建構資料倉儲方法論之研究---探究資料來源與事實網要重疊技術

謝建成
佛光大學 資訊學所
jcshieh@mail.fgu.edu.tw

鍾志明 史孟祥
南華大學 資管所
century_boy@mail2000.com.tw

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，不同資料來源(data sources)彙整所產生的重要性及事實網要重疊(fact schemes overlap)技術對資料結合及系統效益所帶來的種種影響。

關鍵字:資料倉儲、資料探勘、資料來源、資料彙整、事實網要重疊

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]均相關之論述，而資料倉儲系統在建構的過程中，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，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，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]，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]，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。

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，分別從實體-關係(E/R)架構和邏輯關連性

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]，因爲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]。我們將針對在概念化設計當中，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，及重疊所存在的必要性，並提出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，以提升建構倉儲的品質。第二章介紹系統建構中不同資料結合所考慮資料的彙整並從事實綱要重疊技術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，第三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，在第四章做最後結論。

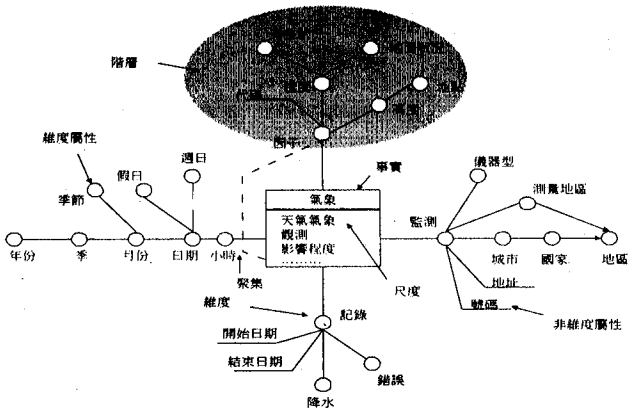
第二章、 資料彙整之重要性

本章節我們就事實綱要重疊技術來說明資料彙整的重要性及分析事實綱要做重疊時有做與沒有做分別會造成怎樣的影響，提出重疊後所得到的資訊是優於原先沒有做重疊考量的，並以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

首先定義 $g=(V,E)$ 是一個經指定、非週期性及弱連結的圖形，g 是一個類似樹狀圖，其根在 $v_0 \in V$ ， $path_{ij}(g) \subseteq g$ 是應有的路徑由點 v_i 開始到點 v_j 結束，且 $(v_i) \subset g$ 根在 $v_i \neq v_0$。其中的事實綱要爲 f=(M, A, N, R, O, S)。其屬性定義爲[29]:

- 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 Ω 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 Ω 而成的。



圖一 天氣氣象的事實綱要

第二節、事實綱要重疊

我們可知資料倉儲所得訊息是未知的，也不一定是有用的。目前資料倉儲的事實綱要重疊是系統根據資料屬性主動化做結合的，若能有效考慮爲何要建構重疊，不但能明確的了解原有資料所蘊涵的資訊，且能夠更正確知道資料倉儲建構系統的最佳狀況，並藉由此重疊過程來延伸出更多相關之資訊。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，並深入考慮這樣因素存在性，看是否能因此幫助系統有效地刪減其屬性與欄位，進而能降低系統資料儲存的容量。事實綱要在做重疊時均有其不同的考量因素，不管來至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，我們知道不同的事實均存在著不同有用的訊息，不但可藉由著做重疊的方式得到其完整資訊，並可幫助強壯資料倉儲架構的完整性。因此在資料倉儲建構時正確做事實綱要重疊過程是極具重要性的，系統會由不同事實綱要做重疊取得的正確訊息，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，來避免不當的多餘資料結合，且可由當中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關連性，幫助減少多餘的屬性及使用者所不需的資料欄位，或是更進一步增加不足的資料屬性。

第三節、完全相容的事實綱要

有二個相容的事實綱要具有共同相關連屬性，分別爲 $f'=(M',A',N',R',O',S')$ 和 $f''=(M'',A'',N'',R'',O'',S'')$，是爲完全兼容的情形，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。$f'\otimes f''=(M,A,N,R,O,S)$：

