摘　要

儘管資料倉儲系統的建立是源自於不同複雜技術的操作系統而來，但並不會因此而影響資料倉儲系統設計方法的一致性與其發展的完整性。由研究得知資料倉儲系統之強固性(或完整性)，並非在於後期資料處理的能力，而是取決於先期相關資料搜集的完善。本文主要是探究資料倉儲系統建構過程中，以天氣氣象資料庫爲實例說明在資料彙整時，事實綱要重疊(fact schemes overlap)技術對資料結合及系統效益所帶來的種種影響。

並提出考慮不同資料來源(data sources)的彙整，來建構另一種不同的資料倉儲系統，從二個在作業系統層面上不相干，但專業知識上卻是被認定有關聯的相異資料庫(天氣氣象(中央氣象局)、空氣污染(行政院環保署))，建構綜合性的資料倉儲，利用專家知識及經驗，智慧的資料整合，然後以類神經網路進行資料挖掘的決策分析探討。其分法是利用天氣氣象因子條件，預測空氣污染的品質，加入區域性的因子來增加預測空氣污染的正確性，並藉由空氣污染預測雛型系統的建立，找出空氣污染的成因及其對空氣污染的影響。

關鍵字:資料倉儲、資料探勘、資料來源、資料彙整、事實綱要重疊

第一章、緒論

資料倉儲系統是一種能適當對資料做整合及管理不同資料來源的技術，此整合性資料儲存體能提供企業解決問題及決策輔助，可供高階主管做查詢擷取、篩選、整合相關資訊。不同於傳統系統的被動式查詢，資料倉儲系統爲主動方式的查詢，當來源資料更動時即做出相對應的反應，許多研究[45, 18, 23]均提出這方面相關之論述，而資料倉儲系統在建構的過程中，我們可知資料倉儲系統中的資料來源與資料彙整，在資料倉儲建構的過程中是最重要的影響步驟，且因不同資料來源所匯整而來的資訊是會相對影響資料倉儲的品質，故本文將針對資料倉儲所強調的完整性問題，從分析不同的資料來源與資料彙整的事實綱要重疊技術二方面去幫助系統的建立，以有效的提高使用者查詢與決策者決策分析。

因資料倉儲本身是一個非常大的資料庫，它儲存著由組織作業資料庫中整合而來的不同資料，特別是從線上即時應用處理(OLAP)所得資料[18]與從分散異質資料資源傳遞而來的資料，例如:不一致資料、不相容資料結構及粒狀資料[22]等等，而資料倉儲系統是必須能自動化轉化這些異質、分散的資料來源，並以遞增的方式使其合併轉化進入資料倉儲系統中[27]，且藉由考慮不同的資料來源間的相互影響之關係來加以強固資料倉儲的完整性，進而將作業層次的資料轉換成有用的策略性資訊，這是整個資料倉儲系統建置的重點所在[19]，故資料倉儲在於初期的資料彙整是不可忽略的。

而建置資料倉儲的完整性是來自於相關資料的搜集及完善的過濾[1]，研究中得知資料倉儲系統之強固性，並非在於後期資料處理的能力，而是取決於先期相關資料搜集的完善，在考量資料來源的重要性是必須藉由了解不同的資料來源所具有之相關性，進而彙整結合成爲使用者所需要的資料資訊。而目前類神經網路在各領域方面均有不同方面的應用與實例，特別是在於預測上更是有很好的成效[32, 33]，我們藉著參考過去資料倉儲與資料探勘的技術[34]，再由資料倉儲

的概念去發現資料探勘的過程[30]，進而了解資料倉儲在建構過程所必須考慮的方法與欄位[29]，並利用知識探索技術在大量的資料庫中定義其演算法及分類規則的項目來進行研究[31, 51]，再從建構的過程中加以專家知識利用智慧代理的方式幫助系統篩選不同的相關性資料來源，完善的考慮此資料倉儲建構時所需的操作資料來源。

我們可知設計一個完整的資料倉儲是需要從不同的操作資訊系統所建構而成的，目前在許多學術文獻中對於資料倉儲的各個步驟上的證明都有特別說明與定義[47, 49, 50]，例如:在於多重資料模型的探討[35, 36, 20]、具體化的概念[37, 38]與索引選擇[39, 40]及特別在於提高資料倉儲系統效率的方法之新索引的技術[21]上。但直到目前為止在發展完全與一致的資料倉儲系統設計的方法上並沒有任何重要性的成果，也無法提出有效的具體研究，且在不同資料倉儲設計步驟的描述是非正式的，但是在沒有 ad hoc 概念化設計是可實際設計的[42]，並且在使用者欲從大量的資料中做查詢時的回應已有很多具體化的效益[23, 24]。而研究中更加深了解到資料倉儲必須是能包含一切不同的資料來源[48]，其可能是從分散及異質資訊的來源所整合儲存而來，並藉由完善資料來源的整合來幫助使用者有效的查詢加以提高資料倉儲建構的品質。

研究除了証明能從一個半自動化技術來完成資料倉儲的初始化概念設計模型[45]，而且 Mcguff[41]曾證明在企業的經營模式中資料倉儲的設計是可根據其實際上的關連資料庫欄位而來，故可分別從實體-關係(E/R)架構和邏輯關連性架構去描述早先現有存在的操作資訊系統[44]，而資料倉儲系統概念化設計的建構過程中資料彙整的事實綱要重疊(overlap)，是在資料處理中佔極大的重要性[45, 46]，因為資訊系統中資料必須能滿足資料屢次更新的情形，故會在即時的環境中改變資訊空間與資料倉儲之間的訊息結合，這樣才能增加資料倉儲的完整性以維持系統效益及執行的穩定性[26]，也因此資料倉儲則利用在於綱要上的修改和原始必須查詢的綱要之間的重疊，來計算及查詢修正後儲存的資料所呈現出

的百分比程度及延伸相關的資訊，以提供決策者做爲分析決策的研究，在於評價這方面的問題，實際上過去研究中也有被提出[25]，並且資料倉儲系統在建構架構的過程已有研究建構的方法論[29]，了解在於因次事實模型中資料倉儲概念的設計是可行的[43]。

目前在於自動化的資料處理上，沒有很有效地完整的定義資料結合的目的與如何去做建製的過程，導致資料倉儲系統沒有發揮其完整的分析效益，因而無法達到使用者原本需求的目的。本研究主要的研究動機則是在於研究資料倉儲在建構的過程中，其最重要的資料來源處理及資料彙整的事實綱要技術部份，從系統如何有效率執行與確實去幫助決策者提高其決策效益問題的方面加以去探討。且因目前在資料倉儲建構過程所提出的方法論中，沒有很明確同時去考量使用者及系統維護者二者之間存在的相依性問題，故我們也將針對在資料倉儲概念化設計當中，技術性動態地在多重的事實綱要架構的階層中利用重疊的方式去增加或減少樹的節點，並適當改變其樹的階層[28]，並在保有原有的屬性下，分別對資料整合的問題去定義，使其能充分了解原有資料含有的意義與不同事實延伸出的相關連性，使得資料倉儲系統更能增加其完整的有效性。

在本篇論文中，我們主要提出在於資料處理中欲強固資料倉儲完整性時，事實綱要做重疊動作所考慮的因素與重疊後所產生的問題，再從資料倉儲建構之相關問題去分析研究事實綱要重疊所存在的必要性，並提出資料倉儲系統建置時先期所考慮的資料來源若是來自於不同資料資源，則如何利用專家知識及專家經驗去做資料的整合，以提升建構資料倉儲的品質。第二章將介紹由 Golfarelli 所提出建構資料倉儲架構的方法論及系統建製過程中整體考慮的問題，並針對其所提的部份去做定義及解釋，第三章介紹系統建構方式中不同資料結合所考慮資料彙整的重要性從事實綱要重疊技術加以分析其最佳使用時機，再提出另外不同於考慮重疊產生的情形與其優點，並在第四章討論不同異質的資料來源利用專家知識在於資料的判斷上所能得到的效益，在第五章做最後結論。

第二章、 資料倉儲之建構

可知有效的資料探勘，是從乾淨的資料倉儲中去挖掘出來的，若能有效地建置乾淨的資料倉儲系統，則能利用資料探勘的技術挖掘出有用的資訊，以明確地提供決策者做決策分析。在這一節中將討論 Golfarelli 對資料倉儲建構其架構過程中所做的相關性問題研究，針對其提出的問題與方法做完整的定義分析，藉以了解如何建立完善的資料倉儲系統。

以下爲 Golfarelli 在研究中提出有效地資料倉儲系統架構的建構過程，主要可分爲[29]:

1、 資訊系統分析:

首先在於資料倉儲系統建構的初步過程中，必須針對原先早已存有的操作資訊系統來收集相關訊息的資料及文件，並整合分散、異質資料來源以結合過去資料庫的大量資料資源，來建構出完整的資料倉儲系統，其中則包含要去考慮各種不同方面的情形，例如從設計者、有關人員、資訊系統管理人員等等各方面的情況去加以衡量。

2、 必需的條件:

建構的重點在於收集及過濾使用者需求，並從資料倉儲的設計者和使用者的角度去做考量和設計。然而基於決策者分析的需求，資料倉儲則必須考慮系統中資料所產生的事實(fact)與系統中工作量(workload)部份。其中的事實是從資訊系統之前的文件所定義而來的，其概念則在於提供決策分析及符合不斷發生的事件，假如其資訊系統文件是來自於資料庫中的 E/R 綱要架構。事實則或許可分別代表爲每一個實體或一個 n 陣列關係，相反的，假如它是代表關連性綱要，事實

即是符合關連的綱要(schema)。而系統中的工作量則是以近似自然語言的方式來表示，它是針對設計者所擁有的權利在概念化設計部份，去定義資料事實中所內含有的維度(dimensions)和尺度(measures)部份。在此每一個事實中我們特別有興趣於其尺度和聚集的部份上。

3、 概念化的設計:

先定義早前現有的操作資訊系統中之綱要及欄位的描述，並考慮事實及系統初始化工作量的部份，再由資料倉儲系統中的事實去創造資料內的因次綱要(dimensional scheme)，此概念化的設計架構是根據定義因次事實模型(dimensional fact model)而來的，其中的因次事實模式是由事實集合所組合而成的，當中的事實是每一個事實綱要的集合。每一個事實綱要的組合過程均是來自於使用者需求，先是根據當中資料屬性的性質，來建立其樹的屬性與圖形化的階層，並對屬性去做修改及轉化，再來定義其中所含有的尺度、維度及階層爲何。而構成樹的屬性是從部份有興趣的原始綱要和實體(關連欄位)所關連而來的，以圖形化的規則去建立類似樹狀(quasi-tree)的表達方式，並用此來呈現整個事實屬性的階層內容。

雖是如此，但是並不是所有的屬性均能充分表示其中所含有的關係性，故必須針對有興趣的部份去做修改及轉化，以保留有用的資訊及屬性，才能有效地定義整個事實的關係性。而其中維度的決定則在於決策分析的過程中所考慮的事實實例(fact instances)的集合，事實實例又來自於維度形式所組合而成的，因而產生所謂的尺度(measure)。尺度是歸因於所有屬性的集合也是建構事實綱要的必要條件，整個事實綱要就由其所組成的維度去定義其階層的架構，最後產生資料倉儲系統中資料的事實。

