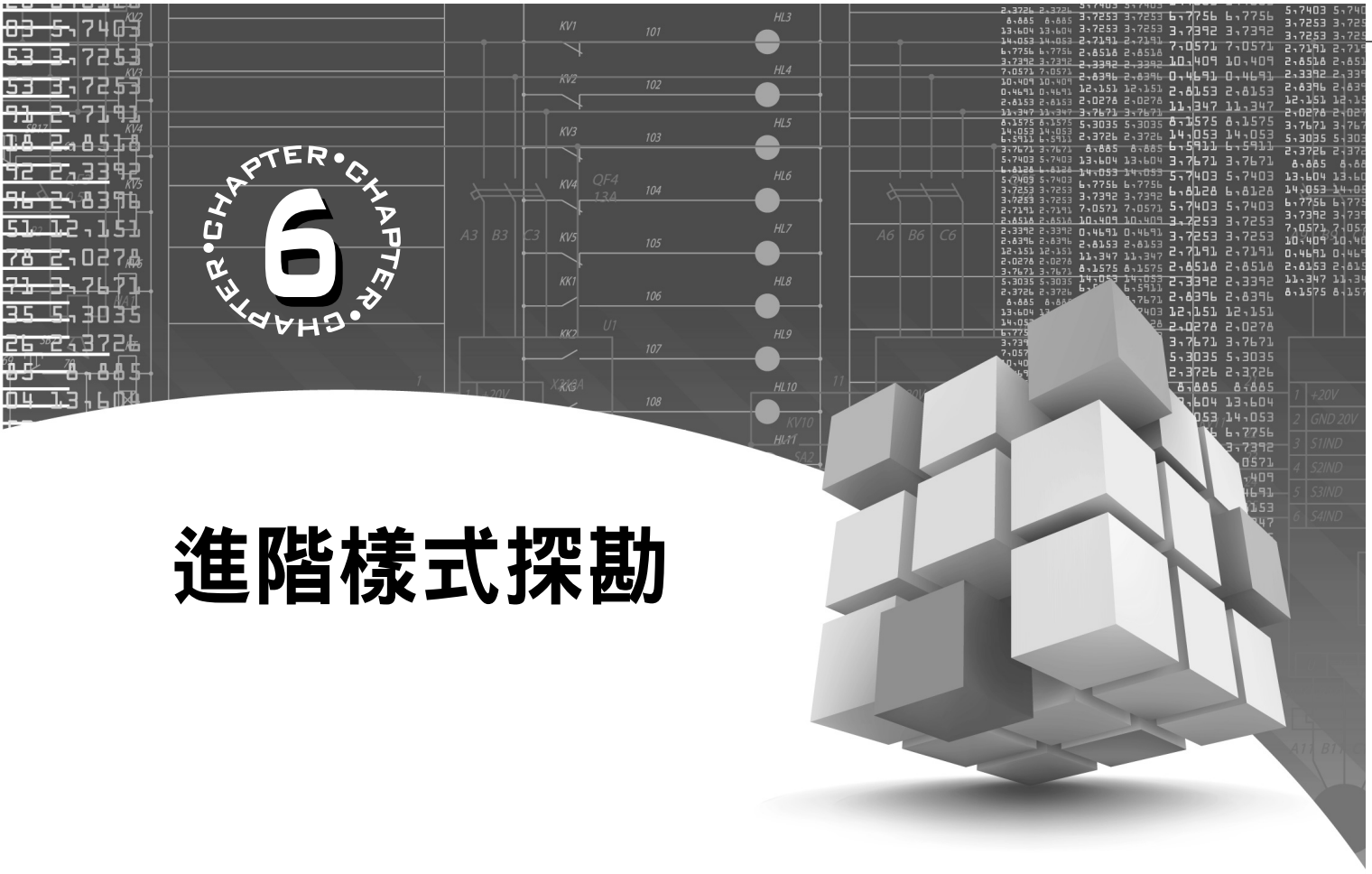

CHAPTER 6

# 進階樣式探勘

由於大量豐富的研究，多面向的問題範圍擴充與廣泛的應用探討，頻繁樣式探勘已經遠遠超越基本的樣式探勘，在本章中，你將學習到進階的樣式探勘方法。一開始，我們先展示樣式探勘的路線圖，介紹探勘多種類型樣式的方法，與探討樣式探勘的延伸應用。我們深入涵蓋多種類型樣式的探勘方法：多層級樣式、多維度樣式、連續資料中的樣式、罕見樣式、負向樣式、受限制的頻繁樣式、高維度資料中的頻繁樣式、巨型樣式、壓縮與近似樣式。其它的樣式探勘主題，包含循序樣式與結構樣式探勘、從時空性資料、多媒體資料與串流資料中探勘樣式，被視為更進階的主題，超出本書的範圍。注意，樣式探勘比頻繁樣式探勘更為一般化，因為前者還涵蓋稀少樣式與負向樣式，然而，在沒有歧義時，這兩個術語可以交換地使用。

Chapter 6

## 6.1 樣式探勘：路線圖

第 5 章使用購物籃分析為範例，介紹頻繁樣式探勘的基本概念、技術以及應用。由於資料類型、使用者的要求與應用問題的多樣性，導致演發出各式各樣多變化的探勘樣式、關聯規則與相互關係的方法，面對該領域豐碩的研究文獻，給出一個清晰的路線圖是很重要的一件事情，它能幫助我們有組織的理解該領域，並為所面對的樣式探勘應用問題，選擇最適當的探勘方法。

圖 6.1 提綱挈領地展示樣式探勘的路線圖，大部分研究主要致力於樣式探勘的三個面向：探勘樣式的類型、探勘方法與應用。然而，某些研究可能整合多個面向，舉例來說，不同的應用問題可能需要探勘不同的樣式，自然而然地演發出嶄新的探勘方法。

根據樣式的多樣性，樣式探勘可以使用以下準則而予以分類：

- **基本樣式**：如同第 5 章所介紹，頻繁樣式有許多不同形式，包含簡單的頻繁樣式、封閉樣式、或最大樣式。回顧之前所介紹的，**頻繁樣式** (frequent pattern) 是滿足最小支持度門檻值的樣式 ( 項目集 )。一個樣式 $p$ 是封閉的，如果不存在與 $p$ 有相同支持度的超樣式 $p'$。樣式 $p$ 是**最大樣式** (max-pattern)，如果它不存在頻繁的超樣式。基於有趣度量測，頻繁樣式可以映射成**關聯規則** (association rule)，或是其他形式的規則。然而有時候，我們可能對於**非頻繁** (infrequent) 或**罕見** (rare) 的樣式也感到興趣 ( 亦即，此樣式很少出現，但它是關鍵重要 )，或是對**負向樣式** (negative pattern) 感到興趣 ( 亦即，能揭露項目間的負相關性的樣式 )。
- **基於樣式涉及的抽象層級**：樣式或關聯規則可能擁有處於高層、低層或多層抽象層級的項目或概念，舉例來說，假設探勘出的關聯規則集合包含下列規則：

$$\text{buys}(X, \text{“電腦”}) \Rightarrow \text{buys}(X, \text{“印表機”}) \qquad \textbf{(6.1)}$$

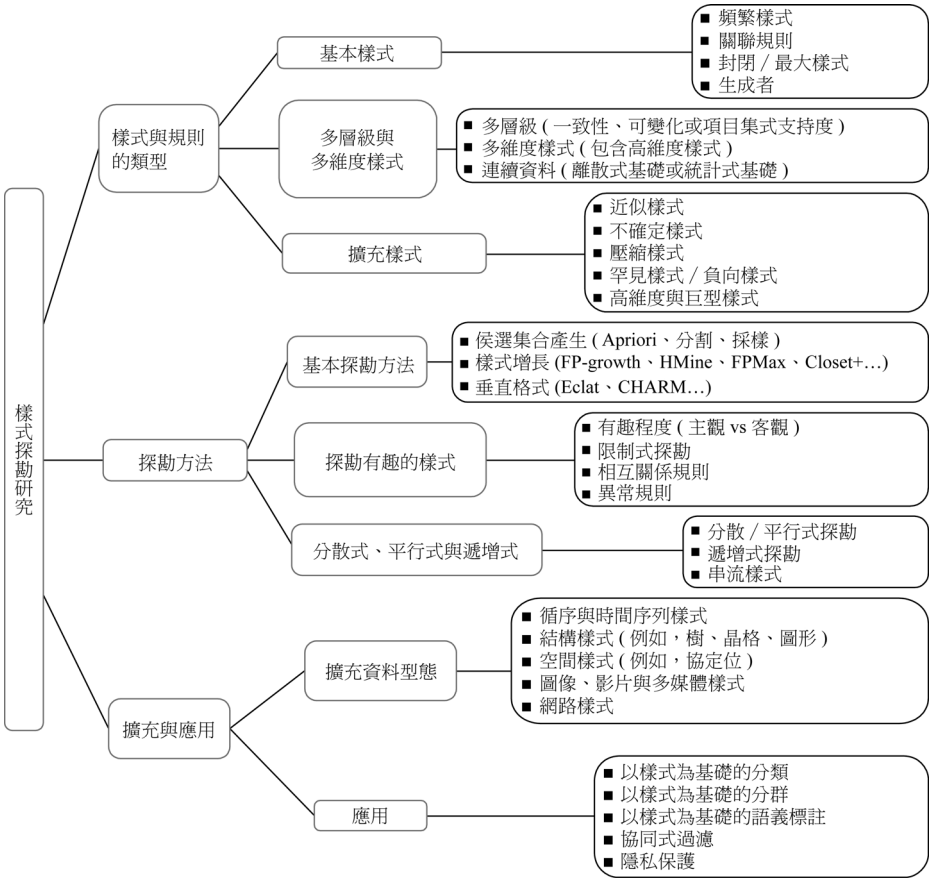


**圖 6.1　樣式探勘的一般路線圖。**

$$\text{buys}(X, \text{“筆記型電腦”}) \Rightarrow \text{buys}(X, \text{“彩色雷射印表機”}) \qquad \textbf{(6.2)}$$

其中 $X$ 是代表顧客的變數。在規則 (6.1) 與 (6.2) 中，購買的商品涉及不同的抽象層級 ( 例如，“電腦”是比“筆記型電腦”位於更高的抽象層級，“彩色雷射印表機”是比“印表機”位於更低的抽象層級 )，我們稱此規則集合是由**多層級關聯規則** (multilevel association rules) 所組成，相反地，如果規則集合不涉及在不同抽象層級的商品項目或屬性，則此集合包含**單層關聯規則** (single-level association rules)。

- **基於規則或樣式所涉及的維度數目**：如果關聯規則或樣式中的項目或屬性只涉及單一維度，它是**單維度關聯規則 / 樣式** (single-dimensional association rule / pattern)，舉例來說，規則 (6.1) 與 (6.2) 是單維關聯規