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

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

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

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

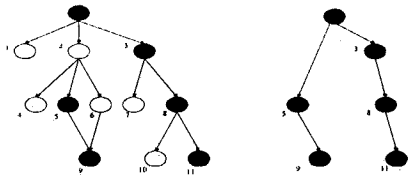
$$R = \{(a_i, a_j) \mid (a_i, a_j) \in \mathrm{cnt}\,(\mathrm{qt}(f'), A)\}$$

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

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

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

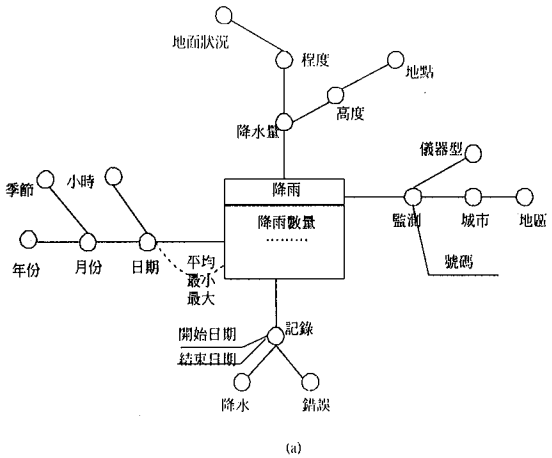
重疊後會針對其相關所共有屬性做結合，從相同的維度找出其相等的屬性來產生新的事實綱要，並因結合的過程而產生新的資訊延伸。



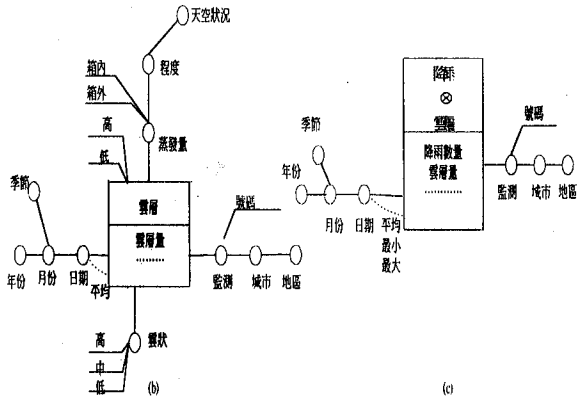
(a) (b)

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

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

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 的範圍中每一個尺度屬性，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。

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。

- 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，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。

第四節、非完全相容的事實綱要

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。定義二個做重疊 f′和f″ 的綱要，$f' \otimes f'' = (M, A, N, R, O, S)$：

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

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

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

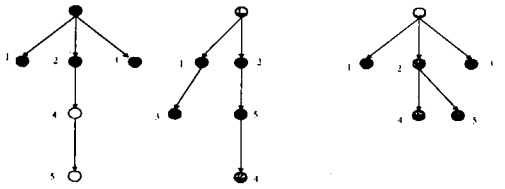
$$N = N' \cap N''$$

$$R = \{(a_i, a_j) \mid \exists p_{ij}(\mathrm{cnt}(qt(f'), A)) \wedge \exists p_{ij}(\mathrm{cnt}(qt(f''), A)) \wedge \forall a_w \neq a_j \mid (\exists p_{wj}(\mathrm{cnt}(qt(f'), A)) \wedge \exists p_{wj}(\mathrm{cnt}(qt(f''), A)))(p_{ij}(\mathrm{cnt}(qt(f'), A)) \subset p_{wj}(\mathrm{cnt}(qt(f'), A)) \wedge p_{ij}(\mathrm{cnt}(qt(f''), A)) \subset p_{wj}(\mathrm{cnt}(qt(f''), A)))\}$$

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

$$S = \{(m_j, d_i, < op >) \mid d_i \in \mathrm{Dim}(f' \otimes f'') \wedge (\exists (m_j, d_k, < op >) \in S' \wedge d_i \in \mathrm{sub}(qt(f'), d_k)) \vee (\exists (m_j, d_k, < op >) \in S'' \wedge d_i \in \mathrm{sub}(qt(f''), d_k))\}$$