而事實的概念主要興趣於支援決策的過程，故會對符合不斷發生變化的事件

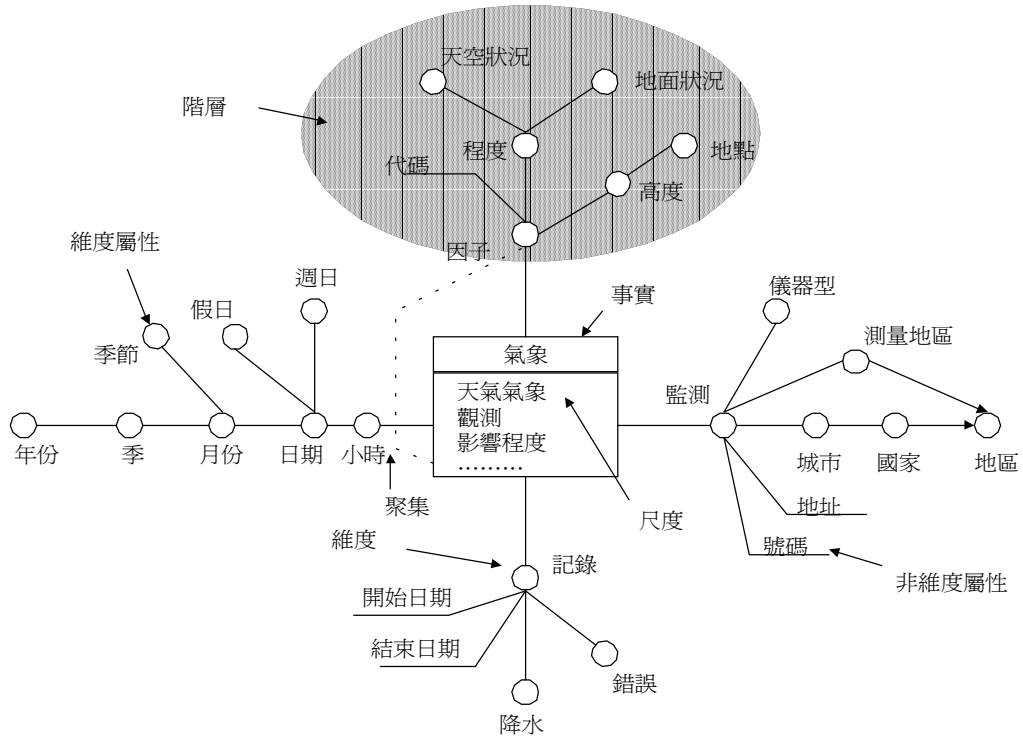
變化而去定義。我們首先定義讓$g=(V,E)$是一個經指定、非週期性及弱連結的圖形，我們說 g 是一個類似樹狀圖(quasi-tree)，其根在$v_0 \in V$，我們定義$path_{ij}(g) \subseteq g$是應有的路徑(假如它存在)則由點v_i開始到點v_j結束，且定義$(v_i) \subset g$樹狀根在$v_i \neq v_0$。

其中的事實綱要我們可定義爲 f=(M, A, N, R, O, S)，內含六個屬性。這裏的屬性定義爲:

- M 是表示尺度的集合，每一個尺度$m_j \in M$定義爲數值或布爾數學表示，其中包含從資料系統所獲得的值。
- A 是維度屬性的集合，每一個維度屬性$a_i \in A$其特徵爲一個不連續區域的值，$Dom(a_i)$。
- N 是非維度屬性的集合。
- R 是一對有條理的集合(a_i, a_j)這裏的$a_i \in A \cup \{a_0\}$和$a_j \in A \cup N(a_i \neq a_j)$，表每一維度的關係性，每一個因素在$Dim(f)$是稱一個維度，然後我們需要強調一個維度屬性$a_i$是一個維度，我們將會表示它爲$d_i$，並稱在維度的階層上$d_i \in Dim(f)$類似樹形的$sub(d_i)$。
- $O \subset R$是一個非必須關係的集合。
- S 是表示聚集的集合，其中每一個集合的組成是來自於(m_j, d_i, Ω)，這裏的$m_j \in M, d_i \in Dim(f)$和$\Omega$是一個集合運算子，說明$(m_j, d_i, \Omega) \in S$表示尺度$m_j$是沿著維度$d_i$聚集$\Omega$而成的。

當然，一個事實符合一個關連性欄位 f，而實體或關連性(關連欄位)則表示屢次更新的檔案及即時更新的資訊環境，其中不同的事實個別具有內含不同的資訊訊息，故我們分別可以由 E/R 綱要架構的方式來表示其不同資料關連性之存在，充分的轉化事實綱要的屬性欄位，從討論不同的組成藉以描述我們欲以彙整

的天氣氣象的事實綱要例子。



圖一 天氣氣象的事實綱要

上圖則是表示一個觀測的天氣氣象事實綱要的實例，天氣氣象是我們具有高度興趣分析的方向，也是將藉了解天氣氣象因子而去探就影響空氣污染的因素。我們可由圖一了解到天氣整體變化的狀況，並可由資料的建立來看出天氣在監測的地區及各因子之間的關係，而我們以圖形化方式去描繪整個天氣氣象所含有的意義，以得到各因子影響程度的正確訊息。

4、 工作量(workload)純化與綱要確認:

主要是針對系統如何去純化資料倉儲中初步工作量設計部份，並以公式化方式去定義其因次綱要中內部的細部過程。其中則包含查詢在於資料倉儲的事實與資料容量的計算，並引導從邏輯與實體的設計方面，使得其概念化的設計能變成

具有成效性的，也因此而能去建構成較佳的資料倉儲結構，其中對於系統中查詢的工作量和資料容量將均會在邏輯及實體的設計部份佔主要的決定因素。

5、 邏輯的設計:

在於邏輯上的設計必須是能針對系統資訊做自動化週期性的更新，要使其操作資訊系統能做到最小查詢回應時間，且在於資料倉儲系統中並不會去影響系統執行上的效能。而我們知道正確的資訊系統在效能的計算上去定義出一個精確的估計方式是佔重要的決定角色。在整個邏輯設計化的過程中，採用一個單純化模式是不適合的，系統要使計算能過於精確則必須是較具複雜性的，須以先前所設計過的資料倉儲再針對其觀點(view)具體化、資料傳送及表格的垂直與水平分隔再重新設計，最後資訊系統設計時，不同的估計考量項目的增加將會是影響所能支援不同設計步驟過程之重要因素。

6、 實體的設計:

在實體設計的部份主要是有關於索引的最理想化選擇，是否根據其邏輯欄位、工作量及要求提供做特殊存取的架構，使其能從資料庫管理系統(DBMS)上去轉化所需的部份。而索引的選擇是一個決定資料倉儲執行效率的關鍵因素，索引的設計可應付資料倉儲的複雜性。它通常用於解決啓發性的問題上，並可利用關連方式來建構出查詢的能力及簡化查詢的複雜度，例如:巢狀迴圈、排序合併及簡單混雜關連，而在於更新查詢的部份上，我們相信一致的邏輯設計，將能有效地降低系統整體更新的時間。

以上爲建構資料倉儲系統的方法論，Golfarelli 提出一套建構過程，針對資料倉儲架構很明確的定義問題，並以 E/R 綱要架構關連方式來表達整個系統資料表示的過程，但實際上在資料倉儲建構時最重要的資料來源中，卻存在著許多技術

面的問題必須去解決，並在於考慮資料處理結合的部份也沒很明確去考量系統建構所需的因素。資料倉儲真正的價值是來自於不同資料資源搜集所得到的有用資訊，其資訊將能有效地提供決策分析。而資料倉儲在做事實綱要重疊時就能延伸出相關連的資訊，所決策出的資訊則會很確實的影響到決策者對整體未來方向的重大判斷。

故本論文中，我們將從使用者需求與系統維護者的角度，針對整個資料倉儲中事實綱要重疊之重要性去做分析及探討。並就在資料倉儲系統建構時，資料如何有效的彙整以強固其完整性的部份，利用事實綱要重疊所產生出的新的事實綱要來做說明，並將根據內部資料彙整在於自動化處理的事實綱要技術對資料結合及系統效益所帶來的種種影響提出研究及討論。最後特別再針對資料倉儲系統建構過程時，來自於不同內含關係的資料來源時，如何利用專家知識智慧式的選擇異質的資料來源，以取得有用的資訊進而產生更有效的決策判斷，並提出實例來加以實驗證明。

第三章、 資料彙整之重要性

我們可知資料倉儲在建構的過程中，最重要的決定因素其一來自於資料在做處理時內部相關資料彙整的完善性，故本章節將從資料倉儲內部的事實綱要(fact schemes)，針對同時考慮使用者需求與系統維護者，就不同事實綱要存在著不同的維度及屬性，從常發生資料結合的事實綱要欄位爲完全相容(strictly compatibles)與非完全相容(non-strictly compatibles)事實綱要的情況去做考量，以完全相容的事實綱要及非完全相容的事實綱要重疊說明及分析系統中事實綱要做重疊時有做與沒有做分別會造成怎樣的影響，並以利用重疊的技術說明重疊後所得到的資訊是優於原先沒有做重疊考量的，從天氣氣象的資料欄位來探討重疊的過程所能延伸之資訊。

第一節、事實綱要重疊

在資料倉儲系統中，不同的事實是表示不同的事實綱要，而所謂重疊則來自於二個相關事實綱要結合成一個新的綱要所產生的過程，其中相關連的內部屬性佔極大的重要性，它必須是不具相衝突的，如此才可根據其事實所需屬性重疊出真正想要的資訊。因此在不同的需求下會產生不同重疊情形，不同的情形相對的就會影響系統在建構時所要考量的因素與效益，並可能改變使用者所能取得資訊的正確與有效地推論原來所存在卻沒發現的問題，所以事實綱要在資料彙整時做重疊所產生新的事實綱要的過程是佔有極大的重要性的。

資料倉儲在概念化的設計必須考慮到的是資料在整合過程中，是否會影響系統整體的效益，首當其衝便是事實綱要重疊的部份，重疊的用處則在於結合於資料倉儲中不同的事實綱要，其可能也來自於不同的領域所產生的新事實綱要。其好處在於能夠使其在資料處理或 OLAP 時，能以較具系統化方式提供使用者做快

速查詢與交叉查詢，容易方便獲得其相關的資訊，並期望在相結合的情形下由原來綱要階層架構中的屬性，去發現二個事實綱要所具有的關連性，幫助我們從中更加能夠了解決策者所欲知的正確訊息。目前資料倉儲的事實綱要重疊均可由系統根據資料屬性主動化做結合的，故在這種情形下，我們試圖根據從原先事實綱要去了解，其是否爲多餘的事實建立，並深入考慮此因素的存在性，看是否能因此幫助系統有效地去刪減其屬性與欄位，進而能降低系統資料儲存的容量。

而不同的事實綱要在做重疊時均有其不同的考量因素，不管來自於建構系統時所需的快速查詢、成本的考量、減少系統資料容量或使用者的需求都是其重要的決策問題，事實上我們知道不同的事實均存在著不同的有用訊息，不但可藉由著做重疊的方式去得到其更完整資訊，並可幫助強壯資料倉儲架構的完整性。故我們要去定義其使用目的或何時才需要做重疊，才能有效地建構一個完整有效益的資料倉儲系統。然而不同情形下的結合會出現不同的事實綱要重疊，資料倉儲從系統建構的出發點去考慮，其不同的因素會造成產生出來的結果不一，故事實綱要重疊的用處除了考慮幫助使用者做即時查詢之外，尚需考慮系統維護的方便性，重疊之後的事實綱要與被重疊的事實綱要是同時存在的，當資料倉儲系統產生不當的綱要欄位重疊時，則資料倉儲系統的建構過程就會是龐大、複雜且不易的，而系統的容量空間也將會因此變大，並且複雜資料階層的結合會導致系統容量的激增，將會造成無法有效正確地取得有用的資訊，也會因此而無從得知不同的事實綱要間所存在的關連性。

