

資料探勘

6.1 操作型資料庫

我們之前就已經知道可以從一或多個操作型資料庫來擷取、收集資料進行分析。而操作型（或稱交易的）資料庫通常設計來用在快速、有效的處理獨立的交易上；因此，這種以交易資料為基礎的互動即被視為是**線上交易處理**（on-line transactional processing），簡稱**OLTP**。

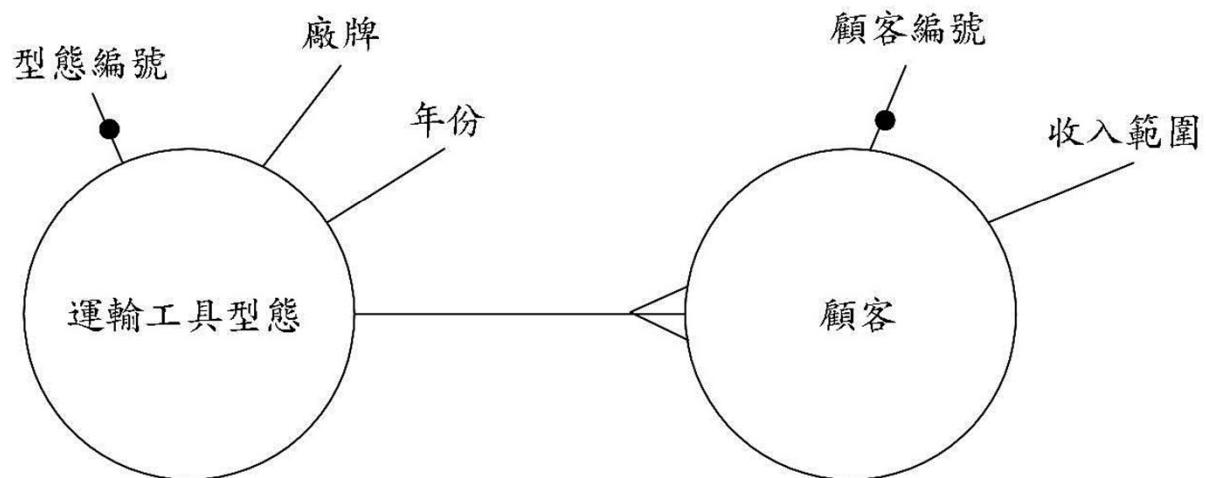
資料探勘

資料模組化與正規化

資料模組（data model）列舉出了資料的結構，以及可以使利用某種系統的結構化資料被拿來獨立使用。實體關聯圖（entity relationship diagram, ERD）是經常用來表示資料模組化的方法，它是由含有 m 個實體的資料、實體關係所組成。實體（entity）的概念就如同是一個類別的人、物或地點。一個實體可以包括一或多個屬性。而這一或多個屬性的結合關鍵，表示出它可以唯一鑑別出每一屬性的差異。

資料探勘

圖 6.1 □ 一個簡單的實體關聯圖



資料探勘

表 6.1a □ 交通工具型態的關聯表格

型態編號	廠牌	年份
4371	雪佛蘭	1995
6940	凱迪拉克	2000
4595	雪佛蘭	2001
2390	凱迪拉克	1997

資料探勘

表 6.1b □ 顧客的關聯表格

顧客編號	收入範圍 (\$)	型態編號
0001	70–90K	2390
0002	30–50K	4371
0003	70–90K	6940
0004	30–50K	4595
0005	70–90K	2390

資料探勘

關聯模組

表 6.2 □ 結合表 6.1a 與表 6.1b

顧客編號	收入範圍	型態編號	廠牌	年份
0001	70-90K	2390	凱迪拉克	1997
0002	30-50K	4371	雪佛蘭	1995
0003	70-90K	6940	凱迪拉克	2000
0004	30-50K	4595	雪佛蘭	2001
0005	70-90K	2390	凱迪拉克	1997

資料探勘

6.2 資料倉儲的設計

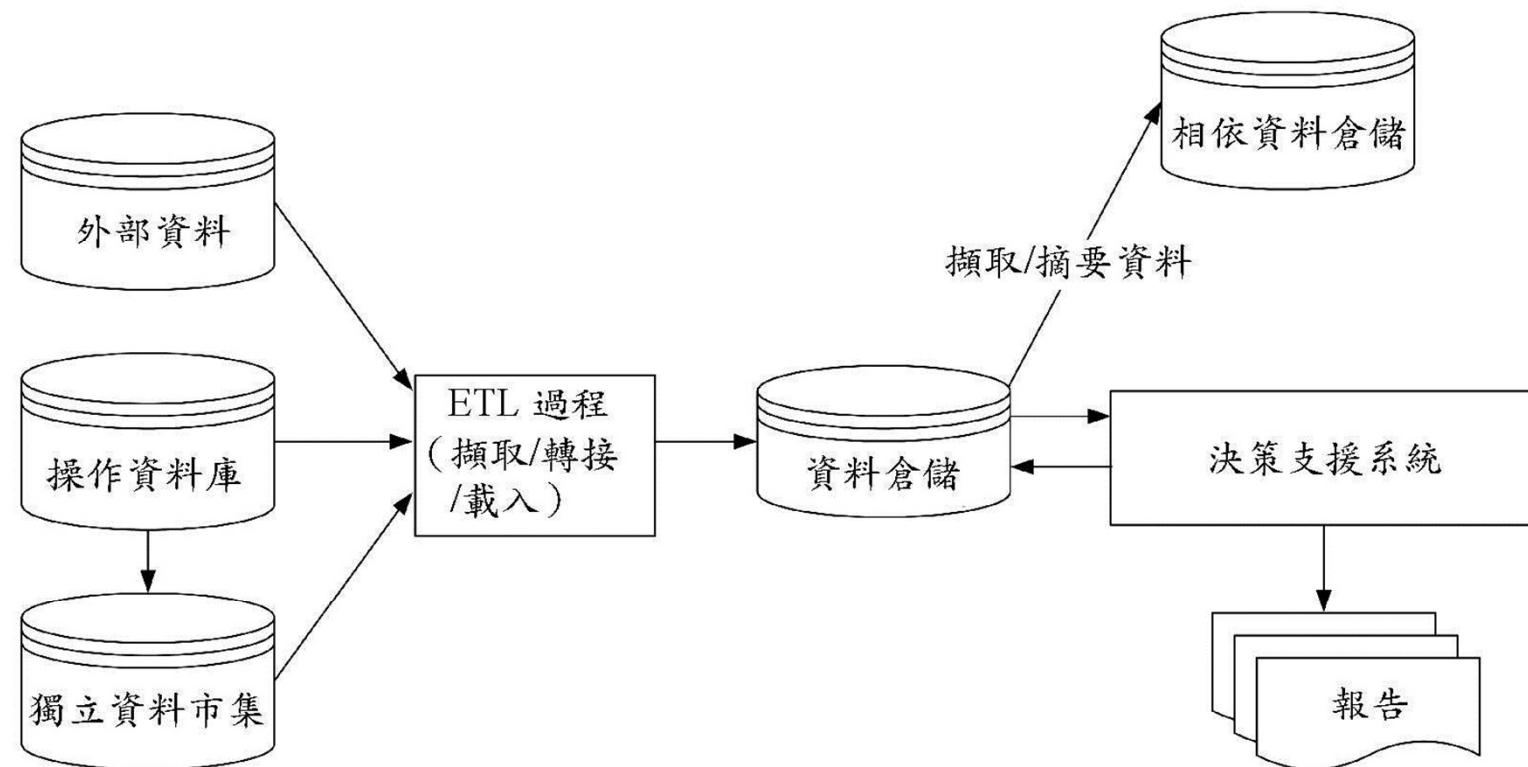
輸入資料到倉儲中

在圖 6.2 中可以看到三種主要的資料來源會將資料置入倉儲之中。一個**獨立資料市集** (independent data mart) 是一種類似倉儲的儲存方式，但它只侷限於一個單一的主題。獨立資料市集會把操作資料視為是一種外部的資料來源。只要組織的其他部分有需要使用到這些資料，則寫在資料市集中的資料就可以被載入到資料倉儲之中。