則，因為它們只涉及單一維度 buys[1]。

如果規則 / 樣式涉及兩個或更多個維度，例如 age, incomes 與 buys，則稱它為**多維度關聯規則 / 樣式** (multi-dimensional association rule)，底下為多維度規則的範例。

$$\text{age}(X, \text{“20...29”}) \wedge \text{income}(X, \text{“52K...58K”}) \Rightarrow \text{buys}(X, \text{“iPad”}) \qquad \textbf{(6.3)}$$

- **基於規則或樣式所處理的值類型**：如果規則涉及項目出現與否之間的關聯性，則它是**布林關聯規則** (Boolean association rule)，舉例來說，規則 (6.1) 與 (6.2) 是從購物籃分析所得到的布林規則。

  如果規則描述數量化項目或屬性之間的關聯性，則它稱為**量化關聯規則** (quantitative association rule)，在這些規則中，項目或屬性的量化值可以分割成數個區間，規則 (6.3) 可視為量化關聯規則，其中量化屬性 age 與 income 已被離散化 ( 成區間 )。
- **基於限制或準則來探勘選擇性樣式**：被發掘的樣式或規則可以是基於**限制的** ( constraint-based，即滿足使用者定義的限制式 )、基於**近似** (approximate)、**壓縮** (compressed) 或**近似匹配** (near-match) 的 ( 即，那些符合支持計數的近似或近乎匹配的項目集 )、top-k ( 即前 $k$ 個最頻繁的項目集，其中 $k$ 是使用者指定的參數 )、**冗餘感知** top-k ( redundancy-aware top-k，即，排除相似與冗餘的 top-$k$ 樣式 )。

另外，樣式探勘可以使用以下的準則，根據涉及資料與應用的類型而予以分類：

- **基於所探勘資料與特徵的類型**：給定關聯資料與資料倉儲資料，大多數人們感興趣的是項目集，因此，在此情況下，頻繁樣式探勘本質上是頻繁項目集探勘，亦即，探勘頻繁的項目集合。然而，在許多應用當中，樣式可能涉及序列或結構，舉例來說，藉由研究那些頻繁出現的商品項目購買順序，我們可能發現顧客傾向先購買 PC，接著購買數位相機，然後購買記憶卡。這導致**循序樣式** (sequential pattern)，亦即在有序事件所

[1] 根據多維度資料庫使用的專業術語，我們稱規則中的不同謂詞為維度。

組成的序列當中，找出頻繁出現的子序列 ( 通常被其他事件隔開 )。

我們也可以探勘**結構樣式** (structure pattern)，也就是在結構資料集合中的頻繁子結構，請注意，結構是一般性概念，可能包含許多種不同形式，例如有向圖、無向圖、晶格、樹、序列、集合、單一項目或這些結構的組合。單一項目是最簡單的結構型式，一般樣式中每一個元素可能包含子序列、子樹、子圖形等等，而且這樣的包含關係可以遞迴地定義下去，所以，結構樣式探勘可視為頻繁樣式探勘最一般化的形式。

- **基於應用領域的特定語義**：資料與應用問題皆可能非常多樣化，因此，探勘的樣式可能會因其應用領域的特定語義而有大幅的差異，不同形態的應用資料包含空間資料、時間資料、時空資料、多媒體資料 ( 例如，圖像、音訊、或視頻資料 )、文字資料、時間序列資料、DNA 與生物序列、軟體程式、化學合成物結構、網站結構、感知網路、社會與資訊網路、生物網路、資料串流等等。明顯地，如此的多樣性導致大量不同探勘方法問世。
- **基於資料分析的用途**：頻繁樣式探勘時常可做為提升資料理解與更強大資料分析的中間步驟。舉例來說，它可作為分類法中的特徵擷取步驟，這通常稱為**以樣式為基礎的分類法** (pattern-based classification)。同樣地，**以樣式為基礎的分群法** (pattern-based clustering) 展示了它在高維度資料分群時的優勢。為了提升資料的理解性，樣式可用來語義註解或是情境分析，樣式分析也可用於**推薦系統** (recommender system) 中，它根據相似使用者的樣式，來推薦使用者可能會感興趣的項目 ( 例如，書本、電影、網頁 )。不同的分析任務可能需要探勘不同類型的樣式。

接下來數個章節中，我們將介紹樣式探勘的進階與延伸方法，6.2 節探討探勘多層級樣式、多維度樣式、具有連續屬性的規則、罕見樣式與負向樣式。6.3 節研究限制式樣式探勘，6.4 節介紹如何探勘高維度與巨型樣式，探勘近似與壓縮樣式將在第 6.5 節介紹。更進階的主題，包含探勘循序與結構樣式，以及在複雜與多樣性的資料類型中探勘樣式，將會在第 12 章簡略介紹。

# 6.2 多層級與多維度空間的樣式探勘

本節聚焦於多層級與多維度空間的探勘方法，尤其是，你將學會探勘多層級關聯規則（6.2.1 節）、多維度關聯規則（6.2.2 節）、量化關聯規則（6.2.3 節）、罕見樣式與負向樣式（6.2.4 節），多層級關聯規則涉及在不同抽象層級中的概念，多維度關聯規則涉及超個一個維度或謂語（例如，將顧客的 buys 關聯到他（或她）的年齡），量化關聯規則涉及到數值屬性，在其值之間有隱含的順序關係（例如，年齡），罕見樣式是推薦稀有但是有趣的項目組合，負向樣式顯示項目之間的負相關性。

## 6.2.1 探勘多層級關聯規則

在許多應用問題中，在高階抽象層級中發掘的強關聯規則，雖然擁有高支持度，但可能只是眾所周知的常識性知識，我們可能想要往更細節的層級向下鑽取，來找出嶄新的樣式。在另一方面，可能有許多在低層或原始抽象層級上散佈的樣式，其中一部分是只高層樣式的簡單特性化。因此，如何發展一個有效率的方法，能夠在多個抽象層級下探勘樣式，並有足夠的靈活性能在不同抽象空間中轉換，是一件很有趣的事情。

**表 6.1** 任務相關資料集 D

| TID | 購買商品 |
|---|---|
| T100 | Apple 17″ MacBook Pro Notebook, HP Photosmart Pro b9180 |
| T200 | Microsoft Office Profesional 2010, Microsoft Wireless Optical Mouse 5000 |
| T300 | Logitech VX Nano Cordless Laser Mouse, Fellowa GEL Wrist Rest |
| T400 | Dell Studio XPS 16 Notebook, Canon PowerShot SD 1400 |
| T500 | Lenovo ThinkPad X200 Tablet PC, Symantec Norton Antivirus 2010 |
| … | … |

## 範例 6.1 探勘多層級規聯規則

假設你被給定與任務相關的交易資料，顯示在表 6.1 中，它們是 AllElectronics 公司的銷售資料，顯示每一筆交易中被購買的商品項目。這些商品項目的概念階層顯示在圖 6.2 中，概念階層定義一系列的從低層級概念至高層且更一般化的概念的映射，藉由將低層級概念被概念階層中更高層級的概念 ( 或祖先 ) 來取代，我們可將資料泛化。

圖 6.2 中的概念階層共有 5 個層級，分別編號 0 至 4，從第 0 層中的節點 All ( 最一般化的抽象層級 ) 開始，這裡，第 1 層包含 computer, software, printer and camera 與 computer accessory，第 2 層包含 laptop computer, desktop computer, office software, antivirus software 等，第 3 層包含 Dell desktop computer, …, Microsoft office software 等，第四層是階層中最具體明細的抽象層，它由原始資料值所組成。

名目屬性的概念階層通常隱含在資料庫綱要中，並且可以透過如第 3 章介紹的方法來自動產生，在我們的範例中，圖 6.2 的概念階層是由產品說明資料來產生。數值屬性的概念階層可以使用離散化技術來產生，第 3 章有介紹許多自動產生概念階層的方法。相對地，概念階層可由熟悉資料的使用者來指定，例如我們的例子中的商店經理。

在表格 6.1 中的商品項目，是位於圖 6.2 概念階層中的最底層，在如此原始層級下的資料中，很難去發掘出有趣的購買樣式。舉例來說，如果“Dell Studio XPS 16 Notebook”或“Logitech VX Nano Cordless Laser Mouse”只在交易資料中出現非常小的比率，如何能找出涉及如此詳盡明細的商品的強關聯規則，是非常重要的。僅有少數人會同時購買這些商品，使得這些商品項目集合要滿足最小支持度門檻值是不太可能的。然而，我們預期要找出這些商品在一般化抽象層級之間的強關聯規則，是比較容易的，例如在“Dell Notebook”與“Cordless Mouse”之間。

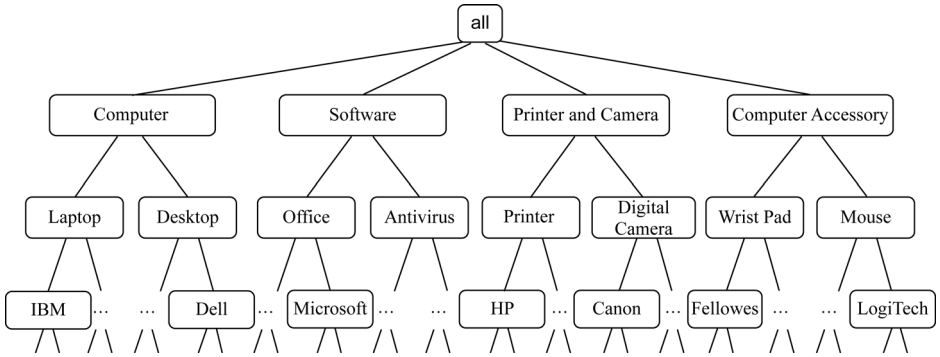


**圖 6.2** AllElectronics 公司電腦商品項目的概念階層。

在多層抽象階層上探勘資料所產生的關聯規則稱為**多層級關聯規則** (multilevel association rules)，多層級關聯規則能夠在支持度－信賴度的框架上，使用概念階層來有效的探勘出，一般來說，它是採用由上至下的策略，其中計數 count 值在計算每一抽象階層的頻繁項目時予以累計，從層級 1 開始，並往下至更特定的概念層級，直至沒有發現頻繁項目集為止。在每一個層級，可以套用任何發掘頻繁項目集的演算法，例如 Apriori 或其變異方法。

接下來會介紹數種不同的方法，這些方法的差異在運用稍微不同的方式來使用支持度門檻值，這些方法闡明在圖 6.3 與 6.4 中，其中節點代表被檢驗的項目或項目集，而使用粗邊框的代表此項目或項目集是頻繁的。

- **對所有層級使用一致的最小支持度** ( 稱為**一致支持度** )：在每一個抽象層級探勘時，使用相同的最小支持度門檻值，例如在圖 6.3 中，貫穿整個階層所使用的最小支持度門檻值皆為 5% ( 例如，從“computer”下至“laptop computer”)，computer 與 laptop computer 皆是頻繁的，但 desktop computer 則不是。

  當使用一致最小支持度門檻值時，搜尋的程序被簡化了，而且使用者只需要指定一個最小支持度門檻值，也讓此方法更容易。根據祖先節點是後代的超集合的知識，我們可以採用類似 Apriori 的最佳化技術，來避免搜尋那些其祖先不滿足最小支持度的項目集。