因此可知在資料倉儲系統建構中，資料彙整時正確做事實綱要重疊過程是極具重要性的，系統會由不同的事實綱要做重疊，因而能取得的正確訊息，此時不但必須由明確的定義使用者需求與系統維護者角度，來避免不當多餘的事實綱要的重疊，且可由當中去相對得到或增加更有用的訊息，並從其中來延伸出不同的事實綱要所存在的相關連性，以幫助減少多餘的屬性增加及使用者不需的資料欄位，或是更進一步增加不足的資料屬性，更加保存原有資訊並藉此強固資料倉儲

系統的完整性。

第二節、完全相容的事實綱要

先定義有二個相容的事實綱要並具有共同相關連屬性關係，當它去做重疊時會創造出一個新的事實綱要產生。例如：有二個事實綱要分別爲 $f' = (M', A', N', R', O', S')$ 和 $f'' = (M'', A'', N'', R'', O'', S'')$，其綱要架構是爲完全兼容的情形，並假設它們最少會具有一個相同維度的屬性，且其內部屬性是不具相衝突性的。

給二個完全相容的綱要 f' 和 f''，我們定義二個做重疊時 f' 和 f'' 綱要的情形，$f' \otimes f'' = (M, A, N, R, O, S)$:

$$M = M' \cup M''$$

$$A = A' \cap A''$$

$$\forall a_i \in A\,(\text{Dom } f' \otimes f''(a_i) = \text{Dom } f'(a_i) \cap \text{Dom } f''(a_i))$$

$$N = N' \cap N''$$

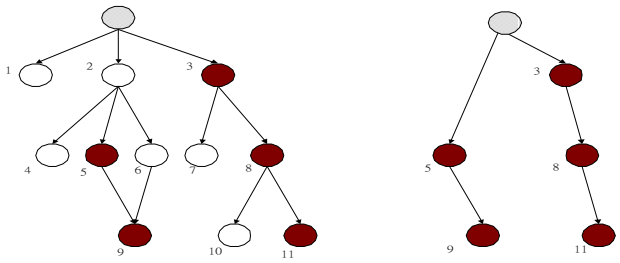
$$R = \{(a_i, a_j) \mid (a_i, a_j) \in \text{cnt}(qt(f'), A)\}$$

$$= \{(a_i, a_j) \mid (a_i, a_j) \in \text{cnt}(qt(f''), A)\}$$

$$O = \{(a_i, a_j) \in R \mid \exists (a_w, a_z) \in O' \mid (a_w, a_z) \in \text{path}_{ij}(qt(f')) \vee \exists (a_w, a_z) \in O'' \mid (a_w, a_Z) \in \text{path}_{ij}(qt(f''))\}$$

$$S = \{(m_j, d_i, \Omega \mid d_i \in \text{Dim}(f' \otimes f'') \wedge (\exists (m_j, d_k, \Omega) \in S' \wedge d_i \in \text{sub}(qt(f'), d_k)) \vee (\exists (m_j, d_k, \Omega) \in S'' \wedge d_i \in S'' \wedge d_i \in \text{sub}(qt(f''), d_k))$$

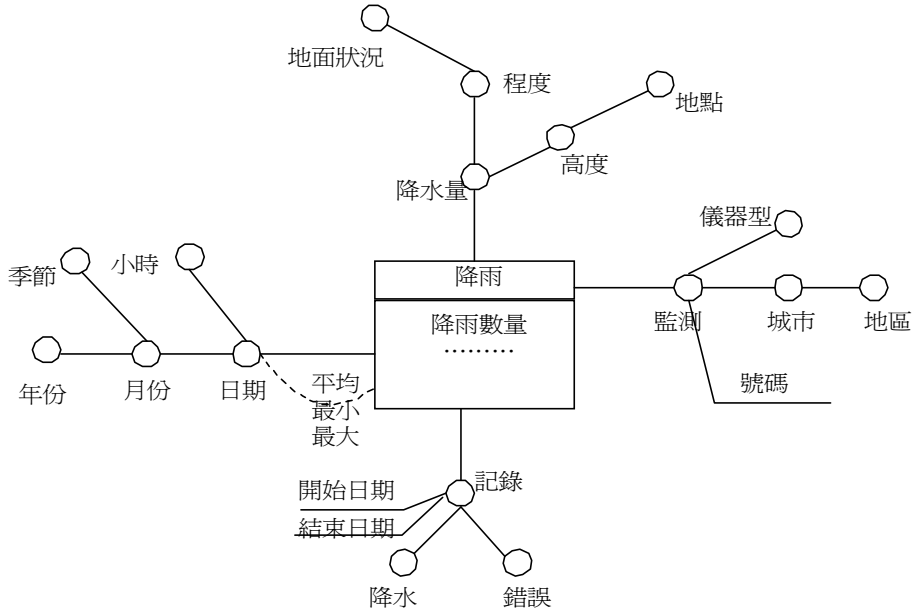
其中 f' 和 f'' 均是完全相容的事實綱要，$\text{cnt}(qt(f), A)$ 則表縮短類似樹狀(quasi-tree)事實綱要的屬性，重疊後則會針對其相關所共有的屬性去做結合，從相同的維度找出其相等的屬性來產生新的事實綱要，並因結合的過程而產生新的資訊延伸，下圖則呈現出一個類似樹狀的圖形以及它的縮減會表示在子節點的屬性上。



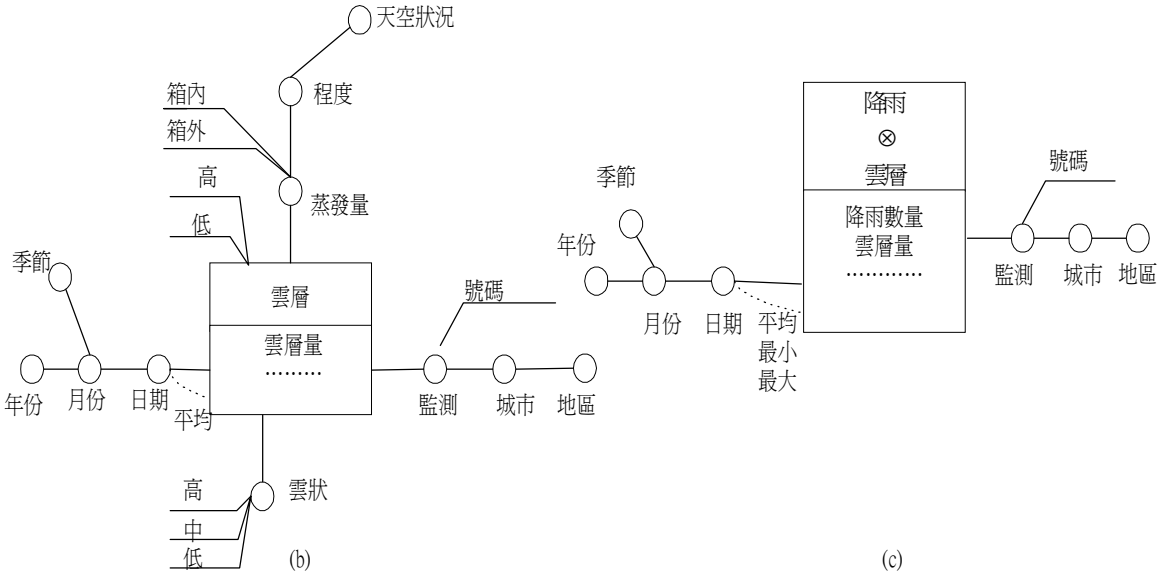
(a) (b)

圖二 呈現一個類似樹狀的圖形(a)縮減在褐色點(b)根爲黑點部份

在這部份呈現出二個單純具有相同關係的事實綱要如何做重疊而產生新的事實綱要的過程，它可能是具有相同的屬性關係，且會分別出現在不同的事實綱要上，或許同時也符合不同的維度，並藉由重疊過程而去產生新的事實綱要。其結果可幫助我們藉此了解系統自動化做重疊所產生的結果。圖三是二個完全相容的綱要做重疊過程的例子，針對降雨及雲層的部份去做探討，圖中其共有分享的資訊是來自於其日期與監測上的維度，這個新產生綱要的結果是從重疊所得來的，圖中可了解到從每一個因子去形成降雨及雲層的數量與其相關方面的資訊。



(a)



圖三 (a)爲降雨的綱要、(b)雲層的綱要及(c)重疊的結果

圖中天氣氣象爲例子: 其事實綱要中降雨屬性{(日期、小時、月份、季節、年份)、(降水量、高度、地點、程度、地面狀況)、(監測、儀器型、城市、地區)、(記錄、降水、錯誤)}和雲層屬性{(日期、月份、季節、年份)、(蒸發量、程度、天空狀況)、(監測、城市、地區)、(雲狀)}，所共有屬性{(日期、月份、季節、年份)、(監測、城市、地區)}，縮減降雨{(小時)、(降水量、高度、地點、程度、地面狀況)、(儀器型)、(記錄、降水、錯誤)}和雲層{(蒸發量、程度、天空狀況)、(雲狀)}屬性，建立新事實綱要降雨⊗雲層{(日期、月份、季節、年份)、(監測、城市、地區)}。

由圖可知，當二個完全相容綱要產生重疊後的結果:

1、f = f′+f″，因此 f 它是儲存包含f′和f″新的 “macro-fact” 。

2、在整個階層架構中，f 保存所有f′和f″共同唯有的屬性。

3、在 f 的範圍中每一個維度屬性，會符合原有f′和f″交叉出範圍所具備的屬性。

4、f 的內部屬性連結是隨意的，只要符合f′和f″的最小路徑。

5、f 即可表示說明f′和f″的集合。

而降雨與雲層的事實綱要中會共享時間及監測二個維度，新產生的事實綱要重疊結果可表示其間之關係，事實上在這二個資料來源的事實綱要階層中，重疊後延伸出更多精確考慮的資訊，不但能更加知道其間具有的關係性存在，並可進以提供決策者決策分析判斷，爲了能有效地建構資料倉儲則必須更明確由使用者需求與系統維護者角度去考量更完整的影響因素。

- 使用者的角度:

我們了解使用者所關心的來自於系統中資料查詢時所能得到的結果，故從使用者對事實綱要欄位查詢的部份舉例來做說明，藉此了解天氣在於降雨及雲層的因素部份其相關方面資訊的取得，以之前所舉天氣氣象在程度上變化的例子來看，假設使用者欲了解在某地區雲層造成在六月份的降雨情形。

例子一 以查詢爲例，假設用公式化的方式去表示:

降雨⊗雲層(日期,地區;日期,月份=“六月”)---降雨查詢---雲層查詢

使用者將不須分別從不同的事實綱要中去了解其所依存的關係，可藉由新的事實綱要快速從中查詢所欲知的資訊，讓查詢能變的更簡單化，並可由事實綱要重疊後從二個不同的事實當中去發現到資料彼此間的相依性或共同存有的綱要欄位，也了解到新的事實所存在的必要屬性爲何，是不會遺漏任何可用的資訊。相同在於做資料的查詢時，若查詢某地區雲層造成降雨量的狀態，也可藉由重疊後所產生的綱要欄位裏去獲得決策者所需要的資訊，同樣可知道其間所形成之影響關係，並可了解到其關連延伸出的資訊訊息。