倉儲中也存入了其他的資料型態，稱之為**解釋資料** (metadata)。解釋資料在技術上被定義為關於資料的資料。它被建立的目的是用來更了解那些自然地包含進倉儲中的資料。

資料探勘

圖 6.2 □ 一個資料倉儲的處理模組



資料探勘

將資料倉儲結構化：星狀圖

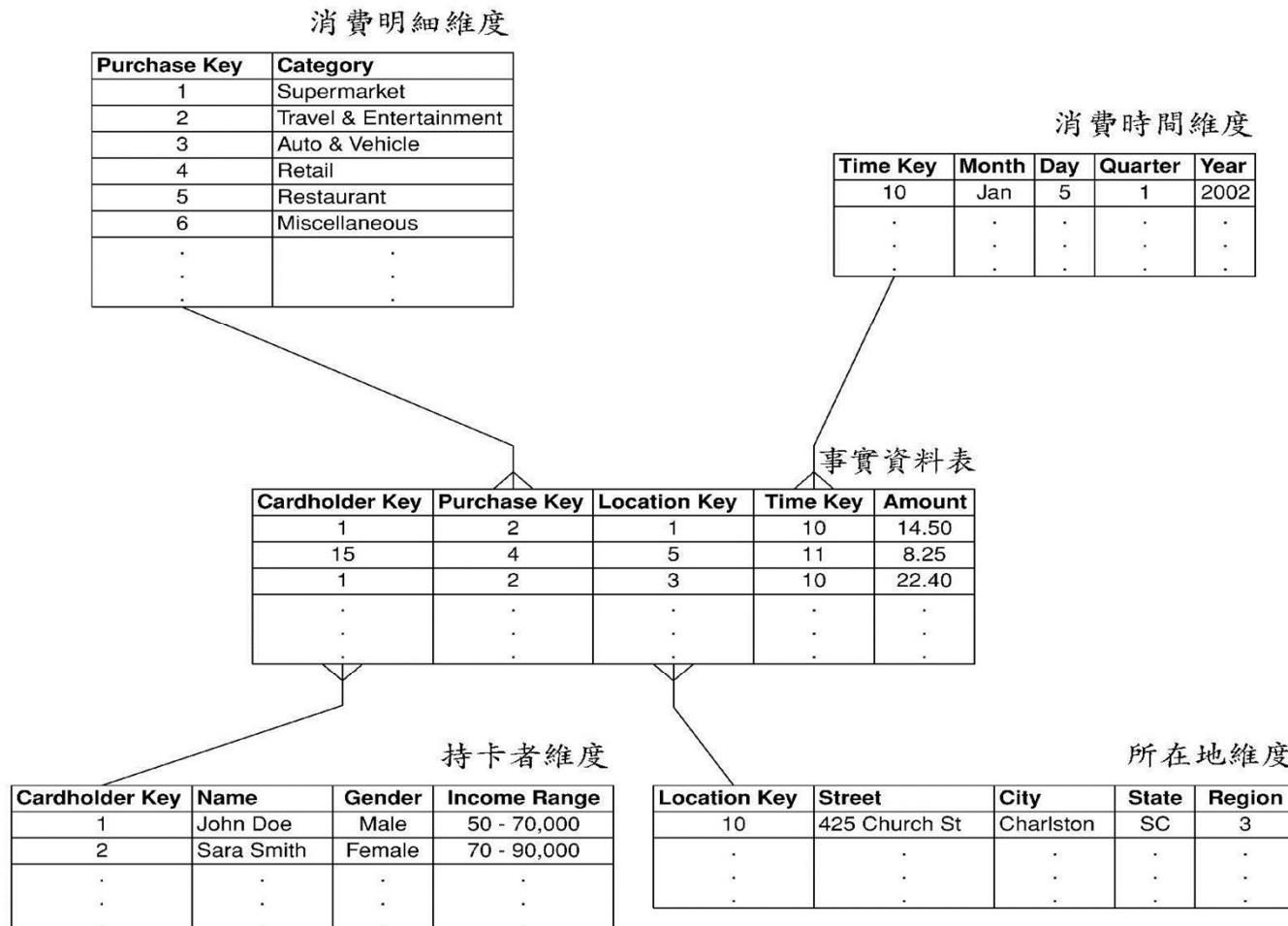
圖 6.3 概述了從圖 5.3 的 Acme 信用卡資料庫所建立的星狀圖。這個星狀圖的主題就是信用卡的購買記錄。因此在星狀圖的中央就是**事實資料表**（fact table）。事實資料表定義了多元空間中的維度。在圖 6.3 中的事實資料表就可以看到四個維度：持卡者、購買明細、地點，與時間。每一筆在事實資料表中的記錄包含了兩種型態的資料——維度鍵與事實。

資料探勘

事實資料表中的每個維度皆可能會有一或多個**維度表** (dimension table) ，維度表組成了星狀圖中的點，而其中心就是事實資料表——因此才稱之為星狀。維度表包含了每個維度的特定資料，而維度表與事實資料表之間的關係是屬於一對多的，因此維度表會顯著的小於中央的事實資料表。最後，星狀圖的維度也經常提到**緩慢改變的維度** (slowly changing dimension) 的問題。這是因為這個維度表屬於先前提過的那種並不時常改變型態的資訊。現在仔細的去看看維度表以及這張由 Acme 信用卡資料庫所建立的星狀圖。