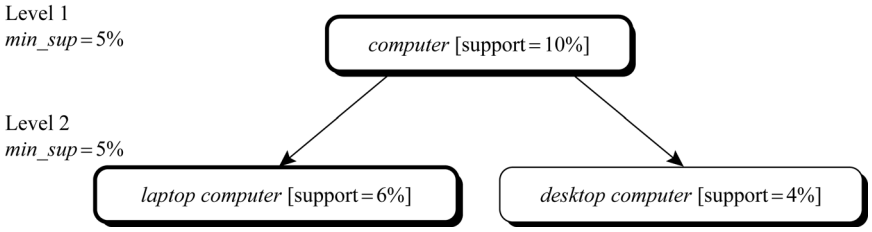


**圖 6.3　具有一致支持度的多層級關聯探勘。**

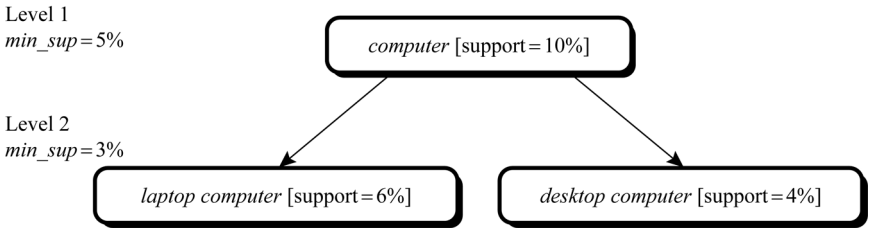


**圖 6.4　具有遞減支持度的多層級關聯探勘。**

然而，一致支持度方法有一些缺點，在較低抽象層級的項目，不太可能如同較高抽象層級的項目那麼頻繁出現，如果最小支持度門檻值設定得太高，可能會遺失一些發生在較低抽象層級中有意義的關聯。另一方面，如果最小支持度門檻值設定得太低，可能會在較高抽象層級中產生過多無趣的關聯規則，這激勵了底下的方法的發展。

- **在較低層級使用遞減的最小支持度**（稱為**遞減支持度**）：每一個抽象層級擁有它自己的最小支持度門檻值，越底層的抽象層級，對應的門檻值就越小，舉例來說，在圖 6.4 中，層級 1 與 2 的最小支持度門檻值分別為 5% 與 3%，使用這個方式，computer、laptop computer 與 desktop computer 皆是頻繁的。
- **使用項目或群組式最小支持度**（稱為**群組式支持度**）：由於使用者或領域專家通常清楚洞悉哪些群組是比其他的更重要，因此有時候，在探勘多層級規則時，能讓使用者指定特定項目或群組的支持度，是更令人滿意的。舉例來說，使用者可以根據產品的價格與項目的有趣程度，來設定最小支持度門檻值。例如對“價格超過 1000 元的 camera”或“Tablet

PC”設定特別低的門檻值，以對這類別的商品投以特別的關注。

對於探勘的樣式包含混合不同支持度門檻值的群組時，通常將所有群組中支持度門檻值最小者，取來當作探勘樣式時的門檻值。這能避免過濾掉那些包含來自支持度門檻最小的群組中的項目，但是卻很有意義的樣式。同時，對各個群組的最小支持度門檻應該保持，能防止產生各個群組中無趣的項目集。在探勘出項目集之後，可以套用其他的有趣度量測，來萃取出真正有趣的規則。

請注意，當使用遞減支持度或群組式支持度時，Apriori 性質不見得會對所有的項目都成立，然而，可以基於此性質的延伸，來發展出有效率的方法，它的細節就留給有趣的讀者做為習題。

探勘多層級關聯規則有一項嚴重的副作用，是它會因為祖先在項目間的關係，在多個抽象層級中產生許多冗餘的規則。舉例來說，考慮底下的規則：

$$\text{buys}(X, \text{“laptop computer”}) \Rightarrow \text{buys}(X, \text{“HP printer”}) \quad [\text{support} = 8\%, \text{confidence} = 70\%] \tag{6.4}$$

$$\text{buys}(X, \text{“Dell laptop computer”}) \Rightarrow \text{buys}(X, \text{“HP printer”}) \quad [\text{support} = 2\%, \text{confidence} = 72\%] \tag{6.5}$$

其中，根據圖 6.2 中的概念階層，laptop computer 是 Dell laptop computer 的祖先，而 $X$ 是代表在 AllElectronics 交易的顧客的變數。

「如果同時探勘出規則 (6.4) 與 (6.5)，則規則 (6.5) 的用途如何？它是否真正地提供任何新穎的資訊？」如果後者這個較不一般化的規則，並沒提供新的資訊，則它應該被移除。讓我們來看這是如何決定的，如果將 R2 中的項目用它概念階層中的祖先取代，可以得到規則 R1 的話，則稱規則 R1 是規則 R2 的祖先。舉例來說，規則 (6.4) 是規則 (6.5) 的祖先，因為“laptop computer”是“Dell laptop computer”的祖先。根據此定義，如果根據此規則的祖先，它們的支持度與信賴度是接近於它們的“預設”值，則此規則被視為冗餘的。

**範例 6.2** 對多層級關聯規則來檢驗冗餘性

假設規則 (6.4) 的信賴度為 70% 且支持度為 8%，而且大約有四分之一的 laptop computer 銷售是 Dell laptop computer，我們可以預期規則 (6.5) 的信賴度大約為 70% ( 因為所有含 “Dell laptop computer” 的資料樣本也含 “laptop computer” )，而支持度大約為 2% ( 即 8%×1/4 )。如果實際上是很接近預設值，則規則 (6.5) 是無趣的，因為它沒有提供更多資訊，而且它的一般性不如規則 (6.4)。

## 6.2.2 探勘多維度關聯規則

至今為止，我們研究僅包含單一謂詞 (predicate) 的關聯規則，也就是謂詞 buys。舉例來說，在探勘 AllElectronics 資料庫時，我們可能發掘到以下布林關聯規則

$$\text{buys}(X, \text{“digital camera”}) \Rightarrow \text{buys}(X, \text{“HP printer”}) \quad \textbf{(6.6)}$$

沿用多維度資料庫的術語，我們稱規則中不同的謂詞為一個維度，因此，我們稱規則 (6.6) 為單維度規則 (single-dimensional)，因為它僅包含一個謂詞 ( 即 buys ) 在多個位置出現 ( 即，謂詞在規則中出現超過一次 )，這類型的規則廣泛使用於探勘交易資料庫中。

除了僅考慮交易資料庫，銷售量與其他相關資訊也常與關聯資料鏈結在一起，或整合至資料倉儲中，這樣儲存的資料本質上就是多維度的。舉例來說，除了持續追蹤在銷售交易中購買的商品項目，關聯資料庫可能記錄其他關聯到此項目與 / 或交易的屬性，例如此交易中商品的描述，或是分公司的位置，購買此商品的顧客的相關資訊 ( 例如，顧客年齡、職業、信用評等、收入與住址 ) 也可能被儲存起來。如果將每一個資料庫的屬性或資料倉儲的維度視為一個謂詞，因此，我們可以探勘包含多個謂詞的關聯規則，諸如

$$\text{age}(X, \text{“20…29”}) \wedge \text{occupation}(X, \text{“student”}) \Rightarrow \text{buys}(X, \text{“laptop”}) \quad \textbf{(6.7)}$$

涉及兩個或更多維度 ( 或謂詞 ) 的關聯規則，稱為多維度關聯規則，規則 (6.7) 包含 3 個謂詞 (age, occupation 與 buys)，每一個謂詞在規則中僅出現一次，因此我們稱他沒有重複謂詞，沒有重複謂詞的多維度關聯規則稱為**維度間關聯規則** (interdimensional association rule)。我們也可以探勘具有重複謂詞的多維度關聯規則，亦即有某些謂詞在多次出現，這些規則稱為**混合維度關聯規則** (hybrid-dimensional association rule)，它的範例如下，其中謂詞 buys 重複出現

$$\text{age}(X, \text{“20…29”}) \wedge \text{buys}(X, \text{“laptop”}) \Rightarrow \text{buys}(X, \text{“HP printer”}) \quad \textbf{(6.8)}$$

資料庫屬性可已是名目的或是量化的，名目屬性 ( 或類別屬性 ) 的值是“事物的名稱”，名目屬性的可能的值是有限的，而且這些值之間沒有順序關係 ( 例如，職業、品牌、顏色 )。**量化** (quantitative) 屬性是數值的，而且屬性值之間有隱含的順序關係 ( 例如，年齡、收入、價格 )。探勘多維度關聯規則的技術可根據量化屬性的處理，而分成兩種基本方法。

第一種方法，使用預先定義的概念階層來將量化屬性離散化，此離散化步驟在探勘前執行，舉例來說，可使用 income 的概念階層，來將該屬性的原始數值用區間 “0..20K”, “21K..30K”, “31K..40K” 等等來取代。在此處，離散化是靜態與預先決定的，第 3 章資料前處理介紹許多離散化數值屬性的技術，那些離散化後的數值屬性，透過它們的區間標籤，可以當作名目屬性來處理 ( 其中，每一個區間視為一個類別 )，我們稱此方法為**對量化屬性使用靜態離散化來探勘多維度關聯規則** (mining multidimensional association rules using static discretization of quantitative attributes)。

在第二種方法中，量化屬性是根據資料的分佈來離散化或群集到不同的“箱子”中，那些箱子可在探勘程序中進一步合併在一起，此離散化程序是動態的，而且是為了滿足某些最小化準則而建立，例如最大化探勘規則的信賴度。由於此策略將數值屬性當數量處理，而不是當作預先定義的區間或是類別處理，由此方法探勘的關聯規則稱為 ( **動態** (dynamic) ) **量化關聯規則** (quantitative association rules)。

讓我們來研討這些探勘多維度關聯規則的方法，為了簡化起見，我們侷限於討論維度間關聯規則，注意，此處我們不是搜尋頻繁項目集（這樣做僅是探勘單維度關聯規則），在多維度關聯規則探勘中，我們尋找頻繁謂詞集合，k-**謂詞集** (k-predicate set) 是包含 $k$ 個謂詞結合的集合，舉例來說，規則 (6.7) 中的謂詞集合 {age, occupation, buys} 是 3-謂詞集，相似於第 5 章所使用的符號，我們可使用符號 $L_k$ 來代表頻繁 $k$-謂詞集合。

## 6.2.3 探勘量化關聯規則

如同前面所討論到的，關聯資料庫與資料倉儲經常涉及量化屬性與量測，我們可以將量化屬性離散化到多個區間，並且在探勘關聯規則時，把它們當作名目屬性對待。然而，如此簡化的離散化方式，可能導致產生大量的規則，而其中有許多規則是無用的。此節，我們介紹能克服此困難而發掘出新穎關聯關係的三種方式：(1) 資料方塊法，(2) 分群式方法，(3) 統計式方法，以發掘出特殊的行為。