- 系統維護者的角度:

多餘的查詢會相對地影響系統資料容量的大小，而系統維護者所關心的來自於在維持於最佳的執行效能上與系統中資料容量及查詢的綱要欄位是否能得到控制，我們從系統在於查詢部份造成資料容量的增減來做說明，當使用者在查詢

降雨的地區時藉而查詢與雲層之間的關係資訊，此時是有必要另建一個新的事實綱要來做查詢存取。

例子二 以查詢爲例:

降雨(日期,監測,地區=“左營”)---降雨查詢

降雨⊗雲層(日期,監測;日期,月份=“六月”)---降雨查詢---雲層查詢

系統正確做事實綱要重疊，不但能減少多餘資料欄位的重複性，並也藉由自動化的重疊因而產生可用的綱要欄位，以減少容量的使用與降低延遲的問題，增加執行上的效能，易於系統維護者做系統的維護及控制。而我們了解系統能自動化依其共有屬性去做重疊，產生出的新事實綱要是有助於了解與查詢二個不同事實綱要之間的關連性。

此時使用者能藉由降雨的事實綱要中去快速取得雲層在某地區監測之訊息，系統不用等到使用者所需求查詢此綱要欄位時，再去從分別由個別複雜的資料綱要中去做產生新的關連，能更加快速的自動化結合所有有用的事實綱要欄位，以達到其系統效率上的提升，並使得使用者能更加方便快速去得到其所需求的資訊。

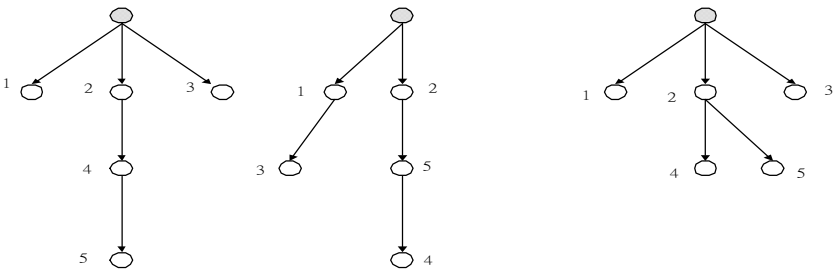
第三節、非完全相容的事實綱要

而當我們了解資料倉儲系統在完全相容的事實綱要重疊的實例後，甚至說還必須了解資料是二個非完全相容的綱要架構做重疊時的情形。在這部份其二個縮短的類似樹形圖是不同的，這裏必須是獨立在二個綱要中有一個或更多衝突的內部屬性，而產生的事實綱要是完全相容的，也因此我們必須去解決其內部相衝突的部份。

同樣給二個非完全相容的綱要f′和f″，我們定義二個做重疊f′和f″的綱要，$f' \otimes f'' = (M, A, N, R, O, S)$：

$$M = M' \cup M''$$
$$A = A' \cap A''$$
$$\forall a_i \in A(Domf' \otimes f''(a_i) = Domf'(a_i) \cap Domf''(a_i))$$
$$N = N' \cap N''$$
$$R = \{(a_i, a_j) \mid \exists p_{ij}(cnt(qt(f'), A)) \wedge \exists p_{ij}(cnt(qt(f''), A)) \wedge \forall a_w \neq a_j \mid (\exists p_{wj}(cnt(qt(f'), A)) \wedge \exists p_{wj}(cnt(qt(f''), A)))(p_{ij}(cnt(qt(f'), A)) \subset p_{wj}(cnt(qt(f'), A)) \wedge p_{ij}(cnt(qt(f''), A)) \subset p_{wj}(cnt(qt(f''), A)))\}$$
$$O = \{(a_i, a_j) \in R \mid \exists (a_w, a_z) \in O' \mid (a_w, a_z) \in path_{ij}(qt(f'))) \vee (\exists (a_w, a_z) \in O'' \mid (a_w, a_z) \in path_{ij}(qt(f'')))\}$$
$$S = \{(m_j, d_i, <op>) \mid d_i \in Dim(f' \otimes f'') \wedge (\exists (m_j, d_k, <op>) \in S' \wedge d_i \in sub(qt(f'), d_k)) \vee (\exists (m_j, d_k, <op>) \in S'' \wedge d_i \in sub(qt(f''), d_k))\}$$

在相同的情形下，我們討論當f′和f″同爲非完全相容的事實綱要時，重疊的過程則會針對其新的f綱要所需的相關性屬性去做結合，並找出其必須所共具有的屬性來產生新的事實綱要，下圖則呈現出二個非完全相容的事實綱要做重疊的過程。

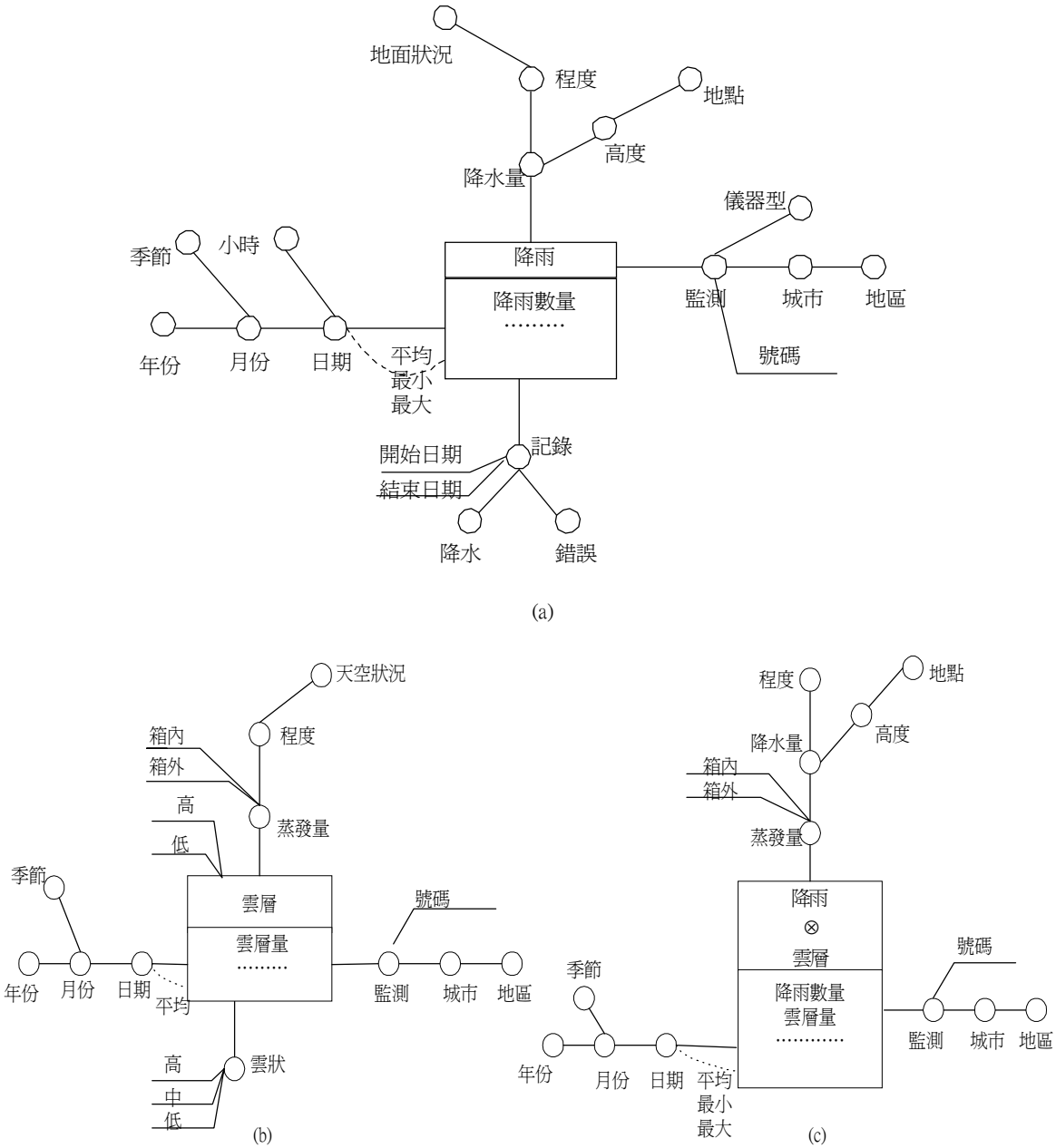


(a) (b)

圖四 二個非完全相容事實綱要 (a)和(b)

上圖是不具有相同的屬性關係只有部份相同屬性的關係，而分別出現在不同的事實綱要上，或許同時也符合不同的領域範圍，同樣也藉由重疊的過程去產生新的事實綱要。圖五是二個非完全相容的綱要做重疊過程的例子，在於天氣氣象

的降雨及雲層的影響程度上，我們可知其共有的屬性是其日期與監測，藉由圖中可了解到每一天雲層數量影響到任何不同的地方降雨的情形，並可進一步了解蒸發量與降水量其相關方面的資訊。



圖五 (a)爲降雨的綱要、(b)雲層的綱要及(c)重疊的結果

圖中降雨屬性{(日期、小時、月份、季節、年份)、(降水量、高度、地點、程度、地面狀況)、(監測、儀器型、城市、地區)、(記錄、降水、錯誤)}和雲層

屬性{(日期、月份、季節、年份)、(蒸發量、程度、天空狀況)、(監測、城市、地區)、(雲狀)}，所共有屬性{(日期、月份、季節、年份)、(監測、城市、地區)}，使用者欲求知雲層的蒸發量導致降雨降水量的程度，縮減降雨{(小時)、(地面狀況)、(儀器型)、(記錄、降水、錯誤)}和雲層{(天空狀況)、(雲狀)}的屬性，建立新的事實綱要降雨⊗雲層{(日期、月份、年份、季節)、(監測、城市、地區)、(蒸發量、降水量、高度、地點、程度)}。

由圖可知，當二個非完全相容的事實綱要產生重疊後的結果:

1、f = f′+ f″，因此 f 它是儲存包含 f′和f″新的 “macro-fact” 。

2、在整個階層架構中，f 選擇性保存 f′和f″所具有的屬性。

3、在 f 的範圍中每一個維度的屬性，也會符合原有 f′和f″交叉出的範圍所具備的屬性。

4、f 的內部屬性連結是給予的，也會符合 f′和f″的最小路徑。

5、f 即可表示說明 f′和f″的集合。

在非完全相容的事實綱要中，我們同樣可藉由重疊來取得我們所需要的資訊訊息，以幫助我們在於了解天氣氣象變化中，在每年或每月蒸發量所造成的雲層及降雨狀況或更多資訊需求。

- 使用者的角度:

在非完全相容的事實綱要時是必須以使用者需求選擇定義，要能完整地保留全部所依存的資料，同樣在於使用者對資料查詢的部份去舉例來做說明，在二個只具部份相同的屬性存在時，我們必須假設使用者的需求爲何，故從使用者了解天氣在降水量及蒸發量的資訊後(其中包含日期、程度、地點……)，可透過新的事實綱要進一步取得其日照形成蒸發量再造成降雨到某地區的資訊。