資料探勘

圖 6.3 □ 信用卡交易的星狀圖



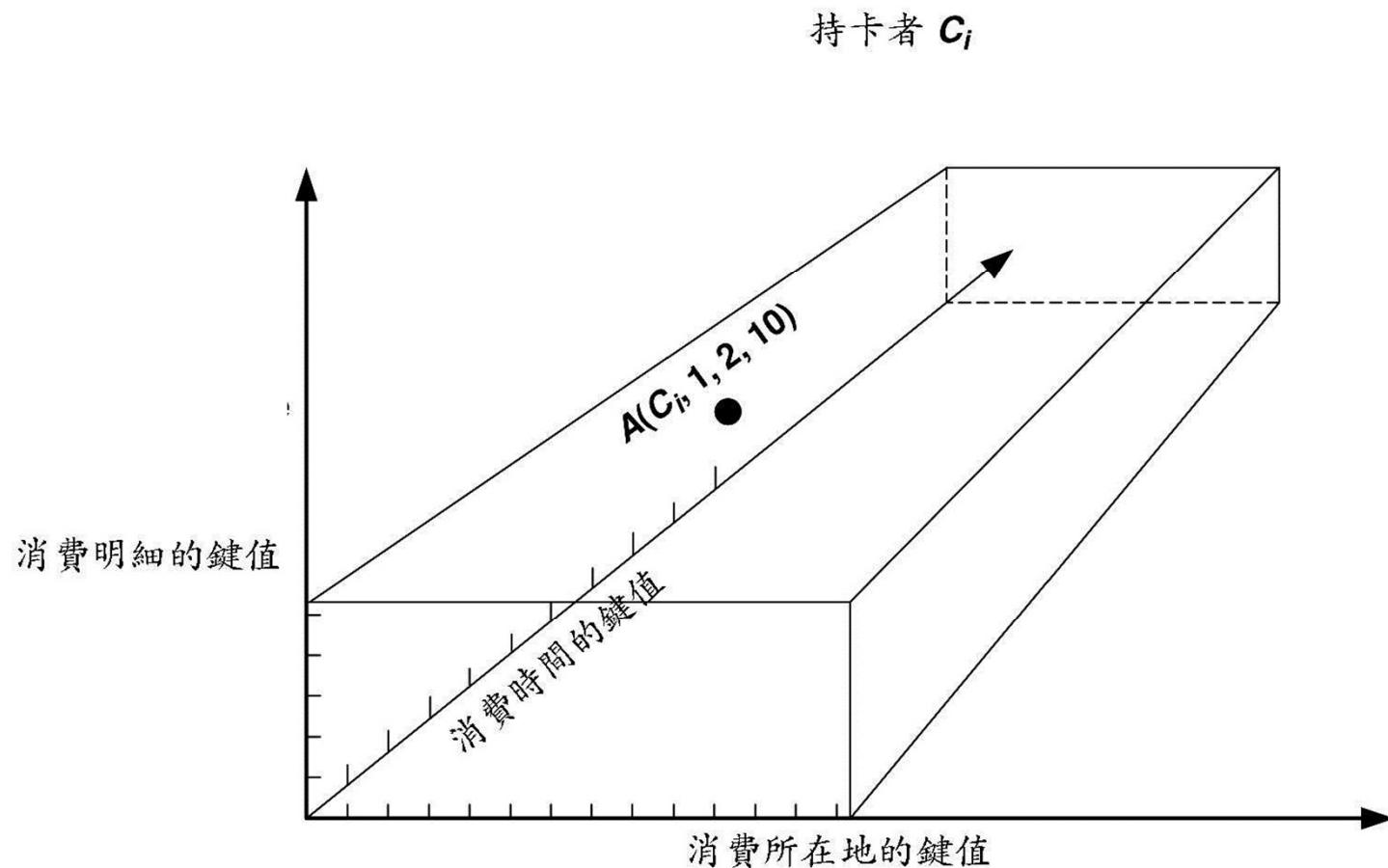
資料探勘

星狀圖的多元性

圖 6.3 的事實資料表定義了四個維度空間，圖 6.4 把其中三個維度（消費明細、地點與時間）對應到一個三度空間的座標系統中。這個三度空間的結構（如圖 6.4 所示）存在於星狀圖中每一個持卡者（第四個維度）。

資料探勘

圖 6.4 □ 圖 6.3 的事實資料表中之各個維度



資料探勘

其他的關聯圖

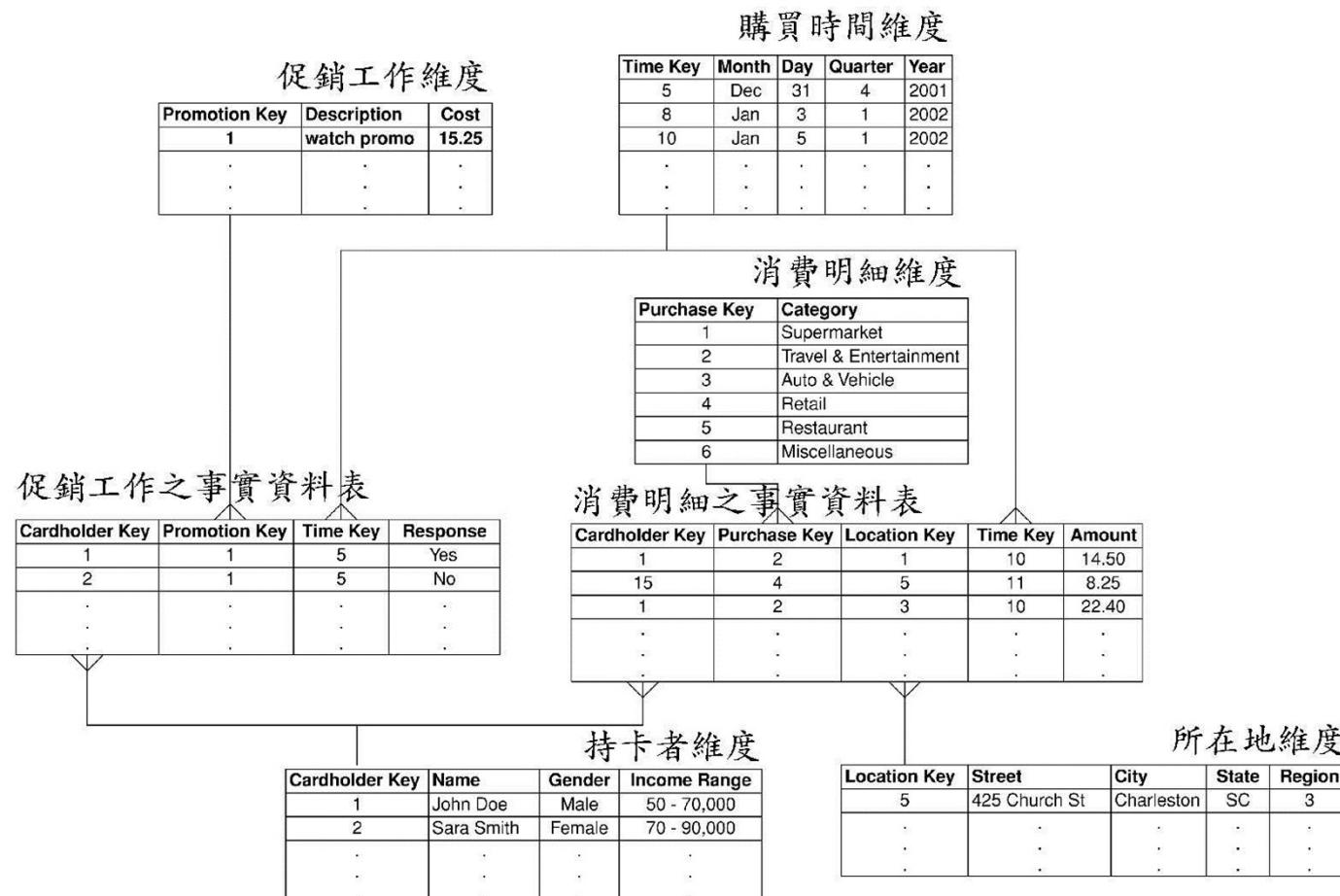
當一些維度表直接關聯到事實資料表時，一種由星狀圖所變形而來，稱之為**雪花圖**（snowflake schema）的概念就會更進一步的切割出來。它允許維度表可以被正規化，這點就代表著總儲存空間的需求降低了。

當模組中需要不只一個的中央事實資料表時，我們就需要另一種星狀圖的變形，稱之為**星座圖**（constellation schema）。圖 6.5 中顯示了一個同時支援信用卡交易與信用卡促銷資料的星座圖結構。

在圖 6.5 中可以看到每一筆促銷事實資料表的記錄把持卡者的鍵值與促銷活動的鍵值、時間鍵值與回應與否寫在一起。

資料探勘

圖 6.5 □ 信用卡交易與促銷的星座圖



資料探勘

決策支援：分析資料倉儲

1. **報告資料**。報告被視為是決策支援的最低階層。
2. **分析資料**。資料的分析通常以某種多元資料分析工具的形式所完成。
3. **發現知識**。發現知識的動作可以很容易的從資料探勘中發生。

除了寫入提供決策支援的資料外，倉儲也是一種資料商店，用來建立小一號的部門倉儲，稱之為**相依資料市集**（dependent data mart）。一個相依資料市集的特色就是單一主題，並且設計來用在特殊的目的之上。另外，資料市集就像是表現比資料倉儲更高階的摘要資訊。

