍵字：數位落差(Digital Divide)、資料倉儲(Data Warehousing)、線上分析處理(On-Line Analytic Processing, OLAP)、資料探勘(Data Mining)

Applying Data Warehousing and Data Mining Techniques To Analyze the Digital Divide of K-12

Student: Hsi-Tsung Hsiao Advisor: Dr. Shian-Shyong Tseng

Degree Program of E-Learning

College of Science

National Chiao Tung University

Abstract

To conduct researches of some specific topics, we should firstly collect related resources. Therefore, how to manage and use such enormous and important resources becomes an issue to deal with. In this thesis, we will bring up the ideas of how to apply Data Warehousing and On-Line Analytic Processing techniques to carry out the framework of this research, and then make use of Data Mining techniques to analyze the data resources of The Digital Divide of K-12.

There are three phases in the research process including the preprocessing phase, the data warehousing phase and the OLAP and Data Mining phase**. In the preprocessing phase,** the raw data will be filtered, transformed and integrated into the suitable format for the data warehouse. **In the data warehousing phase,** a multidimensional data cube will be built based on the preprocessed data from prior phase, and then will be stored in the Data Warehouse of The Digital Divide of K-12. **In the OLAP and Data Mining phase,** after the multidimensional data warehouse has been built, the OLAP and Data Mining procedure can be performed and some meaningful results may be generated. Also, some Data Mining techniques can be applied to perform cluster analyses on the background of the interviewees. Finally, the differences of the clusters can be used to build the Decision Tree that represents the factors which form The Digital Divide of K-12. The effective classification rules extracted from the Decision Tree will help students, teachers or schools for further investigation. These results may be useful for making policy decision for the development of information education in K-12 or other related researches.

Keyword: Digital Divide, Data Warehousing, On-Line Analytic Processing. OLAP, Data Mining.

誌謝

本論文得以順利完成，首先要感謝的是我的指導教授，曾憲雄博士兩年以來在課堂上及課外的指導與教誨，無論在研究方法、論文撰寫，抑或是遇到瓶頸時都不厭其煩的給予指導，適切地引領我進入一個新的學習領域，使我在做學問與待人處世方面都有很大的精進。

更承蒙 莊祚敏教授、黃國禎教授與楊錦潭教授在口試期間不吝指正，並給予許多寶貴的意見，使得本論文更有價值與意義，同時也使我受益良多，不勝感激。此外十分感謝由曾憲雄、張維安、黃國禎教授等，所提供的中小學數位落差之相關研究資料，由於這些相關資料的協助使本論文的研究內容更加充實完整。

也要感謝理學院網路學習專班多位老師的教導，還有實驗室的各位學長陳威州、林耀聰、王慶堯、林順傑、曲衍旭學長的提攜，尤其是蘇俊銘、翁瑞鋒兩位學長的指導與鼓勵，使我在課業及待人處事上，受益匪淺。此外，實驗室中的同窗，哲青、王威、培綺、家瑜、力豪、于彰、建豪、佩琪同學等人的相互勉勵，以及實驗室與專班的助理小姐們、同學及學弟妹們的諸多幫助與鼓勵，還有所任職的學校同事、長官的支持與協助，讓我兩年的碩士生涯能夠充實而愉快地度過，謝謝各位。

最後，我要感謝在背後默默陪著我的家人與妻子靜星對我的支持與鼓勵，有你們的支持與關愛，使我能順利完成論文的研究與撰寫，於此表達無限的感謝，僅將此論文獻給所有關心我的家人、師長與朋友。

蕭斯聰謹識

2004 年 7 月於新竹交通大學知識工程實驗室

目錄

圖目錄

表目錄

第一章　緒論

隨著資訊相關產業的發展，雖然提升了許多民眾在生活上資訊化的便利，但是卻也產生了新的問題：那就是數位落差(Digital Divide)。根據經濟合作發展組織 OECD 的定義，數位落差是指存在於個人、家戶、企業在不同社經背景或居住地理區位上，其接近使用資訊科技及運用網際網路所參與的各項活動的機會差距[7]。這種差異表現在社會面上有資訊取得不易、教育機會少、工作機會少、收入偏低等。而資訊傳播科技所帶來的好處亦並非公平散佈，但其壞處卻往往集中在弱勢群體，造成資訊富人和資訊窮人，尤其是受教育機會的不公平，以及受教育環境的差異[2]。

數位落差現象存在於社會中，造成資訊、知識的吸收與技術利用的不平等，因此，政府必須對於弱勢族群提供資源與協助，以降低數位落差，為了提昇國家競爭力，全民上網、企業上網與政府 e 化服務的應用為必要的措施 [1]。另外，分析數位落差造成之原因，更是提供政府解決數位差之重要決策依據，所以本研究將針對分析的技術與方法，來進行探討。

1.1. 研究動機

世界各進國家致力於數位落差的相關研究時，資料分析方法大多是以問卷及面訪方式進行，針對人口統計變數進行抽樣統計分析[4]。然而在這些大量的資料中，例如：問卷調查資料，常隱藏著極為有用的資訊或知識，在以往的資料分析技術所用的方主要是以統計分析為主，如：敘述統計、機率論、迴歸分析、類別資料分析等皆屬之。然而傳統統計方法往往受限於問卷之設計，而且傳統統計方法是屬於假設、驗證的分析模式，並無法發現超出分析者思考範圍之資訊。

另外國內外目前針對數位落差資料以資料倉儲(Data Warehousing)、線上分析處理(On-Line Analytic Processing, OLAP)與資料探勘(Data Mining)等技術來進行分析與歸納的研究尚不多。設定資料方塊體（Data Cube）仍需資訊技術人員撰寫資料庫程式規範，傳統採問卷調查的研究者很難將分析理念由程式表達，因此需要一種資料處理系統或機制，可以協助領域專家或資料分析者從專業角度直接去分析所感興趣的欄

位量值。然而從問卷調查結果的形成因素來看，如何同時參考多種不同類型的背景資訊或現有的歷史資料，歸納出數位落差形成因素及規則，也是另一項重要議題。

1.2. 研究方法與貢獻

在本篇論文中，主要在提出應用資料倉儲技術之問卷分析架構之設計，其中應用了「資料倉儲」技術去彙集、轉換並整理資料，並搭配使用「資料探勘」技術去發覺出潛在而有用的型樣（Patterns）或規則（Rules）。整體分析架構可分為三個階段：

(1). 資料前處理階段：

在資料前處理階段，使用了資料淨化(cleaning) 處理、資料平滑(smoothing)、聚集(aggregation)與正規化(normalization)等處理，並提出了問卷題目量化轉換(Questionnaire Item Normalization and Measure Transformation, QINMT)演算法，來將問卷中不同的題型答案資料轉為可適用於資料立方體中的量值形式。

(2). 資料倉儲之建置階段：

本研究提出了**多維度概念階層知識擷取(Multiple Dimension Concept Hierarchy Knowledge Acquisition, MDCHKA)** 演算法，來擷取領域專家對問卷中的概念階層知識，其中量值概念階層知識，可以指導量值聚集(Measurement Aggregation)計算的處理,產生廣義化(Generalize)的新量值，經過這些程序後所建立的量值是一個具有較高階層概念的量值資料集合，此階段整理了資料維度與量值，並將之建置成資料倉儲中資料立方體。

(3). 線上分析與資料探勘階段：

透過建立好的資料立方體,利用線上分析工具[9],採用「**由上往下(Top-Down) 階層式的分析**」方法，挑選出各層級重要的資料變項，利用上捲(roll-up)或下探(drill-down)等 OLAP 基本查詢操作，即可進行資料立方體的線上分析。資料探勘[8]分析部分，基於資料探勘輔助系統(Data Mining Assisted System,DMAS)的核心技術基礎下[5]，本研究完成了使用**兩層式資料探勘方法**之 **DMAS 線上資料探勘系統** OLAM(On Line Analytical Mining, OLAM)，這是結合了分群演算法

分析出學校資訊能力類別，然後再使用決策樹演算法針對資訊能力類別建立決策樹，之後則可透過建立好之決策樹模型，進行預測查詢分析。透過此分析系統，分析者可以更便利的進行資料探勘分析。

除了上述的資料倉儲問卷分析架構之設計之外，關於實作部分，我們參考由曾憲雄、張維安、黃國禎教授等，所提供的中小學數位落差之相關研究資料。進行了相關的實作與驗證，由於這些相關資料的協助使本論文的研究內容更加充實完整。

綜合以上所述，本篇論文之主要研究貢獻如下：

- 提出一個可處理不同問卷題型量化問題之資料轉換演算法。
- 提出多維度概念階層知識擷取方法，以利建立資料立方體(Data Cube)。
- 結合現有線上分析處理 OLAP 工具，提出資料分析流程之架構。
- 完成 DMAS 線上資料探勘系統 (DMAS-OLAM)分析系統，輔助分析者更容易進行資料分析。

1.3. 論文架構

本篇論文共分為七章，第一章為緒論，第二章為相關文獻探討，介紹現今問卷分析方法相關的研究。第三章則說明我們提出的**應用資料倉儲技術之問卷分析架構**的設計，第四章為資料前處理之步驟，以及將問卷填答資料及相關資料轉換成量值與維度資料的前處理過程與演算法，在整理了資料維度與量值後，也說明如何將之建置成資料倉儲中資料立方體。第五章則是針對中小學數位學習落差資料立方體，進行線上分析及資料探勘系統之分析方法。第六章為系統實作，介紹我們完成之分析系統並呈現相關結果與分析，最後一章為本篇論文的結論與對未來研究的建議。

第二章　相關文獻探討

本章節主要是對數位落差分析、資料倉儲、線上分析及資料探勘等相關文獻進行探討，將介紹與數位落差分析相關的研究，如：數位落差的定義與因應方案、問卷資料分析方法，以及資料倉儲和決策支援工具的分析與介紹。

在問卷分析相關的研究中，大多是以問卷及面訪方式進行，針對統計變數進行抽樣統計分析，例如:性別、年齡、所得、種族、地區與職業等，來看各種不同族群之間的差異性[3]。但就問卷資料分析方法若再進一步研究，有以下幾種:

(1).傳統推論統計分析法:

這類的分析方法是先針對問題進行假設，經問卷及抽樣設計後進行問卷調查或面訪，根據調查結果資料進行統計分析，驗證假設是否成立。

例如:在台灣地區中等學校學生數位鴻溝差距狀況初研究中[11]，從「家庭收入」以及「居住地區都市化程度」兩項因素，對「資訊科技接近使用」、「資訊內容接近使用」、「資訊素養」三方面加以分析，以了解當前台灣地區中等教育體系中「數位鴻溝」的差距狀況。該項研究首先設定了六項假設，如：家庭收入在「資訊科技接近使用」上有顯著差異、家庭收入在「資訊科技接近使用」上有顯著差異等，在問卷於線上填答完畢後，所有填答資料隨即以 SPSS 8.0 for Windows 統計套裝軟體進行統計分析。而依據研究問題和各變項的性質，採取次數分配(frequencies)、卡方檢定(chi-square test)、單因子變異數分析(one-way ANOVA)：統計方法，分別對各項假設進行考驗，瞭解是否達.05 顯著差異水準。

這類問卷資料分析法，雖可驗證假設，推論結果，可是因缺乏建立問卷階層概念，不易針對不同層級範圍的問題，進行統計分析。

(2).結合分析層級程序(AHP)的統計分析法：

這類問卷資料分析法，主要是應用「層級分析法」(Analytic Hierarchy Process, AHP)[24]，針對問題建立整體層級架構，再邀請學者專家建立各層級問題指標的相對權重，決定問題指標的優先順序，再進行一致性檢定，確定學者專家對權重值有一致性看法。根據指標層級架構設計問卷，再將問卷調查結果，結合層級問題目標的相對權重進行運算，算出各層級的指標評估分數，再進行統計分析。

國內學者曾淑芬教授於 2003 年「台閩地區九十一年數位落差調查」即是採用上述分析法，進行問卷設計及調查，對於調查結果，計算各指標之權重比例，求得各層級指標的評估分數，接著以 SPSS8.0 統計套裝軟體進行分析，內容將包含描述性統計，以及家庭社經地位與各測量構面交叉分析，然後以各指標之權重進行加權，計算出台灣地區數位落差分數[12]。

這類問卷資料分析法，已較傳統單純推論統計較為進步，雖有建立問卷階層概念，也能針對不同層級範圍的問題，進行統計分析，可是因缺乏結合外部相關背景歷史資料，易造成分析結果單調或不足。

(3).結合線上分析(OLAP)技術的統計法:

這類問卷資料分析法，主要是應用線上分析(OLAP)技術，結合網路問卷系統，具有多維度資料結構系統，可對問卷調查資料進行線上統計分析。

例如:經濟合作發展組織 OECD，所開發的統計管理應用軟體(StatWorks)[6]，即是應用上述資料分析技術，具有多維度資料統計分析功能，也可結合資料倉儲進行資料整合成為決策支援工具[9]，但是未見可針對不同問卷題型的填答資料量化轉換處理功能以及資料探勘的機制。

綜合上述分析結果，我們可以了解到現有問卷資料分析方法有下列不足之處：

- 現有問卷分析方法，缺乏結合外部相關背景歷史資料，易造成分析結果單調或不足。
- 現有問卷分析方法，對於問卷資料本身沒有做完善的資料前處理，因此容易造成資料分析結果的誤差。

- 現有問卷分析方法，對於統計欄位沒有建立概念階層(Concept Hierarchy)機制，無法提供不同階層範圍的資料，進行不同層級間的動態分析。
- 現有 OLAP 或 Data Mining 分析缺乏整合性工具，可以讓分析者自由方便的進行分析。

第三章　應用資料倉儲技術之問卷分析架構

為了解決傳統問卷分析在資料維度整合、操作性與累加性等的不足，因此我們提出了一個「應用資料倉儲技術之問卷分析」架構，透過資料倉儲與資料探勘技術，整合其他歷史資料庫，來進行多維度問卷分析。

3.1. 系統架構與資料處理流程

應用資料倉儲技術之問卷分析架構之分析流程主要分三個階段，如圖 3.1。

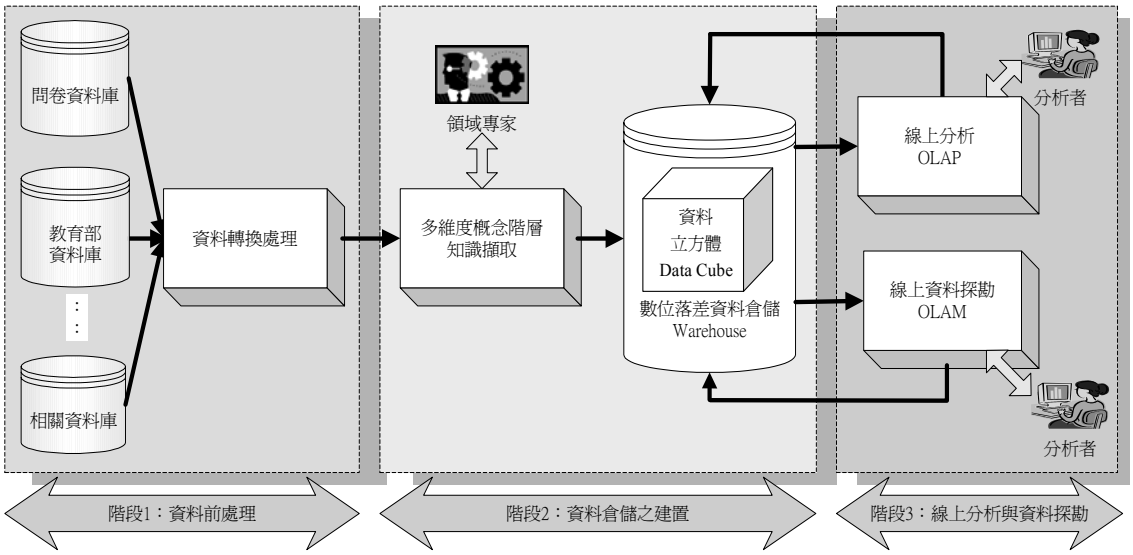

圖 3.1: 應用資料倉儲技術之問卷分析架構

(1). 資料前處理階段：

在此階段我們結合了原始問卷資料庫，以及其他相關歷史統計資料庫，以達到整合更多元化、多面向的方式來進行問卷分析。由於整合了不同來源、不同型態之資料，因此在資料前處理階段，使用了資料淨化(cleaning) 處理、資料平滑(smoothing)、聚集(aggregation)與正規化(normalization)等處理，並提出了**問卷題目量化轉換(Questionnaire Item Normalization and Measure Transformation, QINMT)**演算法，來將

問卷中不同的題型答案資料轉為可適用於資料立方體中的量值形式。

(2). 資料倉儲之建置階段：

在此階段提出了**多維度概念階層知識擷取 (Multiple Dimension Concept Hierarchy Knowledge Acquisition, MDCHKA)** 演算法，來擷取領域專家對問卷中的概念階層知識，其中量值概念階層知識，可以指導量值聚集(Measurement Aggregation)計算的處理,產生廣義化(Generalize)的新量值，經過這些程序後所建立的量值是一個具有較高階層概念的量值資料集合。此階段整理了資料維度與量值，並將之建置成資料倉儲中資料立方體。

(3). 線上分析與資料探勘階段：

透過建立好的資料立方體，利用線上分析工具，即可進行資料立方體的線上分析，並可使用上捲(roll-up)或下探(drill-down)等查詢操作，進行各層級的資料變項分析。經資料立方體視覺化的操作方式觀察分析後的結果，領域專家可以從多種的數位落差資料變項組合中發現重要的分析結果，並可透過參考概念階層對資料立方體的分析維度階層做調整，以取得理想的資料分析結果。資料探勘分析部分，基於 DMAS 核心技術，完成了使用**兩層式資料探勘方法之 DMAS 線上資料探勘系統 (DMAS-OLAM)**，結合分群演算法分析出學校資訊能力類別，然後使用決策樹演算法針對資訊能力類別建立決策樹，之後則可透過建立好之決策樹模型，進行預測查詢分析。

由以上三階段規劃，我們可以了解到應用資料倉儲技術之問卷分析架構，在實作上，我們也可以透過團體分工，依照成員的研究專長進行任務分配，並將研究結果匯集，進而達到整體性分析的目標。

3.2. 架構設計動機與探討

如第二章所述，傳統的研究對於問卷之分析方法，大多以假設、驗證方式來進行，因此如果假設需要更改，往往需要大費周章的處理資料，甚至重頭設計問卷，且容易因問卷答題狀況造成整體分析結果的誤差。

因此在問卷分析的領域中，將會產生以下幾點議題：

- 如何將現有問卷資料，增加結合外部相關資料一起分析之彈性？
- 如何將不同問卷題型，以及不同來源之資料庫，有系統的轉換成可以互相比對分析之欄位？
- 如何幫助分析者，為資料欄位建立概念階層(Concept Hierarchy)，以提供階層式動態分析機制？
- 如何發展符合問卷分析之 OLAP 或 Data Mining 整合性分析工具，可以讓分析者自由方便的進行分析？

為了解決以上幾點議題，我們將資料倉儲問卷分析架構流程分三個階段。分別是*資料前處理階段*，*資料倉儲之建置階段*，還有*線上分析與資料探勘階段*。

(1) 資料前處理階段：

為了整合外部的資料，以期有多元化分析結果，因此本研究主要採取資料倉儲技術[13][14]，透過資料彙整並建立多維度資料欄位來進行問卷分析。然而對於多種題型之問卷來源資料，會造成題目間分析處理之困難，為了解決這個問題，我們提出了**問卷題目量化轉換(QINMT)**演算法，針對一般問卷常見題型，例如：單選、複選、是非、重要程度排序等題型，若含有程度性之意義，則透過資料轉換之技術，將不同題型的問卷資料，進行量化與正規化處理，以提供資料欄位間分析之正確性。而其他非程度性類型之問卷項目則規劃成文字型欄位。

(2) 資料倉儲之建置階段：

由於結合了許多外部資料庫，為了更系統化的提供多維度、階層式之問卷分析功能，讓分析者能以不同的單位顆粒進行分析操作，在此提出使用資料倉儲技術，將問卷與其他資料庫之資料一起整合建構成數位落差資料立方體(Data Cube)。由於問卷資料往往有許多相關之題目對應到相同的概念，而概念間亦有階層性的關係，為了能系統化整理

出題目概念間之關係，本研究則提出了**多維度概念階層知識擷取(MDCHKA)** 演算法，可透過填寫表格方式，來擷取領域專家對問卷中的概念階層知識。此概念階層可以指導量值聚集(Measurement Aggregation)計算的處理,產生廣義化(Generalize)的新量值，並能以較適當之量值階層建置資料立方體，避免原始問卷資料太瑣碎，導致分析效果不佳。

(3) 線上分析與資料探勘階段：

由於一般傳統統計分析方法是以假設、驗證模式進行問卷分析，在此則提出資料探勘方法，使用發現模式進行問卷分析，結合問卷以外之資料進行更多元化之分析，並可以產生問卷設計外之不同結果。然而一般資料探勘(Data Mining)由於缺乏整合性，對分析者來說不容易進行分析操作，基於 DMAS 核心技術，開發完成了使用**兩層式資料探勘方法之 DMAS 線上資料探勘系統 (DMAS-OLAM)**，結合分群演算法分析出學校資訊能力類別，然後使用決策樹演算法針對資訊能力類別建立決策樹，以提供做預測分析。在此系統中並提供使用者依其需求調整分群分析與決策樹分析時，所使用之維度和資料階層來進行線上分析。

第四章　資料前處理

在一般交易性處理(OLTP)的資料庫中常存在不完整的(incomplete)、雜亂的(noisy)及不一致的(inconsistent)資料現象，因此需要資料前處理的流程，提高資料倉儲中的資料品質，進而提昇資料分析及探勘結果的品質。在這個章節中將說明資料彙整、資料淨化處理、資料轉換的作法及流程，並詳述如何將問卷填答資料及相關背景資料轉換為相對的測量值和維度資料。

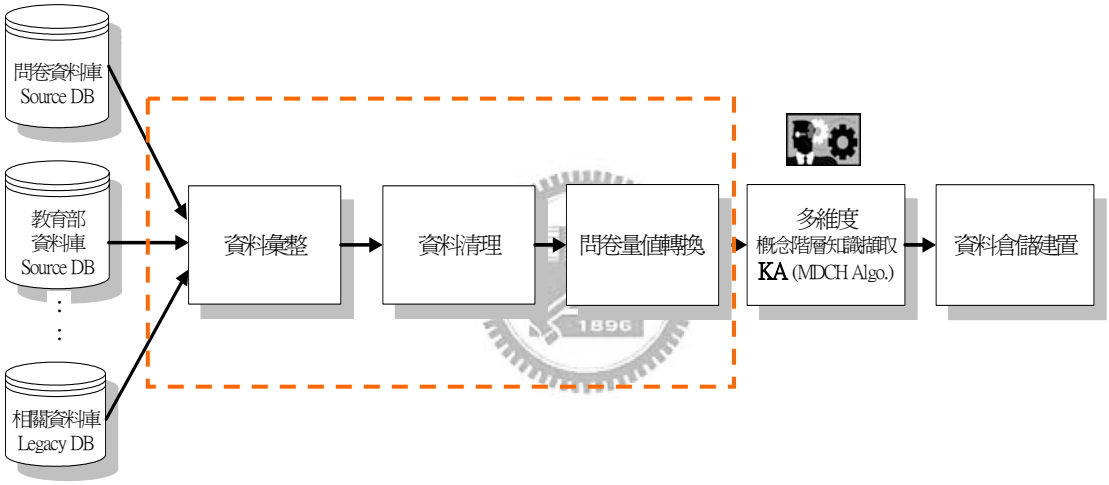

圖 4.1: 資料前處理流程圖

整體資料前處理流程，如圖 4.1 所示，可概分為資料彙整、資料清理、問卷量值轉換等三個階段來進行，其中資料來源有問卷資料庫、教育部資料庫、相關的歷史性資料庫，在經過資料前處理流程後，即可進行多維度概念階層知識擷取階段，擷取出多維度概念階層知識之後就可進入資料倉儲的建置階段。