### 資料方塊式探勘量化關聯規則

在許多情形下，量化屬性可以在探勘之前，先使用預先定義好的概念階層或是資料離散化技術來離散化，其中，數量值被區間標籤取代，如果有需要，名目屬性可以泛化到更高的概念層級。如果任務相關資料是儲存在關聯表中，則我們可以輕易地修改任何之前所介紹的頻繁項目集探勘演算法，來發掘出所有的頻繁謂詞集合。特別是，除了僅搜尋單一屬性，例如 buys，我們必須搜尋所有相關的屬性，將每一個“屬性－值”的配對視為一個項目集。

另外，轉換後的多維度資料可以用來建構資料方塊，資料方塊非常適合於探勘多維度關聯規則，它儲存多維度空間中的聚集值（例如，count 值），這對於計算多維度關聯規則的支持度與信賴度是不可或缺的，資料方塊技術的概述已在第 4 章介紹過了，圖 6.5 顯示由方體的晶格來定義的資料方塊，它涉及維度 age, income 與 buys，此 $n$-維度方體中的單元，可

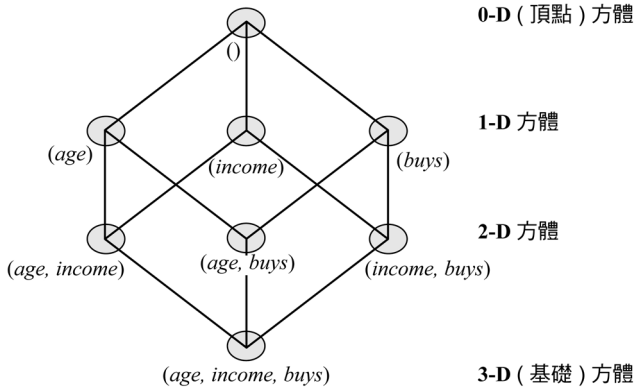


**圖 6.5** 由方體的晶格結構來建構的 3-D 資料方塊，每一個方體代表一個不同的維度分組，基礎方體包含三個謂詞 age, income 與 buys。

用來儲存所對應的 *n*-謂詞集的支持計數 (support count)，基礎方體按照 age, income 與 buys 來聚集任務相關的資料，2-D 方體 (age, income) 則按照 age 與 income 來聚集資料，依此類推，0-D 方體 ( 頂點方體 ) 包含任務相關資料中全部的交易數量。

由於資料倉儲與 OLAP 技術的使用日趨盛行，很有可能包含使用者感到興趣的維度的資料方塊，早已經存在，已全部或部分實體化，如果是此狀況，我們可以輕鬆地取出對應的聚集值，或是使用已實體化的低階聚集值來計算它們，並使用規則產生演算法，來回傳所需要的規則。請注意，即使在此情況，依然可以使用 Apriori 性質來修剪搜尋空間。如果給定的 *k*-謂詞集其支持計數 sup 並未滿足最小支持度門檻值，則可以終止進一步地探索此集合。這是因為此 *k*-項目集的任何更明確化的版本，其支持度都不會大於 sup，因此也不會滿足最小支持度門檻值。

## 探勘群集式量化關聯規則

除了以離散化或資料方塊為基礎來產生量化關聯規則之外，我們也可以透過在量化維度上為資料分群，以產生量化關聯規則 ( 回顧一下，同一群集內的物件是彼此相似，而不同群集間的物件是彼此不相似 )。其基本論述是，有趣的頻繁樣式或關聯規則一般皆是在量化屬性上相對緊密的群

集上發掘。此處，我們介紹由上至下與由下至上的套用分群演算法，來找出量化關聯規則。

典型的由上至下來找出分群式頻繁樣式的方式如下，對每一個量化維度，套用標準的分群演算法 ( 例如 *k*-means 或是密度式分群法，如第 9 章所介紹 ) 來找出在此維度上滿足最小支持度門檻值的群集。對每一個群集，我們接著檢驗透過合併此群集與其他群集，或是與其他維度的名目屬性值合併所建構的 2D 空間，來看這樣的組合是否能夠通過最小支持度門檻值，如果它通過了，我們持續在此 2D 區域搜尋群集，並且逐步往更高維度的組合前進。Apriori 修剪性質仍然可套用在此程序中，如果在任意點上，此組合的支持計數沒有通過最小支持度門檻值，則它的進一步分割，或是與其他維度合併，皆不能通過最小支持度門檻值。

由下而上的來找出分群式樣式的方法如下，首先，在高維度空間中執行分群演算法，來找出滿足最小支持度門檻值的群集，接著將那些群集映射至由較少維度組合建構的空間，並將那些群集合併。然而，在高維度空間中找出高維度群集，本身就是一個棘手的難題，所以此方法是較不實際的。

## 使用統計理論來揭露異常行為

有可能發掘出能夠揭露異常行為的量化關聯規則，其中“異常”是根據統計理論定義的，舉例來說，下述關聯規則可能揭示一個異常行為。

$$\text{性別} = \text{女性} \Rightarrow \text{平均工資} = \$6.90/\text{hr} \quad (\text{整體平均工資} = \$9.02/\text{hr}) \quad \textbf{(6.9)}$$

此規則說明女性的平均工資只有 $6.90 美元 / 小時，此規則是 ( 主觀上 ) 有趣的，因為它揭露有一群人所賺的薪資明顯地低於整體平均工資 $9.02 美元 / 小時 ( 如果整體平均工資是接近 $6.90 美元 / 小時，則女性工資是 $6.90 美元 / 小時則 會是“無趣的”)。

我們的定義的整體方面涉及套用統計檢定來驗證規則的正當性，也就是說，只有當此規則經過統計檢定 ( 此案例中，使用 Z 檢定 ) 驗證它具有高信賴度，能夠推論女性群體的平均工資是確實低於其餘的群體，此規

則才能被接受（此規則是從美國 1985 年人口普查真實資料庫探勘得到的）。

在此新的定義下的關聯規則具有以下形式：

$$\text{Population\_subset} \Rightarrow \text{mean\_of\_vlaues\_for\_the\_subset} \qquad \textbf{(6.10)}$$

其中子集合的平均值是明顯地與它在資料庫中的補集合的平均值是不一樣的（而且通過適當的統計檢定來驗證其正當性）。

## 6.2.4 探勘罕見與負向關聯規則

至今為止，本節所呈現的方法都著重在探勘頻繁樣式，然而有時候，尋找那些罕見，而非是頻繁的樣式，或是找出那些能反映項目之間負相關性的樣式，也是很有趣的。這些樣式分別稱為罕見 (rare) 與負向 (negative) 樣式，在本節中，我們考慮各種定義罕見與負向樣式的方法，這對於探勘它們將有所幫助。

**範例 6.3** 罕見樣式與負向樣式

在珠寶首飾銷售資料中，鑽石手錶的銷售是很罕見的，然而，涉及銷售鑽石錶的樣式可能是很有趣的。在超級市場銷售資料中，我們發現顧客會頻繁地購買可口可樂與健怡可樂，但不會同時購買兩者，則購買可口可樂與購買建怡可樂兩者可視為負向（相互關係的）樣式。在汽車銷售資料中，一個業務員對給定顧客銷售了幾輛高耗油的車輛（例如，SUV），接著又對相同的顧客銷售了油電混合小型車，即便購買 SUV 與購買油電混合小型車是負相互關係的樣式，但是發掘與檢驗此異常的案例可能是有趣的。

一個**非頻繁（或罕見）樣式** (infrequent (or rare) pattern) 是一個擁有支持度小於（或遠小於）使用者設定的門檻值，然而，由於絕大多數的項目集的（出現）支持度都是小於或甚至遠小於最小支持度門檻值，能

夠讓使用者指定其它的條件來定義罕見樣式，是格外令人滿意的。例如，如果我們想要找到的樣式，其包含至少一個商品的價格超過 $500，則我們可以明確地指定此限制。有效地探勘此項目集已經在探勘多層級關聯規則介紹過了 ( 6.2.1 節 )，它的策略就是套用多個 ( 例如，針對項目，或是基於群組 ) 最小支持度門檻值。其它可行的方法將在限制式樣式探勘 (6.3) 中討論，它是將使用者指定的限制嵌入在疊代式探勘程序中。

有許多種方式能夠定義負向樣式，我們將考慮其中三種定義。

**定義 6.1**

如果項目集 $X$ 與 $Y$ 皆是頻繁的，但是很少出現在一起 ( 亦即，$\sup(X \cup Y) < \sup(X) \times \sup(Y)$ )，則項目集 $X$ 與 $Y$ 是負相關的，而樣式 $X \cup Y$ 是負相關樣式，如果 $\sup(X \cup Y) \ll \sup(X) \times \sup(Y)$，則項目集 $X$ 與 $Y$ 是強負相關的 (strong negatively correlated)，而樣式 $X \cup Y$ 是強負相關樣式 (strong negatively correlated pattern)

此定義可以輕鬆的延伸到包含 $k$-項目集的樣式 ( 其中 $k > 2$ )。

然而，此定義的缺點是它不是 null-invariant，也就是說，它可能會被空交易 (null transaction) 誤導，其中，空交易係指不含我們要檢驗的項目集的交易，我們將在範例 6.4 中闡述此問題。

**範例 6.4 定義 6.1 中的空交易問題**

如果資料集合中包含大量的空交易，那麼將強烈影響評估一個樣式是否為負相關的因素，將會是空交易的數量，而不是觀察到的樣式。舉例來說，假設一個縫紉商店銷售針線包 $A$ 與 $B$，此商店賣出針線包 $A$ 與 $B$ 各 100 個，但是其中只有一筆交易同時包含 $A$ 與 $B$，直覺來看，針線包 $A$ 與 $B$ 是負相關的，因為看來購買了其中一個，並不會促進購買另一個。

讓我們來看看定義 6.1 如何處理此情境，如果總共有 200 筆交易，

則我們有 $\sup(A \cup B) = 1/200 = 0.005$ 與 $\sup(A) \times \sup(B) = (100/200) \times (100/200) = 0.25$，所以 $\sup(A \cup B) \ll \sup(A) \times \sup(B)$，而根據定義 6.1 指示，$A$ 與 $B$ 是強負相關的。

但是，如果資料集不是只有 200 筆交易，而是有 $10^6$ 筆交易呢？在此情況下，將會有許多空交易，它們既不包含 $A$，也不包含 $B$。此時，定義 6.1 會怎麼辦？它計算 $\sup(A \cup B) = 1/10^6$ 與 $\sup(A) \times \sup(B) = (100/10^6) \times (100/10^6) = 1/10^8$，所以 $\sup(A \cup B) \gg \sup(A) \times \sup(B)$，這與先前發現的結果式完全相反的，即便 $A$ 與 $B$ 出現的次數都沒有改變，所以定義 6.1 所使用的量測不是 null-invariant 的，而 null-invariant 對高品質的有趣程度量測是不可或缺的，如同 5.3.3 節所介紹的。

**定義 6.2**

如果 $X$ 與 $Y$ 是強負相關的，則

$$\sup(X \cup \bar{Y}) \times \sup(\bar{X} \cup Y) \gg \sup(X \cup Y) \times \sup(\bar{X} \cup \bar{Y})$$