例子三 以查詢爲例:

降雨⊗雲層(日期,程度,地區;日期,月份=“六月”,地區=“左營”)---降雨查詢---雲層查詢

非完全相容的事實綱要所產生的重疊，它是具有部份相同的屬性去做結合，若要從使用者的需求去做明確的定義，則必須要考慮其欄位要如何做重疊才會產生最佳需求的情況。此時重疊後除了能幫助使用者能快速及正確查詢以外，並能了解到獨立在二個事實綱要中，其不同的內部屬性有一個或更多衝突情形予以整合，所會發生的不同整合情形。實例中，我們欲知的是降雨量與雲層的關係，卻能進一步了解蒸發量影響降水量的程度。

能正確的了解使用者需求才能產生整個事實所隱涵的原有意義與其延伸的關連性，而不會造成錯誤的需求分析結果，因而誤導資料倉儲決策分析的方向及能力。原來的事實綱要如果本身沒有相同的關連性屬性存在，則原先的事實綱要只具有部份相同的屬性並無相依性，若能明確的考慮使用者需求則不會模糊掉原有的資料意義，才能有效達到系統的正確性。在二個不同非完全相容的事實綱要下產生新的綱要查詢，不但能完整的保留了原有資料綱要所具有資訊，且可更快速得到所需求的訊息，並可明確延伸出關連二者之間的關係。

- 系統維護者的角度:

在於非完全相容的情形下，系統維護者對資料做查詢時所建立的事實綱要，可從系統容量變化的部份說明其重疊的結果。在此系統維護者是要考量資料屬性複雜度及階層的變化，然後有必要適當去刪減或增加其屬性的節點，以維護系統資料容量空間的控制，保持執行上的有效性，重疊後使用者可藉由新的事實綱要中查詢到其所想要的資訊，以下爲例做說明。

例子四 以查詢爲例:

降雨(日期,程度,地區=“左營”)---降雨查詢

或降雨⊗雲層(日期,程度,地區;日期,月份=“六月”,地區=“左營”)---降雨查詢---雲層查詢

非完全相容的事實綱要可能會因重疊後所產生的結果而提高其複雜度,若不當的綱要重疊是會產生多餘事實綱要的保留,造成系統容量的負載,系統管理者會不易做管理。當不知結合後所產生出來的事實是否有用或何時需要用,此時會使得資料倉儲存有的欄位相形變得複雜,造成執行效率上的遞減與查詢的不易,如此一來反而會延伸出更多資料倉儲系統的問題。正確的做重疊是能有助於系統維護者做資料處理,可去簡化原有事實的複雜性以有效的降低其屬性階層。

重疊可主動依尋存取其相依的關連性,在不會改變原有的資料屬性,對部份相同的屬性去做重疊,並不會影響原有的事實綱要。此系統若能正確定義建構的過程,資料重疊會幫助提供使用者查詢了解到另一不同事實的欄位,能有效存取到有用資訊,建製一個有用資料倉儲系統。且可因正確的事實綱要欄位的重疊,來幫助資料倉儲系統減少龐大容量空間的需求,於不同需求定義下,在事實綱要屬性中增減其所需的屬性,使得原本複雜的綱要變成較具簡單而有用的事實綱要,並使事實綱要簡單化,讓使用者方便查詢,系統相形較易管理。

本章節我們就完全相容的事實綱要及非完全相容的事實綱要在資料結合處理的過程中,探究事實綱要重疊的技術,並針對使用者需求與系統維護者的角度,提出在資料倉儲內的資料彙整過程所必須考慮的問題,再強調事實綱要重疊技術的重要,從提出事實綱要重疊所能得到的好處及效益,了解內部資料處理所產生的資訊需求是優於原先沒有做重疊的資料,且利用重疊的過程強固資料處理的事實綱要,藉以提供資料倉儲在建構過程中,考慮不同資料來源的資料彙整時,所能創造出的相關資訊,以幫助提高資料倉儲系統決策分析的品質,而下章節我們將討論來自於不同資料來源而具有內含關係性資料的重要性。

第四章、資料來源之重要性

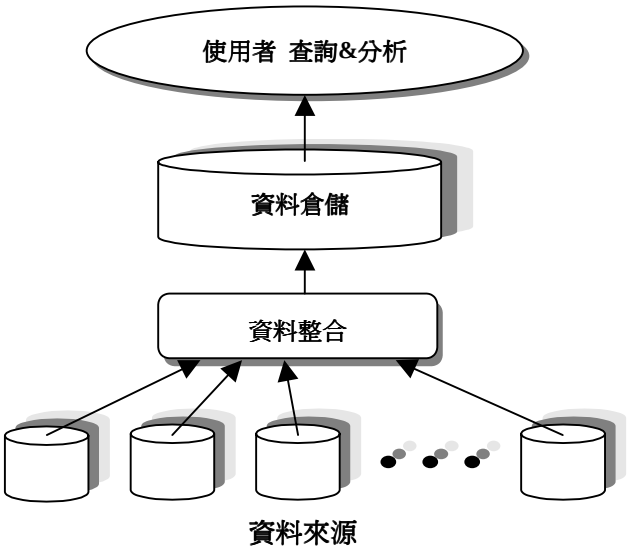
資料倉儲是整合異質或分散式資料資源的資訊儲存體，若其來源爲操作資料層關係所關連而來的資訊，我們會認爲其是有明確的關係性存在。但在另一方面，資料來源也可能會來自於不同的操作資料，我們說它是不明確或具暗示性的內含關係。目前資料倉儲建構均是來自於相同資料資源，所以我們則要去討論來自於不同資料資源的資料倉儲，且其具有高度相依性關係存在的資料，並透過專家知識或專家經驗來組合而成有用的資料倉儲系統[1]，而此資料倉儲系統會是較具能力來強壯資料倉儲的完整性。

我們從例子來說明此論點，嘗試在表面上操作資料階層看似不相干沒有任何關係，但實際上卻存有高度關係的天氣氣象與空氣污染的資料，來建構一個具內含關係的資料倉儲系統，並藉由專家的知識及經驗做資料來源的彙整，以類神經網路的探勘技術，找尋出有用的資訊關連，來建構一個具有效性的資料倉儲之空氣污染預測系統。

第一節、資料倉儲的資料來源

目前資料倉儲系統中的資料是來自於不同異質的資料來源，當使用者欲做特別查詢或分析應用時所需求的資訊資源均可從資料倉儲系統中存取而得。圖六爲描述資料倉儲的基本架構，這是一個基本的架構，我們可知在建構的過程中，其資料來源或許是有關連並具有資訊相關的，故可藉由操作資訊系統透過資料庫語法整合而成，例如:外來 key，做所謂的語法關連以達成資料的索引，使其是具有明確關係性存在的。現階段資料倉儲也是來自於其關連性的資料來源，而目前的操作資訊系統和使用者需求的限制條件卻是在於收集和過濾資料的過程。故必須在資料來源的部份去加強其彙整的完整性，才能是爲具提供決策支援分析能力的

有效資料倉儲系統。



圖六 描述資料倉儲的基本架構

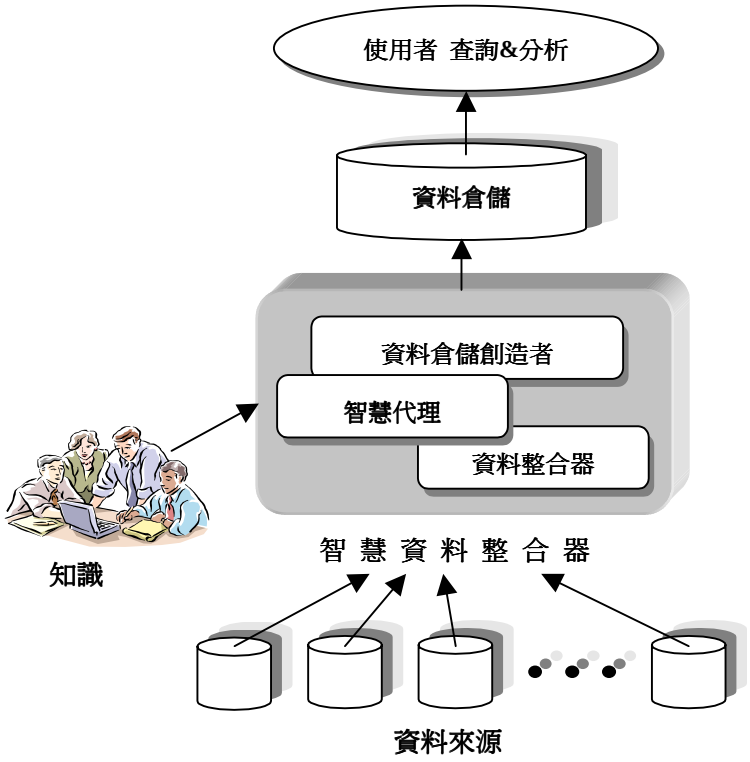
目前我們可從一些複雜的應用上可看出,資料倉儲系統爲了能增加其決策支援或分析能力則必須能包含來自於不同的資料來源才有辨法在於其品質上的提升[48]。而資料資源的來源或許是來自於操作系統的層面上，我們可說它是不明確或具內含關係性的資源，故在於這個部份加強於考慮整合資料資源是必須的，我們則靠專家知識與專家經驗加以篩選資料資訊，並以資料探勘的技術例如:類神經網路(neural network)幫助資料倉儲的挖掘。

第二節、智慧的資料整合

初步證實明確的關連資料來源或許是使用者猜想或一些應用假設所去關連的，雖是如此，但實際上它們可能是存在著高度的關連性，且它們本身也可能是相互具有關係性的。而這個問題早就存在於早先的操作資訊系統中，但目前卻是不好的設計，主要是因爲無法足夠聚集及過濾得到必要的資料和資訊，所以在這

方面我們是必須靠專家知識的幫助，篩選來自於不同的資料來源進而去達成其資訊系統的構成。可知不完整的資料資源會去影響資料倉儲系統的創造，使其容易發生錯誤的資料訊息出現，明確的解決方式則在於資料倉儲中加入專家知識及經驗使其能去取捨整合不同的各類型資料來源。

圖七呈現一個新的智慧型資料倉儲結構的知識架構。這種智慧型的知識整合能創造出新的更具有彈性、精確及智慧的資料倉儲系統。圖中的資料倉儲創造者是必須能去綜合不同欲加以整合性的資料資源，其中再由資料整合器自動產生資料結合，知識和他們之間關連的結果可從智慧代理的部份主動構成，最後組成完整的資料倉儲，這樣的資料倉儲架構是不同於先前單一部份來源而是藉由許多不同的資料來源所構成的。



圖七 智慧型的資料倉儲結構

我們確知外來因素會影響資料來源其間之關係，並造成不同資料來源的模

糊，因而必須需求於專家知識才能去達成資料倉儲的完整性，且根據專家經驗去做收集及過濾資料的過程，即可成爲更完整的整合知識資料來源，以幫助資料倉儲提供更完善的決策分析能力，增加資料倉儲系統的有效性。故若能對於智慧代理部份的知識善加處理，加強自動化的資料處理，則能更正確的考量來自於不同操作資料來源的整合。

而資料是資料倉儲系統建構過程中最重要的考量因素，不同資料來源之間若存有高度相關連的關係，則必須考慮使用者需求及早先操作系統的過程才能正確的從中去搜集及過濾資料訊息，進而創造更完整的資料倉儲系統。而完善的資料倉儲系統是能以智慧知識的方式自動化來達到資料資訊的彙整，以有效的處理任何不同資源而來的資料，並自動利用重疊技術轉化結合來自於過去的資訊，進而延伸更多的相關未知資訊之訊息。