4.1. 資料彙整與資料淨化處理

在應用資料倉儲技術時，常需將多個資料來源進行整合，存放在一個具有一致性資料型態的資料儲存體中，也就是資料倉儲中，在資料整合過程中主要有三個議題需要考慮：

- 綱要整合(Schema integration)，將不同的資料庫實體 (entity) 來源進行資料整合，所以必須調整各資料實體間的資料綱要以取得一致性的資料綱要結構。
- 重複的(Duplication)及多餘屬性值的處理問題，透過屬性間的相關分析(Correlation analysis)偵測出多餘的屬性予以刪減。
- 資料值衝突的問題，來自不同的資料庫可能用不同的單位屬性存放資料值，比如：貨幣單位、重量單、稅制等，亦須加以整合[15]。

在此以中小學數位落差問卷資料、其他相關資料如學校人口統計資料、資訊設備等不同資料格式的整合為例進行說明。我們所收集到的中小學數位落差問卷資料，可分為：高中職學生、國中學生、國小學生及學校行政人員共四類問卷填答資料，以及透過教育部統計處現有的學校相關統計資料及其他相關資料(如:領域專家提供的資料)等，作為整合前的資料來源。本研究主要分析資料來源是參考中小學數位落差相關研究資料[2]，並透過以學校為單位來鏈結其他資料庫之資料。

在資料彙整之後，接下來則需要做資料的淨化處理(Data cleaning)，在此階段的問卷處理過程中，有三種須淨化處理的資料狀況。在此透過刪除此資料來達到資料淨化效果。資料狀況歸納如下：

(1). 不正確的資料：

例如是應填答數字選項內容，卻誤填成選項內容文字內容。

(2). 空白未填答的資料：

例如未完整填答完所有題問或完全未填答。

(3). 不具鑑別度的資料：填答內容太過一致的現象。

例如填答內容太過一致的現象。

在此對於以上三種資料處理方法，以學校問卷為例說明，如範例 1 所示。

範例 1：資料淨化處理範例

表 4.1：資料淨化前之資料範例

Id	Acount	Scode	answer1	answer2	answer3	…	answer rn
21736	C0143260101	014324	1	2	3214	…	1
21147	C0143260102	014326	1	1	2354	…	4
21148	C0143260103	014325	1	1	Null		Null
21140	C0143260104	014328	1	3	234	…	3
21146	C0143260105	014327	Null	Null	Null		Null
21733	C0143260106	014336	1	2	設備不足	…	1
21731	C0143260107	014332	1	2	維修不易	…	1
:	:	:	:	:	:	:	:

(1).不正確的資料：資料 Id=21731 之 answer3 應填答數字選項內容，卻誤填成錯誤格式的文字內容。

(2).空白未填答的資料：資料 Id=21146 未完整填答完所有題問或完全未填答。

(3).不具鑑別度的資料：answer1 填答內容太過一致的現象，則此欄位資料不予取用。

4.2. 資料轉換處理

資料轉換處理主要的目的是將資料轉換成適合資料分析或探勘的形式，在此處理的方式有[15]：

- 平滑(smoothing):消除雜亂的資料，如：分箱法(binning)、迴歸法。
- 彙集(Aggregation):對資料進行彙總運算，例如：總和、平均、最大值等。
- 廣義化(Generalization):以一個較高階層概念屬性項取代多個較低層概念的屬性項集。
- 正規化(Normalization)：將屬性資料按比例縮放，使屬性值對應至的數值區間，例如：為了後續量值的觀察與多維度線性組合的計算，必須對填答量值再進行正規化(normalize)處理，使量值的新值域落於 0 到 1 之間，我們將採極小值-極大值正規化(min-max normalization)計算式：

$$v' = \frac{v - \min_A}{\max_A - \min_A}$$

- 屬性建構(Attribute construction)：因應需求，增加新的屬性項。

在此我們將應用上述**資料轉換處理技術**針對問卷資料進行兩類轉換:

(1). 問卷資料轉換處理方法：

由於問卷中常有不同的題型如:是非題、單選題、複選題、排序題、填充題等，同時也對應對出不同的答案內容(不同屬性值資料型態)，例如：邏輯型(Boolean)、類別型(Symbolic)、數值型(Numeric)、文字型(text)等，為了能將不同的問卷題型將問卷答案轉換成適合多維度資料模式的量值資料或維度資料，在此我們提出了**問卷題目量化轉換(QINMT)演算法**，如圖 4.2 所示。

演算法 1：問卷題目量化轉換(QINMT)演算法

輸入 : 問卷題目資料，問卷填答資料

輸出 : 問卷填答之量化資料

步驟 1：讀取問卷題目資料

步驟 2：讀取問卷填答資料

步驟 3：判別問卷題目資料(Category)題型:

步驟 3.1 : 若問卷題目資料**(Category)=二選一型**

(1). 填(沒有)者給 0 分，填(有)者給 1 分。

步驟 3.2 : 若問卷題目資料(**Category)=程度性單選題(K 個選項)**

(1). 首先將各選項依程度對應到一個整數 t，最高為 K，最低為 1。

(2). 使用正規化計算式，將值域映射到 0～1.0 區間。若填答的值對應到的整數為 t，則設定此題給 t/K 分。

步驟 3.3 : 若問卷題目資料(Category)=**非程度性單選題**(K 個選項)

(1). 將之化成 K 個是非題，沒有選填者給 0 分，有填者給 1 分。

步驟 3.4 : 問卷題目資料(Category)=**複選題**(K 個選項)

(1). 將之化成 K 個是非題，沒有選填者給 0 分，有填者給 1 分。 沒有選填者給 0 分，有選填者給 1 分。

步驟 3.5 : 問卷題目資料(Category)=**排序題**(K 個選項)

(1). 首先將填答項依排序順序對應到一個整數 p,最前面為 K,最後面為 1。

(2). 使用正規化計算式，將值域映射到 0～1.0 區間。若填答的值對應到的位置為 p，則設定此題給 p/K 分

步驟 4：所有題目都轉換結束了嗎?

是:跳到步驟 5。

否:重複步驟 3 直至所有題目之資料處理完畢。

步驟 5：輸出問卷填答之量化資料結果。

圖 4.2: 問卷題目量化轉換(QINMT)演算法

範例 2：二選一型題型之資料轉換

題目：學校網站提供教案或教材分享資料庫

填答： 有□ 沒有□

配分原則：填(沒有)者給 0 分，填(有)者給 1 分，如下表。

表 4.2：二選一型題型問卷填答範例

	選填項 (有/沒有)	轉換後資料
User1	有	1.0
User2	沒有	0
User3	有	1.0
User4	有	1.0

範例 3：程度性單選題題型之資料轉換

題目：學校與縣市網路教育中心的網路連線，未曾斷線的百分比，平均為何?

填答： (1)100% (2)99%~90% (3)89%~80%

(4)79%~60% (5)59%~40% (6)39%以下

配分原則：選(1)者給6分，選(2)者給5分，選(3)者給4分，選(4)者給3分，選(5)者給2分，選(6)者給1分，未選者給0分，即 $t \in \{6,5,4,3,2,1,0\}$，K=6。

值域正規化: m筆平均直接映射到0~1.0區間，每個分數除以6如下表：

表 4.3：程度性單選題題型問卷填答範例

	選填項	轉換後資料
User1	(1) 100%	1.0
User2	(2)99%~90%	0.83
User3	(3)89%~80%	0.66
User4	(4)79%~60%	0.5
User5	(5)59%~40%	0.33
User6	(6)39%以下	0.16

範例4：非程度性單選題題型

題目： 你現在跟誰住在一起？

填答： (1) 父母親 (2) 只與父親 (3) 只與母親 (4) 與父母親以外的人

配分原則：由4個選項產生4個變數項，隨選項而變化，被勾選該項可得1分如下表：

表 4.4：非程度性單選題題型問卷填答範例

	選填項	轉換後資料**1**	轉換後資料**2**	轉換後資料**3**	轉換後資料**4**
User1	(1) 父母親	1.0	0.0	0.0	0.0
User2	(2) 只與父親	0.0	1.0	0.0	0.0
User3	(3) 只與母親	0.0	0.0	1.0	0.0
User4	(4) 與父母親以外的人	0.0	0.0	0.0	1.0

範例 5：複選題題型

題目：你常在哪裡上網？

填答：(1)家裡 (2)學校 (3)網咖(4)校外圖書館 (5)同學或朋友家。

配分原則：由5個選項產生5個變數項，被選項可得1分如下表：

表 4.5：複選題題型問卷填答範例

	選填項	轉換後資料**1**	轉換後資料**2**	轉換後資料**3**	轉換後資料**4**	轉換後資料**5**
User1	12	1.0	1.0	0.0	0.0	0.0
User2	123	1.0	1.0	1.0	0.0	0.0
User3	1234	1.0	1.0	1.0	1.0	0.0
User4	12345	1.0	1.0	1.0	1.0	1.0

範例 6：排序題題型

題目：學校支援資訊教學應用上，常會碰到的狀況有那些，請排序？

(1)校長支持度不高　　(2)教師資訊融入教學能力尚待加強

(3)資訊教學設備不足　　(4)現有資訊教學設備維護不易

(5)資訊教學人力不足。

由5個選項產生5個變數項，按照優先順序給予屬性值，最高5分，最低1分，即 $p \in \{5,4,3,2,1\}$，$K=5$。如下表：

表 4.6：排序題題型問卷填答範例

	選填項	轉換後資料**1**	轉換後資料**2**	轉換後資料**3**	轉換後資料**4**	轉換後資料**5**
User1	12345	1.0	0.8	0.6	0.4	0.2
User2	51234	0.8	0.6	0.4	0.2	1.0
User3	34125	0.6	0.4	1.0	0.8	0.2
User4	23451	0.2	1.0	0.8	0.6	0.4

(2)其他資料庫欄位維度之資料轉換：

維度資料的前處理主是透過離散化(discretization)技術先將連續性數值資料劃分為區間，再透過收集並用較高的概念替代較低層的概念就可以形成概念階層，經由概念階層的建立，便可以有效的簡化資料(data reduction)或廣義化 (generalize)資料，可使大量資料變得容易解釋，也有助於後續的資料分析與探勘工作，因此維度資料前處理流程大致可以兩個步驟完成[15]：

Step1. 離散化

離散化技術主要是將連續性數值資料進行區間劃分，可是對於不連續或雜亂的資料，第一個處理步驟是先進行排序，再進行分區。分區技術主要使用**等深(equal depth)分區法**:每個區間以相等資料筆數進行區分，這樣劃分法，在一區間內值域範圍變化不固定，但區間資料筆數(頻率)可受到控制。

Step2. 階層化

數值型的資料，可以根據資料分佈分析來自動建構概念階層，常見的數值概念階層生成法如：直方圖分析(Histogram Analysis)、分箱(binning)、基於熵值(Entropy base)的離散化等。但對於非數值型資料而言，其概念階層則有其特有的知識，需先擷取這類知識後，才能建置其概念階層，我們將於下個章節解說。

以下將就數位落差研究中小學生人數及教師人數為例，進行數值性概念階層的建立，另以全國地理分區為例，進行非數值性概念階層的建立。

範例 7：建置學生人數維度概念階層

在學生人數規模維度表的概念階層的部分，我們主要是每所學校所擁有的學生人數做為學生規模維度概念階層建立的依據。因為在每所學校之間，所擁有學生人數並不一定連續，所以，首先將所有學校以學生人數做排序。接下來劃分區間，採等深(equal

depth)方式，也就是在排序後的學校順序數中，以等量的學校所數來劃分區間。例如: 在排序後的所有學校所數是 3212 所，劃分 2 個範圍，就是 1606 所為 1 個區間，也就是第 1 到第 1605 所學校為第 1 區，根據統計對應，這個區間中學校裡面的學生人數分佈是從 0~617 人，同理，第 2 區是第 1605 到第 3212 所，這個區間中學校裡面的學生人數分佈是從 619~8340 人，學生人數統計圖如下：

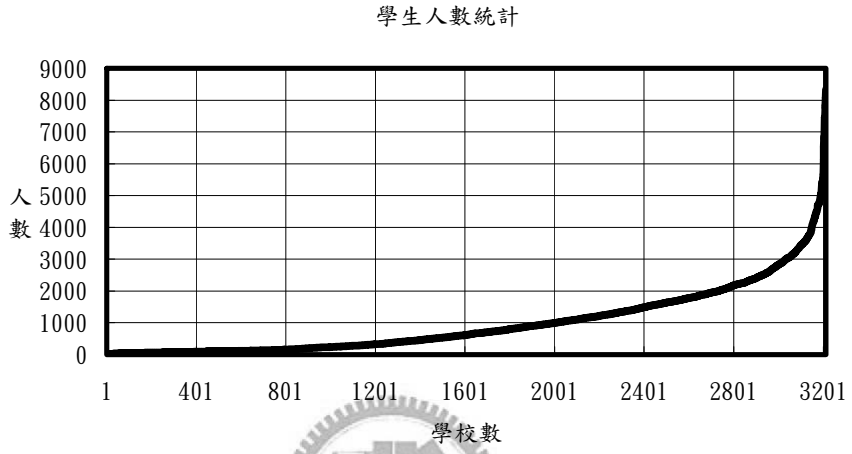

圖 4.3: 學生人數統計圖

接下來進行進一步階層化，同理可再劃分為4個範圍成為下一個層級的成員(members)、以8個範圍成為再下一個層級，每劃分一次，就產生一個層級(level)的維度成員，直到以各校人數為一單位的屬性成員為止，本研究設計為5個階層如下:

學生規模概念階層 ：

- 1個範圍 8340
- 2個範圍 0~617, 619~8340
- 4個範圍 0~159, 160~617, 619~1498, 1499~8340
- 8個範圍 0~86, 87~159, …,2029~8340
- 各校人數 0~8340

相同的做法，也可建立教師人數概念階層如下:

教師規模概念階層:

- 1個範圍 283
- 2個範圍 0~35, 36~283
- 4個範圍 0~13,14~35, 36~80, 81~283
- 8個範圍 0~10, …113~283
- 各校老師人數 0~283

範例8：建置地理位置維度概念階層

由於地理位置有其固有的階層性，只要按其原有的階層關係逐層規劃，即可建立地理位置維度的概念階層，由於本研究是以學校為單位的數位落差分析，所以在地理位置維度概念階層的建置的做法上，自然是以全國學校(中小學3881所)為起點:

步驟1: 以365區碼將學校所在的區碼來廣義化 (generalize)同層區域學校屬性成員。

步驟2: 以25縣市的屬性來廣義化 (generalize) 365區碼屬性成員(member)。

步驟3: 以全國地理分區北, 中, 南, 東4區的屬性來廣義化25縣市同層區域屬性成員。可建立概念階層如下:

地理位置維度概念階層:

- 全國地理分區(1個範圍)
- 全國地理分區(北, 中, 南, 東4區)
- 縣市 (25個縣市)
- 各縣市區域號碼 (365個區碼)
- 各個學校(3881學校)

相同的做法也可用以10種城鄉等級的屬性來廣義化區域學校屬性：

步驟1: 以365區碼將學校所在的區碼來廣義化 (generalize)同層區域學校屬性成員

步驟2: 以10種城鄉等級的屬性來廣義化 (generalize) 365區碼屬性成員(member)。

可建立概念階層如下:

地理城鄉維度概念階層：

- 城鄉分類 (10種城鄉等級)
- 各縣市區域號碼 (365個區碼)
- 各個學校(3881學校)

4.3. 階層性維度與資料倉儲之建置

(1). 階層性維度之建置

在這個章節我們將詳細介紹多維度概念階層知識擷取 (Multiple Dimension Concept Hierarchy Knowledge Acquisition) 演算法，這個演算法主要是用來擷取領域專家(數位落差資料分析學者)的量值或維度概念階層知識，其中量值概念階層知識，可以指導量值聚集(Measurement Aggregation)計算的處理,產生廣義化(Generalize)的新量值，至於維度概念階層知識，則可以指導維度階層的建置，利用維度階層知識我們就可以建立合於學理的概念階層維度。再結合前述量值資料集就可組合成合於專家學理的事實資料表，進而建置一個合乎領域專家學理的資料立方體。

演算法2：多維度概念階層知識擷取(MDCHKA)演算法

輸入：項目資料集。

輸出：概念階層知識表。

步驟1：載入項目資料集，成為候選概念項（Comcept Items）

步驟2：列出所有候選概念項, 詢問使用者目前處理的層級別。

步驟3：隨機挑出一概念項（例如A項）

步驟 4：詢問使用者，勾選出與 A 項相似，可歸類為同一層的概念項。

步驟 5:要求使用者,為(步驟 4)所有已選概念項定義一較高層概念名稱(例如:L2_A1)

步驟 6：排除(步驟 4)、(步驟 5)已選的概念項，重複(步驟 2)~(步驟 6)，直到所有概念項均已挑選及分類完畢。

步驟 7：詢問使用者是否達成目標層級數(是否再 roll-up 一層)。

（1）是，挑出所有在(步驟 5)時所新增的較高層概念項，成為新候選概念項，並回步驟 2。

（2）否，步驟 8。

步驟 8：輸出概念階層知識表

圖 4.4: 維度概念階層知識擷取(MDCHKA)演算法

範例9：應用「多維度概念階層知識擷取演算法」，將5個候選概念項,建構成為3層問卷概念階層知識。

步驟1:列出所有候選概念項:

表 4.7：候選概念項描述表

概念代號	概念描述
Q8	父親會不會上網
Q9	母親會不會上網
Q10	除電腦課外，有無其它科目/領域的老師也會在課堂上使用電腦來協助上課
Q11	除電腦課外，有無其他科目/領域的老師曾經要求使用電腦來完成作業
Q18	老師使用電腦或網路上課時，會讓你更聽得懂老師講的內容？

步驟 2~7:**概念階層組織特徵選擇與命名（Concept Hierarchy Labeling Phase）**

表 4.8：第 1 層概念階層組織特徵選擇與命名

載入第 1 層	隨機出現概念項	相似概念項 1	相似概念項 2	相似概念項 3	較高層概念名稱命名(labeling)
第 1 回	Q11	S10	Q8		課堂資訊教學
第 2 回	Q8	Q9			資訊使用支援

表 4.9：第 2 層概念階層組織特徵選擇與命名

載入第 2 層	隨機出現概念項	相似概念項 1	相似概念項 2	相似概念項 3	較高層概念名稱命名(labeling)
第 1 回	課堂資訊教學	資訊使用支援			資訊環境
第 2 回	結束				

步驟8:輸出問卷概念階層知識

表 4.10：輸出結果問卷概念階層知識

資訊環境量值概念階層知識表		
廣義概念階層		**概念項**
資訊環境	課堂資訊教學	Q10: 課堂上使用電腦來協助上課
		Q11: 使用電腦來完成作業。
		Q18: 使用電腦或網路上課時，會讓你更聽得懂。
	資訊使用支援	Q8: 父親會不會上網
		Q9: 母親會不會上網

有了上表 4.10 的量值概念層知識，建立資料立方體進行分析時，即可以依分析者需求，以較廣義的概念名稱代替細部瑣碎量值。

(2). 資料倉儲之建置

建立了量值與維度之概念階層後，即可經以下步驟進行資料倉儲之建置[16][17][20][21]。

- **選定所欲觀察之測量值 (measures)**

  參考領域專家提供的建議相關資料，我們可定出:9 個量值項：學校資源、社經地位、進階資訊技術、資訊技能、資訊使用支援、資訊近用、資訊應用、網路素養、課堂資訊教學。

- **選定欲觀察之維度 (dimensions)**

  參考上述 9 個量值項後，並考量我們預定探勘的資料維度，可定出以下 12 個維度索引鍵項:**城鄉分類、地理位置、學校類別、教師人數、學生人數、私立學校、資訊種子學校、教師資訊政策、資訊教育方案、資訊教學狀況、與父母親同住、男生比例**，其中**資訊教學狀況**維度表包含了，**校長支持度不高、教師資訊融入教**

學能力尚待加強、資訊教學設備不足、現有資訊教學設備維護不易、資訊教學人力不足等 5 項欄位項。

- **決定所欲觀察之事實表欄位 (fact table)**

考量我們預定探勘的兩個資料立方體功能可定出以下 2 個事實表：

表 4.11：學校及學生實事表

學校與學生事實表
城鄉分類 Key
地理位置 Key
學校類別 Key
教師人數 Key
學生人數 Key
私立學校 Key
資訊種子學校 Key
教師資訊政策 Key
資訊教育方案 Key
資訊教學狀況 Key
與父母親同住 Key
男生比例 Key
學校資源
社經地位
資訊使用支援
課堂資訊教學
資訊應用
資訊近用
網路素養
資訊技能
進階資訊技術

表 4.12：學校實事表

學校事實表
城鄉分類 Key
地理位置 Key
學校類別 Key
教師人數 Key
學生人數 Key
私立學校 Key
資訊種子學校 Key
教師資訊政策 Key
資訊教育方案 Key
資訊教學狀況 Key
與父母親同住 Key
男生比例 Key
學校資源

- **選定所欲建立之資料模式 (例如：星狀綱要、雪花綱要、星系綱要)**

在資料模式方面，採星狀綱要（Star schema）模式，每個資料立方體包含一個事實資料表及一組維度資料表[19][14]，稱為星狀綱要(star schema)如圖 4.5，而且事實表中的維度外來鍵值項只連結一個維度表，且這個維度表不做正規化處理，可以節省查詢時表格轉換(join)時間。

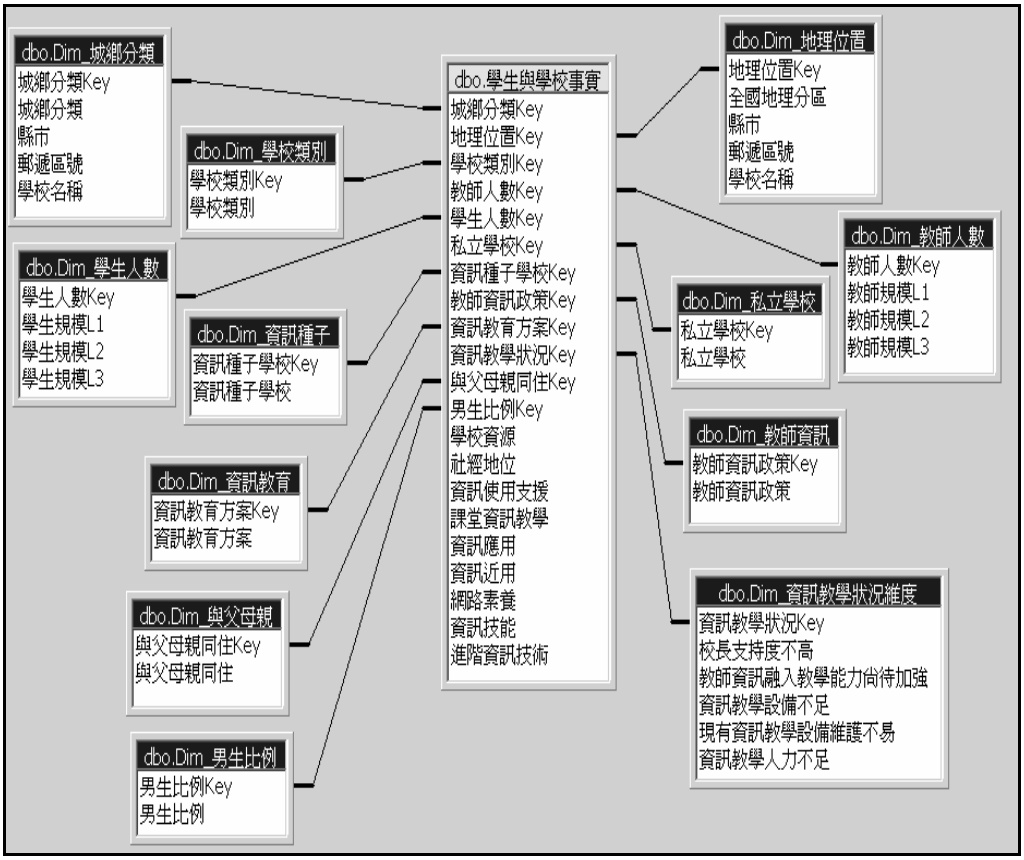

圖 4.5: 資料立方體之星狀綱要

第五章　線上分析與資料探勘

本章主要在介紹如何利用線上分析工具，進行線上分析資料立方體的各種查詢操作，如:上捲(roll-up)或下探(drill-down)等進行資料立方體中各層級的資料變項分析，以資料立方體視覺化的操作方式觀察分析後的結果。如此，我們就可從多種的數位落差資料變項組合中發現重要的分析結果，而且也可以參考概念階層對資料立方體的分析維度階層做調整，以取得理想的資料分析結果，再依分析結果中出現的所有變項，進行資料變項特徵挑選，找出探勘的目標，進行線上分析探勘(On Line Analyze Data Mining)以查出數位學習落差的分類及成因。以下各節將針對線上分析與資料探勘的細節來加以說明。