資料探勘

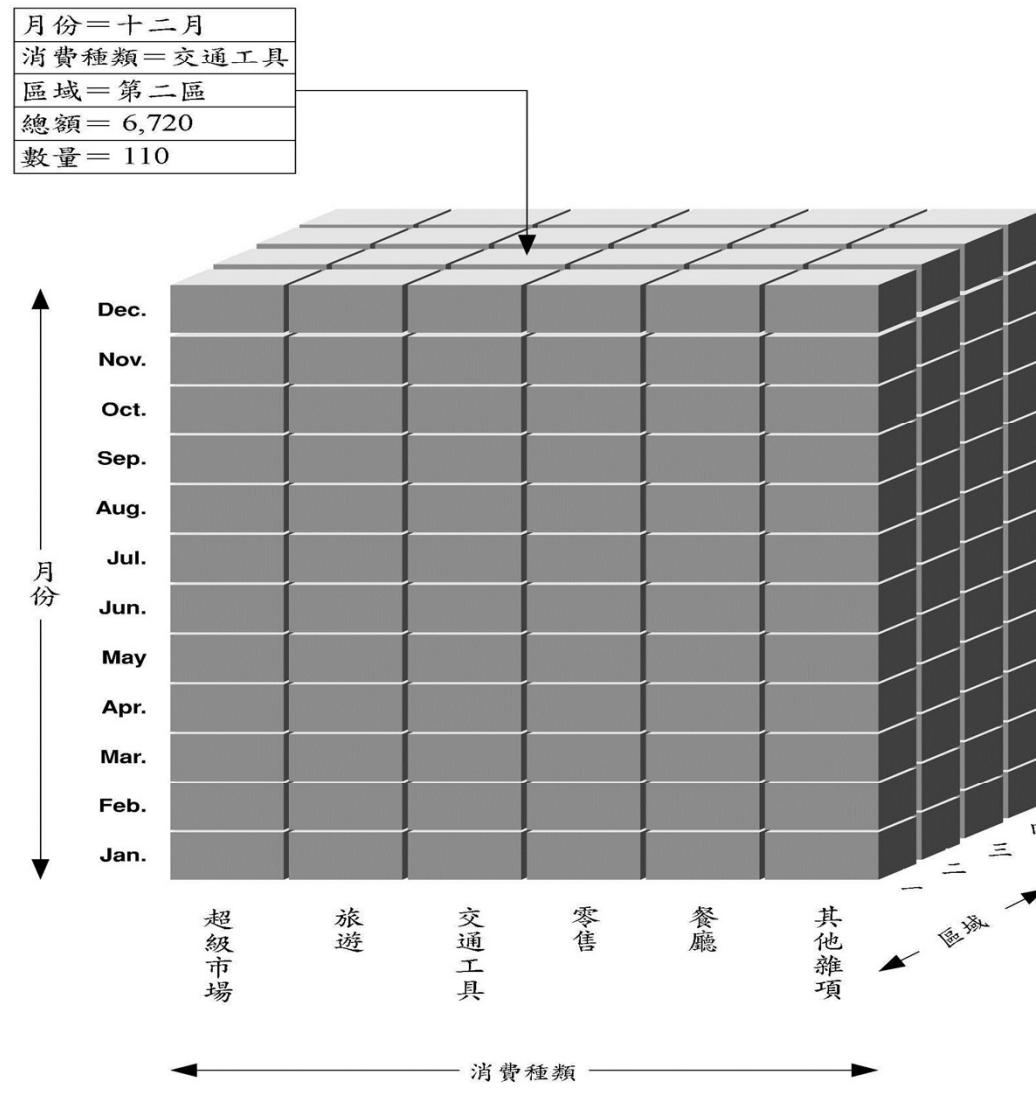
6.3 線上分析處理

線上分析處理（OLAP）是種基於查詢所發展出的方法，且支援在一個多元環境中進行資料分析。OLAP 是一種富有價值的工具，它的價值在於它可以用來證明或反駁人類所建立的假設，以及執行手動的資料探勘。

一個 OLAP 引擎在邏輯上把多元資料架構在一種方塊的型態上，如圖 6.6 所示。這個方塊有三個維度：消費種類、月份，與區域。這些維度是從圖 6.3 的星狀圖中擷取的屬性子集合。圖 6.6 的三維度方塊會比較容易一看就懂，不過要把超過三個維度的資料方塊視覺化也是相當困難的一件事。

資料探勘

圖 6.6 □ 信用卡交易的多維度方塊



資料探勘

OLAP：一個實例

每一個 OLAP 的屬性方塊可以有一個以上的關聯**概念式階層**（concept hierarchy）。一個概念式階層定義了一種可以從不同階層的細節中看到的屬性映對關係。圖 6.7 顯示了所在地屬性的概念式階層。