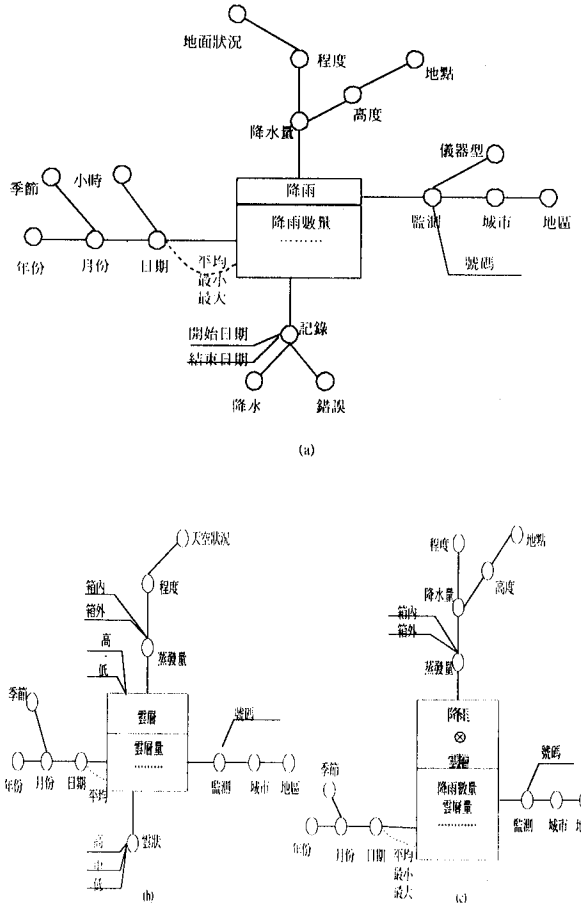
當 f′和f″ 同爲非完全相容的事實綱要時，重疊的過程則會針對其新的 f 綱要所需的相關性屬性去做結合，並找出其必須所共具有的屬性來產生新的事實綱要，下圖則呈現出二個非完全相容的事實綱要做重疊的過程。



(a) (b)

圖四 二個非完全相容事實綱要 (a)和(b)

圖五在於天氣氣象的降雨及雲層的影響程度上，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

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。

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

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。

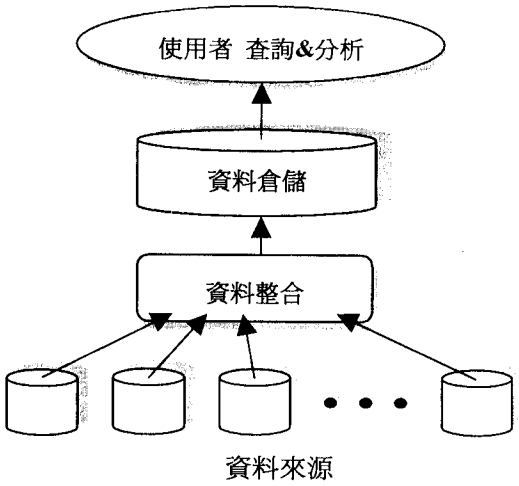
第三章、資料來源之重要性

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。

第一節、資料倉儲的資料來源

圖六爲描述資料倉儲的基本架構，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，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。