5.1. 線上分析處理

透過資料前處理與資料倉儲建置步驟，我們已建置了兩個資料立方體(Data cube)包括了:「學校及學生問卷」和「學校問卷」的資料立方體，共包含了 9 個量值及 14 個分析維度，在本章線上分析(OLAP)處理階段，為了分析各學校間造成數位落差之現況，將會產生以下議題需要探討：

- 如何去找出線上分析(OLAP)主題呢?
- 如何利用線上分析主題去分析這些具有多維度概念階層的倉儲資料呢?
- 如何在數以百計的維度層級切換組合中找出與資料探勘任務相關的組合？
- 用什麼樣的標準來評估維度層級切換結果的適當性呢？

為了解決以上問題，首先，我們提出**由上往下(Top-Down) 階層式的線上分析法**，可針對線上分析(OLAP)主題進行分析並找出與主題相關的維度，做為下一階段資料探勘工作的參考資訊。其次，在線上分析(OLAP)主題方面，我們結合「學校與學生數位落差評估指標架構圖」，對照我們所建置的 14 個維度及 9 種量值，建立了「中

小學數位落差 OLAP 主題分析表」。最後，在找出與資料探勘任務相關的組合及評估維度層級切換結果的適當性方面，由於透過異常值的探查，可以找出與資料探勘任務相關的層級組合，因此本研究根據維度層級資料變項的敘述統計(descriptive statistics)值來評估異常的程度，並以「集中程度」與「離散程度」兩個角度來觀察維度層級切換結果的適當性。

(1) 由上往下(Top-Down) 階層式的線上分析法

一般而言，我們可從最高概念階層往下分析，由較大的顆粒資料集往較小的顆粒資料集分析，或者說從巨觀到微觀，可採一種「**由上往下(Top-Down) 階層式的分析**」方法，分析流程如圖 5.1 所示。

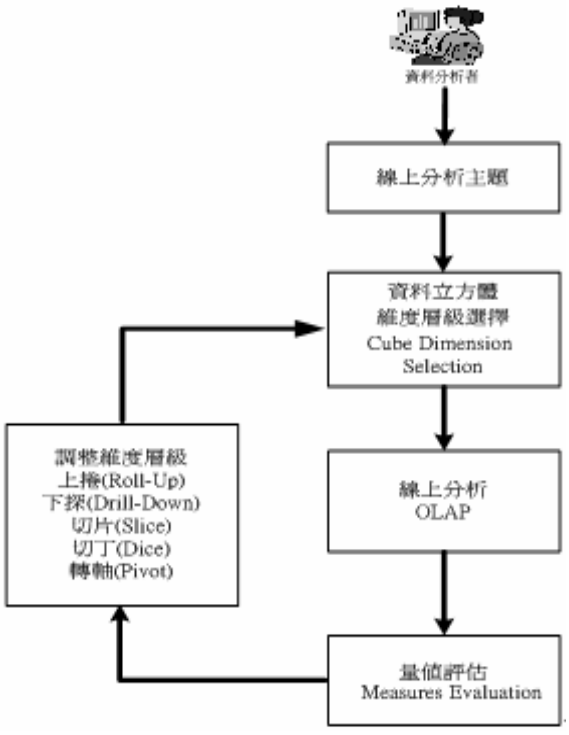

圖 5.1: 由上往下(Top-Down) 階層式的線上分析流程圖

根據圖 5.1 所示分析流程，我們可將整體分析流程細分為，依據資料分析目標，建立一 OLAP 主題分析表、選擇目標量值及相關維度、調整分析維度、進行線上

分析、評估量值結果、完成線上分析目標等 8 個步驟來進行如圖 5.2 所示：

由上往下(Top-Down) 階層式的線上分析法

步驟1　建立一 OLAP 主題分析表。

步驟2　選擇目標量值及相關維度，並填入主題分析表中。

步驟3　依分析主題填入維度分析順序(可以參考領域專家知識)。

步驟4　依分析主題所規畫順序，選擇啟始量值及分析維度。

步驟5　調整分析維度，(Drill-down, Roll-up,Slice,Dice,Pivot)

步驟6　進行線上分析。

步驟7　評估量值結果。

步驟8　是否完成線上分析目標?

是，結束分析。

否，回步驟 5。

圖 5.2: 由上往下(Top-Down) 階層式的線上分析法

其中在步驟 3，我們除了依據「領域專家知識」來決定維度分析順序外，也可以參考資料分析者所感興趣的維度來決定維度分析的順序。

根據上述分析流程，在選擇了目標量值及維度後就可進入線上分析操作，我們可透過各種線上分析基本操作看到不同的維度層級的量值變化，評估量值結果，挑選出我們所需要的分析結果。

例如:當我們從問卷量值概念階層表中最高層挑選 4 個概念主題，配合 5 個維度表，填入分析的順序後，則可組成 OLAP 主題分析(表 5.1)如下：

表 5.1：OLAP 主題分析表

分析維度 / 分析主題	地理位置	學生規模	教師規模	是否為種子學校	是否為私立學校
學校資源	1	4		2	3
資訊技能	1	3		4	2
資訊應用	3	2		4	1
課堂資訊教學	2		1	3	4

從上面 OLAP 主題分析表中，可看出有 5 個維度資料，如果僅取用(地理位置，學生規模，教師規模)這 3 個維度進行線上分析時，可組合出 8 種不同的資料表關係，其關係如圖 5.3 所示，而且，由每一個維度又包含數個屬性階層關係，因此實際組合關係將達到 150 種。

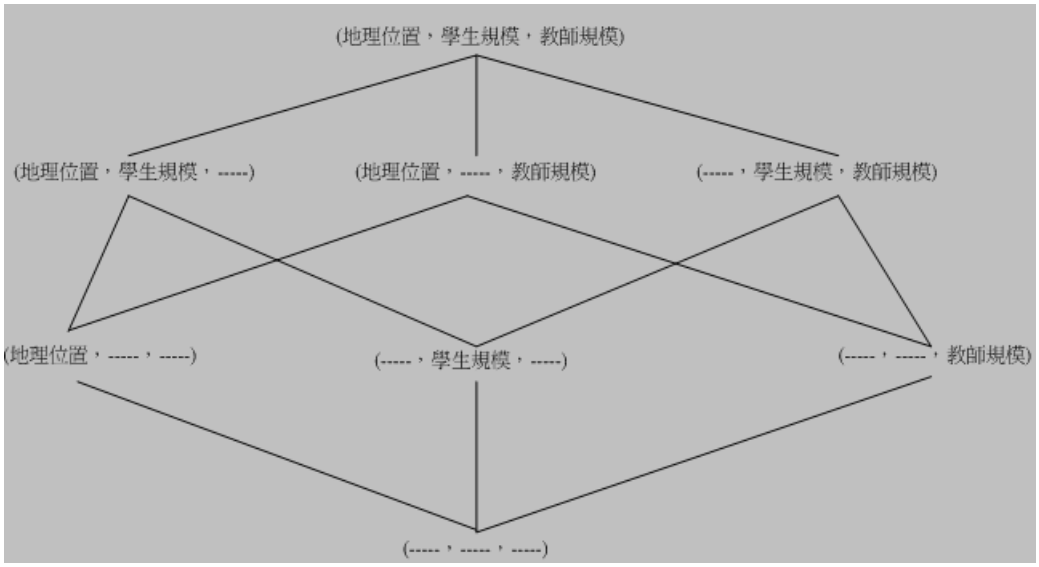

圖 5.3: 地理位置，學生規模，教師規模維度組合圖

範例 10：學校資源分析

以「**由上往下(Top-Down)階層式的分析**」方法，透過「地理位置」、「公私立學校」、「教師資訊政策」分析維度的切換來分析「學校資源」問卷量值。

我們首先以 2004 年台灣地區中小學學校的「學校資源」量值，配合前述維度及分析順序，首先是「地理位置」、其次是「公私立學校」、最後為「教師資訊政策」3 個維度，組成學校資源 OLAP 主題分析(表 5.2)如下：

表 5.2：學校資源 OLAP 主題分析表

分析維度 / 分析主題	地理位置	學生規模	教師規模	是否為種子學校	是否為私立學校	資訊教學人力不足	資訊教學設備不足	教師資訊政策	教師資訊融入教學能尚待加強	現有資訊教學設備維護不易	校長支持度不高
學校資源	1				2			3			

首先是「地理位置」維度的切換分析，在下圖 5.4 可看出學校資源最佳的學校地理位置是在中區及北區。

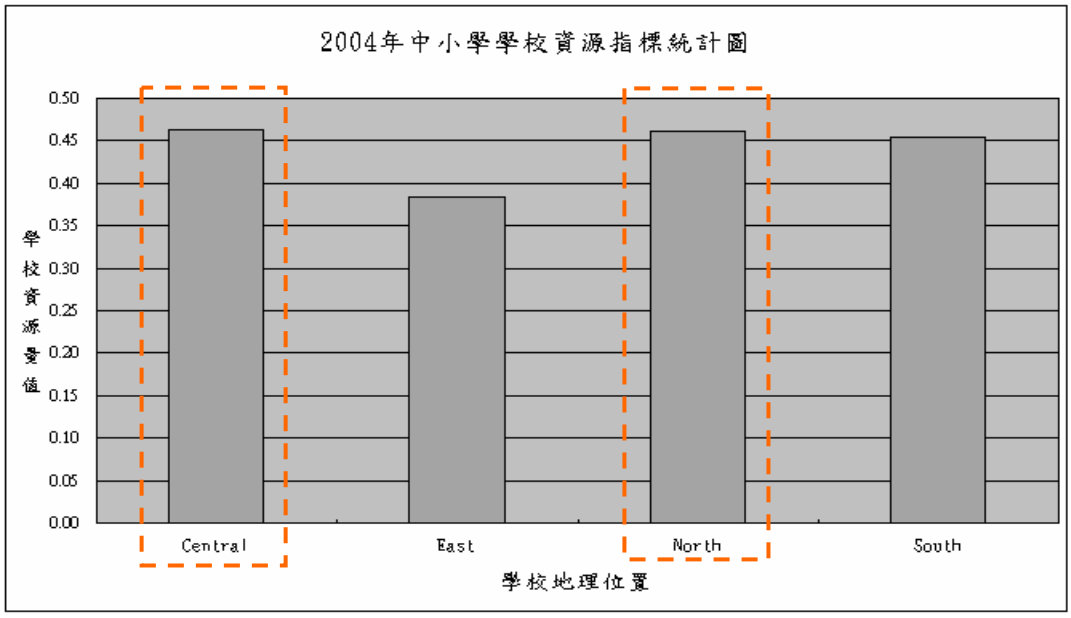

圖 5.4: 全國學校資源最佳地區

我們可再就北區學校進行下探(drill-down)的分析，如下圖 5.5,可看出台北市是北區學校中，資源最佳的學校。

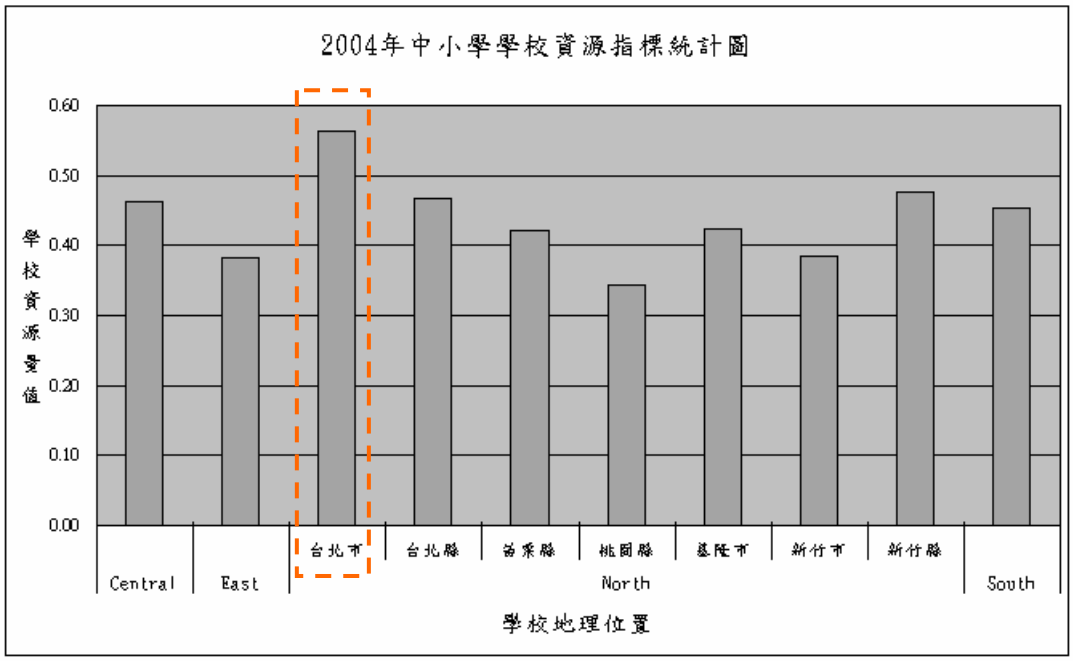

圖 5.5: 北區學校中資源最佳的學校

我們可再就公私立學校、教師資訊政策維度進行下探(drill-down)的分析，如下圖 5.6

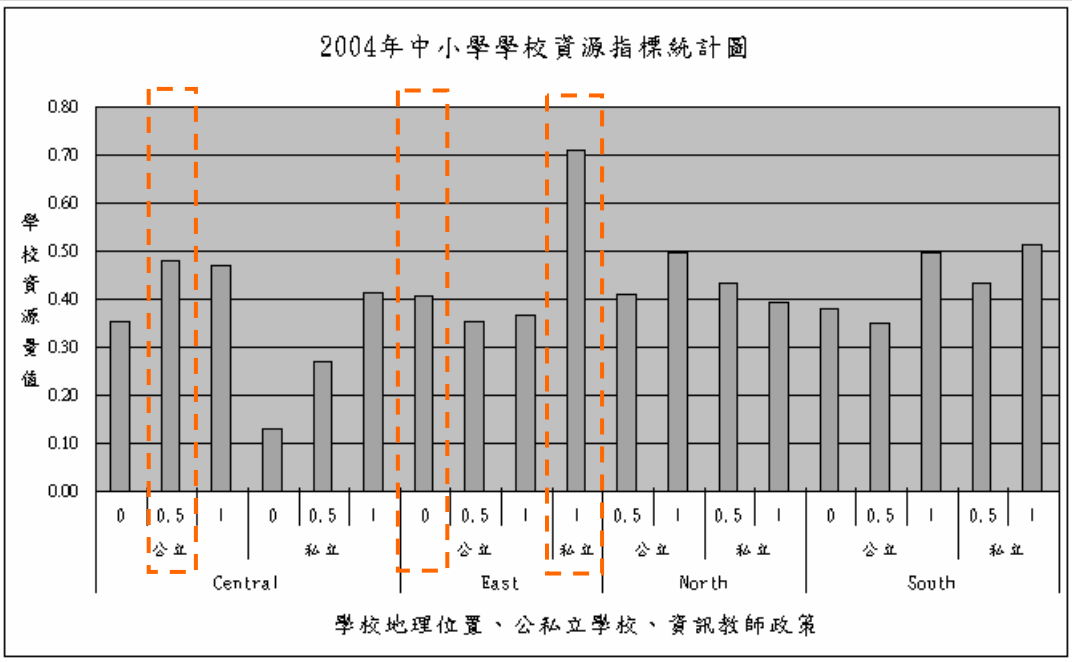

圖 5.6: 教師資訊政策佳，學校資源佳

上圖 5.6，可看出大部分的學校，**「教師資訊政策」維度與「學校資源」量值有**

正相關的趨勢，也就是說，大部分的學校「教師資訊政策」愈好時則「學校資源」也會愈好，由圖 5.6 的中區、北區及南區學校的「學校資源」的 OLAP 統計圖中即可看出，但是必須注意的一點是在大趨勢之下，隨著地理位置的不同，也會有不同的趨勢出現，例如：在上圖 5.6 中，中區「學校資源」較佳的學校是分佈於學校「教師資訊政策為 0.5」之處，與北區公立學校及南區學校分佈於學校「教師資訊政策為 1」之處，這兩類之間是有差異的。

而且，東區公立學校，「教師資訊政策為 0」的「學校資源」略優於「教師資訊政策為 1」，這樣的結果也是不同於其他地理置的「教師資訊政策」對「學校資源」的影響。

另外，從「公私立學校」的維度來分析，也可看出「學校資源」量值隨著地理位置的不同，也會有不同的趨勢出現，例如：中區、北區公立學校的「學校資源」優於私立學校，但是在東區、南區卻是私立學校的「學校資源」較佳。

所以經上述的分析結果，我們可以看出「地理位置」、「公私立學校」、「教師資訊政策」這三個維度是影響「學校資源」量值相關的維度。

由上例中我們可了解到藉由調整維度的分析階層，就可以觀察出不同的量值資訊的趨勢，這些結果，將是下一階段資料探勘工作的重要參考資訊。我們也可以利用上述的分析流程，對其他主題進行線上分析，這些相關分析結果將詳述於第六章的實作部分。

(2) 中小學數位落差 OLAP 主題分析表

下表 5.3 為 2004 年台灣地區中小學校數位落差分析維度及量值名稱表，表中分析維度及量值是根據前述 2 個事實資料表中 9 個量值及 12 個維度表所建立出來的，共建立了 2 個資料立方體(Data cube)，「學校及學生問卷」和「學校問卷」的資料立方體，其中包含了 14 個分析維度及 9 個量值。

表 5.3：2004 年台灣地區中小學校數位落差分析維度及量值名稱表

	維度名稱	值域	量值名稱	值域
1.	地理位置(Location)	0~3881	社經地位	0~1
2.	學生人數規模(Student scale)	0~8340	進階資訊技術	0~1
3.	教師人數規模(Teacher Scale)	0~283	資訊技能	0~1
4.	是否為私立學校(Private School)	0 .or.1	資訊使用支援	0~1
5.	是否為資訊種子學校(Seed)	0 .or.1	資訊近用	0~1
6.	資訊教育方案執行程度	0~1	資訊應用	0~1
7.	教師資訊政策執行程度	0~1	網路素養	0~1
8.	校長支持度不高狀況	0~1	課堂資訊教學	0~1
9.	教師資訊融入教學能力尚待加強	0~1	學校資源	0~1
10.	資訊教學設備不足狀況	0~1		
11.	現有資訊教學設備維護不易狀況	0~1		
12.	資訊教學人力不足狀況	0~1		
13.	學校男女比例	0~1		
14.	學生與父母同住之比例	0~1		

參考中小學數位落差相關資料中有關於「學校與學生數位落差評估指標架構圖」，對照我們所建置的 14 個維度及 9 種量值，可得對照表 5.4 如下:

表 5.4：學校_數位落差量值維度與評估指標架構圖對照表

構面	次構面	說明	相對應之量值或維度
資訊近用 Information Access	資訊基礎建設	衡量學校內部資訊相關建設程度。	學校資源
	資訊經費	衡量學校資訊設備經費運用情形	學校資源
	網路服務建設	衡量學校網站與相關服務應用建設與維護程度。	學校資源
資訊素養 Information Literacy	教師資訊素養	衡量學校教師之資訊相關人力資本	教師資訊政策
	資訊教育方案	衡量學校內行政部門推動資訊教育的相關政策	資訊教育方案
資訊教育應用 Application	教學應用	衡量學校電腦與網路應用在教學的情形	課堂資訊教學
	網路應用	衡量學校提供的網路服務功能與應用情形	資訊應用
	合作學習	衡量學校內教師運用資訊科技互動與校際間的交流合作情形	課堂資訊教學

表 5.5：學生_數位落差量值維度與評估指標架構圖對照表

構面	次構面	說明	相對應之量值或維度
資訊近用	網路近用	衡量學生在網路使用上的廣度	資訊近用
	網路使用行為	衡量學生在網路使用上的深度	資訊應用
資訊素養	資訊技術能力	衡量學生資訊技術方面的應用能力與知識	進階資訊技術
	資料處理與分析能力	衡量學生其基礎的資料處理與分析能力的程度	資訊技能
	網路應用能力	衡量學生對網路的使用能力之程度	資訊應用
	網路素養	衡量學生對網路規範及倫理的理解程度	網路素養
資訊學習環境	科技融入教學	衡量學生上課時老師運用資訊科技融入教學的情形	課堂資訊教學
資訊應用	課業學習	衡量學生在課業上應用資訊科技的程度，包括主動學習、合作學習與創意學習	資訊應用
	人際關係	衡量學生應用資訊科技於人際關係上的程度	資訊應用

整理上述對照表，可得中小學數位落差OLAP主題分析表如下所示:

表 5.6：中小學數位落差 OLAP 主題分析表

分析主題	相關量值及意含	相關維度
資訊學習環境含資訊教育應用	課堂資訊教學 (教師) 資訊使用支援 (父母) 社經地位 (父母)	1. 地理位置 2. 學生人數 3. 教師人數 4. 私立學校 5. 資訊種子學校 6. 資訊教育方案 7. 教師資訊政策 8. 校長支持度不高 9. 教師資訊融入教學能力尚待加強 10. 資訊教學設備不足 11. 現有資訊教學設備維護不易 12. 資訊教學人力不足 13. 學校男女比例 14. 學生與父母同住之比例
資訊近用	學校資源 (學校教材資源) 資訊近用 (家中設備)	
資訊應用	資訊應用 (上網的習慣及時數)	
資訊素養	資訊技能 (電腦網路技能) 進階資訊技術 (進階技能) 網路素養 (道德)	
整體綜合分析	綜合上述量值	

從「中小學數位落差OLAP主題分析表」(表5.6)中，可看出分析主題可概分為5個。特別是資訊學習環境分析、資訊近用分析、資訊應用分析、資訊素養及整體綜合分析、表中詳列每個分析主題的相關量值及相關維度，同時我們可從相關量值欄中的資料了解到這些量值相關的內在意含，例如: 在資訊學習環境分析主題中「課堂資訊教學」量值是與教師資訊融入教學相關的量值項，「資訊應用」量值是與學生與同儕使用網路的習慣或機率相關的量值項，而「資訊使用支援」及「社經地位」量值是學生的父母提供上網的能力支援及社經地位相關的量值項，也就是說這五個主題分析含蓋了教師資訊融入教學、學生同儕使用網路的習慣、父母上網的能力支援、學校教材

資源設備、學生家中資訊設備、學生資訊基本技能、學生資訊進階技能、學生資訊道德的分析意含。

維度是分析量值的角度，其中地理位置維度，分為六個層階供分析者作不同地理範圍及組合的分析，學生人數、教師人數均分為五個層階供分析者作不同人數範圍及組合的分析，私立學校及資訊種子學校均分為{是,否}的範圍供分析，其他還有屬於資訊教學活動的資訊教育方案維度、獎勵教師的資訊教師政策維度，除此之外尚有關於學校政策的維度如：校長支持度不高、教師資訊融入教學能力尚待加強、資訊教學設備不足、現有資訊教學設備維護不易、資訊教學人力不足等。在建立 OLAP 主題分析表後，搭配應用前述的「**由上往下(Top-Down)階層式的分析**」，我們將可從維度來分析量值，找出與量值相關的維度資訊，以做為下一階段資料探勘工作的參考。

(3) 維度層級切換與評估

由於在資料倉儲中的資料集合非常大，因此在進行線上分析時，將面臨兩個問題: 就是如何在數以百計的維度層級切換組合中找出與資料探勘任務相關的組合，以及用什麼樣的標準來評估維度層級切換結果的適當性。

首先說明的是關於找出最佳維度層級切換組合的問題，因為是屬於資料立方體探查(Exploration of data cubes)問題，而這類問題有兩種處理法[15]:

(a).假設驅動的探查 (Hypothesis-driven exploration):

資料分析師可 OLAP 基本操作例如如:上捲(roll-up)、下探(drill-down)、 切片(slice)、切丁(dice)、樞紐(pivot)或稱轉軸分析，協助資料分析者，從不同角度、機動地進行量值資料觀查與驗證統計資料，可挑出所有最佳層級組合性。