第三節、實例驗證

科技發達帶來科技文明的進步，然而人類在享受文明時已面臨到環境的加速改變，全球的空氣污染目前已嚴重的影響到環境保護，人類賴以生存之空氣、土壤、水源、海洋等自然環境，全因環境的遭受破壞、侵蝕，而不知覺已全然不同，所造成之污染嚴重影響生態及自然環境，實質帶給人類極大的衝擊。而現今工業化的社會造成目前空氣嚴重的污染，長期對人體的健康產成不良的後遺症，使得環境保護已開始成爲全世界世人所共同關心的焦點，這是一個國人所必須去正視的一個問題，如果能有效預測空氣品質的好壞與破壞的程度，再加以分析決策出問題的所在與變因的發生，不但可做適當合理的預防，且可由其中找尋出相關連的變數而去做有效的改善。

根據過去許多研究指出[5]，從南高屏地區研究臭氧高值中發現空氣污染物

臭氧的高值區位置易受風向之影響，而空氣污染的懸浮微粒及揚塵微粒濃度會因降雨量來抵制灰塵使其揚塵變小[6]，並發現隨著季節的變化，地形對降雨的影響有顯著的改變[9]，全省區域性地形其地形特徵程度的不同，降雨也會明顯的改變[10]，因此我們可由此大膽的假設天氣氣候是空氣污染的最大影響因素。而污染物中的 SO_2 與 NO_2 是造成雨水酸化最主要之污染物，研究後了解二者之間是具有高度相互關係[12]。

並由 SO_2 和 NO_2 之逐年及逐月趨勢分析，發現 SO_2 濃度受固定污染源的分佈影響較大[13]，其中每日最大臭氧濃度有高度相關的氣象因子，由最高溫與最低溫之差的變化做改變，氣溫使得光合反應更加旺盛，造成臭氧濃度不斷累積升高，形成高污染事件[7, 8]。臭氧濃度也因地形、季節及大氣的環境而有所不同[11]，因此天氣變化對空氣品質會有明顯的改變。研究實驗也發現車行揚塵微粒濃度與車流量、坋土負荷量與風速之間呈現正相關，其中在車流量的影響上，以汽機車所帶來影響最爲嚴重[6]。

又因台灣的地理位置相當特殊連帶著其氣候型態也是與眾不同，氣流的流動會對空氣污染會造成極大的影響[2]。另一方面研究則認爲，隨著季節的變化，地形對降雨的影響有顯著的變化，空氣污染物質在大氣邊界層中的擴散作用與氣象、地形及污染源特性等等息息相關[3]。由此可知天氣氣象與空氣污染有明顯的相依關係，故可藉由天氣氣候與空氣污染的資料提供去做預測分析，了解相關要素的變化[4, 15]。空氣污染產生的原因很多，這些原因的急遽變化，皆會影響預測的準確性，然而預測的精確度與預估能力是否能達到污染防治的要求，則必須藉由實際的模擬試驗來給予肯定。

我們就相關專家領域知識可以非常確定，天氣氣象是會實質影響空氣污染的程度。而影響天氣的主要因素爲平均氣壓、平均氣溫、相對濕度水氣壓、平均風、降水量及降水紀錄，加以配合空氣污染指數一氧化碳、臭氧、二氧化氮、二氧化

硫及懸浮微粒，利用天氣氣象的條件來預測空氣污染因子以準確了解空氣品質污染程度。本文首先改變由過去傳統單一或原本相關聯資料源的建構方式，改以由二個於作業系統階層不相關聯的資料庫(天氣氣象資料、空氣污染資料)，建構綜合性的資料倉儲，然後利用專家知識及經驗的相關研究做資料資源的整合，並使用資料探勘的技術來分析資料及建立預測模式。

首先我們先分析天氣氣象與空氣污染各項主要因素及其間影響與關聯:

1、 天氣氣象的因素:

由[15]可知天氣氣象原始資料的欄位:

測站	時間				氣壓	氣溫		濕度		露點	雲狀		
號碼	年	月	日	時	0.1 hpa	乾 0.1^{0}C	濕 0.1^{0}C	絕 0.1^{0}C	相 0.1^{0}C	0.1^{0}C	高	中	低

低雲量	總雲量	風		降水		日照時數	能見度	天空狀況	地面狀況
		向 16	速 0.1m/s	量 0.1mm	時 0.1hr	0.1 hr	0.1 km		

天空及視障				地中溫度 0.1 ^{0}C					降水記錄	錯誤記錄
雷暴龍捲	液體降水	固體降水	視障	0 cm	5 cm	10 cm	20 cm	30 cm		

並在幾個研究報告中發現氣候與空污的相關資訊[6, 7, 9, 16]:

1. 在同時考慮溫室氣體的增溫作用與氣溶膠的冷卻作用情況下，所有氣候模式皆可推估台灣鄰近地區的平均氣溫將持續上升。在二氧化碳增爲 1.9 倍時，溫度將上升 0.8-2.4^{o}C，但冬夏季增溫的程度並無明顯區別。
2. 溫室氣體增加的同時，大氣中的懸浮微粒相對也增加，並且具有冷卻的作用。
3. 降水量的增加可有效降低揚塵，使得懸浮微粒物大量地減少。

繼之，與台大大氣科學所專業研究員 Eric Ma 先生研討得知另外有 5 個影響因素：

1. 風向的影響: 風向正確(吹向外海)將有效排除空氣污染物。
2. 氣壓: 若有高氣壓盤旋該地將使得廢氣停滯，造成空氣污染的加劇。
3. 例假日: 城市大氣中的 NOx 其中 2/3 來自汽車等流動源的排放，1/3 來自固定污染源，人們若大都待在家中休息工廠也不開工則會降少污染物的產生。
4. 空氣中的浮游物可幫助霧氣凝結，形成雨雲增加降雨的機會。
5. 日照量充足將有效地幫助光化合反應。

故根據學者專家的意見，初步決策分析出來天氣氣象影響主要因子有以下:

時間			平均氣壓	平均氣溫	相對濕度水	平均風	降水量	降水
年	月	日	0.1hpa	0.1m/s	氣壓 0.1hpa	0.1m/s	0.1mm	紀錄

以上這些是主要影響空氣污染的天氣因子，本研究將以此爲主要空氣污染的影響變數，再藉由從不同地區(台北、花蓮)來做預測模擬的比較。

2、**空氣污染的因素:**

空氣污染物中之二次污染物臭氧爲高氧化性污染物，在別的研究中提出它不是由污染源直接排放而來的，而是在低層的大氣中由各個污染源(包括固定污染源、移動性污染源等)所排放出氮氧化物、碳氫化物及反應性有機化合物等，再藉著複雜的光化反應所形成而產生的。

空污又可分爲一次污染物、二次污染物的產生及危害:

- **一次污染物:** 一次污染物是指直接從污染源排放出的污染物質，如二氧化硫、一氧化氮、一氧化碳及顆粒物等等。

- **二次污染物:** 二次污染物是指由一次污染物在大氣中互相作用經化學反應

或光化學反應形成的，與一次污染物的物理及化學性質完全不同的新的大氣污染物，其毒性比一次污染物還強。最常見的二次污染物如硫酸、硫酸鹽氣溶膠、硝酸、硝酸鹽氣溶膠、臭氧及光化學氧化劑 O_x，以及許多不同壽命的活性中間物（又稱自由基），如 HO_2、HO 等。

因而可藉由[15]中所表示，監測的各污染物之成因及危害空氣污染物的影響因子:

時 間			一氧化碳	臭氧	二氧化氮	二氧化硫	懸浮微粒
年	月	日	(ppm)	(ppb)	(ppb)	(ppb)	(ug/m^3)

以上這些均是空氣污染主要決定因子，我們可知地面氣象觀測站之儀器的安裝和位置，對數據的正確與否有著關鍵性的影響。故必須準確的藉由中央氣象局與行政院環保署所收集的資料做資料的建置，進而去探討空氣污染有效預測的可行性，並找出影響空氣污染的成因爲何，期望此系統能幫助政府相關單位在對於防治空氣污染方面提供一些有用的幫助。

空氣污染預測雛型系統開發的原則是以天氣氣象條件來預測空氣污染因子，其建置步驟如下：

1、 網頁氣象資料的收集：在中央氣象局的網頁上，收集84~87年逐時的天氣氣象的資料，再統計分析出逐時、逐日的連續性資料。

2、 網頁空氣污染的收集：在環保署的網頁上，84~87年逐時、逐日的天氣氣象的資料來結合天氣氣象的資料來做有效的決策分析。

3、 資料倉儲的建置：多維度資料分析包括地區（台北、花蓮）、天氣氣象（平

均氣壓、平均氣溫、相對濕度水氣壓、平均風、降水量、降水紀錄）、空氣污染（一氧化碳、臭氧、二氧化氮、二氧化硫、懸浮微粒）。

4、 資料準備的階段：解讀資料欄位與轉換資料，將原始資料中的特有表示方法轉換成可運算的數值。同時處理資料淨化包括空值及無效值處理，均以月平均值來代替。最後整合資料將各個欄位中的資料儘量以統一單位來表示。

5、 資料分析的階段：首先過濾欄位，將氣象資料欄位依照專家發表之論文來過濾欄位，目的是避免不相干欄位的影響，而遮罩了真正相關欄位與因果的關係。其次再統計資料，將逐時的氣象資料轉換成逐日的資料，相對應空氣污染值逐日平均值。最後合併資料，利用年、月、日的欄位，將氣象資料與空氣污染資料結合在一起。

6、 知識探索階段：決定資料探勘的方法爲類神經網路，因爲輸入的資料及結果爲數值，如以決策樹(decision tree)及分類(classification)分析則不適合；且因爲多個輸入欄位(變數)如以迴歸(regression)分析亦不適合。

類神經網路是在於建立資料倉儲時很好的探勘技術，其自動學習的方式與相關影響的特性，均能有效模組化並控制資料的型態與系統的正確，而在於許多文獻中都有理論上的依據與實際上的探討，例如:可辨識印刷字、手寫字的字元、語音輸入法及股票預測或者可選擇地理區域去向特定消費者來進行行銷策略等等，這些都是很有效的應用例子。

而本文依系統資料的型態，選擇類神經網路做爲模擬預測的方法，其中利用類神經網路在資料探勘技術中的監督學習（Supervised learning）的學習方式的特性，建立提供輸出入的相關特徵因子，加以在動態學習或相互抑制的神經網路中，從問題領域中取得訓練（有輸入資料，也有輸出變數），並從中學習輸入變

數與輸出變數的內在對映規則，然後再考慮相互影響的情況，期以得到具相互關係的神經元，監督學習後模擬輸出再加以預測，藉以快速開發新的預測空氣污染雛型系統。

本系統運用類神經網路的學習方式，建置空氣污染網路的預測模式，利用資料挖掘的技術，藉以準確預測空氣污染值，來有效的提供決策分析，以下爲其學習的過程:

1、選擇輸入的預測變數，本系統以專家的知識及經驗，建製天氣氣象爲主要的影響因子，並挑選出空氣污染的主要變數因子，並爲了驗証其假設則加入區域性因子(台北、花蓮)，而此輸入初選的天氣變數因子爲（平均氣壓、平均氣溫、相對濕度水氣壓、平均風、降水量及降水紀錄）在類神經網路的輸入層中來做爲其模型中輸入變數的資料。