資料探勘

圖 6.7 □ 所在地的概念式階層

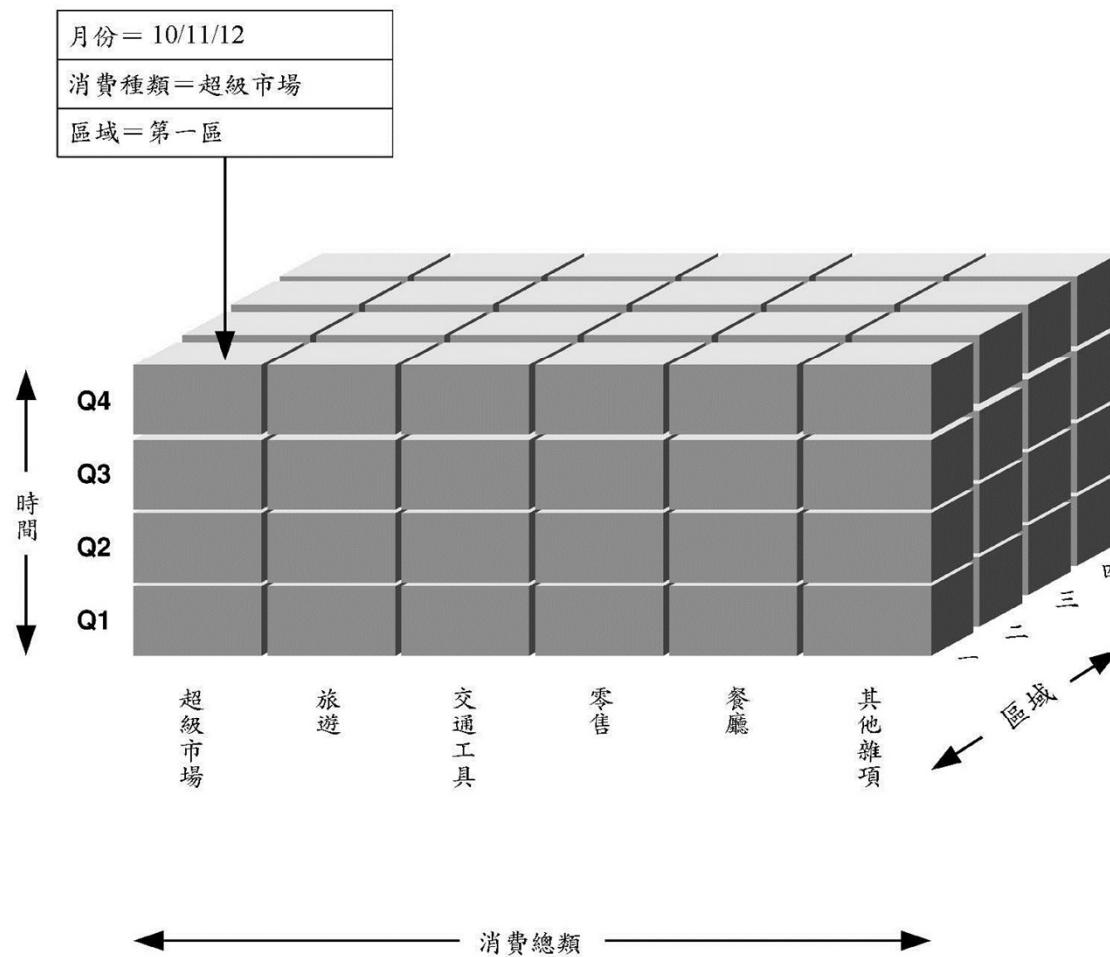


資料探勘

1. **切割** (slice) 的動作是從 OLAP 方塊中的一個維度來選取資料。
2. 在兩個以上的維度中**隨機選取** (dice) 一個，從原來的方塊中擷取出一個子方塊。
3. **滾上** (roll-up)，或稱聚集，用來結合定義在一個方塊之中的一或多個維度的資料集。一個滾上的形式是把概念式階層中的一個維度組織成更高階層的方式來儲存。
4. **鑽下** (drill-down)，與滾上相反的，在一些階層中加入更多細部的檢查資料。
5. **旋轉** (rotation)，或稱樞紐，允許我們從一個新的觀點來看資料。

資料探勘

圖 6.8 □ 從月到季進行的滾上動作



資料探勘

一般性考量

MS Excel 提供一個介面，可以從關聯資料庫中建立 OLAP 方塊，包含在方塊中的資訊可以在 Excel 中的**樞紐表**（pivot table）顯示且操作。樞紐表可以很簡單的上手，並且也提供了許多等同於其他更完整的 **OLAP** 介面工具的特色。樞紐表也提供我們分析表格資料，它的特色包含了總結資料的能力、以不同的方法分群資料、用不同的方法來顯示資料。在下一節中將會談到樞紐表的應用。

資料探勘

6.4 應用 Excel 樞紐表來進行資料分析



建立一個簡單的樞紐表

我們用一個應用在信用卡促銷資料庫的簡單例子來顯現樞紐表如何對屬性『收入範圍』總結整理資料。

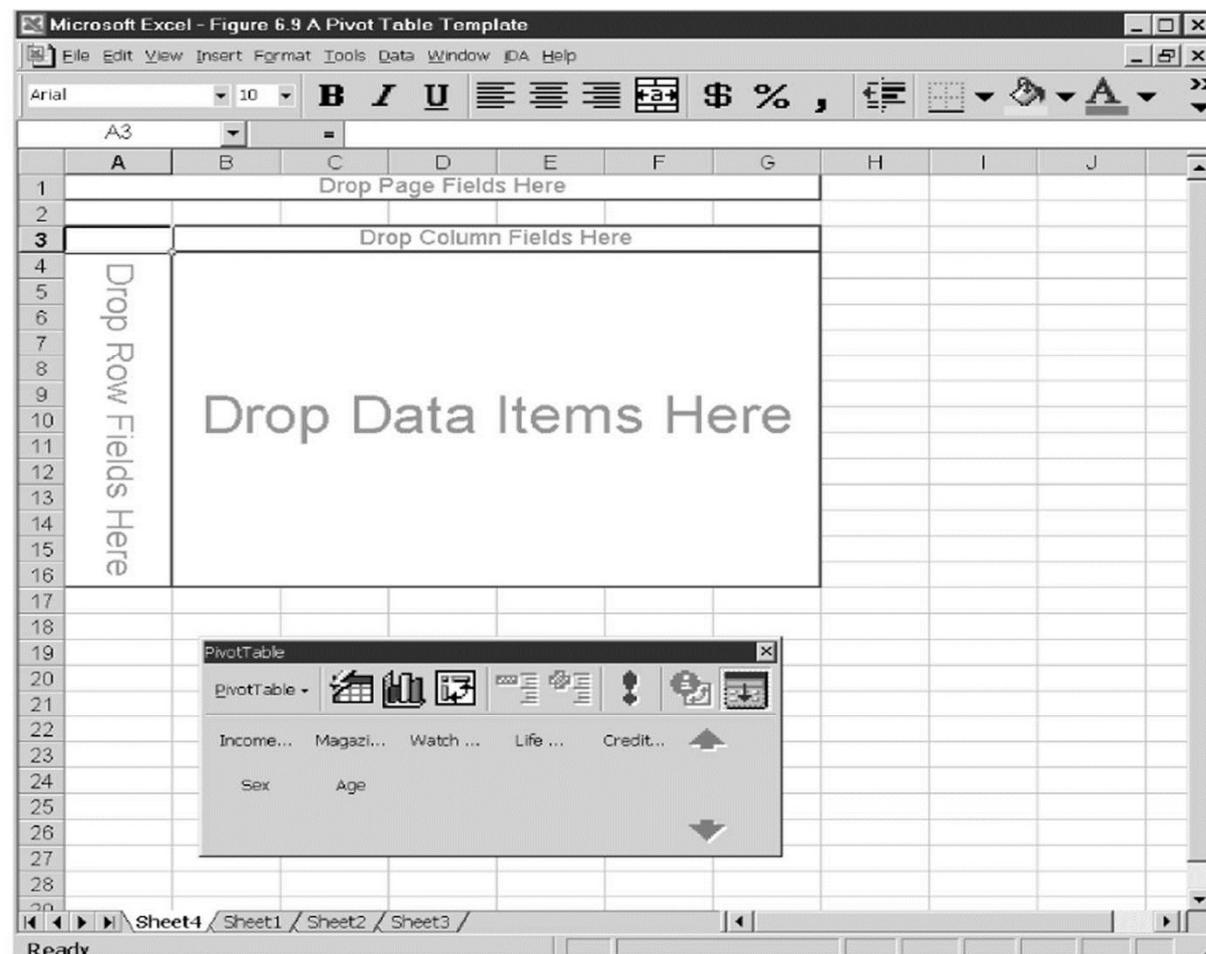
1. 一開始，先在 Excel 中載入名為 CreditCardPromotion.xls 的檔案，把資料載入到一個新的表格中以便於完整的保留舊有的資料。
2. 刪除表格第二與第三列的資料，因為它們跟我們的分析並不相關。
3. 確定游標的位置在資料集其中之一上，繼續在『資料』下拉選單中選取『樞紐分析表與圖報表』。此時將會出現一個三步驟的『樞紐分析表與圖報表精靈』。

資料探勘

4. 選取『Excel 清單』或『外部資料庫』連結按鈕。它指出被分析的資料將會被放到 Excel 表格中。此時可以選擇建立樞紐表或樞紐圖。選取『樞紐分析表』選項並按『下一步』繼續。
5. 在步驟 2 中，對話方塊會問到用來建立樞紐表的資料參數的範圍。不過當一開始就把游標放到資料集時，這個資料的範圍就應該要是正確的了，按『下一步』進入步驟 3。
6. 在步驟 3 中，指定樞紐表的位置，選取『新的工作表』連結按鈕，並且按下『完成』繼續。