(b). 發現驅動的探查 (Discovery-driven exploration):

這種是為資料立方體中所有異常(exception)單元值加上標識，例如:不同的背景顏色，它有三種標識法，SelfExp:相較於同層級單元值異常的程度，InExp:在指示

單元值下層級某個單元有異常的程度。PathExp: 在指示單元值下層級每條下探(drill-down)路徑異常的程度。

透過異常值的探查，可以找出與資料探勘任務相關的層級組合。但是須定訂評估異常的程度的標準。由於維度層級切換組合中所對應的量值必須與資料探勘任務相關，所以可以根據這些量值的敘述統計資料來評估維度層級切換與資料探勘任務的適切性。敘述統計(descriptive statistics)值，主要用以描述母體，一般而言會以「集中程度」與「離散程度」兩個角度來觀察。並以平均數、變異變、標準差..等來代表母體特徵[10]。

範例 11：維度階層切換與評估

當我們在進行了某一階層的線上分析時，如圖 5.7 與表 5.7，可以透過以敘述統計數據的計算如：平均數、變異變、標準差..等統計數據來代表母體特徵如表 5.8。

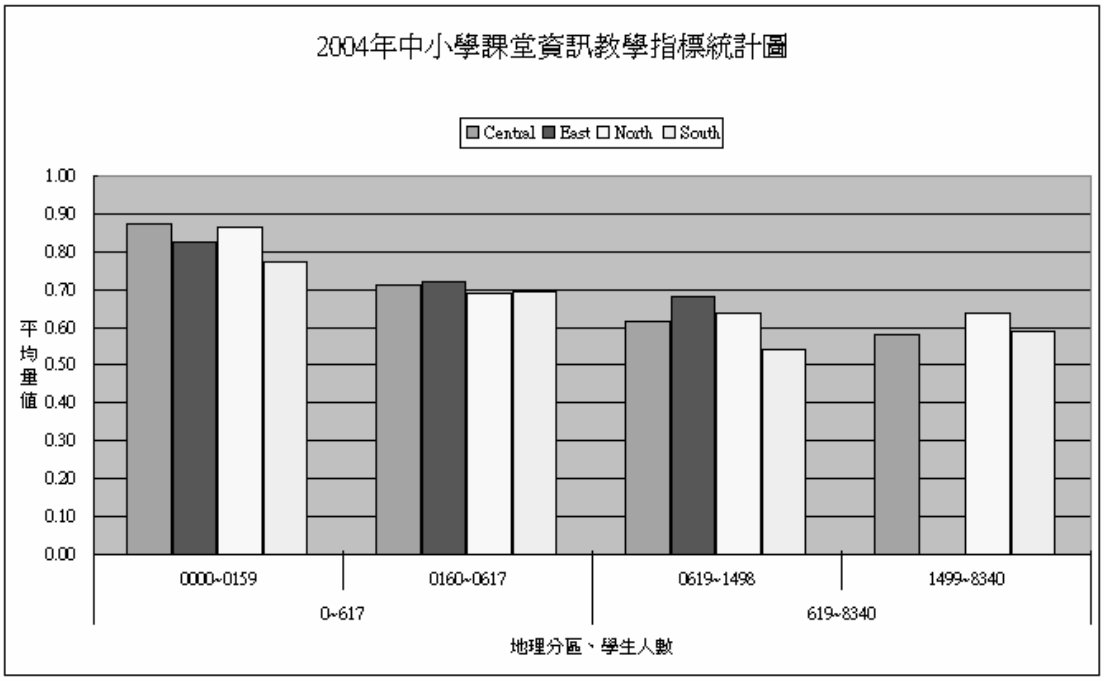

圖 5.7: 課堂資訊教學量值對應學生規模與地理分區之分析圖

表 5.7：課堂資訊教學量值對應學生規模與地理分區之分析表

課堂資訊教學		全國地理分區				
學生規模 L1	學生規模 L2	Central	East	North	South	總計
0~617	0000~0159	0.87	0.83	0.86	0.77	0.82
	0160~0617	0.71	0.72	0.69	0.69	0.70
619~8340	0619~1498	0.62	0.68	0.64	0.54	0.60
	1499~8340	0.58		0.63	0.59	0.61

表 5.8：線上分析之評量參考值

學生規模 L2	0000~0159	0160~0617	0619~1498	1499~8340
平均數	**0.833705594**	**0.7042**	0.619038	0.600871
標準誤	0.023268173	0.007242	0.029014	0.016737
中間值	0.844538462	0.7014	0.627214	0.59
標準差	**0.046536347**	**0.014483**	0.058029	0.02899
變異數	0.002165632	0.00021	0.003367	0.00084
峰度	-0.142489011	-2.85117	1.109959	0
偏態	**-1.000253849**	**0.556834**	-0.77708	1.450223
範圍	0.103345455	0.03	0.138276	0.054837
最小值	0.7712	0.692	0.541724	0.578889
最大值	0.874545455	0.722	0.68	0.633725
總和	3.334822378	2.8168	2.476153	1.802614
個數	4	4	4	3

表 5.8 是以 Excel 試算表軟體，針對「學生規模 L2」維度階層所計算出的敘述統計數據，可供線上分析之評量參考值，從表中所列數據可知在學生人數規模 0~0159 的「課堂資訊教學量值」平均數最高(0.833705594)，在不同地區的「課堂資訊教學量值」的分佈有點分散(標準差 0.046536347)，由偏態為-1.000253849 可知量值分佈並不完全對稱，呈左偏分配。相較之下，學生規模 0160~0617 的「課堂資訊教學量值」平均數次之(0.7042)，但卻有較集中的量值分佈(標準差 0.014483)或可觀察變異數為 0.00021。透過上述統計數值的觀察，我們可看出不同學生人數規模的層級，有不同的量值分佈型態，也代著不同的「課堂資訊教學」趨勢。

5.2. 資料探勘分析

在資料倉儲中所建立的資料立方體(Data cube)，透過線上分析(OLAP)技術，可以清楚的查詢顯現各個學校在不同指標中所表現的量值。為了能進一步分析各學校間造成數位落差之成因，將會產生以下議題需要探討：

- 如何客觀評定一所學校之資訊能力高低？
- 如何鑑別學校間資訊能力高低之成因？
- 如何使用分析結果，提供決策單位參考資訊？

為了解決以上問題，在此開發了使用**兩層式資料探勘方法之 DMAS 線上資料探勘系統**，透過結合資料倉儲與資料探勘技術，使用數位落差問卷與其他歷史統計資料庫，進行多維度的資料探勘分析。多層次資料探勘系統流程圖如下圖 5.8。

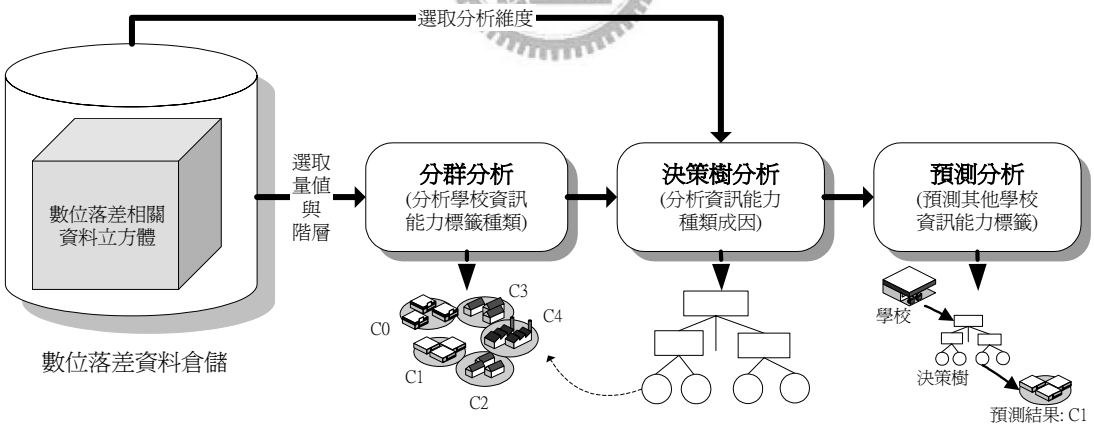

圖 5.8: 兩層式資料探勘方法流程圖

資料探勘分析步驟如下：

(1). 分群分析：

透過分群演算法[18][22]，以學校中學生的能力為量值，並使用分群演算法，將表現類似的學校分到同一群，以此分析學校資訊能力標籤種類。

(2). 決策樹分析：

透過使用決策樹演算法[23]，對於不同分群結果之資訊能力種類，加入其他各項維度欄位來建立決策樹，分析各群造成之原因。

(3). 預測分析：

透過決策樹的建立，可以提供決策單位了解不同環境背景與學校政策造成之資訊能力，並預測其他未做問卷學校之資訊能力標籤。

為了能分析出資訊能力類別，以及類別之成因，所以在此使用了資料探勘技術中的分群演算法與決策樹演算法，層次的搭配分析，並將結果建立決策樹模型(Decision Tree Model)，以利進行之後的預測分析。以下各節將介紹各層資料探勘之分析方法與結果。

5.2.1. 分群分析

透過對學生之資訊能力調查問卷資料，可以依學校為單位彙整出數值型態指標，其中包含：***學校資源***、***社經地位***、***資訊使用支援***、***課堂資訊教學***、***資訊應用***、***資訊近用***、***網路素養***、***資訊技能***、***進階資訊技術***等欄位。這些指標值皆為正規化到 0~1.0 之小數，數字越大代表學校在此指標能力越高，透過這些指標值之高低，可鑑別比較出每所學校中學生在各項指標的平均能力表現。因為在此需要將所有欄位一起納入評斷學校之考慮，然而，如果只是用能力值總分來評定各學校之資能力等級，將會造成以下問題：

- 如何設定各等級之分數範圍是很困難的，且容易因資料變動而造成等級間之臨界值設定不客觀或不準確。
- 對於總分數表現相同的學校，雖被分到同一等級，但是卻喪失細部描述其特性之資訊。

為了能更客觀使用所有欄位一起評斷學校間之資訊能力差異，本研究提出資料探勘技術中之分群演算法(Clustering Analysis)，依資料先將表現相同的學校群聚到同一群，然後再依各群之屬性，來做特徵分析，以顯示出各群對應的指標與特性。

(1). 分群分析

由於在建置資料倉儲時，已經將問卷題目轉換成數值型態欄位值，因此在分群演算法的選擇方面，則使用運算效率較高之 K-Means 演算法，然後分群數由小到大進行多次分群分析，搭配特徵分析、各群群數分配等資訊，以取得最適當之分群數。

(2). 特徵分析

在進行完分群分析之後，不同的群代表不同的資訊能力，接下來則是針對個別的群去進行特徵分析，以瞭解群與群之間的差異點，並可以針對各群表現之資訊能力特徵，使用不同補救數位落差政策。

在分群分析與特徵分析進行完後，會依分群結果產生不同的群與其描述標籤，來代表中小學資訊能力歸納後的類別，有了這些類別標籤，我們即可用較客觀的屬性，來描述學校中學生的資訊能力等級。

5.2.2. 決策樹分析

在分群分析之後，每個學校會依其資訊能力值被歸納到某一群，並產生對應的描述標籤可以敘述此群之特性。接下來則是要分析不同的環境背景，以及不同的學校資訊政策方針，對於這些各個資訊能力類別之關係，並分析各類別之成因。透過之前資料倉儲所彙集的資料中，描述一個學校的環境背景與政策方針，可以整理成下面維度：

歷史統計資料庫相關維度：*地理分區、學生規模、教師規模、是否為私立學校、是否為資訊種子學校。*

學校問卷維度：*資訊教育方案、教師資訊政策、校長支持度不高狀況、教師資訊融入教學能力尚待加強狀況、資訊教學設備不足狀況、現有資訊教學設備維護不易狀況、資訊教學人力不足狀況。*

為了能分析出哪些因素可能為學校資訊能力高低之成因，因此在資料倉儲階段時，我們引用並延伸了問卷以外的其他歷史統計資料庫，如此才能更多元化的分析出數位落差與環境的可能成因，但是結合了許多的資料，即會產生下面的議題：

- 如何從將這許多維度，一起分析出維度對於資訊能力高低的影響？
- 如何有系統的歸納分析結果，以供決策者能更容易更清楚的瀏覽結果？

為了能達到以上之需求，在此透過資料探勘技術中的決策樹分析演算法(Decision Tree Induction)，將各種可能成因欄位加入當作鑑別維度，然後針對分群結果建立決策樹。

範例 12：對各資訊能力類別標籤建構決策樹模型

在此使用 ID3 演算法，主要針對之前分群結果歸納之類別標籤，搭配其他欄位資料建構決策樹。圖 5.9 表示為決策樹之部分結果範例。

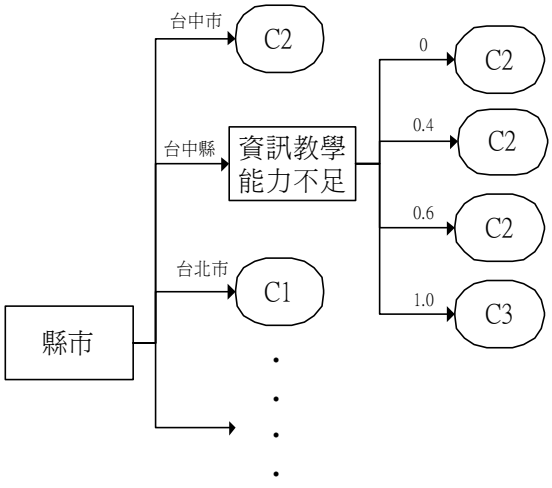

圖 5.9: 學校政策環境對資訊能力之決策樹

決策樹演算法會計算各個維度對於資訊能力類別之鑑別資訊高低，依次將較具鑑別力之維度挑選出來，來將現有學校做分類，並且將分析結果系統化的以樹狀結構來表達分類過成，**因此每個由根節點(root)到葉節點(leaf)的路徑，即可視為一條分類的規則。**以圖 5.8 的決策樹為例:

學校分類規則 ： (縣市=台中縣 and 資訊教學能力不足=1) ➔ C3

對於路徑：**(縣市=台中縣 而且 資訊教學能力不足=1 導致 分類到結果 C3)**，這裡便可以解讀為:在台中縣、資訊教學能力不足情況嚴重的學校，其學校之學生能力大多是分類到 C3 類別，其對應之標籤為：***學校資源最低、課堂資訊教學最低、資訊技能最低***。

為了防止決策樹過度資料依賴(Over fitting)的問題，在此我們使用 Pre-pruning 方法，當某一分枝資料亂度值小於一個門檻值(Threshold)時，我們即輸出分類結果，以防決策樹建立的結果只針對這份培訓資料有用，而無法提供較客觀的分類結果。

透過決策樹的建立，即可系統化的從許多對學校相關的描述維度中，分析出較具資訊與鑑別力維度，並可建立成樹狀分類結構，以提供進一步的預測分析使用。

5.2.3. 預測分析

結合之前分群分析所歸納之資訊能力等級標籤，與決策樹分析中所建立之決策樹，可以產生學校資訊能力之分類器(Classifier)。透過此分類器，即可從一所學校之環境，與教學策略等資訊來推估其學生可能之資訊能力表現，例如範例 13。

範例 13：使用決策樹分類器進行預測分析

從資料中任選一高中進行預測分析，由該學校問卷相關資料: 縣市=台中縣、資訊教學人力不足=1.0，配合決策樹分類器進行分類預測，如圖 5.10，即可知道該學校可能屬於 C3 ，該類別標籤為：***學校資源最低、課堂資訊教學最低、資訊技能最低***。

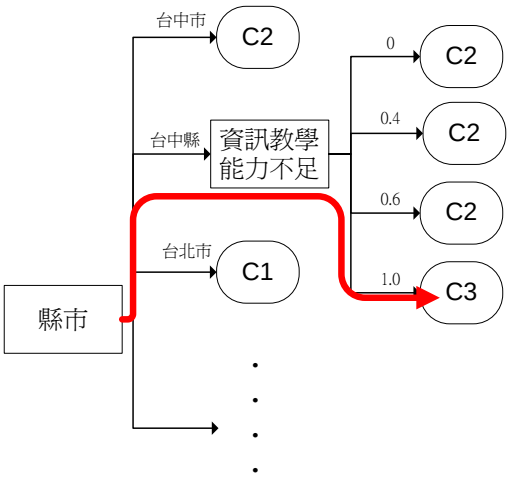

圖 5.10: 透過決策樹進行預測分析

由於分類器是由部分有做問卷統計之學校資料培訓出來，並系統化的建構成樹狀結構，或儲存成一條條規則格式，因此用在預測分析時，可以快速的幫助決策單位提供符合之前資料規則之分類訊息。

第六章　系統實作

本章節將介紹透過實際建立之資料立方體，結合本研究所提出之分析流程，線上分析的實作結果，以及資料探勘的實作結果。透過結合現有系統與我們開發之 DMAS 線上探勘系統，讓使用者更容易透過各種面向進行多維度動態問卷資料分析。

6.1. 線上分析流程實作

本研究中提出的線上分析流程，主要是透過 MS-SQL Server 2000 之線上分析系統 Analysis Service 建構出資料立方體，資料分析者可直接在線上分析系統上瀏覽、編輯及修改資料立方體結構，其使用者介面如圖 6.1 所示。

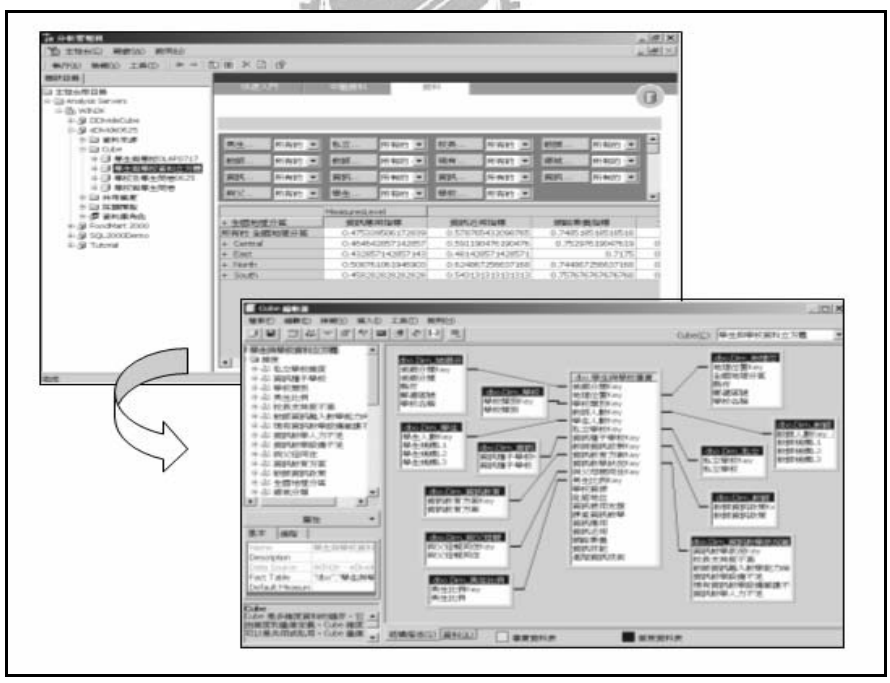

圖 6.1：線上分析系統之資料立方體

在透過資料倉儲與線上分析系統之建立資料立方體後，資料分析者也可以使用試算表軟體，例如:Excel 來連接資料立方體，利用表格或圖表進行樞紐分析，分析者也可以透過表格或圖表顯現方式的切換，任意加入或刪除分析之維度欄位，進行線上分析之上捲(Roll up)、下探(Drill down)、切片(Slice)、切丁(Dice)、轉軸(Pivot)等分析操作，線上分析系統畫面如下圖 6.2。

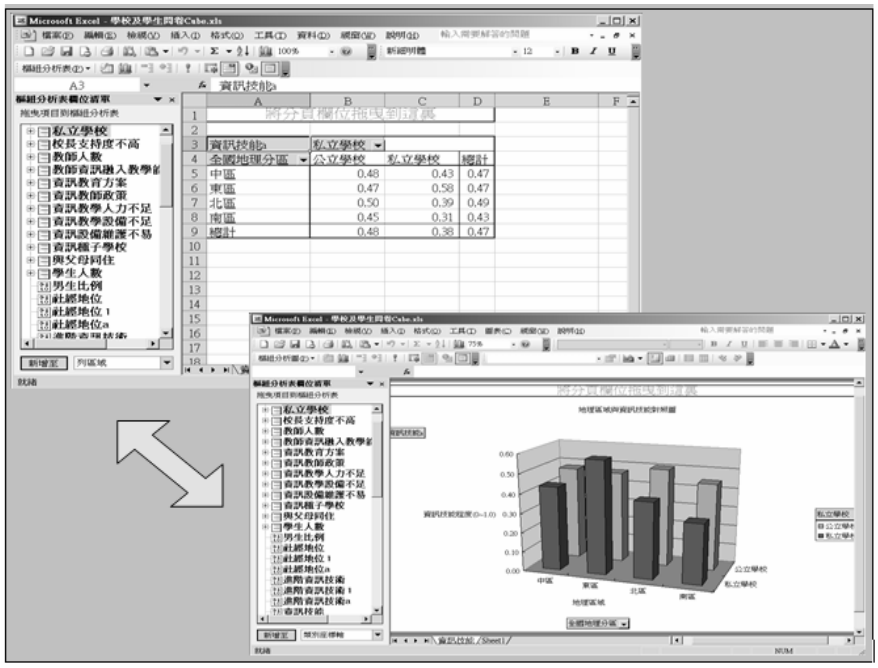

圖 6.2: Excel 樞紐分析畫面

從第五章中的「中小學數位落差 OLAP 主題分析表」(表 5.6)中，可看出分析主題可概分為五個分別是「資訊學習環境」、「資訊近用」、「資訊應用」、「資訊素養」及「整體綜合」分析主題、表中詳列每個分析主題的相關量值及相關維度，同時我們可從相關量值欄中的資料了解到這些量值相關的內在意含。

本章將以此五個分析主題相關的 9 個量值指標，並參考領域專家建議後，搭配問卷量值概念階層表中的 14 個相關維度表，組成 OLAP 主題量值與維度之分析順序表如(表 6.1)後進行線上分析實作，並詳列重要結果如後：

表 6.1：OLAP 主題量值與維度之分析順序表

分析維度 / 分析主題	地理位置	學生規模	教師規模	是否為種子學校	是否為私立學校	資訊教學人力不足	資訊教學設備不足	教師資訊政策	教師資訊融入教學能尚待加強	現有資訊教學設備維護不易	校長支持度不高	與父母親同住比例	男生比例	資訊教育方案
社經地位	1	4		5	2	3								
資訊使用支援	1	2		3	4	5						6		7
課堂資訊教學	1		2	3	4		5	6	7	8	9		10	11
資訊近用	1	2		3	4	5					6	7		
學校資源	1		2	3	4	5		6	7		8			9
資訊應用	1	2		3	4	5	6	7				8	9	
資訊技能	1	2	3	4	5		6	8	9	7	10	11	12	
進階資訊技術			1		2		3	4	5	6		7		
網路素養	1	4	6	5	2	3							7	