2、選擇全部的輸出空氣污染變數，即爲空氣污染的排放有害物質，就是產生預測的空氣污染因子（一氧化碳、臭氧、二氧化氮、二氧化硫及懸浮微粒）的資料欄位變數，這是在輸出層其欲預測的變數值，在類神經網路中定義爲輸出的變數影響因子，其變數也是我們欲查詢及預測的值，並經由建立的類神經網路模型做模擬預測。

3、再來選擇全年歷史資料在此預設爲民國 84~87 年的天氣氣象與空氣污染完整的歷史性資料，以前 3 年爲類神經網路學習訓練的資料，藉以幫助建立完整的預測模型，此目的則是爲能準確的預測正確值，且爲了不失其正確性，故必須是近年的歷史性資料，再以 87 年整年的歷年資料做爲預估測試的資料，來比較其預測的準確並分析其變數的影響。

4、此網路利用倒傳遞類神經網路(Back-propagation Network，BPN)的監督學習方

式，使用最陡坡降法(Gradient Steepest Descent Method)的觀念將修正網路連結上的加權值，使誤差函數予以最小化，並且爲了讓其值趨近於目標輸出值，則必須控制其學習的功能函數變化，讓輸出值會是最佳的收斂值，而我們選擇輸出的正確預測資料位置，爲預則空氣污染值資料存放所在，來建立完整的類神經網路探勘模型。

5、定義要預測模式的型態，再依據類神經網路所能解決的問題型態作選擇 prediction，以做爲預測模擬系統的建立，目的在於能表示其輸出的資料是爲連續數字範圍或抽象的數字順序，因輸出的預測值會是數值型態，而此時的資料輸出會是單一固定的值，即爲透過計算推論輸出的空氣污染預測值。

6、根據輸出與輸入資料的雜訊影響，挑選處理雜訊的等級，其因資料型態則爲較高的雜訊，故選擇 very noisy data，非常不一致的行爲資料來做爲空氣污染的預測，目的則在於要消除資料雜訊問題，而在選擇較高雜訊處理等級時，較可避免因資料的雜訊而產生過適化問題，故利用 Kalman 學習法來處理雜訊的問題，且在隱藏層單元數設定的部份，因其爲問題雜訊很高的資料學習型態，故爲了避免發生過度學習的現象，因而隱藏層處理單元就不宜過多，不然就很難收斂了。

7、分析天氣氣象輸入的資料欄位並且決定其欄位及轉換的型態，定義其爲數值資料型態 comprehensive data transformation，使預測出的資料欄位在類神經網路中能有效使用，且內設給予網路學習的加權值，使其學習速率能適當學習，以達到良好的收斂性，因探討對象爲單一方向變動之測試，故轉換函數採用雙曲函數(transfer function)，目的把作用函數之輸出值轉換成爲運算元之輸出值，而其間的輸出值域爲[0,1]。

8、輸入變數分類的選擇則是使用基因演算法(Genetic Algorithms)爲期初步訓練

的方法，再從資料分析及轉換過程中去建立全部欄位以找出較佳的分類及合理的集合，以做分堆的資料處理。而使用基因演算法時，由於不同的母體將產生不同的結果，所以當使用兩個不同資料去建立相同模式時，會有不同的變數選擇，選擇 exhaustive variable selection 以建立複雜的資料模式，使其爲適當的變數，並在輸入過程後則再加入類神經網路進一步的分析。

9、爲了要能建立 exhaustive network，當建立模式時就必須藉由類神經網路自動學習的方式，在收斂激發的過程中，並允許選擇訓練的困難度，以建立自動訓練的學習網路，且在網路的訓練過程通常以學習循環(learning cycle)的方式，逐步的學習以達到有效的收斂爲止，並適當的學習循環數目下，才能使其避免收斂誤差的出現。

10、最後系統會依資料內容的定義、變因選擇性及歷史時間的不同而改變，完成建立類神經網路的空氣污染預測模型，而我們以輸入天氣氣象的變數因子有效的預測即時的空氣污染的輸出預測值，並分別建立台北及花蓮二個地區的預測系統，利用 client-server 存取的方式提供使用者做系統的查詢，以得到相關資訊的訊息。

資料倉儲預測模擬系統建立後，我們以 87 年 6 月 1~7 日 00 時定時爲例，分別在台北、花蓮二區做預測模擬，將其預測結果顯示在以下的表 1、2 中，由表中可見平均誤差均在 25%左右，二個地區中會有某誤差率偏高，但注意其誤差變異數之平均值接近於 0，代表其有可能因爲外來因素改變或天氣因子的不明確性，如單日預測不準的因素就可能來自於外來因素所導致(如:汽車廢氣、工程施工、氣候驟變等等)，或者是外來影響因素考慮不足等因素造成。本研究期望就根據專家的建議經由專家的知識及經驗透過這樣的研究分析過程，能發掘出更多具實際影響的因子。

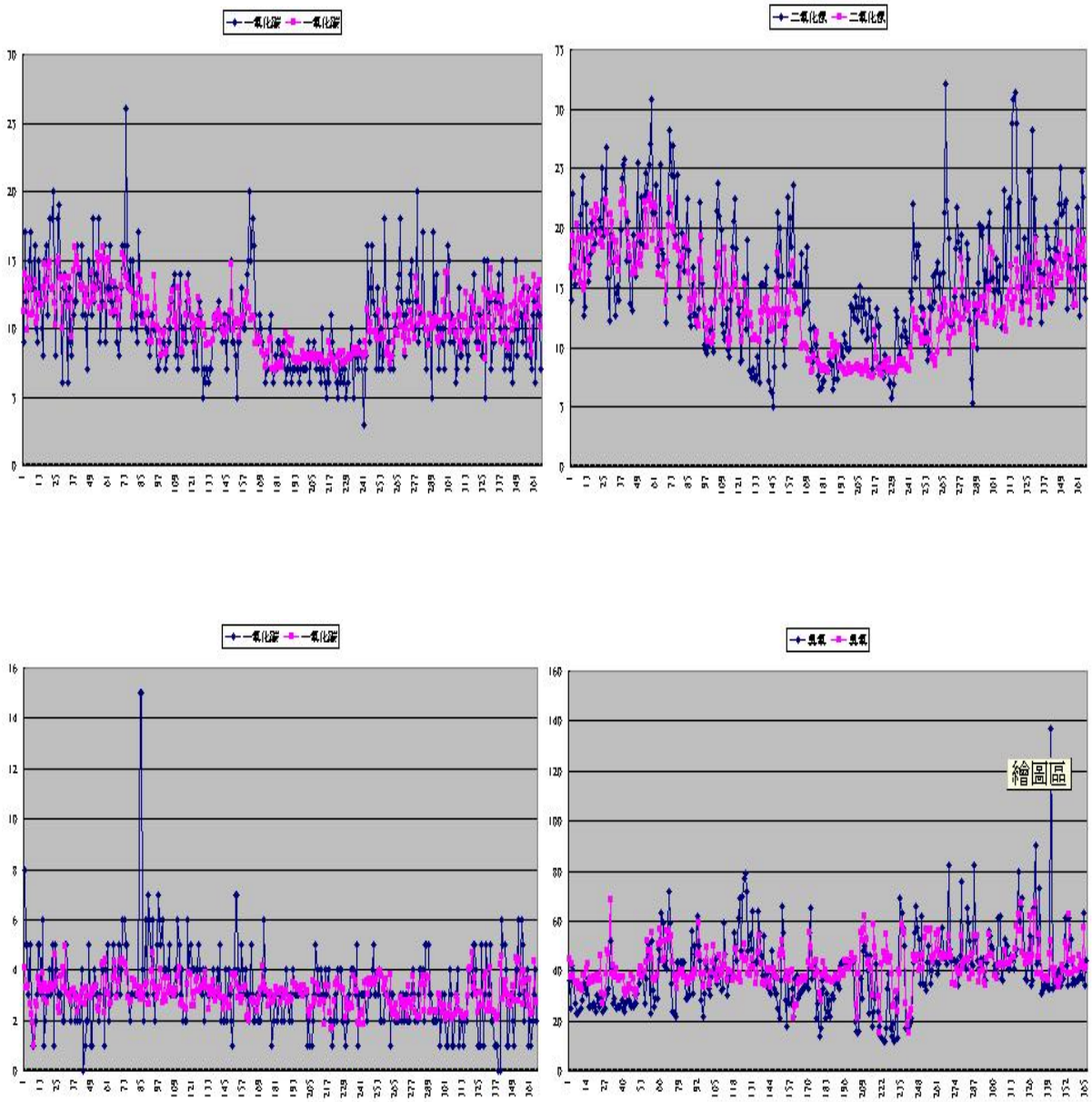
表 1 台北空氣品質預測

污染物 / 日期	一氧化碳(ppm)		臭氧(ppb)		二氧化氮(ppb)		二氧化硫(ppb)		懸浮微粒(ug/m^3)	
	預測值	真實值	預測值	真實值	預測值	真實值	預測值	真實值	預測值	真實值
87/06/01	0.4394	0.567	47.0610	39.465	4.9063	6.938	4.3794	5.828	21.2248	46.971
87/06/02	0.2897	0.355	27.019	19.683	4.113	5.782	2.0989	3.96	14.6928	22.926
87/06/03	0.32355	0.266	26.7866	16.332	4.2073	5.697	2.0942	2.053	14.0861	14.781
87/06/04	0.2703	0.334	30.2714	26.466	4.3826	6.493	2.0735	1.596	15.8147	16.016
87/06/05	0.2924	0.294	36.0178	25.272	4.0213	5.536	1.7707	3.242	13.8416	18.263
87/06/06	0.3322	0.255	31.362	23.335	4.5741	5.835	2.9447	3.69	14.999	13.276
87/06/07	0.3329	0.315	30.219	21.201	4.6657	5.780	2.8446	2.051	15.149	16.085
平均誤差	16.87%		36.334%		26.436%		29.722%		19.954%	

表 2 花蓮空氣品質預測

污染物 / 日期	一氧化碳(ppm)		臭氧(ppb)		二氧化氮(ppb)		二氧化硫(ppb)		懸浮微粒(ug/m^3)	
	預測值	真實值	預測值	真實值	預測值	真實值	預測值	真實值	預測值	真實值
87/06/01	0.91806	0.581	25.2333	15.942	12.2589	8.439	0.3088	0.255	39.9026	30.304
87/06/02	0.94006	0.590	27.7991	19.707	11.7901	11.778	0.5586	0.691	38.7219	35.725
87/06/03	1.00798	0.745	20.2868	23.272	13.2254	22.642	0.4101	0.521	34.3654	34.868
87/06/04	1.05019	0.781	30.1849	26.614	14.0223	16.096	0.4295	0.344	35.1652	26.824
87/06/05	0.93864	0.727	26.2205	23.685	16.2072	20.887	0.264	0.350	29.2661	36.334
87/06/06	0.92258	0.618	30.3855	41.202	17.7412	18.372	0.3324	0.258	38.8518	35.706
87/06/07	1.04834	0.878	27.3427	30.602	18.4371	23.577	0.2012	0.300	31.7126	40.653
平均誤差	39.625%		24.742%		21.068%		24.675%		17.551%	