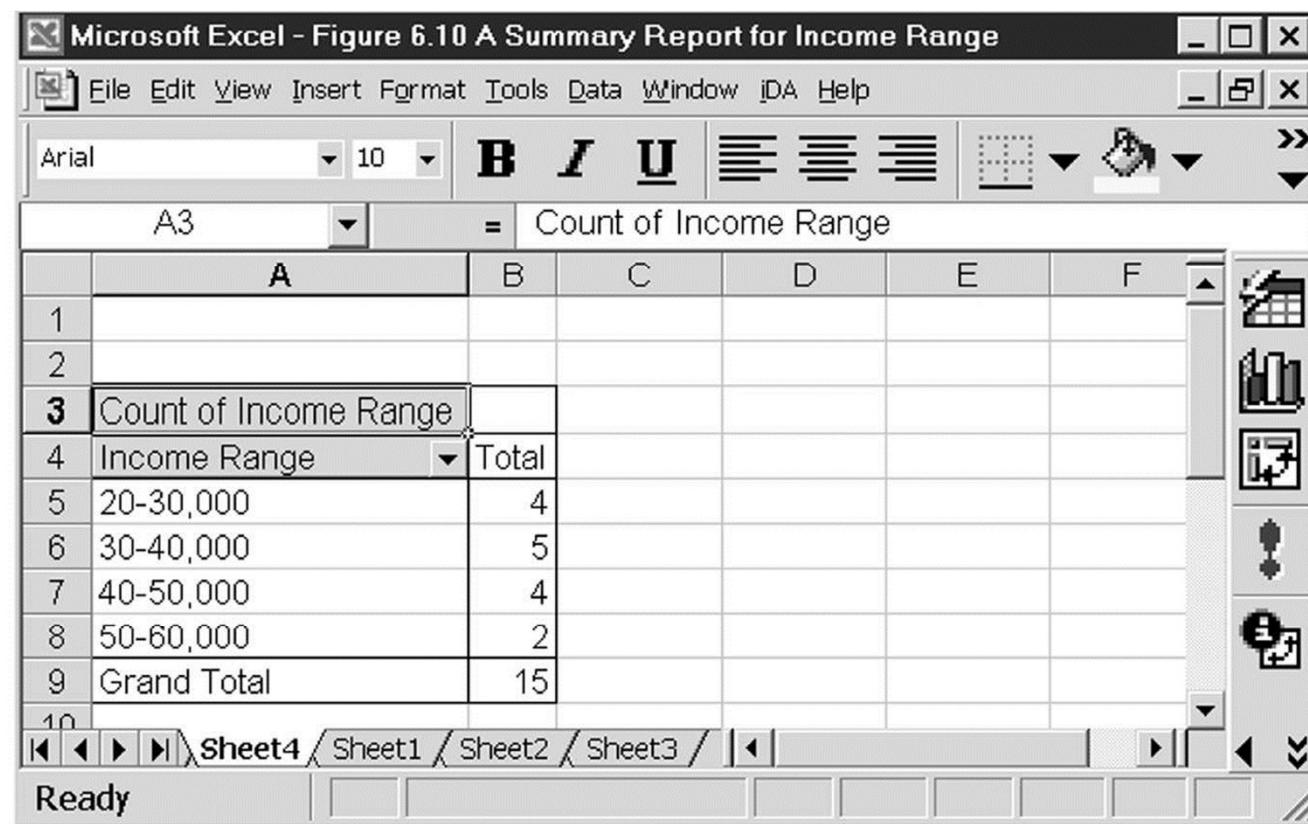
資料探勘

圖 6.9 □ 一個暫時的樞紐表



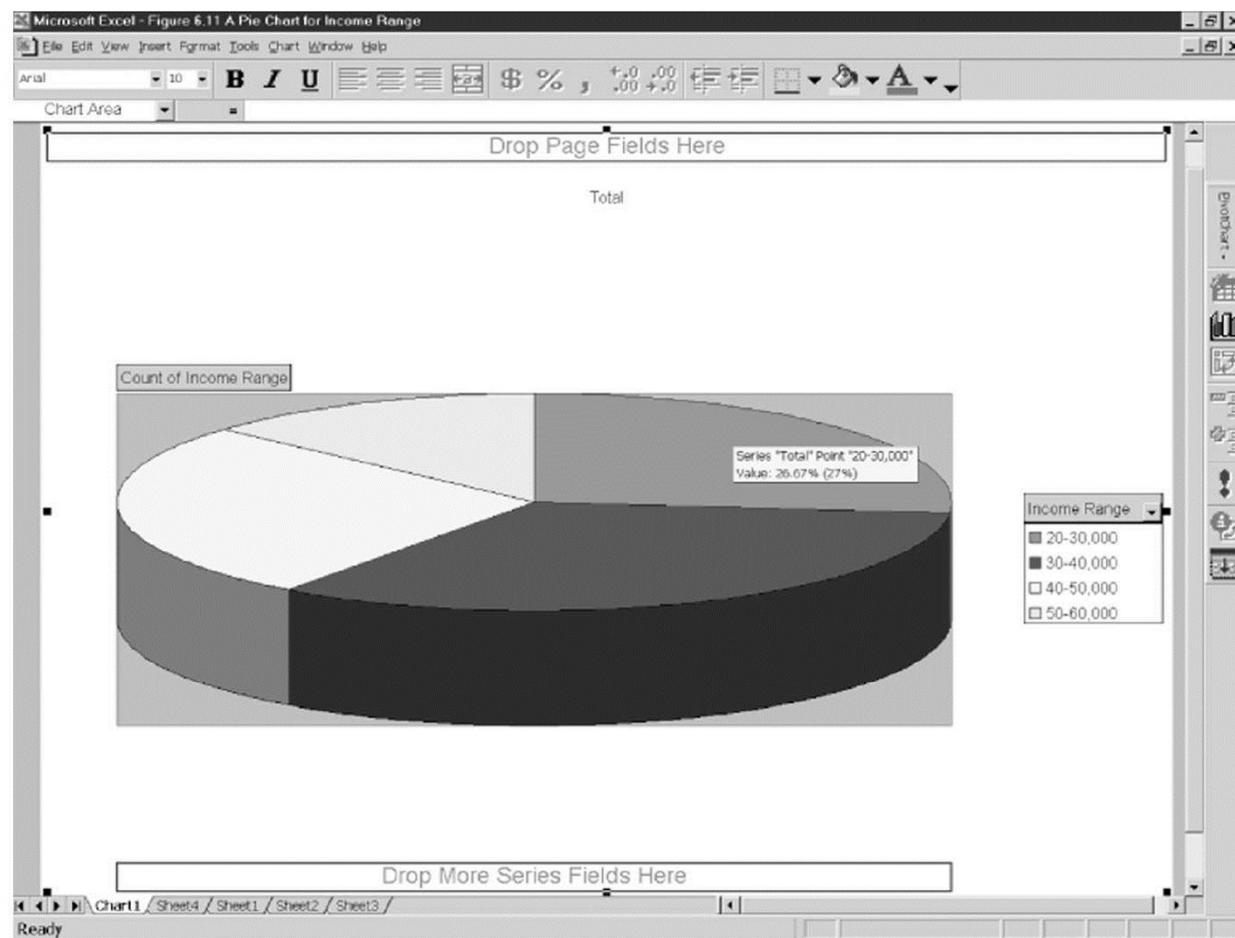
資料探勘

圖 6.10 □ 一個簡單的收入範圍報告



資料探勘

圖 6.11 □ 一個收入範圍的簡圖



資料探勘

用於假設檢定的樞紐表

Acme 信用卡公司已經決定用打電話的方式去吸引那些持卡未滿一年，且在持卡最初並沒有購買信用卡保險的持卡者。在他們的資料分析中相信，持卡者的年齡與持卡者是否擁有信用卡保險之間有特別的關係存在。另外，分析師想要檢定較年輕的持卡者購買信用卡保險的意願會比年紀較長的人高。如果假設成立，那就只有在這一個年齡層的持卡者會被選出做為電話促銷的對象。

為了檢定這個假設，我們使用樞紐表與我們的想像力，假設信用卡促銷資料庫含有非常多的持卡者樣本。接下來的步驟將會檢定這個假設：主張年齡與信用卡保險狀態的關係。

資料探勘

1. 確定現在的視窗是在包含了資料的『工作表 1』上。按下『資料』下拉選單並按下『樞紐分析表與圖報表』並按『完成』。
2. 把『年齡』拖曳到『將列欄位拖曳到這裡』的區域，再將『信用卡保險』拖曳到『將欄欄位拖曳到這裡』的區域。
3. 將『信用卡保險』拖曳到『將資料欄位拖曳到這裡』的區域，最後的結果如圖 6.12 所示。

資料探勘

圖 6.12 □ 一個表示年齡與信用卡保險的選擇之樞紐表

The screenshot shows a Microsoft Excel window titled "Microsoft Excel - ~creditCardPromotion". The menu bar includes File, Edit, View, Insert, Format, Tools, Data, Window, and Help. The toolbar contains font and style buttons. The formula bar shows "A3 = Count of Credit Card Ins.". The PivotTable is located in the range A3:G17. It has three columns: Age (No, Yes, Grand Total). The data rows show counts for various ages: 19, 27, 29, 35, 38, 39, 40, 41, 42, 43, 45, and 55. The Grand Total row sums up to 15. The PivotTable ribbon on the right side of the table shows icons for PivotTable, PivotChart, PivotFilter, PivotSort, and PivotOutline.