此量測是否為 null-invariant 的呢？

**範例 6.5 定義 6.2 中的空交易問題**

使用裁縫商店的問題，假設資料庫中共有 200 筆交易，我們有

$$\sup(\bar{A} \cup B) \times \sup(A \cup \bar{B}) = \frac{99}{200} \times \frac{99}{200} = 0.245$$
$$\gg \sup(A \cup B) \times \sup(\bar{A} \cup \bar{B}) = \frac{199}{200} \times \frac{1}{200} \approx 0.005$$

而這根據定義 6.2，得知 $A$ 與 $B$ 是強負相關的，但如果資料庫中有 $10^6$ 筆交易時，又將如何呢？該量測計算

$$\sup(\bar{A} \cup B) \times \sup(A \cup \bar{B}) = \frac{99}{10^6} \times \frac{99}{10^6} = 9.8 \times 10^{-9}$$

$$\ll \sup(A \cup B) \times \sup(\bar{A} \cup \bar{B}) = \frac{199}{10^6} \times \frac{10^6 - 199}{10^6} \approx 1.99 \times 10^{-4}$$

在此時，此量測代表 $A$ 與 $B$ 是正相關的，得到矛盾的結果，所以此量測不是 null-invariant。

作為第三個方案，定義 6.3 是以 kulczynski 量測為基礎 ( 即，條件機率的平均值 )，他遵循第 5.3.3 節介紹的有趣度量測的精神。

**定義 6.3**

如果項目集 $X$ 與 $Y$ 皆是頻繁的，也就是說，$\sup(X) \geq \min\_sup$ 與 $\sup(Y) \geq \min\_sup$，其中 min_sup 是最小支持度門檻值，如果 $(P(X|Y) + P(Y|X))/2 < \epsilon$，其中 $\epsilon$ 是**負樣式門檻值** (negatively correlated pattern)，則樣式 $X \cup Y$ 是負相關樣式。

**範例 6.6** 根據 kulczynski 量測，使用定義 6.3 的負相關樣式

讓我們再次審視針線包的範例，並令 min_sup 為 0.01%，與 $\epsilon = 0.02$，當資料集合中有 200 筆交易時，我們有 $\sup(A) = \sup(B) = 100/200 = 0.5 > 0.01\%$，而且 $(P(B|A) + P(A|B))/2 = (0.01 + 0.01)/2 < 0.02$，所以 $A$ 與 $B$ 是負相關的。但如果當我們有大量筆交易時，這樣的結果還是成立嗎？當資料庫中有 $10^6$ 筆交易時，此量測計算 $\sup(A) = \sup(B) = 100/10^6 = 0.01\% \geq 0.01\%$ 與 $(P(B|A) + P(A|B))/2 = (0.01 + 0.01)/2 < 0.02$，再一次，這表明 $A$ 與 $B$ 是負相關的，這與我們的直覺相符合，此量測不像前面考慮的兩個定義會有 null-invariant 的問題。

讓我們檢視另一個案例，假設在 100,000 筆交易中，縫紉商店賣出

1000 個 $A$ 針線包，而 $B$ 針線包只賣出 10 個，然而，每當賣出一個 $B$ 針線包時，也會賣出一個 $A$ 針線包（亦即，它們出現在同樣的交易中），在此情況下，此量測計算 $(P(B|A)+P(A|B))/2=(0.01+1)/2=0.505\gg 0.02$，這代表 $A$ 與 $B$ 是正相關的，而不是負相關的，這同樣與我們直覺是一致的。

使用此全新的負相關的定義，可以輕鬆的推導出有效率的方法，來從大型資料庫中探勘出負向樣式，這將留給有興趣的讀者做為習題。

## 6.3 基於限制的頻繁樣式探勘

資料探勘程序能夠從給定資料集合中，發掘出上千條規則，其中絕大多數的規則，到最後評估都是不相關的，或是使用者不感興趣的規則。通常，使用者有很好的辨別力，能夠知道該往哪個“方向”探勘，才能得到有趣的樣式，或是知道我們想要找出的樣式或規則的“形式”，使用者也同時知道對於這些規則的“條件”，這能夠避免挖掘出那些已知不會感到興趣的樣式。所以，如果能夠讓使用者詳細明述這樣的直覺或期望，作為侷限搜尋空間的限制式，將是很好的啟發式方法。此策略稱為基於限制式的探勘，其限制式可以包含

- **知識型態的限制式**：這些限制指定要被探勘的知識形態，例如是關聯規則、相互關係、分類或分群。
- **資料限制式**：這些限制指定任務相關的資料。
- **維度 / 層級限制式**：這些限制指定期望探勘的資料維度（屬性）、抽象層級或是概念階層中用來探勘的層級。
- **有趣程度限制式**：這些限制指定對於規則有趣程度的統計量測的門檻值，例如支持度、信賴度或相互關係。
- **規則限制式**：這些限制式指定要被探勘的規則的形式或是條件，這些限

制式可能表達元規則（metarule，規則的模板），在規則前項與後項部分內謂詞的最大或最小數目，或是屬性、屬性的值、與 / 或聚集之間的關係。

這些限制式可以用高階資料探勘描述式查詢語言與使用者介面來指定。

前面四種限制式已經在本書與本章節前面部分討論過，在本節，我們介紹使用規則限制式來聚焦探勘任務，這類型的限制式探勘允許使用者描述他所想要發掘的規則，因此，讓探勘程序變得更有效。此外，複雜的探勘查詢優化程序能用來開發使用者指定的限制式，使得探勘程序更有效率。

基於限制式的探勘激勵了互動探索式探勘與分析，在 6.3.1 節，你將學會元規則引導式探勘，其中規則限制語法是用來指定規則模板的形式，6.3.2 節討論使用樣式空間修剪（修剪要被探勘的樣式）與資料空間修剪（它修剪資料空間的片段，而那些資料空間片段即便進一步探索，也無法發掘出滿足限制的樣式）。

對於樣式空間修剪，我們介紹三種能有利於限制式搜尋空間修剪的性質：反單調性、單調性與簡潔性。我們同時探討一類特殊的限制式，稱為可轉換的限制式，透過適當的資料排序，此限制式能夠深納入疊代式探勘程序中，並與單調式與反單調式限制具有相同的修剪能力。對於資料空間修剪，我們介紹兩種性質：資料簡潔性與資料反單調性，並探討如何將它們整合到資料探勘程序中。

為了容易介紹，我們假設使用者在搜尋關聯規則，此程序可以藉由在支持度－信賴度框架中增加使用相互關係來量測有趣程度，而輕鬆的延伸到探勘相互關係規則。

## 6.3.1 元規則導引來探勘關聯規則

「元規則的用途是什麼？」元規則 (metarule) 允許使用者指定它們有興趣探勘的規則的語法形式，規則形式可做為限制式來提升探勘程序的

效率，元規則可以使用者的經驗、期望、或對資料的直覺為基礎，或是根據資料庫綱要來自動地產生。

**範例 6.7** 元規則導引探勘

假設你做為 AllElectronics 的市場分析師，你必須存取顧客的描述資料 ( 例如，顧客年齡、住址、信用評等 ) 以及顧客的交易列表，你對於找出顧客特質與他們購買商品之間的關聯規則很感到興趣。然而，並非要找出所有反映那些關係的關聯規則，你感到興趣的只是，那一對顧客的特質能促進他們購買辦公室軟體，元規則能用來指定你所感興趣要尋找的規則的形式，這樣的元規則範例為

$$P_1(X,Y) \wedge P_2(X,W) \Rightarrow \text{buys}(X, \text{“辦公室軟體”}) \quad \textbf{(6.11)}$$

其中 $P_1$ 與 $P_2$ 是**謂詞變數** (predicate variables)，$X$ 是代表顧客的變數，$Y$ 與 $W$ 分別為指派到屬性 $P_1$ 與 $P_2$ 上的值，使用者能指定可考慮為 $P_1$ 與 $P_2$ 的例證的屬性列表，否則，可使用預設集合。

一般來說，元規則構成使用者有興趣探索或證實的關係的假說，資料探勘系統能接著搜尋與給定元規則匹配的規則，例如，規則 (6.12) 匹配 ( 或遵從 ) 元規則 (6.11)

$$age(X, \text{“30..39”}) \wedge income(X, \text{“41K..60K”}) \Rightarrow \text{buys}(X, \text{“辦公室軟體”}) \quad \textbf{(6.12)}$$

「如何使用元規則來引導探勘程序？」讓我們進一步檢驗此問題，假設我們要探勘範例 6.7 中的多維度關聯規則，此元規則的規則模板形式為

$$P_1 \wedge P_2 \wedge \ldots \wedge P_l \Rightarrow Q_1 \wedge Q_2 \wedge \ldots \wedge Q_r \quad \textbf{(6.13)}$$

其中 $P_i (i = 1, \ldots, l)$ 與 $Q_j (j = 1, \ldots, r)$ 是謂詞範例或謂詞變數，令在元規則中的謂詞數目為 $p = l + r$ ，為了發掘滿足此模板的維度間關聯規則 (interdimensional association rule)

- 我們必須發掘所有頻繁 $p$-謂詞集，$L_p$。
- 我們也必須有 $L_p$ 內 $l$-謂詞子集合的支持計數，以計算從 $L_p$ 推導的規則的信賴度。

這是探勘多維度關聯規則的典型範例，藉由使用以下章節描述的限制推進技術來延伸這些方法，我們可以推導出有效率的元規則引導探勘方法。

### 6.3.2 基於限制式的樣式產生：樣式空間修剪與資料空間修剪

規則限制指示探勘規則內的變數中期望集合 / 子集合的關係，變量的常數初始值，聚集函數上的限制，或是其它形式的限制。使用者通常套用他們對應用問題與資料的知識，來指定探勘任務中的規則限制。這些規則限制能與元規則導引一同使用，或是做為他的替代方案。在本節，我們檢驗如何使用規則限制來使得探勘程序更有效率，讓我們研討以下範例，其中規則限制被用來探勘混合維度關聯規則。

**範例 6.8** **探勘關聯規則的限制**

假設 AllElectronics 公司擁有多維度的銷售資料庫，它包含以下相關的關係

*Item*(*item_ID, item_name, description, category, price*)
*Sales*(*transaction_ID, month, year, store_ID, city*)
*Trans_item*(*item_ID, transaction_ID*)

此處，商品項目 *Item* 表格包含屬性 *item_ID, item_name, description, category, price*，而 *Sales* 表格包含屬性 *transaction_ID, month, year, store_ID, city*，這兩個表格是由 *Trans_item* 表格的外鍵 (foreign key) 屬性 *item_ID, transaction_ID* 來鏈結。

假設我們的關聯探勘查詢為“從芝加哥 2010 的銷售資料中，找出購買哪幾種便宜商品項目 ( 總價少於 $10 )，能夠促進購買 ( 即，出

現在同一筆交易 ) 哪幾種昂貴商品項目 ( 價格最小為$50 ) 的樣式或規則。”