(1)「資訊學習環境」主題分析

在這個分析主題中主要包括「課堂資訊教學」、「資訊使用支援」、「社經地位」3 個指標量值的分析。

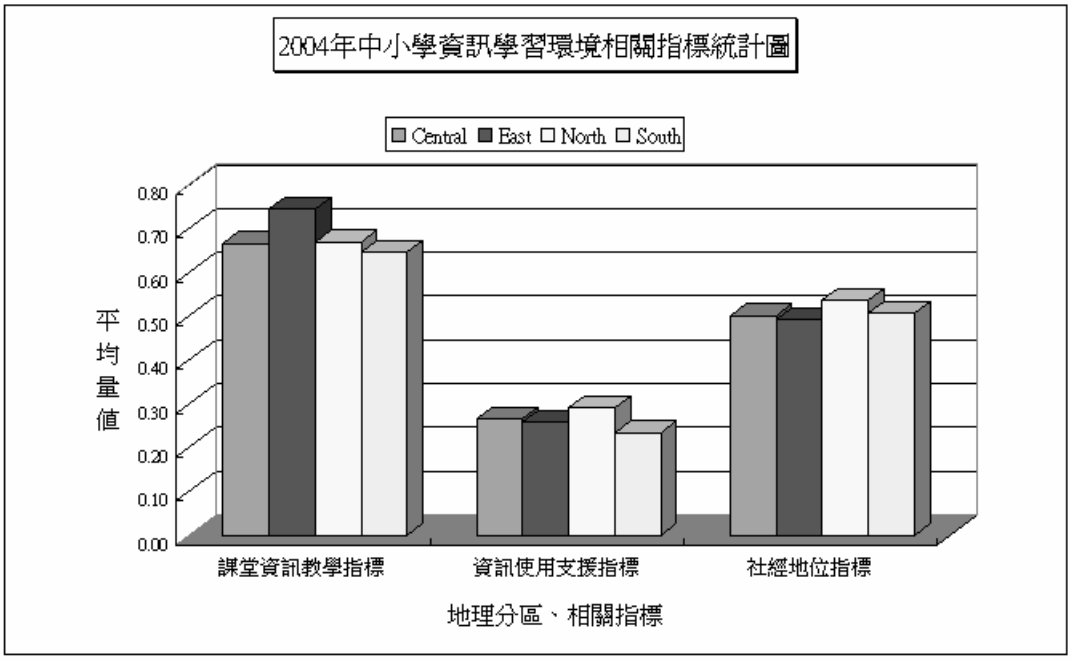

圖 6.3: 學生資訊學習環境相關指標統計圖

在資訊學習環境主題的分析中，我們可看出以**課堂資訊教學**量值較高，代表各區學校在課堂資訊教學表現最佳，顯示各區學校在**課堂資訊教學方面相當支持**，其中東區學校表現最理想。但在資訊使用支援量值最不理想，顯示各區學生**家長在資訊使用支援的部分有困難**。在社經地位量值部分整體而言，相差不大但仍以北區學校為佳。

由上圖 6.3 中我們可了解到中小學學生資訊學習環境的現況是：家長的**社經地位**中等，但**資訊使用支援**的部分有困難，主要的**資訊學習環境**仍偏重於學校。為了解各相關量值細部的差異，我們可參考主題分析表，進行下探(Drill down)分析，來了解各區中小學在資訊學習環境的差異特徵如下：

- **「課堂資訊教學」量值分析**

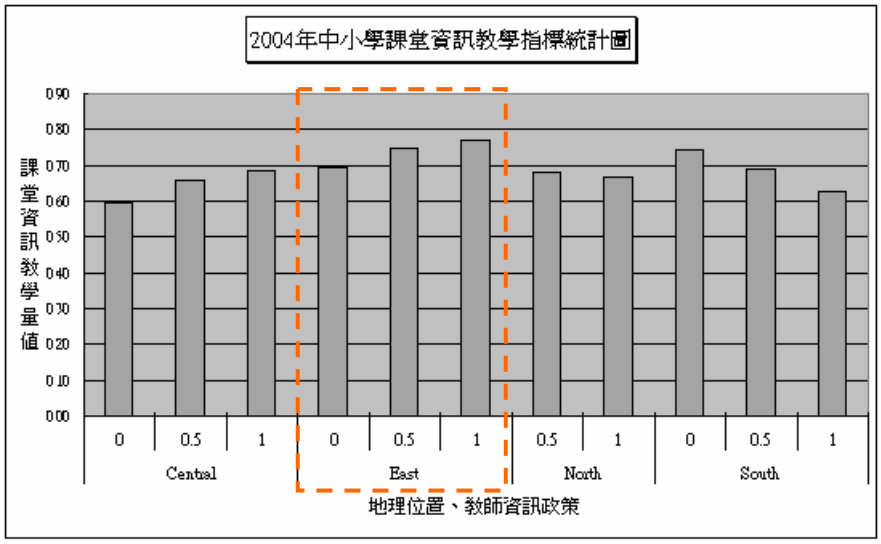

圖 6.4: 學校地理位置及教師資訊政策維度分析課堂資訊教學之量值統計圖

從上圖 6.4 學校地理位置及教師資訊政策維度的分析中我們可看出，**中區及東區學校的「課堂資訊教學」量值與「教師資訊政策」維度有正相關的趨勢**，北區及南區則是相反的趨勢。同時東區學校的「課堂資訊教學」的平均量值也是最佳，代表**中區及東區學校中「教師較在課堂上使用電腦來協助學生上課及完成作業」與由「學校提供教師資訊教學應用研習或對資訊組長或網管人員獎勵」有正比例的趨勢**，相較於北區及南區學校這種情況較少，也顯示「教師資訊政策」對「課堂資訊教學」的影響在東區及北區學校有較大的差異。

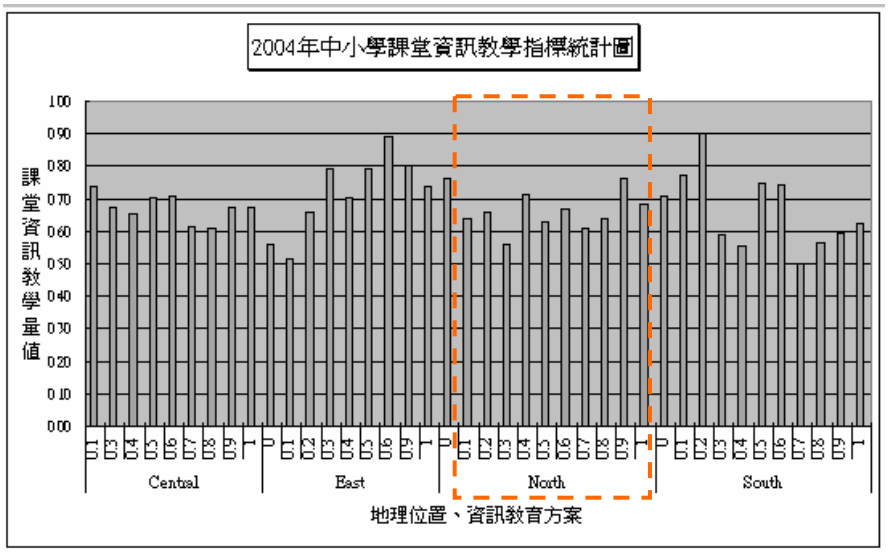

圖 6.5: 學校地理位置及資訊教育方案維度分析課堂資訊教學之量值統計圖

從上圖 6.5 學校地理位置及資訊方案維度的分析中我們可看出，**北區學校的「課堂資訊教學」量值與「資訊教育方案」維度有正相關的趨勢**，北區學校資訊方案在 0.9~1 時，有最佳的「課堂資訊教學」量值，代表**北區學校的「教師在課堂上使用電腦來協助學生上課及完成作業」**與**「舉辦學生資訊應用相關競賽或舉辦資訊融入教學觀摩」有正比例的趨勢。**這種以學校的資訊方案而提昇課堂資訊教學的強度會因地區而不同，東區學校資訊教育方案在 0.6 時，南區學校資訊教育方案在 0.2 時，有較佳的「課堂資訊教學」量值。

我們也可針對北區學校進行下探(Drill down)分析，如圖 6.6 所示，可看出北區學校大部在資訊教育方案在 0.9~1 時，有較佳的「課堂資訊教學」量值，但是台北市的學校則在資訊教育方案在 0.5 及 0.9 時，有較佳的「課堂資訊教學」量值。

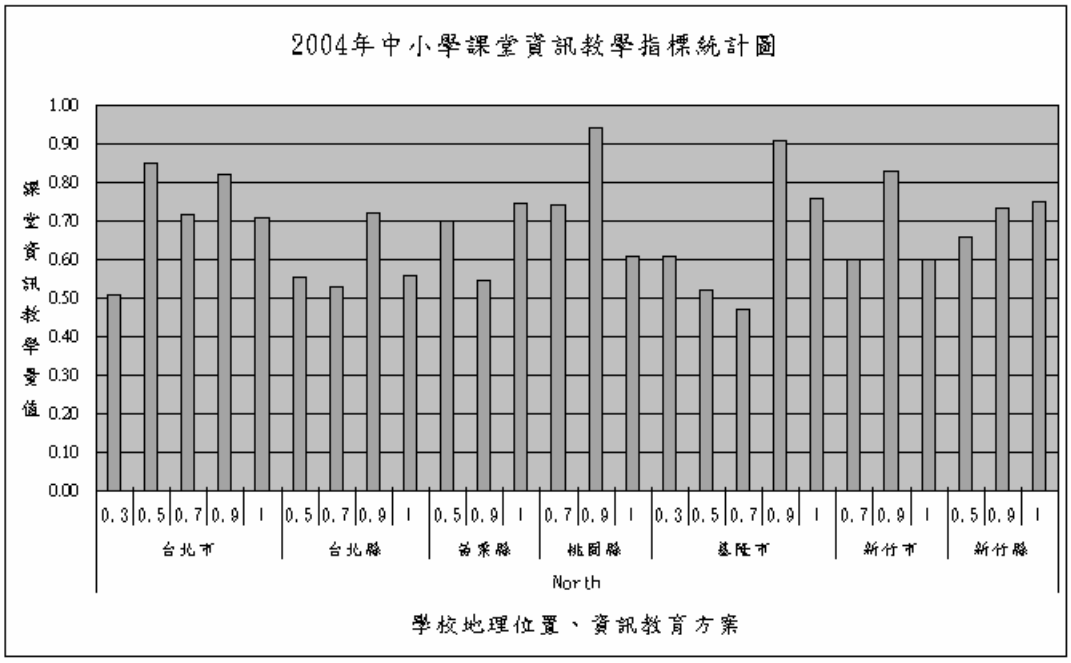

圖 6.6: 北區學校及資訊教育方案維度分析課堂資訊教學之量值統計圖

- **「資訊使用支援」量值分析**

在「學生資訊學習環境相關指標統計圖」(圖 6.3)中的「資訊使用支援」量值部分可看出各區學生家長在這方面普遍是有因難的，其中表現較佳的地區是北區學校，所以我們針對北區學校來分析。從下圖 6.7 學校地理位置及資訊方案維度的分析中我們可看出，**北區學校的「資訊使用支援」量值與「資訊教育方案」維度**

有正相關的趨勢，代表**北區多數的學校在「學校舉辦學生資訊應用相關競賽或舉辦資訊融入教學觀摩」**與「**學生家長使用網路的能力」有正比例增加的趨勢。**其中桃園縣學校趨勢不同於北區其他縣市，這種特殊情形我們可透過下探(Drill down)分析找出個案所在地點，再進行個案更進一步的研究。

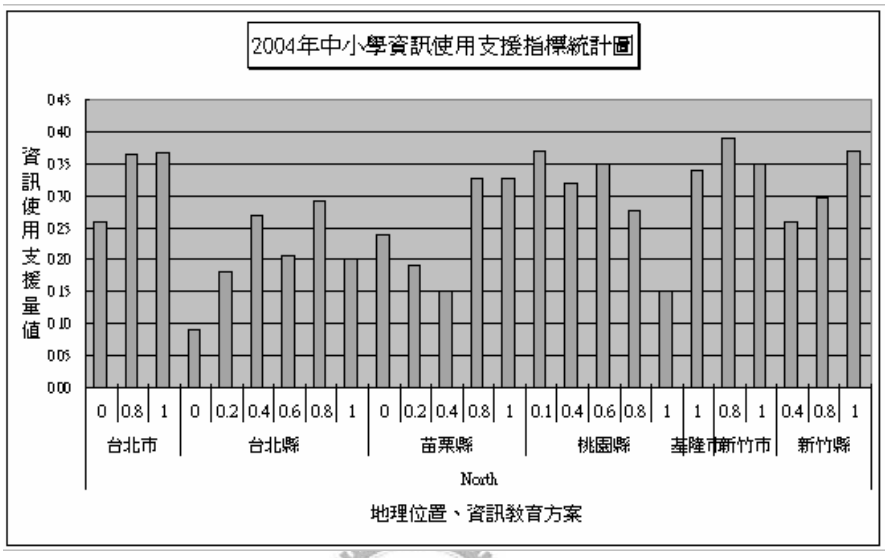

圖 6.7: 北區學校及資訊教育方案維度分析資訊使用支援之量值統計圖

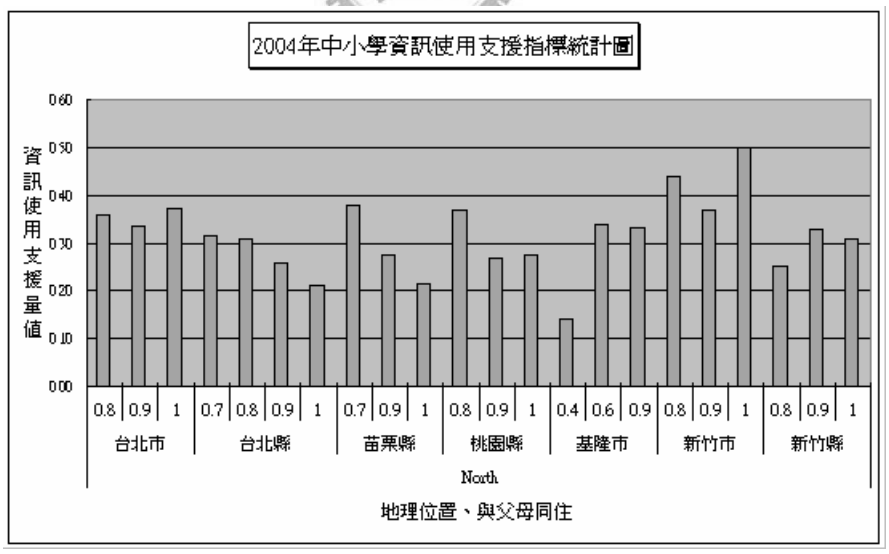

圖 6.8: 北區學校及與父母同住維度分析資訊使用支援之量值統計圖

從圖 6.8 北區學校及與父母同住維度的分析中我們可看出，北區學校的「資訊使用支援」量值較佳的分佈情形，**台**北市、基隆市、新竹市、新竹縣是分佈於「與

父母同住」比例為 0.9~1,代表**與父母親同住**的學生有較佳「資訊使用支援」量值，其他的縣市，台北縣、苗栗縣、桃園縣是分佈於「與父母同住」比例為 0.7~08, 代表與**僅與父親同住**的學生有較佳「資訊使用支援」量值。

- **「社經地位」量值分析**

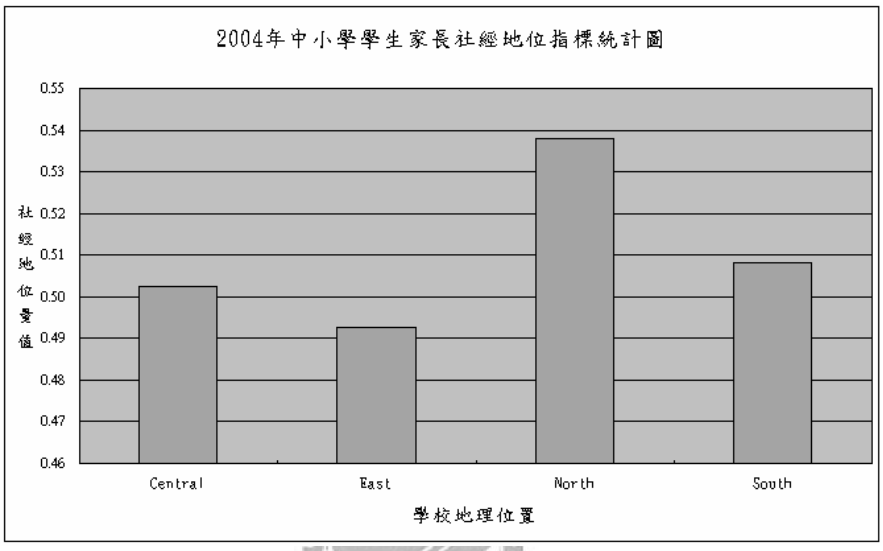

圖 6.9: 學校地理位置維度分析社經地位之量值統計圖

在上圖 6.9 學校地理位置維度分析社經地位之量值統計圖中，各區學生家長的社經地位的量值，其中以**北區學生家長「社經地位」量值較高**。

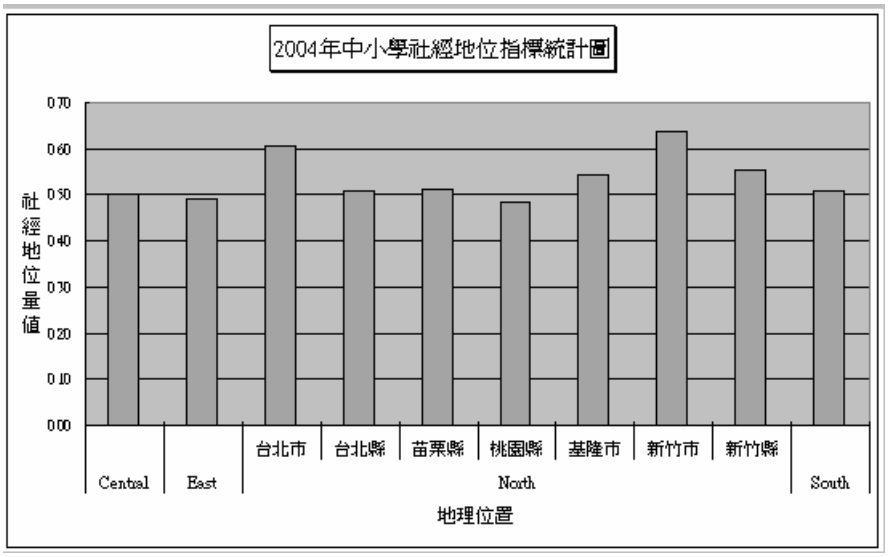

圖 6.10: 北區學校維度分析社經地位之量值統計圖

針對北區學校進行下探(Drill down)分析後，由圖 6.10 可看出其中**新竹市及台北市**

的學生家長「社經地位」量值較高，代表新竹市及台北市的學生家長平均而言，有較高的學歷及經濟收入。

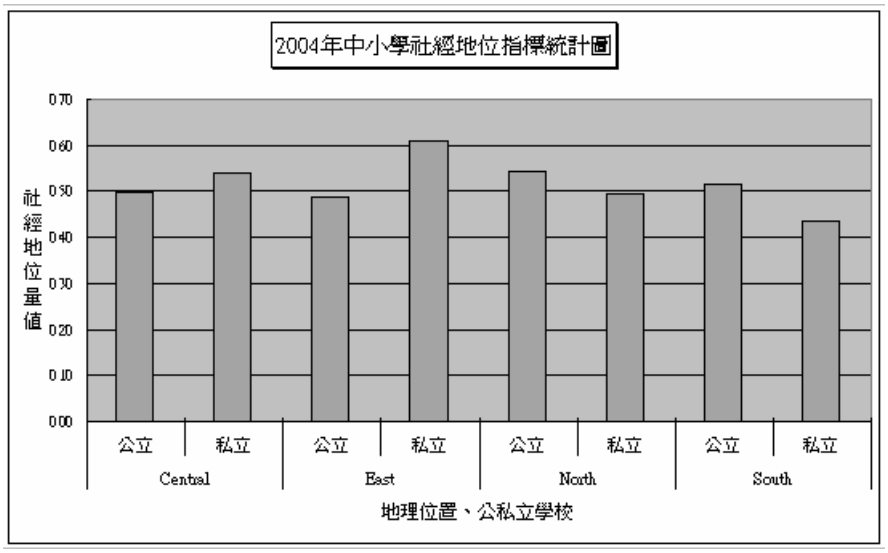

圖 6.11: 加入公私立學校維度分析社經地位之量值統計圖

在加公私立學校維度後，如圖 6.11 可看出中區及東區**私立學校**的學生家長有較佳的社經地位，北區及南區則是**公立學校**的學生家長有較佳的社經地位。

(2)「資訊近用」主題分析

在這個分析主題中主要包括「學校資源」、「資訊近用」2 個指標量值的分析。

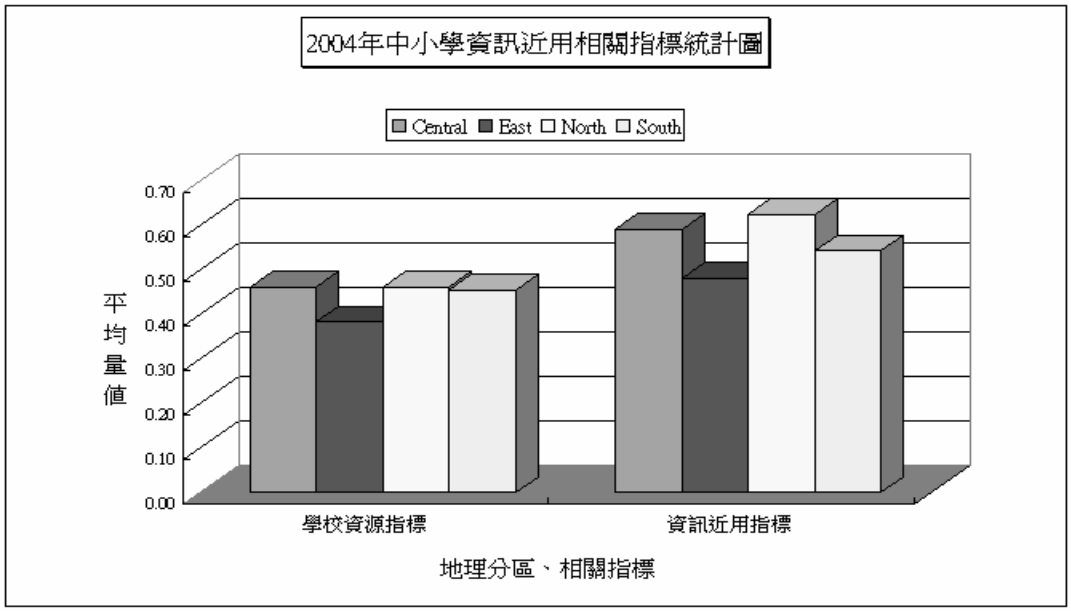

圖 6.12: 學生資訊近用相關指標統計圖

在資訊近用主題的分析中，由圖 6.12 中我們可看出以「資訊近用」量值較高，代表各區學生在資訊近用(家中設備)是中等的表現，其中以中區、北區學生資訊近用表現較佳，代表學生在資訊近用(家中設備)的情況是會隨地理位置而出現差距。

各區學校在「**學校資源」量值表現上並不理想**，顯示各區學校在學校網頁、數位教學(材)資源的部分有待加強。其中東區學校學校資源表現較不理想。

我們可透過相關維度的線上分析找出主題量值的特徵如下：

- **「學校資源」量值分析**

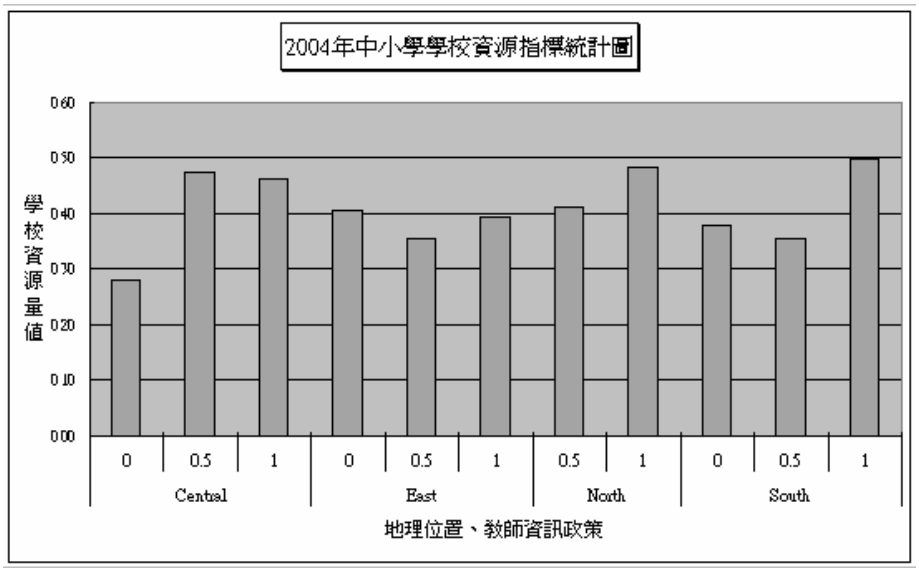

圖 6.13: 學校地理位置及教師資訊政策維度分析學校資源之量值統計圖

從圖 6.13 學校地理位置及教師資訊政策維度的分析中我們可看出，大部分學校的「教師資訊政策」與「學校資源」都有正相關的趨勢，代表**「學校提供教師資訊教學應用研習或對資訊組長或網管人員獎勵」與「學校網頁、數位教學(材)資源」有成正比例增加的趨勢**。現階段在「學校資源」量值表現方面，則以中區及北區學校較佳。