	A	B	C	D	E	F	G
1							
2							
3	Count of Credit Card Ins.	Credit Card Ins.					
4	Age	No	Yes	Grand Total			
5	19		1	1			
6	27	1		1			
7	29	1		1			
8	35		1	1			
9	38	1		1			
10	39	1		1			
11	40	1		1			
12	41	1		1			
13	42	1		1			
14	43	2	1	3			
15	45	1		1			
16	55	2		2			
17	Grand Total	12	3	15			
18							
19							
20							
21							
22							
23							
24							
25							
26							
27							
28							
29							

資料探勘

這個樞紐表將會顯示，目前購買信用卡保險的人數非常少。但是年齡的分配（像是年齡與信用卡保險的關係）將會是非常難以下結論的。我們可以使用『群組』功能來歸納一個更乾淨的觀點，該觀點是在兩個屬性中任何可能的關係。其方法如下：

1. 在樞紐表上單擊『年齡』屬性。
2. 單擊『資料』下拉選單。
3. 滑鼠移動到『群組及大綱』然後移動到『群組』。單擊『群組』。一個群組對話方塊會允許你選取群組的『開始點』、『結束點』及『間距值』。
4. 按下『完成』，使用預設的設定就可以了。

資料探勘

新的樞紐表如圖 6.13。雖然資料集太小而不能做出合理的結論，但是群組年齡的資料可以更清楚地觀察在兩個屬性間的關係。若想要復原群組，單擊『資料下拉選單』按『群組及大綱』，再按『取消群組』即可。

資料探勘

圖 6.13 □ 由年齡群組信用卡促銷資料

The screenshot shows a Microsoft Excel window titled "Microsoft Excel - ~creditCardPromotion". The menu bar includes File, Edit, View, Insert, Format, Tools, Data, Window, JPA, Help. The toolbar includes font style (B, I, U), alignment, and number formats (\$, %, ,). The formula bar shows "A3 = Count of Credit Card Ins.". The PivotTable is located in the range A3:G28. It has three columns: Age (row 4), Credit Card Ins. (row 3), and Grand Total (row 9). The data rows (5-8) show counts for age groups 19-28, 29-38, 39-48, and 49-58. The Grand Total row shows a total of 15. The PivotTable ribbon on the right side of the Excel window indicates that a PivotTable is currently selected.

	A	B	C	D	E	F	G
1							
2							
3	Count of Credit Card Ins.	Credit Card Ins.					
4	Age	No	Yes	Grand Total			
5	19-28		1	1	2		
6	29-38		2	1	3		
7	39-48		7	1	8		
8	49-58		2		2		
9	Grand Total		12	3	15		
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							
23							
24							
25							
26							
27							
28							
29							

資料探勘

第二個決定年齡與信用卡保險關係存在的方法，是去計算那些擁有與沒有信用卡保險的人之平均年齡。取代先前從信用卡促銷資料庫開始的工作，利用呼叫『樞紐表精靈』來修改現有的樞紐表，作法如下：

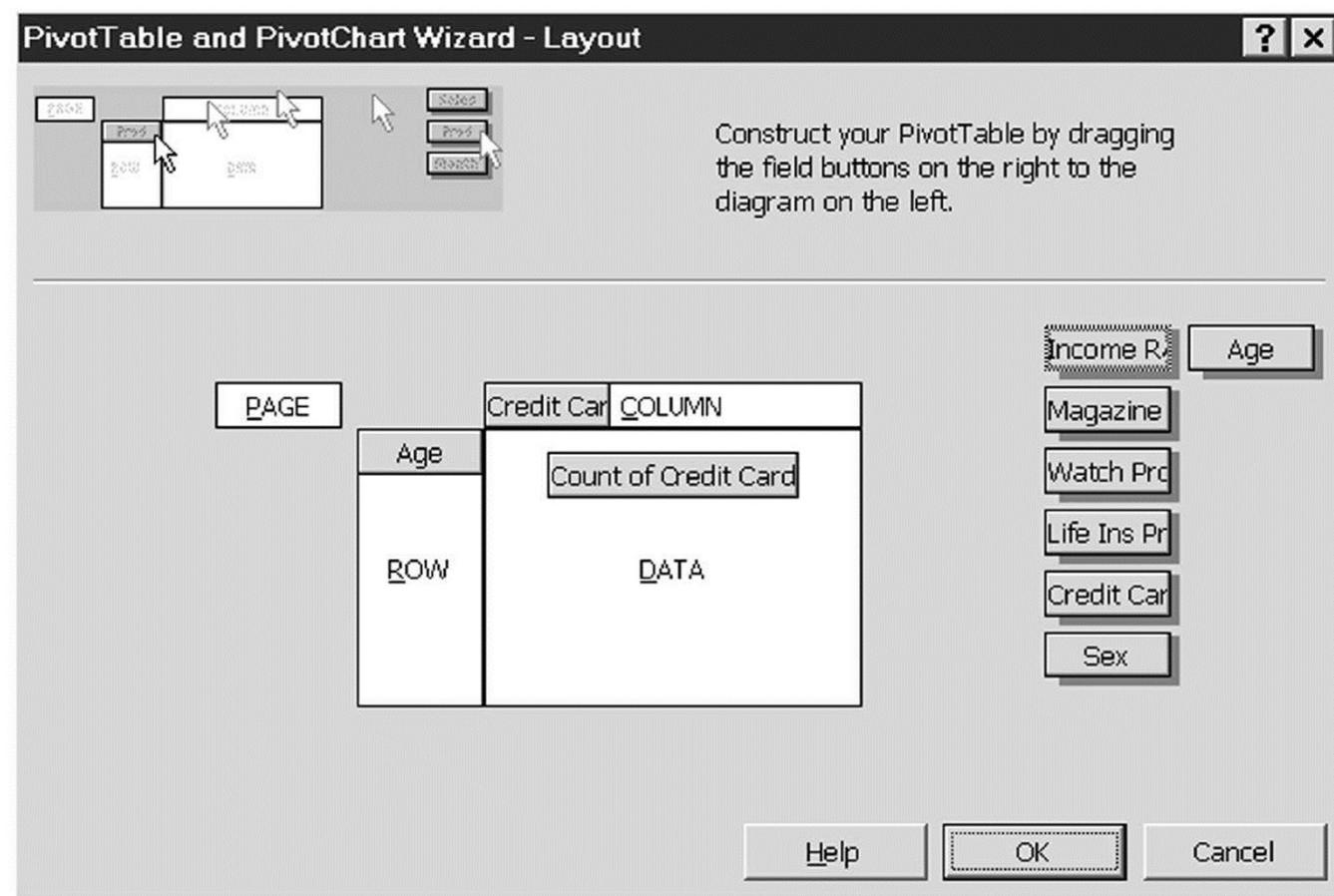
1. 把『樞紐表精靈』放到工具列上。
2. 單擊精靈按鈕，這個動作會呼叫出步驟 3 的『樞紐表精靈』。
3. 按『版面設定』按鈕。現在的樞紐表版面配置會顯現在『精靈視窗』中，圖 6.14 顯示了現在的設定值。在 Excel 97 之中也是一樣類似的圖形。
4. 將滑鼠從『列』欄位拖曳屬性『年齡』到右邊去放，再把『信用卡保險』欄位從『欄』欄位拖曳到『列』欄位區。