此查詢包含下列四個限制：(1) *sum*(*I.price*) < $10，其中 *I* 代表便宜商品的 *item_ID*；(2) *min*(*J.price*) ≥ 50，其中 *J* 代表昂貴商品的 *item_ID*；(3) *T.city* = 芝加哥；與 (4) *T.year* = 2010，其中 *T* 代表 *transaction_ID*，為了簡潔起見，此處我們不顯示詳細的探勘語，然而，從探勘查詢的語義能清楚明瞭限制的情況。

我們可以在探勘規則過後，套用維度 / 層級限制與有趣程度限制來過濾那些規則，然而，一般來說，在探勘過程中使用它們來幫忙修剪搜尋空間，是更有效率與成本低廉的。維度 / 層級限制在 6.2 節已經介紹過，諸如支持度、信賴度、相互關係量測等有趣程度限制，已經在第 5 章介紹過了，本章節聚焦在規則限制上。

「如何使用規則限制來修剪搜尋空間？更明確地說，哪些類型的規則限制能夠嵌入在探勘程序中，並確保對於探勘查詢所回傳的答案是完整的」

一般來說，在探勘過程中，頻繁樣式探勘處理程序可以使用以下兩種主要方式來修剪搜尋空間：樣式空間修剪與資料空間修剪。前者檢驗候選樣式，並決定此樣式是否可被移除，套用 Apriori 性質，如果該樣式在剩餘的探勘過程中，不會產生此樣式的超樣式，則它將刪除此樣式。後者檢驗資料集，來決定在後續的探勘過程中，某特定的資料空間片段對其後產生的符合樣式 ( 某特定樣式 ) 是否會有貢獻，如果不會，則此資料空間片段將予以刪除，不再進一步的探索。能夠有助於樣式空間修剪的限制，稱為樣式修剪限制，而能夠用於資料空間的限制，稱為資料修剪限制。

## 使用樣式修剪限制來修剪樣式空間

根據限制式如何與樣式探勘程序互動，樣式探勘限制可以分成 5 類：(1) 反單調；(2) 單調；(3) 簡潔的；(4) 可轉換的；(5) 不可轉換的，對於

每一個類別，我們使用一個範例來顯示它的特徵，並解釋該類型限制如何使用於探勘程序中。

第一類限制是**反單調的** (antimonotonic)，考慮範例 6.8 中的規則限制 "$sum(I.price) \leq \$100$"，假設我們使用 Apriori 框架，他在第 $k$ 次疊代時探索出長度為 $k$ 的項目集，如果在一候選項目集內項目的總價格不少於 $100，則此項目集可以從搜尋空間予以刪除，因為增加更多項目到此集合中 ( 假設項目的價格不少於 0 )，將會使商品項目集合更加昂貴，因此永遠不會滿足此限制。換句話說，如果某一項目集不滿足此規則限制，則它的超集合沒有一個能滿足此限制，如果一個規則限制遵從此性質，則它是反單調的。可以在具有 Apriori 風格的演算法的每一次疊代中，套用反單調性來修剪，來提升整體探勘的效率，同時確保探勘任務的完整性。

Apriori 性質表示，頻繁項目集內的所有非空子集合皆是頻繁的，這也是反單調特性，如果給定的項目集不滿足最小支持度門檻值，則它的超集合也不可能滿足。在 Apriori 演算法內每一次疊代套用此性質，能夠縮減要被檢驗的候選項目集的數量，因此修剪了關聯規則的搜尋空間。

其它的反單調性限制範例包含 "$min(J.price) \geq 50$", "$count(I) \leq 10$" 等等，任何違反這些限制的項目集均可以被捨棄，因為添加新的項目至這些項目集，也永遠不會滿足限制式。請注意，像 "$avg(I.price) \leq 10$" 的限制，它並不是反單調性的，給定一個違反此限制的項目集，藉由添加某些 ( 便宜的 ) 項目所建構的超集合，結果可能會符合此限制式，因此，把此限制式嵌入在探勘程序中，將不能保證資料探勘任務的完整性。表格 6.2 第一行顯示基於 SQL 的限制的列表，這些限制的反單調性顯示在第二行中，為了簡化我們的討論，只給出存在性運算符 ( 例如，$=$、$\in$，但沒有 $\neq$、$\notin$ ) 與帶等號的比較性 ( 或包含 ) 運算符 ( 例如，$\leq$、$\subseteq$ )。

第二類限制是**單調的** (monotonic)，如果範例 6.8 中的規則限制為 "$sum(I.price) \geq \$100$"，此限制式處理方法將會相當不同，如果某項目集 $I$ 滿足此限制，則此項目集內商品的總價格將不少於 $100，進一步添加其他商品至此項目集中，將會增加它的總價格，並永遠會滿足此限制式。因此，進一步對項目集 $I$ 檢驗此限制式將會變為冗餘，換句話說，如果一個項目

集滿足此規則限制，則它所有的超集合也會滿足，如果一個規則限制符合此性質，則它是單調性的。相似的單調性規則限制包含 "$min(J.price) \le 10$", "$count(I) \ge 10$" 等等，基於 SQL 的限制的單調性顯示在表格 6.2 第三行中。

第三類的限制是**簡潔的** (succinct)，對於此限制類別，我們可以列舉出所有滿足該限制的集合，也就是說，如果一個規則限制是簡潔的，我們甚至在支持計數開始前，就可以直接明確地產生所有滿足它的集合，這避免了"產生與測試"架構的龐大負擔，換句話說，這種限制是計數前可修剪的。舉例來說，範例 6.8 中的限制 "$min(J.price) \ge \$50$" 是簡潔的，因為我們可以詳盡與明確地產生所有符合此限制的項目集。

更明確地說，這樣集合是為由價格不低於 $50 的商品所組成的非空集合，它的形式為 $S$，其中 $S \ne \varnothing$ 是所有價格不低於 $50 的商品所組成的子集合，由於有明確的"公式"來產生所有滿足簡潔限制的集合，我們不需要在探勘過程中疊代地檢驗規則限制，基於 SQL 的限制的簡潔性顯示在表格 6.2 第四行中。[2]

第四類是**可轉換的** (convertible) 限制，某些限制不屬於前面三類限制，然而，如果項目集內的項目用特定的順序排列，此限制式變成單調或反單調。舉例來說，限制式 "$avg(I.price) \le 10$" 既不是單調的，也不是反單調的。然而，如果交易內的項目以價格的遞增順序來添加入項目集內，此限制變成反單調的，因為如果一個項目集 $I$ 違反此限制式 ( 即項目集內的平均價格大於 $10 )，則進一步添加更昂貴的項目至此項目集內，永遠不會讓此項目集滿足限制。同樣地，如果交易內的項目以價格遞減順序來添加入項目集內，則此限制變成單調的，因為如果此項目集滿足此限制 ( 即項目集內的平均價格不大於 $10 )，則加入更便宜的商品項目到此項目集中，將讓此項目集的平均價格仍然不大於$10。除了表格 6.2 中的

[2] 對於限制 $count(S) \le v$ ( 以及相似的 $count(S) \ge v$ )，我們有基於基數的成員產生函數 ( 即，$\{X \mid X \subseteq \text{Itemset} \wedge |X| \le v\}$ )，使用此方法產生的成員具有不同風格，所以稱之弱簡潔性。

**表 6.2** 常用的基於 SQL 樣式修剪限制的特徵

| 限制 | 反單調的 | 單調的 | 簡潔的 |
|---|---|---|---|
| $v \in S$ | 否 | 是 | 是 |
| $S \supseteq V$ | 否 | 是 | 是 |
| $S \subseteq V$ | 是 | 否 | 是 |
| $min(S) \le v$ | 否 | 是 | 是 |
| $min(S) \ge v$ | 是 | 否 | 是 |
| $max(S) \le v$ | 是 | 否 | 是 |
| $max(S) \ge v$ | 否 | 是 | 是 |
| $count(S) \le v$ | 是 | 否 | 弱 |
| $count(S) \ge v$ | 否 | 是 | 弱 |
| $sum(S) \le v \quad (\forall a \in S, a \ge 0)$ | 是 | 否 | 否 |
| $sum(S) \ge v \quad (\forall a \in S, a \ge 0)$ | 否 | 是 | 否 |
| $range(S) \le v$ | 是 | 否 | 否 |
| $range(S) \ge v$ | 否 | 是 | 否 |
| $avg(S)\theta v, \theta \in \{\le, \ge\}$ | 可轉換的 | 可轉換的 | 否 |
| $support(S) \ge \xi$ | 是 | 否 | 否 |
| $support(S) \le \xi$ | 否 | 是 | 否 |
| $All\_confidence(S) \ge \xi$ | 是 | 否 | 否 |
| $All\_confidence(S) \le \xi$ | 否 | 是 | 否 |

“$avg(S) \le v$” 與 “$avg(S) \ge v$” 外，還有許多可轉換的限制，例如 “$variance(S) \ge v$” 與 “$standard\_deviation(S) \ge v$” 等等。

請注意，上述的討論不代表所有限制都是可轉換的，舉例來說，“$sum(S)\theta v$” 其中 $\theta \in \{\le, \ge\}$，而且 $S$ 中每一個元素可以是任意實數，此限制是不可轉換的。因此，還有第五類的限制，稱為**不可轉換的** (inconvertible) 限制。好消息是，雖然存在某些不好處理的不可轉換限制，大多數透過 SQL 聚集來建構的 SQL 表達式，屬於前四類別的限制，可以有效的套用在限制式探勘方法中。

## 使用資料修剪限制來修剪資料空間

在限制式頻繁樣式探勘中，第二種可以修剪搜尋空間的方法是修剪資料空間，如果某資料片段對於後續產生的符合樣式沒有貢獻的話，此策略便將該資料片段修剪，我們考慮兩個性質：資料簡潔性與資料反單調性。

如果在樣式探勘程序開始前，就能夠刪除不滿足此限制的資料子集合，則此限制是**資料簡潔的** (data-succinct)。舉例來說，如果要探勘包含“數位相機”的樣式，則任何沒有包含數位相機的交易資料，便可以在探勘程序開始前，予以刪除，這可以有效地縮減要檢驗的資料集合。

有趣的是，許多限制是**資料反單調性的** (data-antimonotonic)，意旨在探勘過程中，如果基於目前的樣式，一個資料實體不能滿足資料反單調性的限制，則它可予以刪除，因為在後續的探勘過程中，它不會對由目前樣式所產生的超樣式有任何貢獻。

**範例 6.9** 資料反單調性

假設一個探勘查詢要求為 $C_1 : sum(I.price) \geq \$100$ ，也就是，要求在探勘出樣式中，其中項目的總價格不少於 $100 元，假設目前的頻繁項目集 $S$ 並不滿足限制 $C_i$ ( 例如，目前項目集 $S$ 中的商品總價是 $50 )，如果在交易 $T_i$ 中剩餘的頻繁項目為 ( $i_2$ ：價格 = \$5； $i_5$ ：價格 = \$10； $i_8$ ：價格 = \$20)，則 $T_i$ 不會使得 $S$ 滿足此限制，所以， $T_i$ 不會對從 $S$ 探勘出的樣式有所貢獻，因此可以把它修剪掉。