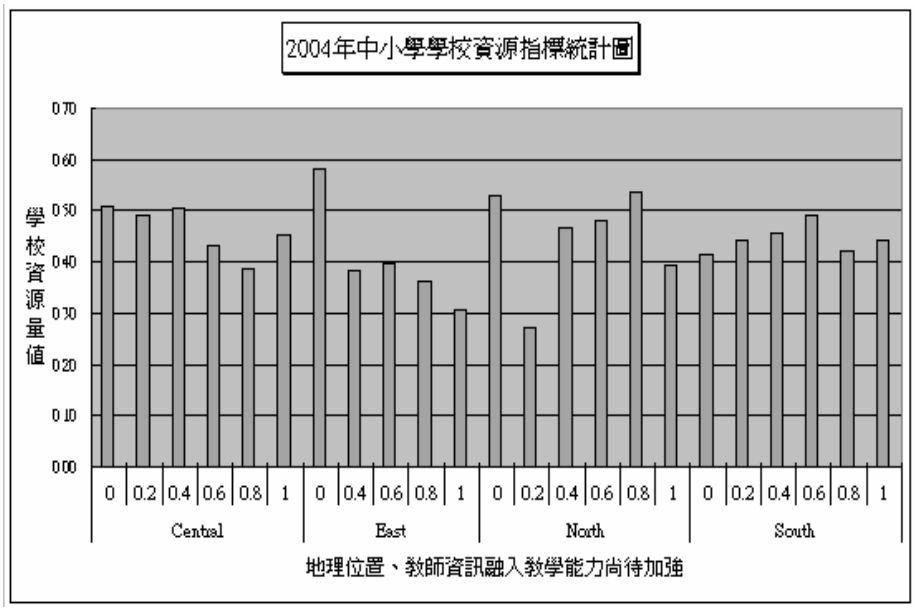

圖 6.14: 資訊融入教學能力尚待加強維度分析學校資源之量值統計圖

從圖 6.14 學校地理位置及教師資訊融入教學能力尚待加強維度的分析中我們可看出，大部學校的「學校資源」與「教師資訊融入教學能力尚待加強」都有負相關的趨勢，代表**「教師資訊融入教學能力尚待加強」與「學校網頁、數位教學(材)資源」有成反比例增加的趨勢**。換句話說，也就是「教師資訊融入教學能力尚待加強」的量值愈低，「學校網頁、數位教學(材)資源」量值有愈高的趨勢，

值得注意的是，上圖中北區學校有「教師資訊融入教學能力尚待加強」量值在 0.2 時學校資源也最低，這代表北區學校中存在著特殊趨勢的學校，也就是「教師資訊融入教學能力愈好」，「學校網頁、數位教學(材)資源」卻沒有正比例增加的趨勢，這些情形我們可透過下探(Drill down)分析找出個案所在地點，再進行個案更進一步的研究。

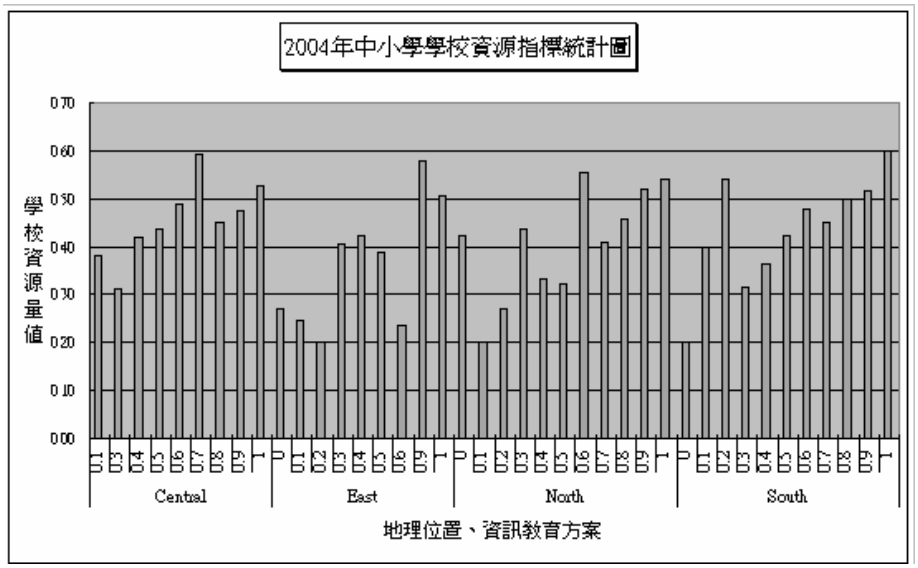

圖 6.15: 學校地理位置及資訊教育教育方案維度分析學校資源之量值統計圖

從上圖 6.15 學校地理位置及資訊教育方案維度的分析中我們可看出，學校的「學校資源」與「資訊教育方案」維度有正相關的趨勢，大部分學校的「資訊教育方案」在 0.9~1 時，有較佳的「學校資源」量值，代表**北區學校「舉辦學生資訊應用相關競賽或舉辦資訊融入教學觀摩」**與**「學校網頁、數位教學(材)資源」有正比例增加的趨勢**。值得注意的是，在大趨勢下也存在一些各區特有的情況，如中區學校在「資訊教育方案」維度選值為 0.7，北區學校在「資訊教育方案」維度選值為 0.6，南區學校在「資訊教育方案」維度選值為 0.2 時也有很好的「學校資源」量值，這些情形我們可透過下探(Drill down)分析找出個案所在地點，再進行個案研究。

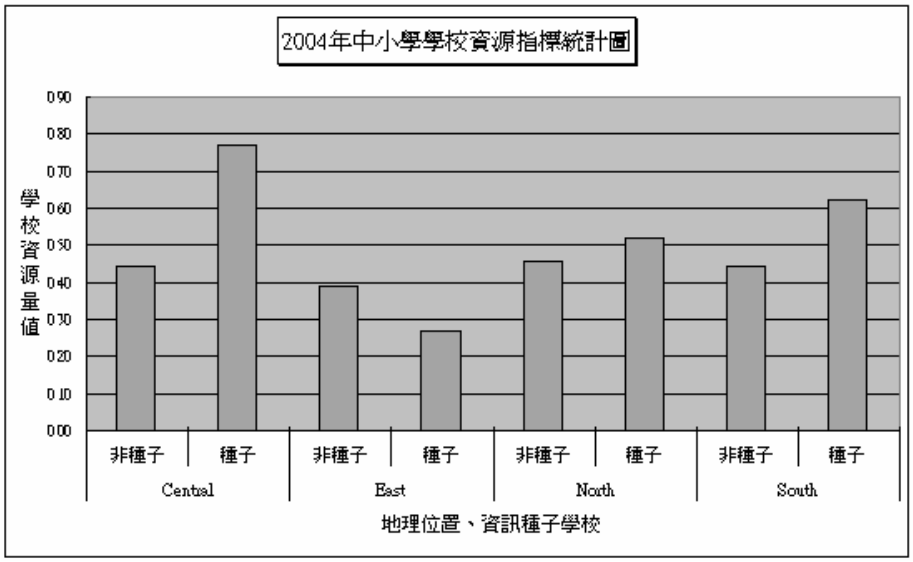

圖 6.16: 學校地理位置及資訊種子學校維度分析學校資源之量值統計圖

從上圖 6.16 學校地理位置及資訊種子學校維度的分析中我們可看出，中區、北區及南區學校的「學校資源」與「資訊種子學校」維度有正相關的趨勢，代表**中區、北區及南區資訊種子學校具有較佳「學校網頁、數位教學(材)資源」的趨勢。**

至於**東區資訊種子學校「學校網頁、數位教學(材)資源」較不理想**的情況，我們可透過下探(Drill down)分析找出個案所在地點，另外進行個案研究。

- **「資訊近用」量值分析**

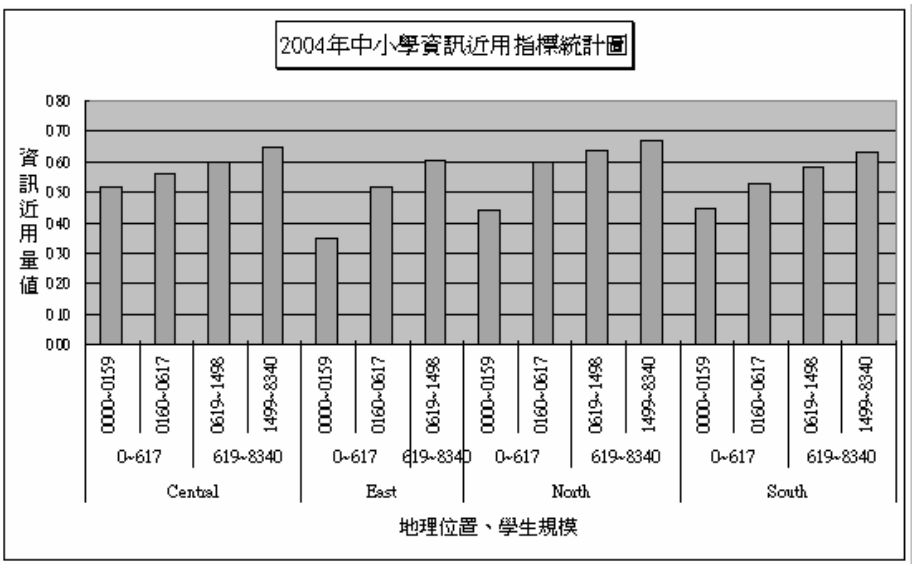

圖 6.17: 學校地理位置及學生人數維度分析學生資訊近用之量值統計圖

從上圖 6.17 學校地理位置及學生人數維度的分析中我們可看出，各區**學校的「學生人數」與學生的「資訊近用」有正相關的趨勢**，代表**各區學校的「學生人數」與「家裡的電腦擁有率，上網頻寬及時數」成正比率增加的趨勢**。可推測為，較大型的學校中的學生家中的電腦及上網設備較佳，具有較佳的「資訊近用」量值。

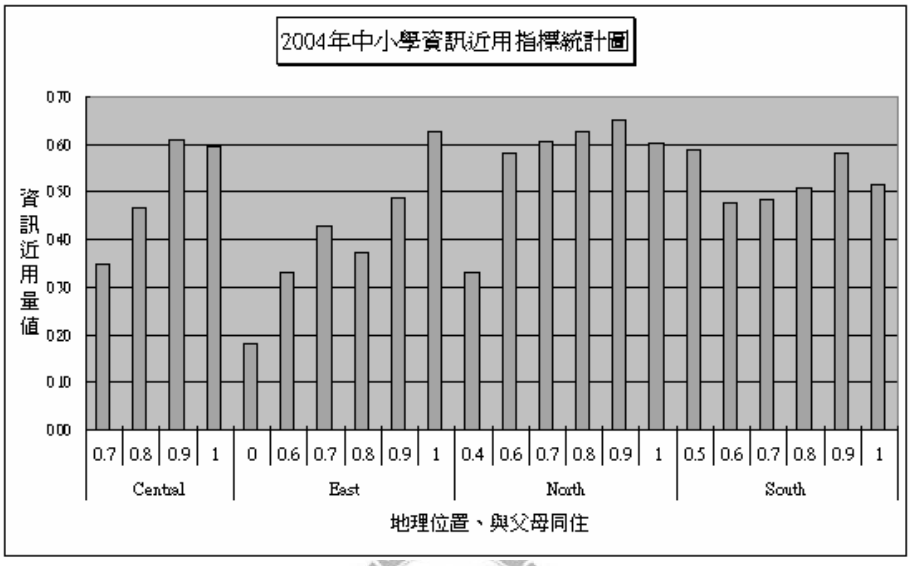

圖 6.18: 學校地理位置及與父母同住維度分析學生資訊近用之量值統計圖

從上圖 6.18 學校地理位置及與父母同住維度的分析中我們可看出，各區**「與父母同住」維度與學校的「資訊近用」有正相關的趨勢**，代表各區學校學生的「與父母同住」比例與「家裡的電腦擁有率，上網頻寬及時數」成正比率增加的趨勢。可推測為，與父母雙親同住的學生家中的電腦及上網設備較佳，具有較佳的「資訊近用」量值。

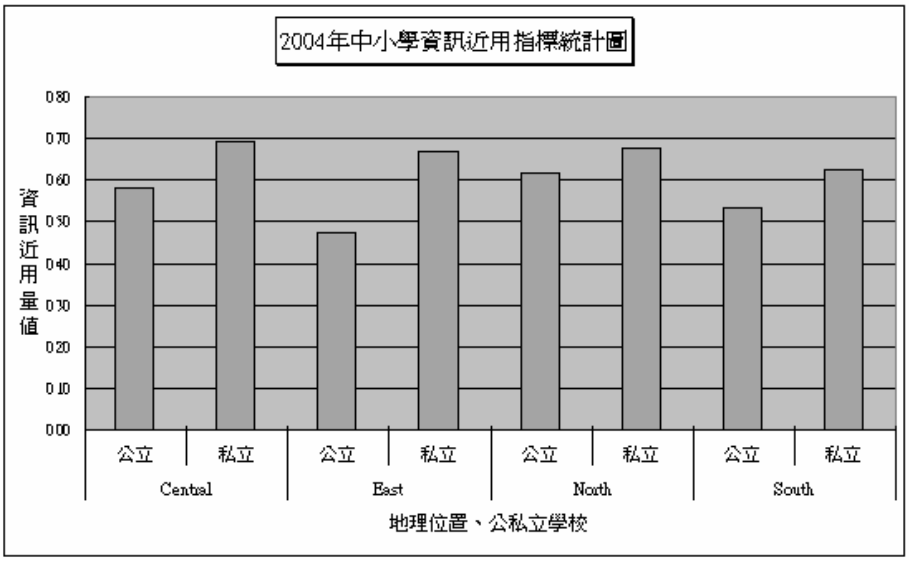

圖 6.19: 學校地理位置及公私立學校維度分析學生資訊近用之量值統計圖

從上圖 6.19 學校地理位置及與公私立學校維度的分析中我們可看出，各區**私立學校學生具有較佳的「資訊近用」量值**，代表各區私立學校學生的「家裡的電腦擁有率，上網頻寬及時數」量值較佳。可推測為，私立學校學生家中的電腦及上網設備較佳，具有較佳的「資訊近用」量值。

(3)「資訊應用」主題分析

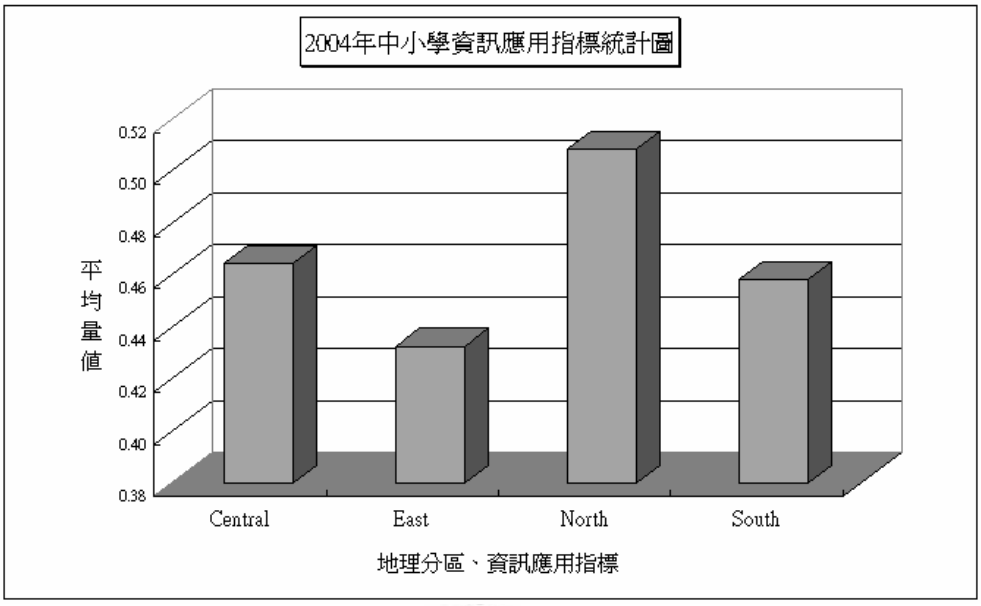

圖 6.20: 學生資訊應用指標統計圖

在資訊應用主題的分析中(圖 6.20)，我們可看出以北區中小學學生資訊應用量值較高，代表**北區學生**在資訊應用(上網的習慣及時數)表現較佳，其次是中區學生，最後是東區學生資訊應用表現較不理想，可見在不同地區學生在**資訊應用**情況仍存在著不同的差距，其中以北區與東區中小學學生的資訊應用情況差距最大。

- **「資訊應用」量值分析**

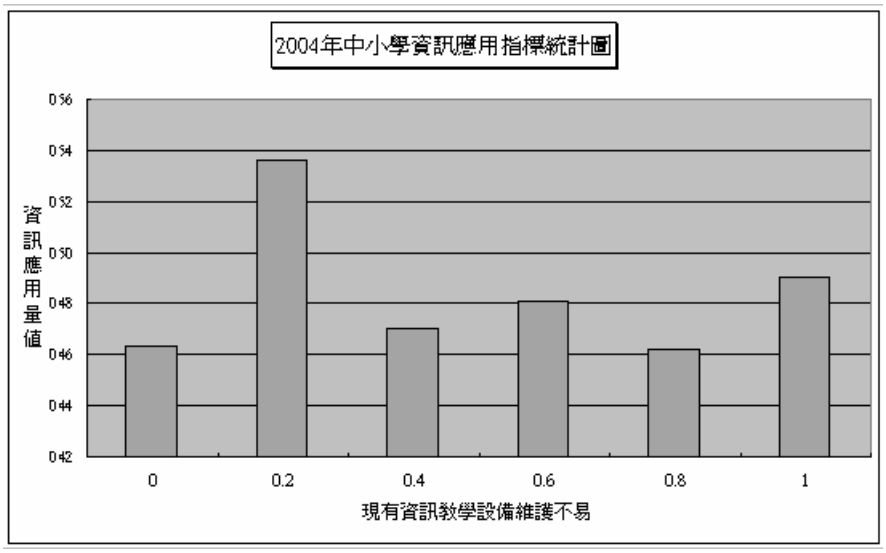

圖 6.21: 現有資訊教學設備維護不易維度分析學生資訊應用之量值統計圖

從上圖 6.21 學校地理位置及與現有資訊教學設備維護不易維度的分析中我們可看出，各區**「現有資訊教學設備維護不易」維度選值為 0.2 時，學校學生具有較佳的「資訊應用」量值**，可推測為，各區學校的「現有資訊教學設備」不是很老舊，而且維護不易的情形不嚴重時學生有較佳的上網找資料、和其他同學透過網路合作收集資料完成作業及上網路跟朋友或同學討論問題等的表現。

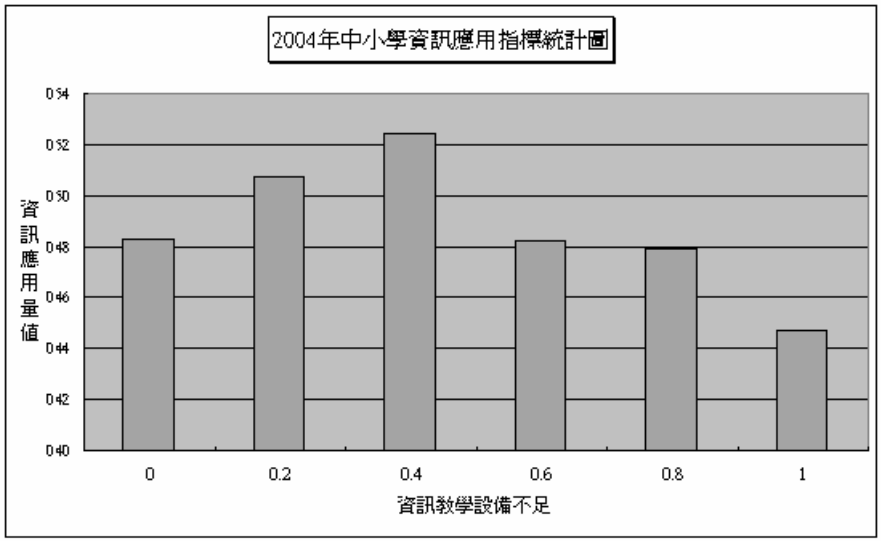

圖 6.22: 現有資訊教學設備不足維度分析學生資訊應用之量值統計圖

從上圖 6.22 學校地理位置及與現有資訊教學設備不足維度的分析中我們可看出，各區**「現有資訊教學設備不足」維度選值為 0.4 時，學校學生具有較佳的「資訊應用」量值「現有資訊教學設備不足」維度選值為 1 時，學校學生具有最差的「資訊應用」量值**，可推測為在各區學校的「現有資訊教學設備不足」的情形不嚴重時學生會有較佳的上網找資料、和其他同學透過網路合作收集資料完成作業及上網路跟朋友或同學討論問題等的表現。

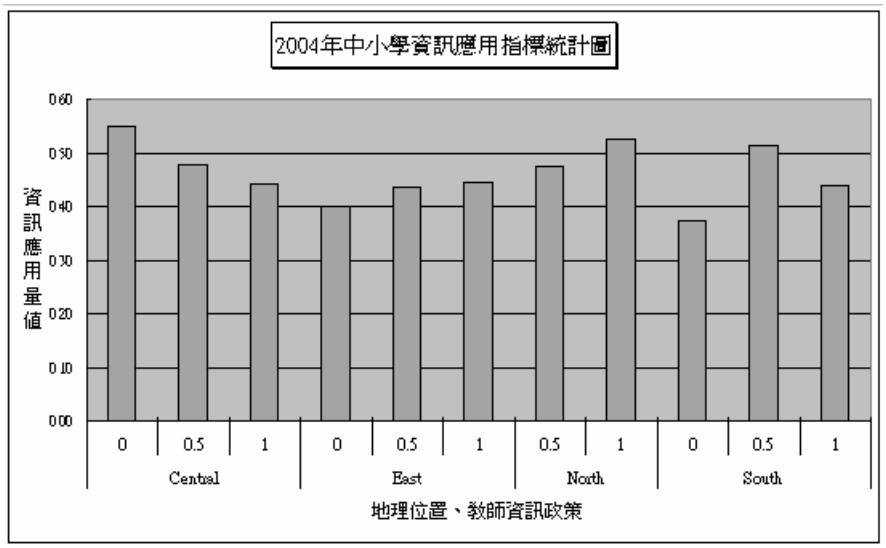

圖 6.23: 學校地理位置及教師資訊政策維度分析學生資訊應用之量值統計圖

另外透過加入[**學校地理位置**]、[**教師資訊政策**]，進行線上分析可以得到上圖 6.23。**東區及北區學校的學生「資訊應用」量值與「教師資訊政策」維度有正相關的趨勢，中區學校的學生「資訊應用」量值與「教師資訊政策」維度有負相關的趨勢。**根據此圖中可以解釋為東區、北區若學校的「學校提供教師資訊教學應用研習或對資訊組長或網管人員獎勵」值較高時，也就是推動較多時，則學生的上網找資料、和其他同學透過網路合作收集資料完成作業及上網路跟朋友或同學討論問題等的表現也會較佳，而中區則是相反的情形，至於南區則是在教師資訊政策中等時有較佳的學生之資訊應用。