利用此模式產生87年的預測資料，並與87年真正資料比對並繪製成圖。如圖以下爲台北地區(一氧化碳、二氧化氮)，花蓮地區(一氧化碳、臭氧)比較預測圖的圖例:



圖八 預測結果比較圖

我們成功地將原本於作業系統階層不相關的資料資源(資料庫)，經過領域分析與專家知識，透過智慧資料整合的方式，提供知識相關的資訊關連，建構出綜合性的資料倉儲，然後再藉由重疊的技術欲以幫助資料彙整，並進而以類神

經網路的技術，發展出相關之預測雛型系統。事實上類似這種“看似不相干”的資料資源，在我們許多決策支援應用當中，才是扮演決定性的角色。然而由於其範圍廣泛，複雜度太高，常被忽略以致於所建構之資料倉儲無法真正支援決策分析，再加上使用者種種不穩定的知識與需求，才會導致資料倉儲應用不彰的結果。

因此如何有效率地發掘出真正所需的資料，是我們研究努力的課題，若能明確的由專家知識智慧的彙整資料來源並加強考慮不同的資料來源其間之相關性，再利用事實綱要重疊的技術，在衡量各種的情形下，以延伸出更多相關有用的資訊訊息，這樣才能幫助更多相關資訊的取得，以有效的增加資料倉儲的分析能力，提高建構出的資料倉儲系統之品質。

第五章、結論

在本篇論文中，特別提出一種不同的資料倉儲建構的方式，針對資料來源與資料彙整二個方面的重要性，去探討來自於不同的資料來源，是必須加強明確的考量資料彙整的完整性。文中得知智慧式的專家知識及經驗的資料整合型態，可進而關連資料間之相關連性，並能從事實綱要重疊的技術去得到有用的資訊結合，最後由實例模擬結果得到驗證。例子證明在二個不同"看似不相干"的資料庫中，若實際存有高度關係性的話，且其中內含的關係是有資訊關連的，則能利用這樣的關係性從中去挖掘出有用的資訊。

而我們從早先資料倉儲在於資料資源的重要性去考量，加強整合不同異質資料來源的關係性，利用專家知識與經驗幫助不同資料來源的彙整，由類神經網路的探勘技術建立一個有效的摸擬預測系統，並從使用者需求與系統維護者角度就資料倉儲系統建構過程中，事實綱要重疊的目的、存在的必要性與其建構之優缺點，作深入分析探討，然後再分別從完全相容的事實綱要和非完全相容的事實綱要二種情況去分析其使用的時機，提出重疊後的資訊是優於原先沒有考慮重疊的，並且利用這樣子的資訊產生進而去推論原先初步欄位建構時是否正確，期望藉此能幫助強固資料倉儲系統的完整性，以增進使用者資訊存取的效益，減少系統負擔，進而提供決策者更有效益的決策支援。

參考文獻

[1] 謝建成，史孟祥，李修宇，謝馥安，“非關聯性資料庫之資料倉儲建立—以天氣氣象與空氣污染爲例”，第六屆資訊管理研究暨實務研討會，2000。

[2] 王君賢，“臺灣地區氣流軌跡之氣候統計”，國立中央大學大氣物理研究所碩士論文，1991。

[3] 陳幼麟，“臺灣區域氣候之研究”，國立臺灣大學大氣科氣系碩士論文，1992。

[4] 中央氣象局網站 http://www.cwb.gov.tw/index.html。

[5] 張育銜，“南高屏地區高臭氧事件日之研究”，國立中興大學環境工程學系碩士論文，1998。

[6] 林煜棋，“鋪面道路車行揚塵特性與排放係數之建立”，國立中興大學環境工程學系碩士論文，1998。

[7] 胡婷堯，“桃園地區每日最大臭氧濃度之預測”，國立中央大學大氣物理研究所碩士論文，1995。

[8] 鄭佳芳，“南高屏地區伴隨臭氧污染事件之環流場特徵分析”，國立中央大學大氣物理研究所碩士論文，1998。

[9] 吳明進，“臺灣北部地區降雨氣候之研究”，國立臺灣大學大氣科學研究所碩士論文，1994。

[10] 蔡素芬，“臺灣地區道路塵粒特性之研究”，國立清華大學原子科學系碩士論文，1998。

[11] 謝國發，“台中盆地邊界層大氣結構對臭氣濃度之探討”，東海大學環境科學系碩士論文，1998。

[12] 林能暉，彭啓明，陳進煌，陳靖沅，“台灣酸雨之研究：源與受體關係”，第六屆全國大氣科學學術研討會論文集。

[13] 周經文，“中部地區空氣品質監測系統代表性之探討”，東海大學環境科學系碩士論文，1998。

[14] 吳明進，“臺灣區域之數值模擬”，第六屆全國大氣科學學術研討會論文集。

[15] 行政院環境保護署 http://www.epa.gov.tw/。

[16] 周昌宏，許晃雄，陳正達，柯文雄，鄒治華，“台灣環境變遷與全球氣候變遷衝擊之評析---氣候”，行政院國家科學委員會專題研究計畫成果報告。

[17] 李永安，吳思儀，“長期氣候變遷的主要時空演化結構”，第六屆全國大氣科學學術研討會論文集。

[18] Shanmugasundaram, J., Fayyad, U., Bradley, P. S., Compressed Data Cubes for OLAP Aggregate Query Approximation on Continuous Dimensions. KDD, 1999.

[19] Extraction, C. S., Transformation for The Data Warehouse. SIGMOD, 1995.

[20] Agarwal, S. R., Agarwal, P. M., Deshpande, A., Gupta, J. F., Ramakrishnan, N. R., Sarawagi, S., On the Computation of Multidimensional Aggregates, in Proc. 22nd Int, VLDB, Conf, Mumbai (Bombay), 1996.

[21] Sarawagi, S., Indexing OLAP data. Bulletin of technical Committee on data Engineering, 20-1, (1997).

[22] Zhuge, Y., Molina, G. H., Wiener, J. L., The Strobe Algorithms for Multi-Source Warehouse Consistency, in Proc. Conference on Parallel and Distributed Information Systems, Miami Beach, FL (1996).

[23] Choen, S., Nutt, W., Serbrenik, S., Algorithms for Rewriting Aggregate Queries Using Views, in Proc. Int, Workshop on Design and Management of Data Warehouses, Heidelberg, Germany, 1999.

[24] Yan, W. P., Larson, P., Eager and Lazy Aggregation, in Proc. 21st Int, Conf, On Very Large Data Base, Zurich Switzerland, pp, 345-357, 1995.

[25] Amy J. L., Andreas K., Anisoara N., Rundensteiner, E. A., Data Warehouse Evolution: Trade-offs between Quality and Cost. Technical Report WPI-CS-TR-98-2, WPI, 1998.

[26] Amy J. L., Andreas K., Anisoara N., Rundensteiner, E. A., Data Warehouse Evolution: Trade-offs between Quality and Cost of Query Rewritings. Institute of

Electrical and Electronics Engineers, Inc, 1998.

[27] Wiener, J. L., Gupta, H., Labio, W. J., Zhuge, Y., Molina, C. H., Widom, J., The WHIPS Prototype for Data Warehouse Creation and Maintenance. Institute of Electrical and Electronics Engineers, Inc, 1997.

[28] Martin, G., Jessie, B. K., Chris, H., The Challenge of Visualizing Multiple Overlapping Classification Hierarchies. Institute of Electrical and Electronics Engineers, Inc, 1998.

[29] Golfarelli, M., Rizzi, S., A Methodological Framework for Data Warehouse Design. In Proceedings ACM Fist International Workshop on Data Warehousing and OLAP (DOLAP), Washington, 1998.

[30] Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., Zanasi, A., Discovering Data Mining from Concept to Implementation. Prentice Hall, p12, 1997.

[31] Agrawal, T., Imielincki, T., Swami, A., Mining Association Rules Between Sets of Items in Large Database. ACM, 1993.

[32] Vellido, A. P., Lisboa, J. G., Vaughan, J., Neural Networks In Business: A Survey of Applications (1992-1998). Expert Systems with Applications, 17, p51-70, 1999.

[33] Bo, K. W., Thomas A., Bodnovich, Y. S., Neural Network Applications In Business: A Review and Analysis of The Literature (1988-95). Decision Support Systems, 19, p301-320, 1997.

[34] Inmon, W. H., The Data Warehouse and Data Mining. Communications of The ACM, Vol. 39, No. 11, November 1996.

[35] Agrawal, R., Gupta, A., Sarawagi, S., Modeling Multidimensional Batabase. IBM Research Report, 1995.

[36] Gyssens, M., Lakshmanan, V. S., A Foundation for Multi-dimensional Database, in Proc. 23rd, VLDB, p106-115, (Athens, Greece 1997).

[37] Cardenas, A. F., Analysis and Performance of Inverted Database Structures.

Comm, ACM, 18, 5, p253-263, 1975.

[38] Harinarayan, V., Rajaraman, A., Ulman, J., Implementing Data Cubes Efficiently, in Proc. of ACM Sigmod Conf, (Montreal, Canada, 1996).

[39] Gupa, H., Harinarayan, V., Rajaraman, A., Index Selection for OLAP, in Proc. Int, Conf, Data Engineering, (Binghamton, UK, 1997).

[40] Johnson, T., Shasha, D., Hierarchically Split Cube Forests for Decision Support: Description and Tuned Design. Bulletin of Technical Committee on Data Engineering, 20, 1, 1997.

[41] Mcguff, F., Data Modeling for Data Warehouse. http://menbers.aol.com/fmcguff/dwmodel/dwmodel.htm, 1996.

[42] Caibbo, L., Torlone, R., Un quadro metodologico per la costruzione e l'uso di data warehouse, in Proc. Sesto Convegno nazionale sui Sistemi Evoluti per Basi di dati, 1, p123-140, (Ancona, Italy, 1998).

[43] Golfarelli, M., Rizzi, S., Designing The Data Warehouse: Key Steps and Crucial Issues. Journal of Computer Science and Information Management, vol. 2, n. 3, 1999.

[44] Golfarelli, M., Maio, D., Rizzi, S., Conceptual Design of Data Warehouse from E/R Schemes, in proc. HICSS-31, VII, p334-343, (Kona, Haeaii, 1998).

[45] Golfarelli, M., Maio, D., Rizzi, S., The Dimensional Fat Model: A Conceptual Model for Data Warehouses. Invited Paper, International journal of cooperative information systems, vol. 7, n. 2&3, 1998.

[46] Golfarelli, M., Maio, D., Rizzi, S., Vertical Fragmentation of Views In Relational Data Warehouse. Proceedings of Settimo Convegno Nazionale su Sistemi Evoluti Per Basi Di Dati, Como, p19-33, (Italy, 1999).

[47] Srivastava, J., Chen, P., Warehouse Creation-a Potential Roadblock to Data Warehousing, IEEE Trans. on Knowledge and Data Engineering, Vol. 11, No. 1, 1999.

[48] Widom, J., Research Problems In Data Warehousing, in Proc. 4th Int, Conf on Information and Knowledge Management, 1995.

[49] Inmon, W.H., Hackathorn, R. D., Using The Data Warehouse. John Wiley and Sons, 1994.

[50] Inmon, W. H., The Data Warehouse and Data Mining. Communication of ACM, Vol. 39, No. 11, 1996.

[51]Nicolas, P., Yves, B., Rafik, T., Lotfi, L., Efficient Mining of Association Rules Using Closed Internet Lattices. Information System, Vol. 24, No. 1, p25-46, 1999.