資料探勘

5. 把『信用卡保險總數』從資料區移走，再把『年齡』放進去。
6. 雙擊資料區的『年齡』按鈕，此時會出現一個『樞紐分析表』欄位的對話方塊。
7. 從對話框中單擊『平均值』取代原有的『總和』，然後按『確定』。回到『精靈視窗』。
8. 從『版面設定精靈』視窗中按『確定』，再從『樞紐表精靈』中按『完成』。

資料探勘

圖 6.14 □ 樞紐表版面設定精靈



資料探勘

建立一個多維度樞紐表

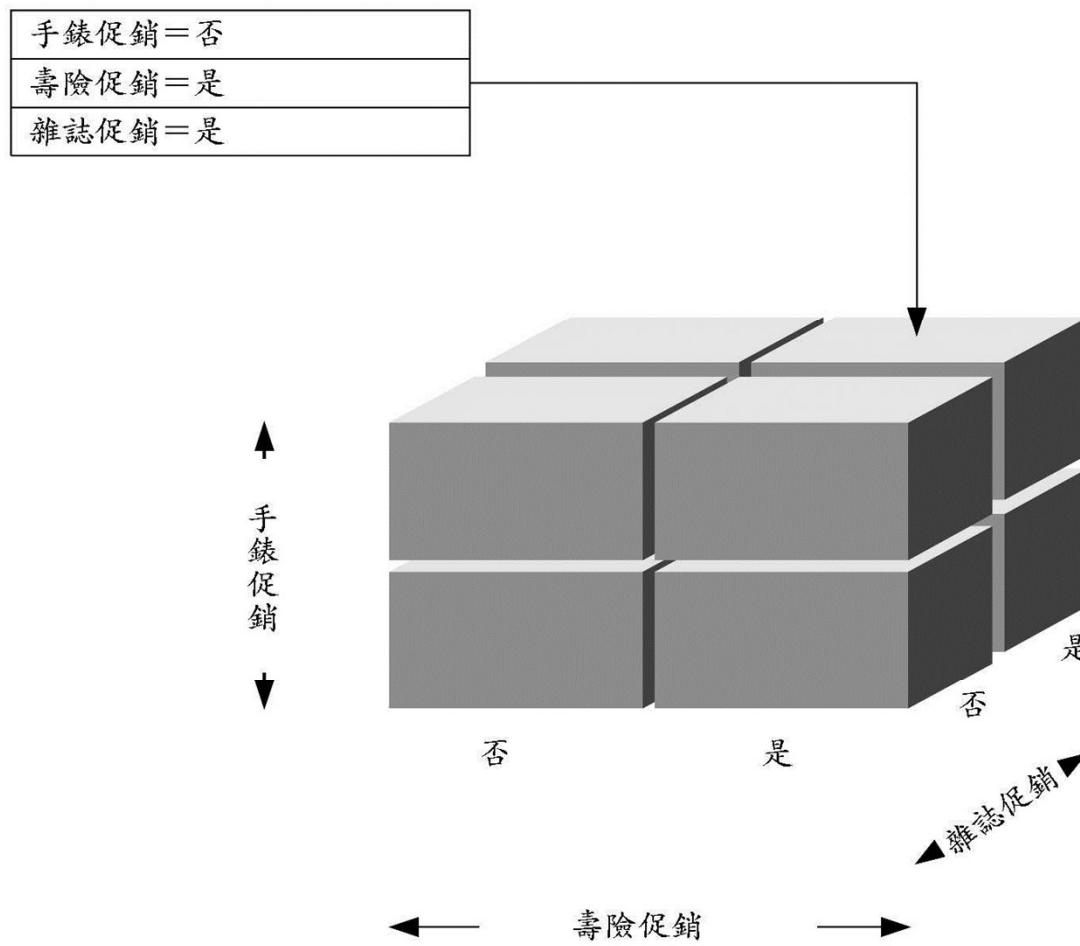
在這個例子中，將使用樞紐表來調查雜誌、手錶與人壽保險的促銷之間的關係，該促銷工作也關係到顧客的性別與收入範圍。藉由建立三個維度的方塊，如圖 6.15 所示，來完成這個工作。每個方塊的資料集包含了接受促銷與沒有接受的顧客數量之計算。圖 6.15 的箭頭指出了資料集包含的顧客總數，其中這些顧客都被人壽保險與雜誌的促銷所吸引，但是並沒有收到手錶促銷的訊息。把性別與收入範圍藉由標出這些屬性為分頁變數而包含進分析之中。以下是工作的程序：

資料探勘

1. 一開始，確定游標與畫面在含有資料的那個『工作表』上，按下『資料』下拉選單，選取『樞紐分析表與圖報表』再按『完成』。
2. 用滑鼠拖曳『手錶促銷』與『壽險促銷』到『將列欄位拖曳到這裡』。再把『雜誌促銷』拖曳到『將欄欄位拖曳到這裡』。
3. 把『手錶促銷』、『壽險促銷』、『雜誌促銷』都拖曳到『把資料欄位拖曳到這裡』的區域。
4. 最後，把『性別』與『收入範圍』拖曳到『把分頁欄位拖曳到這裡』的區域。

資料探勘

圖 6.15 □ 一個信用卡促銷方塊



資料探勘

圖 6.16 □ 一個具有分頁變數的信用卡促銷樞紐表

The screenshot shows a Microsoft Excel window with the title bar 'Microsoft Excel - creditCardPromotion'. The menu bar includes File, Edit, View, Insert, Format, Tools, Data, Window, iPA, Help. The ribbon tabs are 'Home', 'Insert', 'Page Layout', 'Formulas', 'Data', 'Review', 'View'. The font is set to Arial, size 10. The table is a pivot table with the following structure:

		Income Range				
1	Income Range	(All)				
2	Sex	(All)				
4	Magazine Promo					
5	Life Ins Promo	Watch Promo	Data	No	Yes	Grand Total
6	No	No	Count of Life Ins Promo	2	2	4
7			Count of Watch Promo	2	2	4
8			Count of Magazine Promo	2	2	4
9		Yes	Count of Life Ins Promo	2		2
10			Count of Watch Promo	2		2
11			Count of Magazine Promo	2		2
12	No Count of Life Ins Promo			4	2	6
13	No Count of Watch Promo			4	2	6
14	No Count of Magazine Promo			4	2	6
15	Yes	No	Count of Life Ins Promo	1	2	3
16			Count of Watch Promo	1	2	3
17			Count of Magazine Promo	1	2	3
18		Yes	Count of Life Ins Promo	2	4	6
19			Count of Watch Promo	2	4	6
20			Count of Magazine Promo	2	4	6
21	Yes Count of Life Ins Promo			3	6	9
22	Yes Count of Watch Promo			3	6	9
23	Yes Count of Magazine Promo			3	6	9
24	Total Count of Life Ins Promo			7	8	15
25	Total Count of Watch Promo			7	8	15
26	Total Count of Magazine Promo			7	8	15
27						
28						