(4)「資訊素養」主題分析

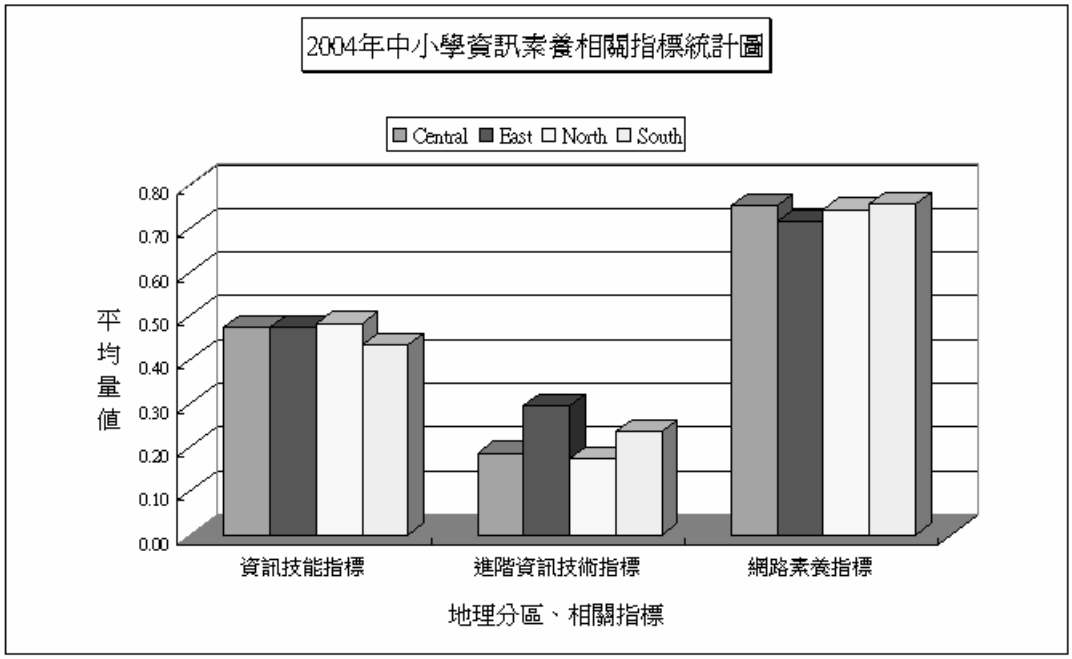

圖 6.24: 學生資訊素養相關指標統計圖

在這個分析主題中主要包括「資訊技能」、「進階資訊技術」、「網路素養」3 個指標量值的分析。由上圖 6.24 中我們可了解到全國中小學學生資訊技能(電腦網路技能)趨向中等，但進階資訊技術(進階技能)的部分並不理想，顯示學生在資訊技術深度的不足。但是相對而言，各區中小學普遍有不錯的網路素養(道德)。

- **「資訊技能」量值分析**

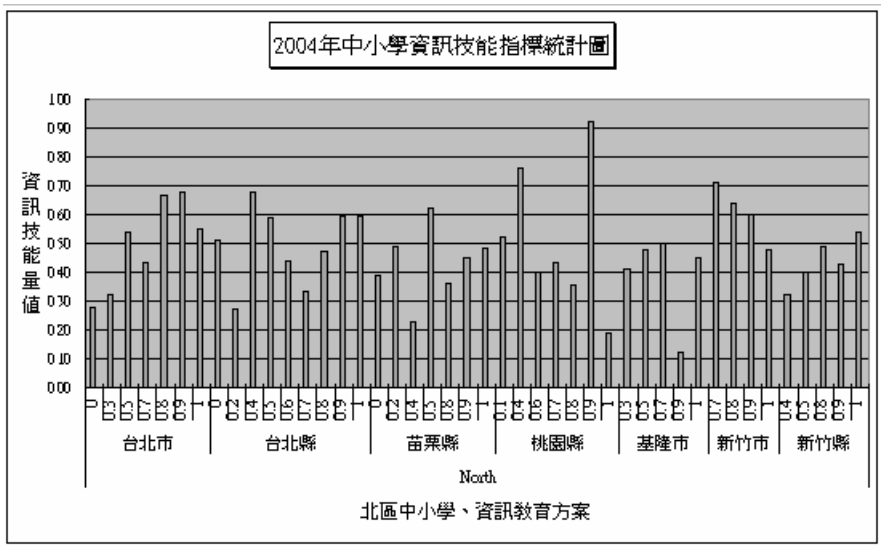

圖 6.25: 北區學校及資訊教育方案維度分析學生資訊技能之量值統計圖

在「學生資訊素養相關指標統計圖」(圖 6.1)中我們可看出北區學校的學生有

較佳的資訊技能量值，因此先針對北區學校進行分析。

從上圖 6.25 學校地理位置及資訊教育方案維度的分析中我們可看出，**大部分的北區學校的「資訊教育方案」值與學生的「資訊技能」量值有正相關的趨勢**，大部分學校的「資訊教育方案」在 0.8~1 時，有較佳的「資訊技能」量值，可解釋為**大部分的北區學校「舉辦學生資訊應用相關競賽或舉辦資訊融入教學觀摩」與「學生基本電腦及網路功能的使用能力」有成正比例的趨勢。**值得注意的是，有些縣市在大趨勢下兼具一些各縣市特有的情況，如台北縣學校在「資訊教育方案」維度選值為 0.4~0.5，苗栗縣學校在「資訊教育方案」維度選值為 0.5，新竹市學校在「資訊教育方案」維度選值為 0.7 時也有很好的「資訊技能」量值，這些情形我們可透過下探(Drill down)分析找出個案所在地點，再進行個案研究。

- **「進階資訊技術」量值分析**

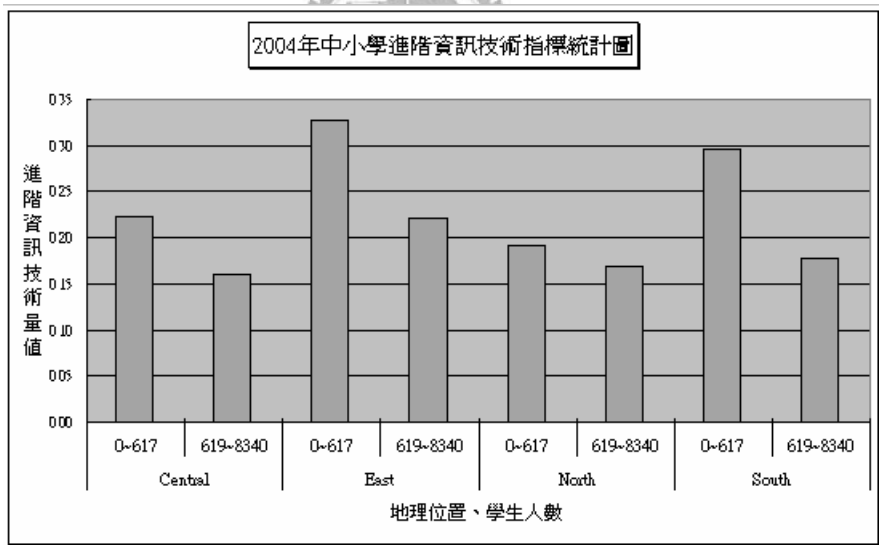

圖 6.26: 學校地理位置及學生人數維度分析學生進階資訊技能之量值統計圖

從上圖 6.26 學校地理位置及學生人數維度的分析中我們可看出，各區**學校的「學生人數」與學生的「進階資訊技能」有負相關的趨勢**，代表**各區學校的「學生人數」與「自己維修電腦，參與電腦相關比賽活動」有成反比例的趨勢。**可推測為，較小型的學校中的學生自己維修電腦的能力以及參與電腦相關比賽活動的情況較佳，具有較佳的

「進階資訊技能」量值。

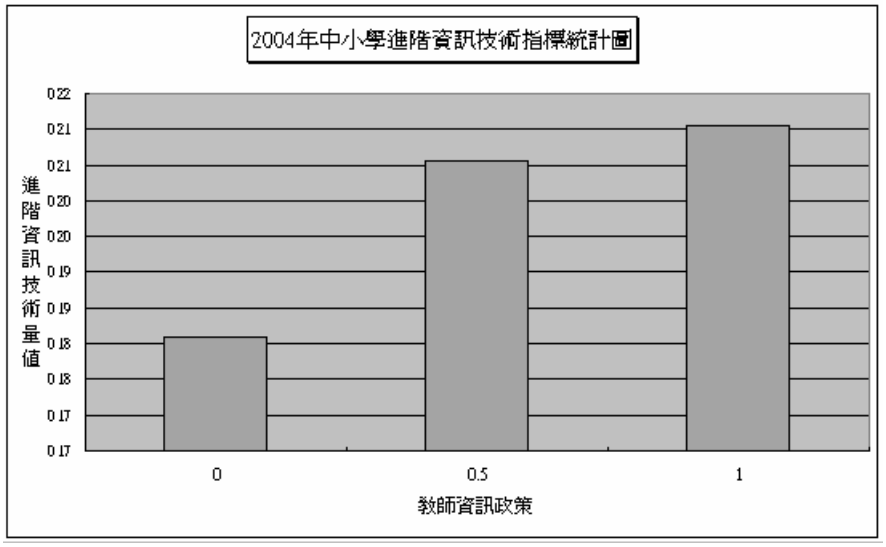

圖 6.27: 教師資訊政策維度分析學生進階資訊技能之量值統計圖

從上圖 6.27 教師資訊政策維度的分析中我們可看出，各區**學校的「教師資訊政策」與學生的「進階資訊技能」有正相關的趨勢**，

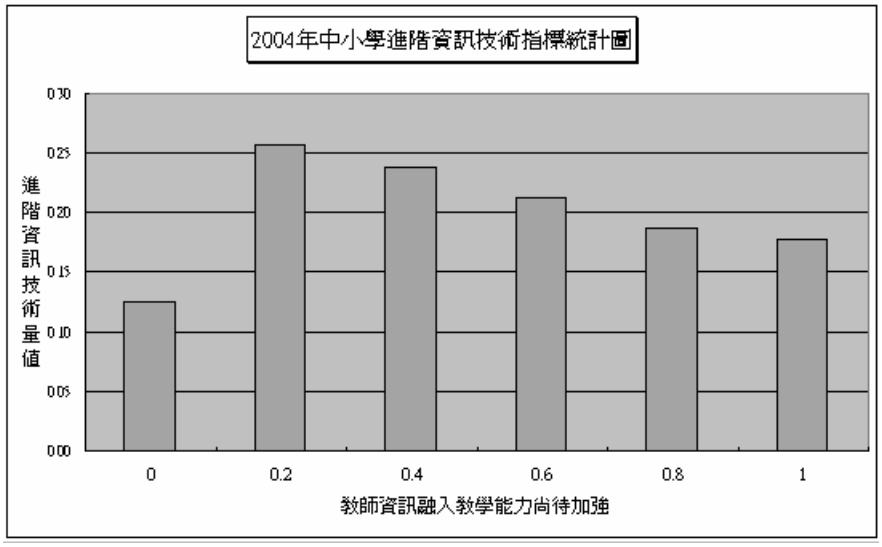

圖 6.28: 資訊融入教學能力尚待加強維度分析進階資訊技能之量值統計圖

從上圖 6.28 的分析中我們可看出，各區**學校的「教師資訊融入教學能力尚待加強」與學生的「進階資訊技能」有負相關的趨勢**。

- **「網路素養」量值分析**

透過選取鑑別度較高之欄位，選取[**教師資訊融入教學能力尚待加強狀況**]欄位(數值越大表示教師越需要加強資訊融入教學)時，可以發現當狀況越嚴重的，其對應之網路素養越差，也就是「**教師資訊融入教學能力尚待加強狀況」與學校之網路素養表現有負相關的趨勢**如下圖 6.29。

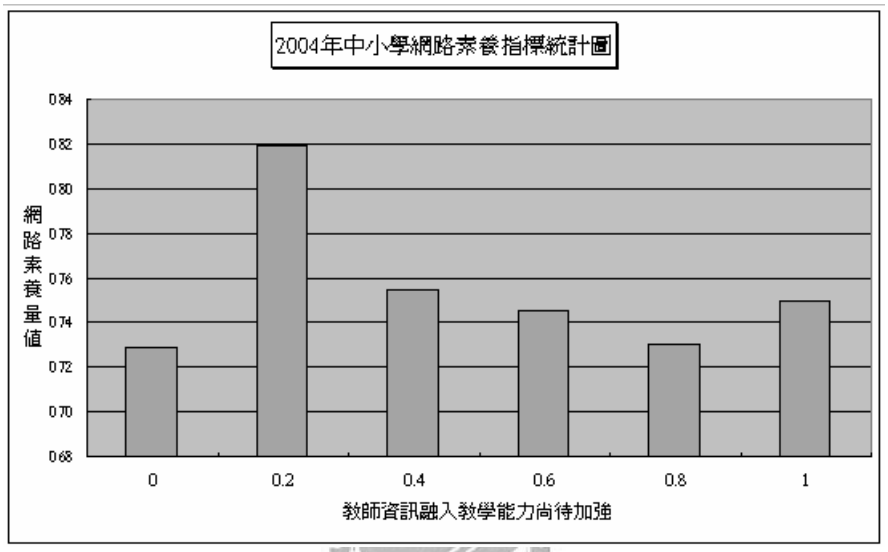

圖 6.29: 資訊融入教學能力尚待加強維度分析學生網路素養之量值統計圖

接著透過加入[**私立學校**]欄位進行 Drill down 深入分析，則可發現**私立學校大部分之網路素養較公立學校低**(圖 6.30)。

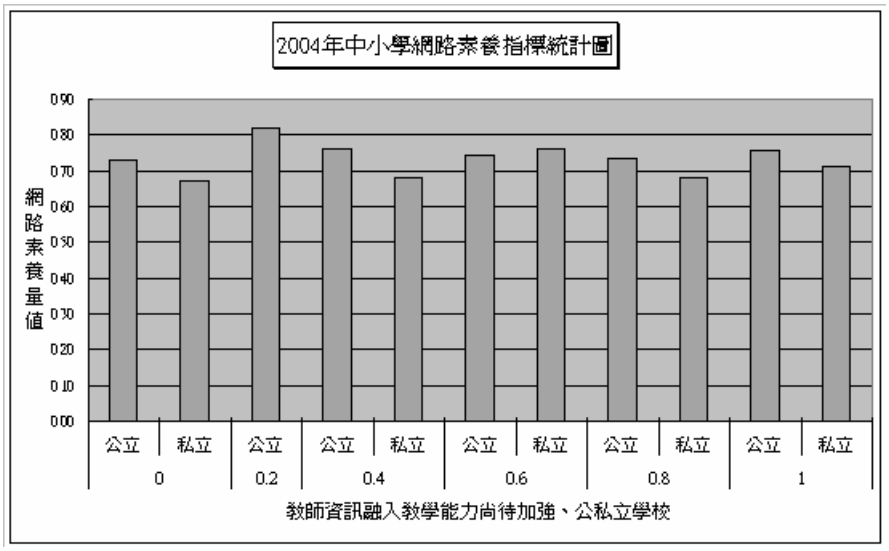

圖 6.30: 加入私立學校維度分析學生網路素養之量值統計圖

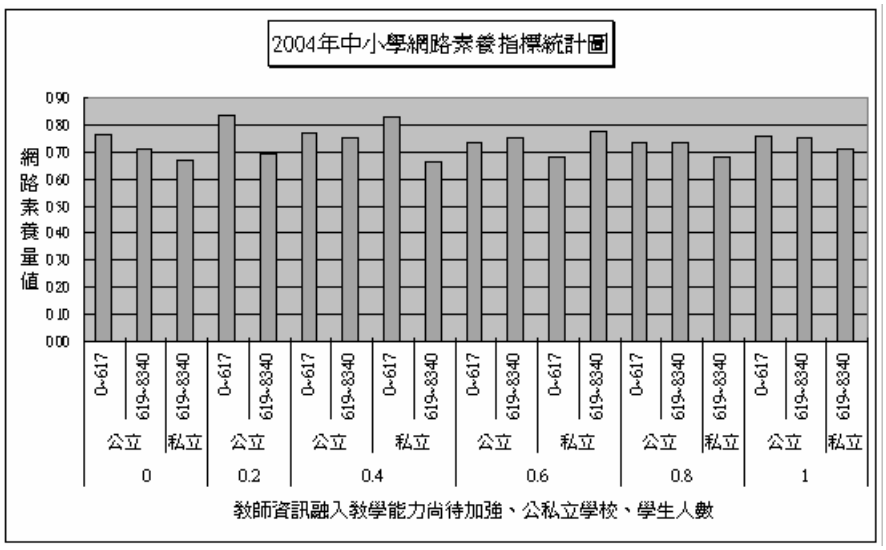

圖 6.31: 加入學生人數維度分析學生網路素養之量值統計圖

再透過加入[**學生人數**]欄位 Drill down 一層，則可發現「**「學生人數」與學生之「網路素養」表現有負相關的趨勢**，也就是說，學生人數少的學校之學生網路素養表現較佳 (圖 6.31)。透過此分析，可以推測「教師資訊融入教學能力尚待加強」狀況、「是否為私立學校」與「學生人數」會影響學校「學生之網路素養」。

(5) 中小學學生數位學習落差綜合分析

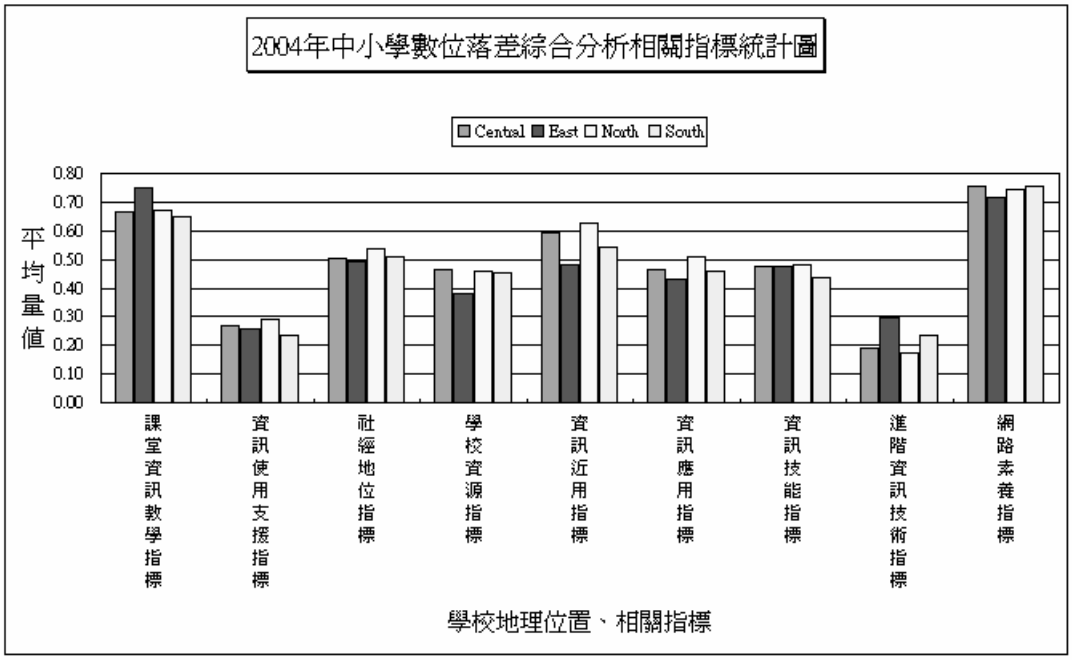

圖 6.32: 中小學數位落差綜合分析之相關指標統計圖

綜合上述分析，在學校資訊學習環境及學生數位學習表現方面，從圖 6.32 我們可看出整體趨勢，顯示我國中小學學生在**網路素養**(道德)，有中上的表現，其次則是課堂資訊教學量值，代表各區學校在課堂資訊教學表現也不錯，顯示各區學校在**課堂資訊教學**方面相當支持，因此學生在一般的資訊技術也有中上的表現。但是在學校資源量值部分並不理想，顯示各區**學校在學校網頁、數位教學(材)資源的部分有待加強**，必須多加努力，同時在**進階資訊技術(**進階技能)的部分並不理想，顯示現階段的學生在資訊技術深度的不足。

在學生的家中資訊學習環境方面，其中資訊使用支援量值最不理想，顯示各區學生家長在**資訊使用支援**的部分有困難，在學生家長的社經地位量值部分，整體而言，由上圖中我們可了解到整體中學學生家長的**社經地位**有中等的趨勢，相差不大但仍以北區學校為佳，而且在資訊近用主題的分析中，我們可看出以資訊近用量值較高，代表各區學生在**資訊近用**(家中設備)表現較佳，其中仍以北區學生資訊近用表現較佳，同時北區中小學學生的**資訊應用**量值也較高，代表北區學生在資訊應用表現較佳，顯示學生家長的**社經地位(**學歷及收入**)**與學生在**資訊近用**(家中設備)及**資訊應用**(上網的習慣及時數)的表現有正相關的趨勢。

如果就前述的中小學學生數位落差綜合分析後的特徵可整理如下：

- 中區及東區學校「課堂資訊教學」量值與「教師資訊政策」維度有正相關的趨勢。
- 北區學校的「資訊使用支援」、「課堂資訊教學」量值與學校的「資訊教育方案」維度有正相關的趨勢。
- 北區學生家長「社經地位」量值較高。
- 中區及東區「私立學校」的學生家長有較佳的社經地位，北區及南區則是「公立學校」的學生家長有較佳的社經地位。
- 大部分學校的「教師資訊政策」與「學校資源」量值有正相關的趨勢。
- 大部學校的「教師資訊融入教學能力尚待加強」與「學校資源」量值有負相關的趨勢。
- 大部分學校的「資訊教育方案」維度選值較高時(0.9~1)，有較佳的「學校資源」量值。
- 中區、北區及南區的「資訊種子學校」維度與「學校資源」量值有正相關的趨勢。

- 學校的「學生人數」與學生的「資訊近用」量值有正相關的趨勢。
- 學生「與父母同住」比例與學校的「資訊近用」量值有正相關的趨勢。
- 私立學校學生具有較佳的「資訊近用」量值
- 「現有資訊教學設備維護不易」維度選值低時(0.2)，學校學生具有較佳的「資訊應用」量值，
- 「現有資訊教學設備不足」維度選值低時(0.4)，學校學生具有較佳的「資訊應用」量值。
- 「現有資訊教學設備不足」維度選值高時(1)，學校學生具有最差的「資訊應用」量值。
- 東區及北區學校的學生「資訊應用」量值與「教師資訊政策」維度有正相關的趨勢，中區學校的學生「資訊應用」量值與「教師資訊政策」維度有負相關的趨勢。
- 大部分的北區學校的「資訊教育方案」值與學生的「資訊技能」量值有正相關的趨勢
- 學校的「學生人數」與學生的「進階資訊技能」有負相關的趨勢，
- 學校的「教師資訊政策」與學生的「進階資訊技能」有正相關的趨勢，
- 學校的「教師資訊融入教學能力尚待加強」與學生的「進階資訊技能」有負相關的趨勢。
- 「教師資訊融入教學能力尚待加強狀況」與學生之「網路素養」表現有負相關的趨勢。
- 「學生人數」與學生之「網路素養表」現有負相關的趨勢。

6.2. DMAS 線上資料探勘分析系統實作

本研究中提出的資料探勘分析流程，主要連接之前資料倉儲中所建立的資料立方體，以利分析時可以依資料之分佈情形或是領域專家的意見，來調整資料來源中欲分析的之欄位與資料階層，在此研究所使用分析工具為我們所開發之 DMAS v2.1 API 為資料探勘分析程式核心，並使用 JAVA SDK 1.4.1 來整合使用者介面，開發而成之 DMAS 線上資料探勘系統 (DMAS-OLAM)，下圖 6.33 為系統之選取分析欄位畫面。

圖 6.33: DMAS 線上資料探勘系統 (DMAS-OLAM)

在本研究所提出的**兩層式資料探勘系統**中，透過結合了資料探勘技術中的分群分析與決策樹分析，來建構數位落差之預測模型，並搭配此 DMAS 線上資料探勘系統來進行分析實作，以下為實驗分析之結果：

(1). 分群分析

透過對學生之資訊能力調查問卷資料，可以依學校為單位彙整出以下 9 項指標值：

學校資源、社經地位、資訊使用支援、課堂資訊教學、資訊應用、資訊近用、網路素養、資訊技能、進階資訊技術。

這些指標值皆為正規化到 0~1.0 之小數，數字越大代表學校在此指標能力越高。透過這些欄位資料，在此首先使用 DMAS 線上資料探勘系統，選入要進行分群的學生資訊能力問卷相關指標維度，然後進行分群分析，如下圖 6.34。