請注意，此修剪不能在探勘開始前執行，因為在那個時間點，我們並不知道納入 $T_i$ 中所有的項目，是否能讓總價格超過 $100 ( 例如， $T_i$ 中可能有 $i_3$ ：價格 = \$80 )，然而，在疊代式探勘程序中，我們可能發現某項目 ( 例如 $i_3$ ) 並不是頻繁的，所以它們應該被刪除。因此，此檢查與修剪方法應該在每一次疊代時實施，以修剪資料搜尋空間。

請注意，限制 $C_1$ 對於樣式空間修剪是單調的，如同我們所見到的，此限制式在修剪樣式搜尋空間的能力是非常有限的，然而，同樣的限制式，可有效地修剪資料搜尋空間。

對於反單調性限制，例如 $C_2 : sum(I.price) \leq \$100$，我們可以同時修剪樣式與資料搜尋空間，根據我們在樣式修剪的研究，我們已經知道，如果目前項目集內的總價格超過 $100 的話，此項目集應該予以刪除 ( 因為在此項目集做進一步的擴展，也不會滿足限制 $C_2$ )。在同一時間，我們也可以將交易 $T_i$ 中不能符合限制 $C_2$ 的剩餘項目，給予以刪除。舉例來說，如果目前項目集 $S$ 中的商品項目總價為 $90，則交易 $T_i$ 中任何價格超過 $10 的其餘頻繁項目都應當被刪除，如果交易 $T_i$ 中沒有剩餘項目能夠讓此限制成立，則整個交易 $T_i$ 都應當被刪除掉。

考慮既非是反單調，也非是單調的樣式限制，例如 "$C_3 : avg(I.price) \leq \$10$"，它們有可能是資料反單調的 (data-antimonotonic)，因為如果交易 $T_i$ 中剩餘的項目不能使得此限制成立，則交易 $T_i$ 也可被刪除。因此，資料反單調性限制對於限制式資料空間修剪是非常有幫助的。

注意，藉由資料反單調性來修剪搜尋空間，僅侷限於樣式增長 (pattern-growth) 的探勘演算法，因為要決定是否應刪除一個資料實體，取決於它是否對特定樣式有所貢獻。如果使用 Apriori 演算法，則不可使用資料反單調性來修剪資料空間，因為資料與目前所有活耀的樣式相關聯，在每一次疊代中，通常會有許多活耀的樣式，不能對給定樣式的超樣式有所貢獻的資料實體，仍然有可能對其他活耀樣式的超樣式有所貢獻。所以，對於非樣式增長的演算法而言，資料空間修剪方法的能力是非常侷限的。

## 6.4 探勘高維度資料與巨型樣式

迄今為止，我們所探討的頻繁樣式探勘方法，都侷限在處理維度數量較少的資料集合，然而，某些應用可能需要探勘高維度資料 ( 即，擁有數百或數千個維度 )，我們可以用之前探討的方法來探勘高維度資料集嗎？

不幸的是，答案是否定的，因為那些典型探勘方法的搜尋空間會隨著維度數目增加而指數成長。

研究專家沿著兩個方向來克服此困難，其中一個方向是進一步藉由探索使用垂直資料格式 (vertical data format)，來擴充樣式增長式 (pattern-growth) 方法，以處理擁有大量維度 ( 也稱特徵、或項目，例如基因 )，但擁有少量的列 ( 也稱交易、或值組，例如，樣本 ) 的資料集。這對於許多應用都是很有用的，例如分析生物資訊中的基因表達 (gene expression) 資料，其中我們通常需要分析微陣列資料，它包含大量的基因 ( 例如，10,000 至 100,000 個 )，但只有少量的樣本 ( 例如，100 至 1000 )。另一個方向是發展新的探勘方法，稱為**樣式融合** (pattern-fusion)，它探勘**巨型樣式** (colossal pattern)，也就是，長度非常長的樣式。

讓我們簡略地檢驗第一個方向，特別是樣式增長式列枚舉方法，其基本原理是探索垂直資料格式，如同 5.2.5 節所描述的，也稱為列枚舉 (row enumeration) 方法。列枚舉方法不同於傳統的行 ( 即，項目 ) 枚舉方法 ( 也稱為水平資料格式 )。在傳統行枚舉方法中，資料集 $D$ 視為列 (row) 的集合，其中每一列包含一個項目集。在列枚舉方法中，相反地，資料集 $D$ 視為一個項目集，每一個包含一組 row_ID，代表此項目出現在 $D$ 中傳統視圖的位置。原始資料集 $D$ 可以輕易的轉換至轉置 (transposed) 資料集 $D$，擁有列數較少，但是含有大量維度的資料集，便可轉換成擁有較少維度與大量列數的轉置資料集。可以在此相對低維度的資料集上發展有效的樣式增長方法，此方法的詳盡細節留給有興趣的讀者做為習題。

本節剩餘的部分聚焦在第二個方向，我們介紹一種新的探勘方法，稱為樣式融合，它探勘巨型樣式 ( 即，非常長的樣式 )，此方法在樣式搜尋空間中跳耀，得到巨型頻繁樣式完整集合很好的近似解。

## 藉由樣式融合來探勘巨型樣式

雖然我們已經研究在各種情況下探勘頻繁樣式的方法，但是，許多應用問題含有難以探勘的隱藏樣式，主要歸因於它們巨大的長度或尺寸。考

慮生物資訊學為例子，其常見的任務是分析微陣列與 DNA 資料，這涉及映射與分析非常長的 DNA 或蛋白質序列，相較於找出小型樣式，研究者對於找出大型樣式 ( 即，長序列 ) 是更感興趣的，因為較長的樣式通常攜帶更多重要的意義，我們稱這些大型樣式為巨型樣式 (colossal pattern)，以和具有大型支持集合的樣式做區隔。發現大型樣式是一件艱困的挑戰，因為在找出候選的大型樣式之前，遞增式探勘會趨向於被爆炸性數量的中型樣式給困住而無法前進，此問題將在範例 6.10 中闡述。

**範例 6.10** **探勘巨型樣式的挑戰**

考慮一個 40×40 的表格，每一列包含整數 1 至 40，以遞增順序排列，接著移除在對角線上的數字，得到一個 40×39 的表格，然後增加 20 個相同的列到此表格的底部，每一個列包含整數 41 至 79，以遞增順序排列，得到一個 60×39 的表格 ( 如圖 6.6 )。我們將每一列考慮為一筆交易，並設定最小支持計數門檻值為 20，此表格包含長度為 20 的中型封閉 / 最大頻繁項目集的數量是指數的 ( 即，$\binom{40}{20}$ 個 )，但是它只有一個長度為 39 的巨型樣式，即 $\alpha = (41, 42, \ldots, 79)$，目前我們所介紹的頻繁樣式探勘演算法都無法在合理的時間內執行完成，其樣式搜尋空間如圖 6.7 所示，其中型樣式數量大過於巨型樣式。

迄今為止，我們所探討的樣式探勘策略，例如 Apriori 或 FP-growth 演算法，其天性都是採用遞增成長的策略，每一次將候選集合的長度增加 1。像 Apriori 這種寬先搜尋的方法，無法避免要產生大量的中型樣式，使得它不可能到達巨型樣式，即便是像 FP-growth 這種深先搜尋的方法，也很容易在到達巨型樣式之前，陷入龐大數量的子樹的陷阱。明顯地，需要發展一種全新的探勘方法來克服此障礙。

一種稱為樣式融合 (pattern-fusion) 的新探勘策略因應而生，它將少量的較短樣式融合，來產生巨型樣式。因此，它在樣式搜尋空間中跳耀，避

| row/col | 1 | 2 | 3 | 4 | ... | 38 | 39 |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | ... | 39 | 40 |
| 2 | 1 | 3 | 4 | 5 | ... | 39 | 40 |
| 3 | 1 | 2 | 4 | 5 | ... | 39 | 40 |
| 4 | 1 | 2 | 3 | 5 | ... | 39 | 40 |
| 5 | 1 | 2 | 3 | 4 | ... | 39 | 40 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 39 | 1 | 2 | 3 | 4 | ... | 38 | 40 |
| 40 | 1 | 2 | 3 | 4 | ... | 38 | 39 |
| 41 | 41 | 42 | 43 | 44 | ... | 78 | 79 |
| 42 | 41 | 42 | 43 | 44 | ... | 78 | 79 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 60 | 41 | 42 | 43 | 44 | ... | 78 | 79 |

**圖 6.6** 簡單的巨型樣式範例，此資料集包含指數數量的中型樣式（長度為 20），而只有一個巨型樣式，即為 (41, 42, ..., 79)。

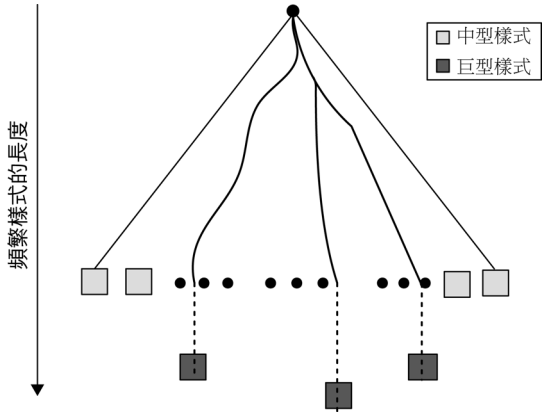


**圖 6.7** 在此人造資料集中，包含指數數量的中型樣式，但只包含一些巨型樣式。

免了寬先搜尋與深先搜尋中的陷阱，此方法可以找到巨型樣式完整集合的很好近似解答。

樣式融合方法有以下主要特徵，首先，他以有限寬度 (bounded-breadth) 的方式在樹中搜尋，只有在有限尺寸候選池 (pool) 中的固定數量的樣式，可以做為起始點，往下搜尋樣式樹。這樣，它避免了指數搜尋空間的問題。

第二點，樣式融合方法具有判別可能的捷徑的能力，每一次的樣式增長，不是透過添加一個項目來執行，而是將池中數個樣式聚集而成，這些捷徑領導樣式融合法能更加快速的沿著搜尋樹往下到達巨型樣式，圖 6.8 概念性的闡述此探勘模式。

由於樣式融合方法是設計來產生巨型樣式的近似解，需要導入一個品質評估模式，來衡量演算法傳回的樣式。研究實驗證實樣式融合方法能夠有效地傳回高品質結果。

讓我們更加詳細的檢視資料融合方法，首先，我們介紹核心樣式 (core pattern) 的概念，對於一個樣式 $\alpha$，項目集 $\beta \subseteq \alpha$ 被稱為 $\alpha$ 的 $\tau$-核心模式，如果 $|D_\alpha|/|D_\beta| \geq \tau$，$0 < \tau \leq 1$，其中 $|D_\alpha|$ 是資料集 $D$ 中包含 $\alpha$ 的樣式數目，$\tau$ 稱為核心比率 (core ratio)，樣式 $\alpha$ 是 $(d, \tau)$-強健的，如果 $d$ 是項目的最大數目，使得將這些項目從 $\alpha$ 中移除後，所得到的樣式仍然是 $\alpha$ 的 $\tau$-核心模式，也就是說，

$$d = \max_{\beta}\{|\alpha| - |\beta| \,\big|\, \beta \subseteq \alpha \quad \text{且} \quad \beta \text{是} \alpha \text{的} \tau\text{-核心樣式}\}$$