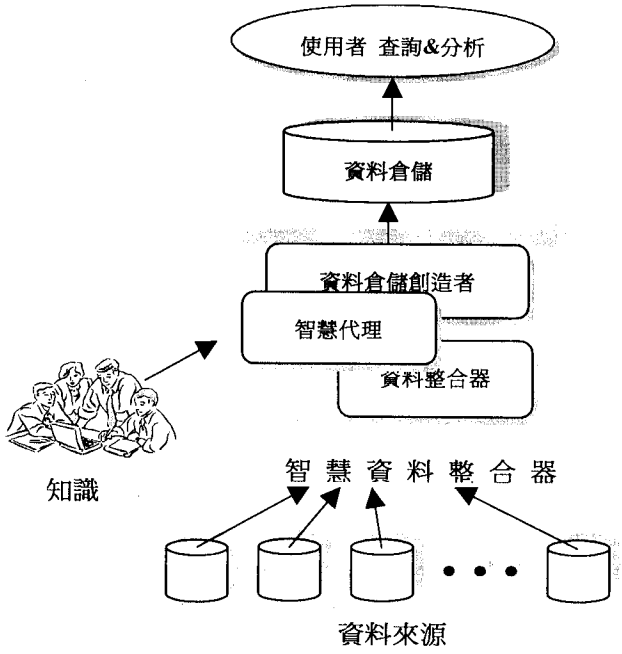


圖六 描述資料倉儲的基本架構

而資料的來源或許是來自於操作系統的層面上，我們說它是不明確或具內含關係性的資源，這個部份加強於考慮整合來源是必須的。

第二節、智慧的資料整合

圖七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，故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。

第三節、實例驗證

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]，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。

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:

- **天氣氣象的因素**:由表 1 可知天氣氣象原始資料的欄位[15]: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°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。

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。

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(年、月、日)、平均氣壓 0.1hpa、平均氣溫 0.1m/s、相對濕度水氣壓 0.1hp、平均風 0.1m/s、降水量 0.1m、降水紀錄。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。

- **空氣污染的因素**:藉由[15]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(年、月、日)、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)。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。

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，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、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]。

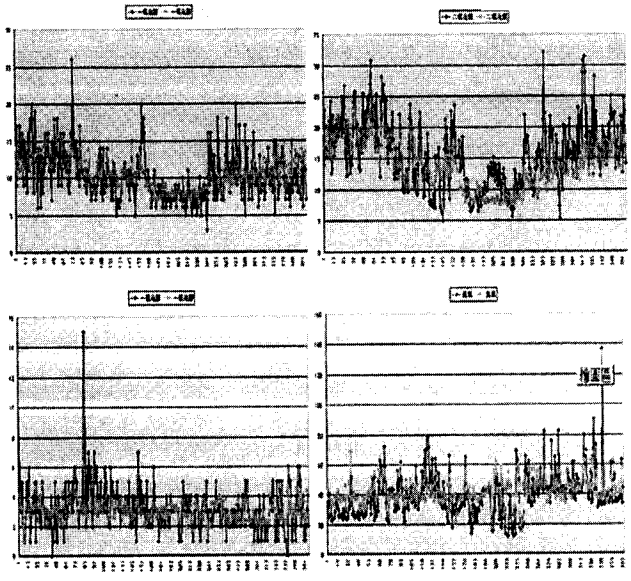
8、輸入變數分類的選擇則是使用基因演算法(Genetic Algorithms)爲期初步訓練的方法，再從資料分析及轉換過程中去建立全部欄位以找出較佳的分類及合理的集合，以做分堆的資料處理。而使用基因演算法時，由於不同的母體將產生不同的結果，所以當使用兩個不同資料去建立相同模式時，會有不同的變數選擇，選擇 exhaustive variable selection 以建立複雜的資料模式，使其爲適當的變數，並在輸入過程後再加入類神經網路進分析。

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。

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表 2、3 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)。圖下爲台北地區(一氧化碳、二氧化氮)，花蓮地區(一氧化碳、臭氧) 87 年的預測資料比較預測圖:



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，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，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。

第四章、結論

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探究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.

表 1 天氣氣象原始資料欄位

測站號碼	時間				氣壓 0.1 hpa	氣溫		濕度		露點	雲狀		
	年	月	日	時		乾 0.1°C	濕 0.1°C	絕 0.1°C	相 0.1°C	0.1°C	高	中	低

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

天空及視障				地中溫度 0.1 °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		

表 2、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/m³)	
	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

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
87/06/06	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	5.78	2.8446	2.051	15.149	16.085
平均誤差	16.87%		36.334%		26.436%		29.722%		19.954%	

表 3、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%	