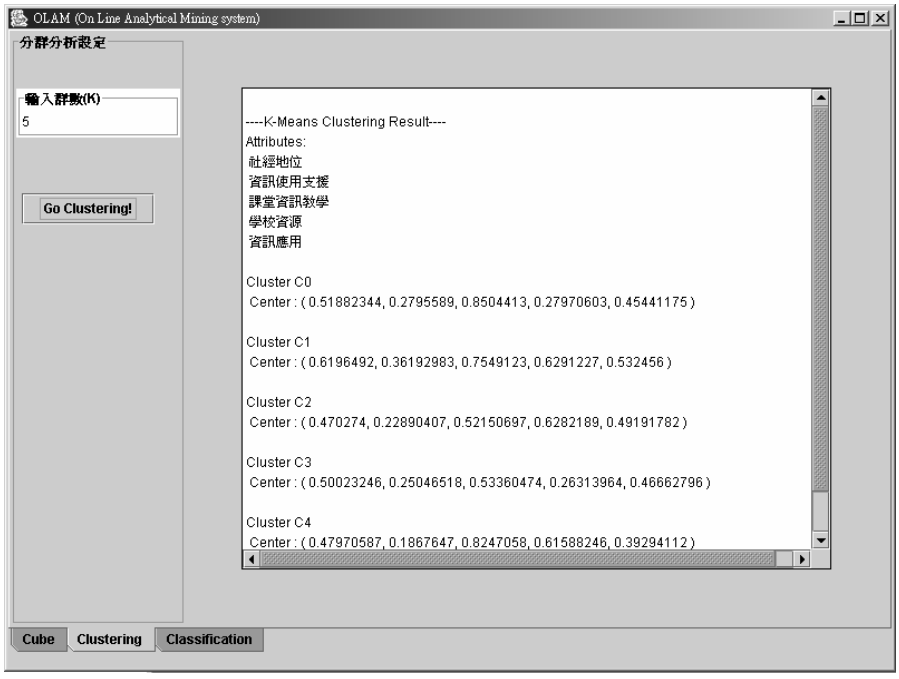

圖 6.34: DMAS 線上資料探勘系統中分群分析畫面

為了分析現有學校中，其學生或環境所表現之資訊能力指標高低有幾種不同類別，因此選擇學生能力指標相關屬性欄位進行分群分析。經測試後，將學校分為 5 群不同資訊能力標籤，各群之群中心座標值如表格 6.2 所示

表 6.2：各群群中心

Cluster	學校資源	社經地位	資訊使用支援	課堂資訊教學	資訊應用	資訊近用	網路素養	資訊技能	進階資訊技術
第一群編號 C0	0.42	0.56	0.32	0.81	0.49	0.54	0.76	0.6	0.63
第二群編號 C1	0.57	0.62	0.36	0.75	0.48	0.63	0.8	0.67	0.25
第三群編號 C2	0.63	0.48	0.25	0.55	0.5	0.6	0.72	0.39	0.14
第四群編號 C3	0.28	0.51	0.26	0.58	0.47	0.6	0.73	0.39	0.16
第五群編號 C4	0.45	0.46	0.18	0.86	0.41	0.44	0.8	0.45	0.11

對資訊能力分群結果之特徵分析部分，在此針對各群之群中心座標進行特徵分析，各群依指標項目值所繪成之長條圖如下圖 6.35 所示。

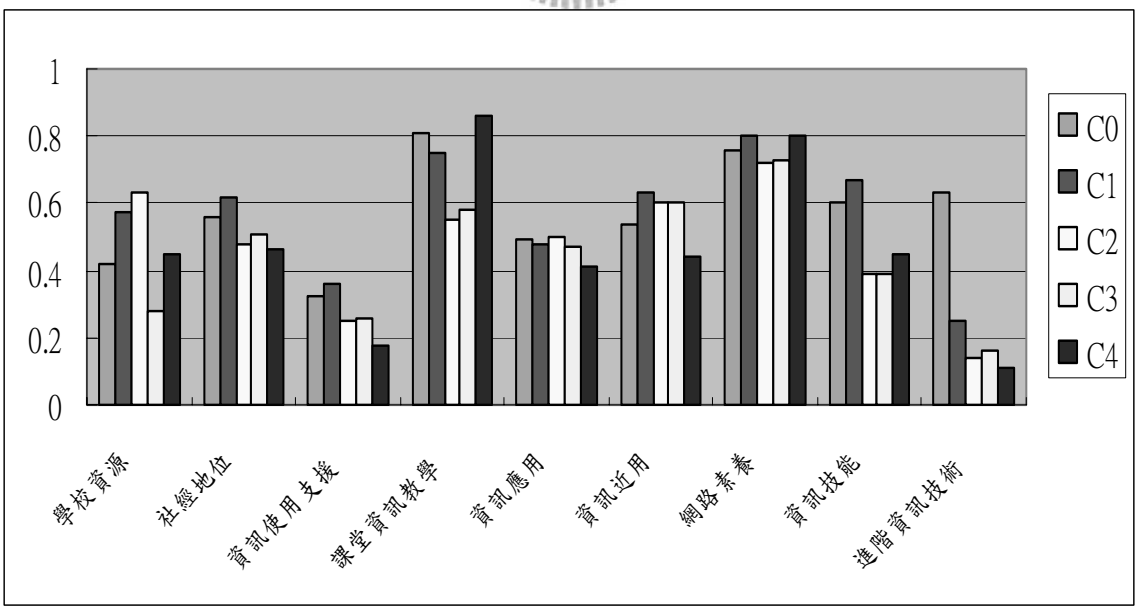

圖 6.35: 各分群之資訊能力指標比較圖

透過對各群的特徵分析，由學校分群結果歸納出各群之資訊能力標籤，分別為

C0 到 C4：

- *第一群 C0：進階資訊技術最高*
- *第二群 C1：社經地位最高、資訊使用支援最高、資訊近用最高,網路素養高、資訊技能最高*
- *第三群 C2：學校資源最高、資訊應用最高、社經地位最低、網路素養最低、資訊技能最低*
- *第四群 C3：學校資源最低、課堂資訊教學最低、資訊技能最低*
- *第五群 C4：課堂資訊教學最高、網路素養高、資訊使用支援最低、資訊應用最低、資訊近用最低*

由此分群結果看來，可以知道 C0 類別的學校是進階資訊技術較突出，而 C1 類別的學校一般來說資訊環境較好，學生資訊近用與網路素養也比較好。C2 類別的學校則表示資訊環境較好，但是資訊技能與素養卻比較差。C3 類別學校資訊環境與資源比較不足，而 C4 則是學校老師資訊教學好，但家庭與資訊應用比較差。

(2)決策樹分析

透過之前資料倉儲所彙集的資料中，描述一個學校的環境背景與政策方針，可以整理成下面 12 個維度：

- 歷史統計資料庫相關維度：

  地理分區、學生規模、教師規模、是否為私立學校、是否為資訊種子學校。

- 學校問卷維度：

  資訊教育方案、教師資訊政策、校長支持度不高狀況、教師資訊融入教學能力尚待加強狀況、資訊教學設備不足狀況、現有資訊教學設備維護不易狀況、資訊教學人力不足狀況。

在此使用我們的線上資料探勘系統，選入學校背景與政策相關維度，然後對分群結果進行決策樹分析，如下圖 6.36。

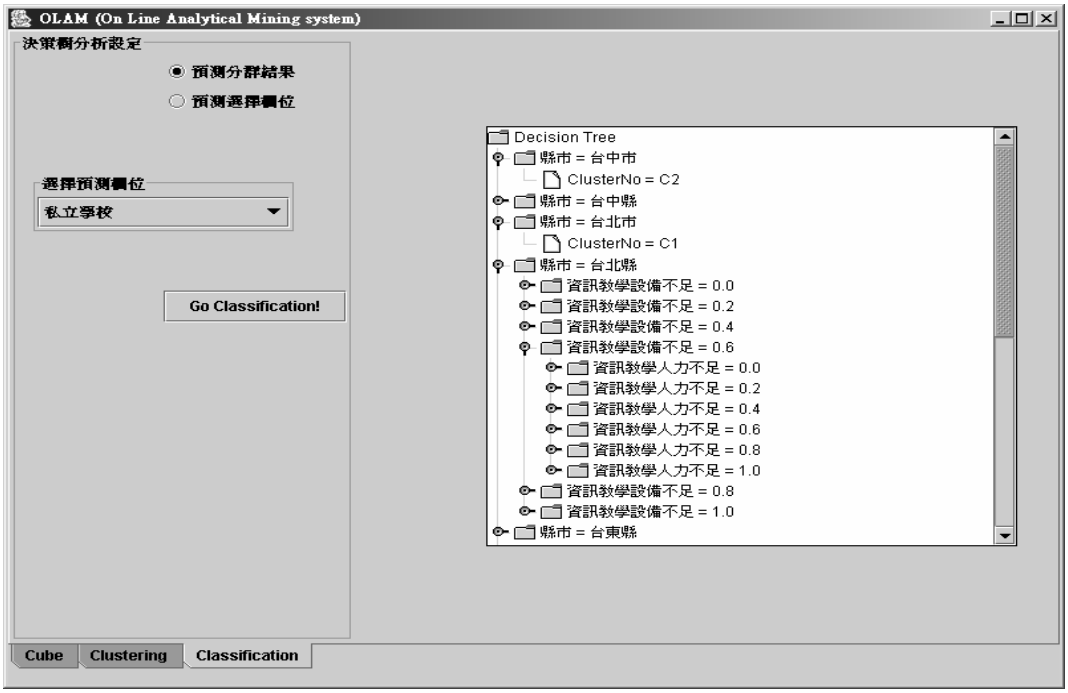

圖 6.36: DMAS 線上資料探勘系統中決策樹分析畫面

建構出來的決策樹，則是如圖 6.37 之樣式，下圖只列出決策樹模型的一部分。

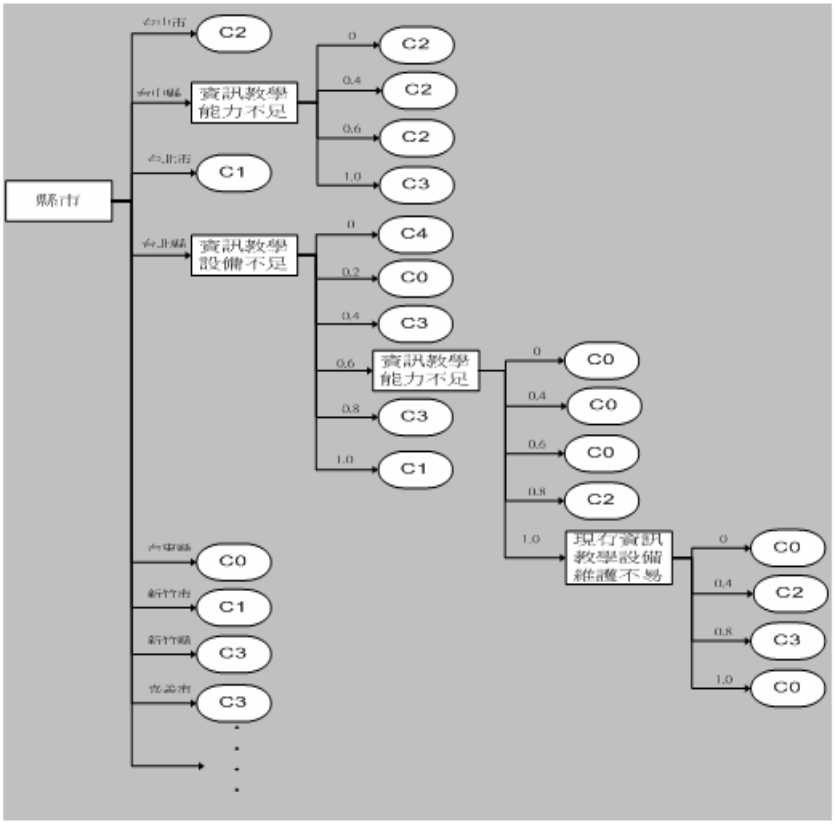

圖 6.37: 學校資訊能力決策樹模型

透過決策樹分析結果，發現學校所在*縣市*欄位為第一個被挑選出來之欄位，接著則是*資訊教學人力不足*、*資訊教學設備不足*或*現有資訊教學設備維護不易*等等，表示一般造成學校學生資訊能力等級的不同，除了地區性的差別為，學校資訊相關資源，如人力或設備等，是造成差異較重要的因素之一。若進一步將決策樹之結果透過依存性網路分析(圖 6.38)，將會發現主要對於分類結果之影響維度分別為：

(1). 資訊教學設備不足。

(2). 教師資訊政策(是否獎勵資訊教師或常辦教師資訊研習)。

(3). 教師資訊融入教學能力尚待加強。

(4). 現有資訊教學設備維護不易。

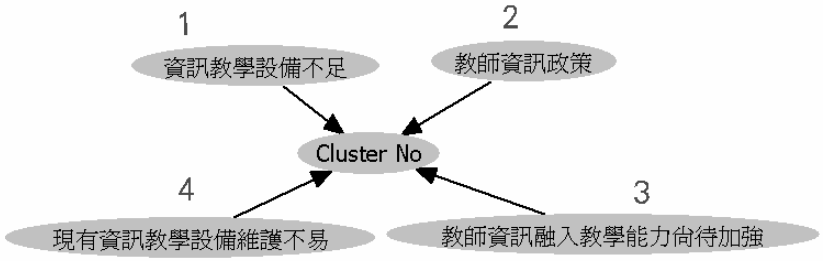

圖 6.38：決策樹中對分類較有影響力之欄位

若進一步將決策樹之結果透過統計可得下列結果：

表 6.3：決策樹規則之分類統計

規則	資訊教學設備不足	教師資訊政策	資訊融入教學能力尚待加強	教學設備維護不易	C0	C1	C2	C3	C4	Sp%
1	<= 0.65		<= 0.65		1%	**89%**	2%	5%	22%	20.1%
2	<= 0.65		> 0.65		0%	4%	**54%**	3%	**52%**	19.8%
3	>0.65	<=0.625		<=0.65	0%	0%	15%	**19%**	13%	8.8%
4	>0.65	<0.625		>0.65	9%	0%	0%	**65%**	0%	15.1%
5	>0.65	>0.625	<=0.85		**90%**	7%	12%	8%	13%	32.7%
6	>0.65	>0.625	>0.85		0%	0%	**17%**	0%	0%	3.5%
					100%	100%	100%	100%	100%	100.0%

將上表統計結果，另以文字方式來表示此**6 項學校分類規則**如下:

規則１：(資訊教學設備不足 <= 0.65 and 教師資訊融入教學能力尚待加強 <= 0.65) ➔C1 (89%)

規則 2-1：(資訊教學設備不足 <= 0.65 and 教師資訊融入教學能力尚待加強 > 0.65)➔C2 (54%)

規則 2-2：(資訊教學設備不足 <= 0.65 and 教師資訊融入教學能力尚待加強 > 0.65)➔C4 (52%)

規則 3：(資訊教學設備不足 > 0.65 and 教師資訊政策<= 0.625 and 資訊教學設備維護不易<= 0.65) ➔ C3 (19%)

規則 4：(資訊教學設備不足 > 0.65 and 教師資訊政策<0.625 and 資訊教學設備維護不易> 0.65) ➔ C3 (65%)

規則 5:(資訊教學設備不足 > 0.65 and 教師資訊政策< 0.625 and 教師資訊融入教學能力尚待加強 <= 0.85) ➔ C0 (90%)

規則 6:(資訊教學設備不足 > 0.65 and 教師資訊政策> 0.625 and 教師資訊融入教學能力尚待加強 > 0.85) ➔ C2 (17%)

至於**規則的解讀**方式，我們將以規則 1~規則 3 為例解說如下：

規則１：

(資訊教學設備不足 <= 0.65 and 教師資訊融入教學能力尚待加強 <= 0.65)

➔C1 (89%)

解讀：資訊教學設備不足情況較不嚴重(不大於 0.65)，而且教師資訊融入教學能力尚待加強情況較不嚴重(不大於 0.65)的學校，其學校之學生能力大多是分類到 C1 類別，其對應之標籤為：***社經地位最高、資訊使用支援最高、資訊近用最高,網路素養高、資訊技能最高***，*(在以這 6 個規則對 C1 進行分類的條件下，有 89%機會分配到 C1)*。

規則 2-1：

(資訊教學設備不足 <= 0.65 and 教師資訊融入教學能力尚待加強 > 0.65)

➔C2　　(54%)

解讀：資訊教學設備不足情況較不嚴重(不大於 0.65)，而且教師資訊融入教學能力尚待加強情況較嚴重(大於 0.65)的學校，其學校之學生能力大多是分類到 C2 類別，其對應之標籤為：***學校資源最高、資訊應用最高、社經地位最低、網路素養最低、資訊技能最低，****(在以這 6 個規則對 C2 進行分類的條件下，有 54%機會分配到 C2)*。

規則 2-2：

(資訊教學設備不足 <=　0.65　and　教師資訊融入教學能力尚待加強 > 0.65)

➔C4　　(52%)

解讀：資訊教學設備不足情況較不嚴重(不大於 0.65)，而且教師資訊融入教學能力尚待加強情況較嚴重(大於 0.65)的學校，其學校之學生能力大多是分類到 C4 類別，其對應之標籤為：***課堂資訊教學最高、網路素養高、資訊使用支援最低、資訊應用最低、資訊近用最低，****(在以這 6 個規則對 C4 進行分類的條件下，有 52%機會分配到 C4)*。

規則 3：

(資訊教學設備不足 > 0.65　and 教師資訊政策<= 0.625　and 資訊教學設備維護不易<= 0.65)　➔ C3　　(19%)

解讀：資訊教學設備不足情況較嚴重(大於 0.65)，而且教師資訊政策情況較不嚴重(不大於 0.625)，而且資訊教學設備維護不易情況不太嚴重(不大於 0.65)的學校，其學校之學生能力大多是分類到 C3 類別，其對應之標籤為：***學校資源最低、課堂資訊教學最低、資訊技能最低，****(在以這 6 個規則對 C3 進行分類的條件下，有 19%機會分配到 C3)*。

第七章　結論與未來展望

隨著資訊相關產業的發展，雖然提升了許多民眾在生活上資訊化的便利，但是卻也產生了新的問題：那就是數位落差(Digital Divide)。世界各進國家致力於數位落差的相關研究時，資料分析方法大多是以問卷及面訪方式進行，針對人口統計變數進行抽樣統計分析。然而在這些大量的資料中，例如：問卷調查資料，常隱藏著極為有用的資訊或知識，在以往的資料分析技術所用的方主要是以統計分析為主，然而現有問卷資料分析方法仍有許多不足之處，例如：缺乏結合外部相關背景歷史資料之彈性、對於問卷資料本身沒有做完善的資料前處理、對於統計欄位沒有建立概念階層機制以及現有線上分析與資料探勘分析缺乏整合性工具，可以讓分析者自由方便的進行分析等等。

為了解決傳統問卷分析方法在資料維度整合、操作性與累加性等的不足，因此本研究提出了一個**資料倉儲問卷分析架構**方法，透過資料倉儲與資料探勘技術，整合其他歷史資料庫，來進行多維度的問卷分析。**資料倉儲問卷分析架構**包含了三個處理階段：首先是**資料前處理**階段，主要使用了資料淨化處理、資料平滑、聚集與正規化等處理，並提出了**問卷題目量化轉換**演算法，來將問卷中不同的題型答案資料轉為可適用於資料立方體中的量值形式。其次是**資料倉儲之建置**階段，在此提出了**多維度概念階層知識擷取**演算法，來擷取領域專家對問卷中的概念階層知識，並可以產生廣義化的新量值，此階段整理了資料維度與量值，並將之建置成資料倉儲中資料立方體。最後是**線上分析與資料探勘**階段中，透過建立好的資料立方體，在此完成發展了使用**兩層式資料探勘方法**之 **DMAS 線上資料探勘系統**，透過此分析系統，分析者可以更便利的進行資料探勘分析。

除了上述的資料倉儲問卷分析架構之設計之外，關於實作部分，我們參考由曾憲雄、張維安、黃國禎教授等，所提供的中小學數位落差之相關研究資料。進行了相關的實作與驗證，由於這些相關資料的協助使本

論文的研究內容更加充實完整。

綜合以上成果，本篇論文之主要研究貢獻如下：

- 提出了一個**資料倉儲問卷分析架構**方法。
- 提出一個可處理不同**問卷題型量化問題之資料轉換演算法**。
- 提出**多維度概念階層知識擷取方法**，以利建立資料立方體(Data Cube)。
- 結合現有 OLAP 工具，提出資料分析流程之架構。
- 完成資料探勘 OLAM 分析系統，輔助分析者更容易進行資料分析。

由於資料倉儲之設計具有資料累加的特性，因此在未來可朝三個方向改進：

(1) 廣度性:

可加入全球地理資訊系統等，進而擴展資料倉儲之資料維度，增加讓問卷分析的廣度。

(2) 深度性：

另外可以持續增加每年度之新的問卷調查資料，將前一年度與下一年度之資料加以彙整，增加分析之深度。

(3) 系統完善性:

未來並可依使用者對分析工具之意見，改善分析系統，使其更加完善。

參考文獻

[1] 江政達,“數位落差的一些省思” ,資策會 ACI-FIND, http://www.find.org.tw/focus_disp.asp?focus_id=193, 2001/5/5.

[2] 曾憲雄、張維安及黃國禎等,”建立中小學數位學習指標暨城鄉數位落差之現況調查、評估與形成因素分析網站”, http://e-divide.nctu.edu.tw/, 2004/2.

[3] “研究數位落差方法”,台灣電子國際商務中心, http://www.nii.org.tw/cnt/info/Report/20020305_4.htm, 2004/5.

[4] “Falling Through the Net: Toward Digital Inclusion”, http://www.ntia.doc.gov/ntiahome/fttn00/contents00.html, 2000/10.

[5] “Introduction to DMAS”,Data Mining 網站,知識工程實驗室,交通大學, http://rss.cis.nctu.edu.tw/course/dm04/doc/DMAS-class.ppt, 2004/5.

[6] “StatWorks – an IT toolkit for Statistical data management”, http://www.oecd.org/dataoecd/50/38/18247342.pdf.

[7] “Understanding the Digital Divide“, http://www.oecd.org/pdf/M00002000/M00002444.pdf, 2001.

[8] 尹相志. (2002). *SQL2000 Analysis Servic 資料採礦服務*. 台北,維科圖書有限公司.

[9] 林傑斌等. (2002). *資料採掘與 OLAP 理論與實務*. 台北,文魁資訊股份有限公司.

[10] 林宏諭., & 萬衛華等. (2001). *Excel 2002 完整學習-資料分析與市場調查*. 台北, 博碩文化.

[11] 陳敬如. (2000). *台灣地區中等學校學生數位鴻溝差距狀況初探*. 碩士論文, 教育研究所,台灣師範大學.

[12] 曾淑芬., 陳啟光., &吳齊殷等. (2003). *台閩地區九十一年數位落差調查報告*. 行政院研究發展考核委員會.

[13] Chaudhuri, S., & Dayal, U. (1997). *An Overview of Data Warehousing and OLAP Technolog.* SIGMOD Record 26(1): 65-74.

[14] Golfarelli, M., & Rizzi, S. (1999).*Designning the data warehouse：key steps and crucial issue*s. Journal of Computer Science and Information Management, vol. 2, n.3.

[15] Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques*. London, Academic Press.

[16] Inmon, W. H. (1996). *Building the Operational Data Store.* John Wiley & Sons Inc.

[17] IBM Redbooks. (1998). *Data Modeling Techniques for Data Warehousing.* U.S.A., IBM Corporation.

[18] Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: an Introduction to Cluster Analysis.* New York: John Wiley & Sons.

[19] Kortnik, M. A. R., & Moody, D. L. (1999). *From Entities to Stars, Snowflakes, Clusters, Constellations and Galaxies: A Methodology for Data Warehouse Design.* 18th.International Conference on Conceptual Modelling. Industrial Track Proceedings.

[20] Kimball, R. (1996). *The Data Warehouse Toolkit.* John Wiley & Sons Inc.

[21] Kimball, R. L., Reeves, Ross, M., & Thornthwaite, W. (1998). *The Data Warehouse Lifecycle Toolkit.* Wiley Computer Publishing.

[22] MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observation in proc.* 5th Berkeley Symp. Math. Statist, Prob., 1:281-297.

[23] Quinlan, J. R. (1986). *Induction of decision trees*. Machine Learning, pp.81-106.

[24] Saaty, T. L. (1980). *The Analysis Hierarchy Process.* New York.