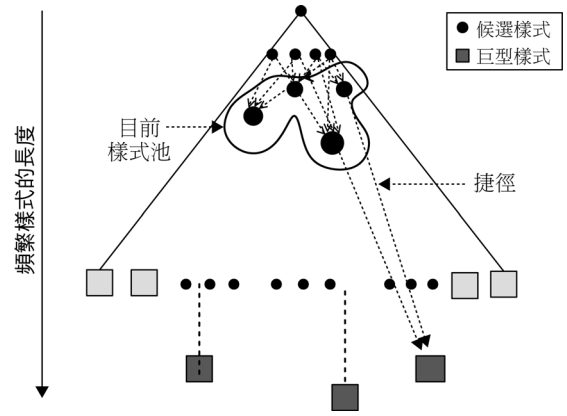


**圖 6.8** 樣式樹搜尋：從樣式池中取出候選者，得到在樣式空間中通往巨型樣式的捷徑。

**範例 6.11** 核心樣式

圖 6.9 顯示一個簡單的交易資料集，它包含 4 個不同的交易，每個重複 100 次：$\{\alpha_1 = (\text{abe}), \alpha_2 = (\text{bcf}), \alpha_3 = (\text{acf}), \alpha_4 = (\text{abcfe})\}$，如果我們設定 $\tau = 0.5$，則 (ab) 是 $\alpha_1$ 的核心樣式，因為 (ab) 僅被 $\alpha_1$ 與 $\alpha_4$ 所包含，所以 $|D_{\alpha_1}|/|D_{(ab)}| = 100/200 \geq \tau$。$\alpha_1$ 是(2, 0.5)-強健的，而 $\alpha_4$ 是(4, 0.5)-強健的，此表格同時顯示較長的樣式 ( 例如，(abcfe) ) 比較短的樣式 ( 例如，(bcf) ) 包含更多的核心樣式。

從範例 6.11 中，我們可以推斷出大型或巨型樣式比小型樣式擁有更多的核心樣式，所以巨型樣式是更為強健的，意旨為如果從樣式中移除少量的項目，所得到的樣式會擁有相似的支持集，樣式的尺寸越大，它的強建性越明顯。巨型樣式與它對應的核心樣式之間的強健性關係可以延伸到多層級架構下，巨型樣式在較低層級下的核心樣式，稱為**核心後代** (core descendants)。

給定一個較小的 c 值，相較於小型樣式，巨型樣式通常擁有更多長度為 c 的核心後代，這代表如果我們從長度為 c 的完整樣式集合中隨機抽取的話，我們有較多的可能選取到巨型樣式的核心後代，而非是小型樣式的核心後代。在圖 6.9 中，考慮長度為 2 的完整樣式集合，它總共包含 $\binom{5}{2} = 10$ 個樣式，為了方便闡述，我們假設 abcef 為巨型樣式，隨機挑選到為 abcef 的核心後代的機率是 0.9，相反地，挑選到小型 ( 非巨型 ) 樣式的核心後代的機率至多為 0.3。因此，可以藉由適當地合併它的核心樣式來得到巨型樣式，舉例來說，abcef 可以透過合併它的核心樣式 ab 與 cef 來得到，而不用去合併它全部的 26 個核心樣式。

現在，讓我們來看看這些觀察如何幫助我們在樣式空間中跳耀，更直接地到達巨型樣式。考慮以下方案，首先，依據使用者指定的尺寸，產生不超過該尺寸的頻繁樣式的完整集合，然後隨機選取一個樣式 $\beta$，$\beta$ 有很

| 交易（交易數量） | 核心樣式 ($\tau = 0.5$) |
|---|---|
| (abe)(100) | (abe), (ab), (be), (ae), (e) |
| (bcf)(100) | (bcf), (bc), (bf) |
| (acf)(100) | (acf), (ac), (af) |
| (abcef)(100) | (ab), (ac), (af), (ae), (bc), (bf), (be), (ce), (fe), (e),<br>(abc), (abf), (abe)(ace), (acf), (afe), (bcf), (bce), (bfe), (cfe),<br>(abcf), (abce), (bcfe), (acfe), (abfe), (abcef) |

**圖 6.9** 交易資料集：包含重複的交易，與每一個不同交易的核心樣式集合。

高的機率是屬於某個巨型樣式 $\alpha$ 的核心後代。在此完全集合中，判別出所有 $\alpha$ 的核心後代，並且合併它們，將會產生更大的 $\alpha$ 的核心後代，給我們有能力在核心樣式樹 $T_\alpha$ 中沿著通往 $\alpha$ 的路徑往下跳躍。使用相同的方式，我們選取 $K$ 個樣式，所產生的較大核心後代集合，是下一次疊代的候選池。

現在有一個問題：給定一個巨型樣式 $\alpha$ 的核心後代 $\beta$，我們要如何找出 $\alpha$ 的其他核心後代？給定兩個樣式 $\alpha$ 與 $\beta$，它們之間的樣式距離定義為 $Dist(\alpha,\beta) = 1 - |D_\alpha \cap D_\beta| / |D_\alpha \cup D_\beta|$，樣式距離滿足三角不等式。

對於一個樣式 $\alpha$，令 $C_\alpha$ 為他所有核心樣式的集合，我們可以證明 $C_\alpha$ 被尺度空間中的一個球圍住，其中球的直徑為 $r(\tau) = 1 - 1/2\sqrt{\tau - 1}$，這代表給定一個核心樣式 $\beta \in C_\alpha$，我們可以藉由提出一個範圍查詢，來判別出所有在目前候選池中 $\alpha$ 的核心後代，注意，在探勘演算法中，每一個隨機選取的樣式，可能不僅是單一個巨型樣式的核心後代，因此，藉由合併在球中發掘的樣式，可以不只有一個的較大的核心後代會產生。

根據這些討論，樣式融合方法分成下列兩個階段：

1. **池初始化**：樣式融合方法假設一個由小型頻繁樣式組成的初始池是可用的，這是小型頻繁樣式（例如，尺寸小於 3）的完整集合，此初始池可以由任何現存的探勘演算法來挖掘出。

2. **疊代式樣式融合**：樣式融合採取使用者指定的參數 $K$ 做為輸入，他是要探勘出樣式的最大數目，此探勘程序是疊代式的，在每一次疊代時，從目前池中隨機選取 $K$ 個種子樣式 (seed pattern)。對每一個種子樣式，我們找出在直徑由 $\tau$ 決定的球內的所有樣式，在每一個球中的所有樣式接著融合在一起，來產生一組超樣式集合。這些超樣式集合組成新的池，如果池包含超過 $K$ 個樣式，下一次疊代從池開始新一輪的隨機挑選。隨著每一次疊代，每一個超樣式的支持集收縮，此疊代程序終止。

注意，樣式融合方法將小型樣式合併成為大型樣式，而不是遞增地添加單一項目來擴充樣式，這讓此方法有一個優點，它可以繞過中型樣式，沿著可能的路徑往巨型樣式前進。我們在圖 6.10 中闡述此概念，在尺度空間中的每一個點，代表一個核心樣式。相較於小型樣式，大型樣式有更多的核心樣式，它們彼此接近，所有核心樣式皆被一個球圍住，此球用虛線表示。當從初始樣式池中隨機選取時，我們有更大的機率來得到大型樣式的核心樣式，因為大型樣式的球是更為密集得多了。

理論已證明樣式融合能得到巨型樣式很好的近似解，此方法已在人造與現實資料 ( 程式追蹤資料與微陣列資料 ) 中測試過，實驗證實此方法能很有效率的發掘大部分的巨型樣式。

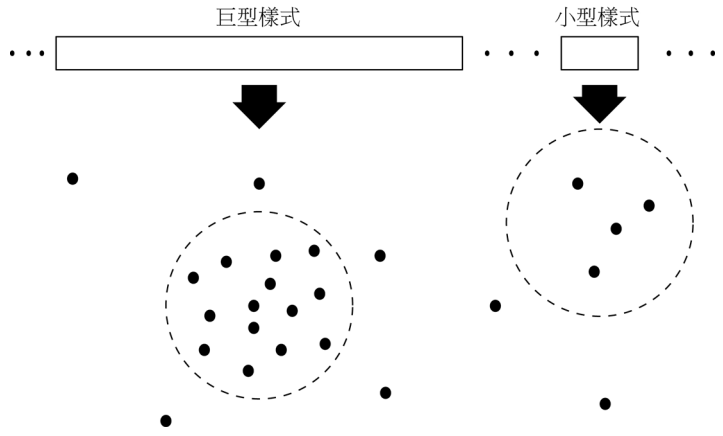


**圖 6.10** 樣式尺度空間：每一個點代表一個核心樣式，相較於小型樣式，巨型樣式內的核心樣式是比較稠密的，我們用虛線來表示。

## 6.5 探勘壓縮或近似樣式

頻繁樣式探勘的一項艱鉅挑戰是所發掘的樣式數量十分龐大，使用最小支持度門檻值來控制樣式的數目，它的效果是有限的。如果門檻值過低，會產生爆炸性多量的輸出樣式，如果門檻值過高，則會導致只能發掘出常識性的樣式。

為了縮減在探勘過程中產生龐大數量的頻繁樣式，同時維持高品質的樣式，我們可以轉而由探勘壓縮或近似樣式來代替。有學者提出 Top-$k$ 最頻繁封閉樣式探勘方法，使得探勘程序僅著重在前 $k$ 個最頻繁樣式，雖然它們是令人感興趣的，但它們通常不是象徵 $k$ 個最有代表性的樣式，因為這些樣式的頻率分佈並不均勻。限制式頻繁樣式探勘 ( 課本 6.3 節 ) 融入使用者指定的限制來過濾掉不感興趣的樣式，量測樣式 / 規則的有趣度與相關度 ( 課本 5.3 節 )，也能用來侷限搜尋有趣的樣式 / 規則的空間。

在本節中，我們介紹兩種頻繁樣式的“壓縮”格式，這是根據封閉樣式與最大樣式的概念來建構的，回顧在 5.2.6 節所提到，封閉樣式是頻繁樣式集的無失真壓縮 (lossless compression)，而最大樣式是頻繁樣式集的有失真壓縮 (lossy compression)。具體來說，6.5.1 節探索分群式壓縮頻繁樣式，它根據樣式之間的相似度與支持頻率，來將樣式群組在一起。6.5.2 節使用“匯總”方式，致力於推導出冗餘感知 top-$k$(redundancy-aware top-$k$) 代表性樣式，來涵蓋整個 ( 封閉 ) 頻繁樣式集合，此方法不僅考慮樣式的代表性，還考慮它們的相互獨立性，以避免產生的樣式集合具有冗餘性。此 $k$ 個具有代表性的樣式提供整個頻繁樣式集合的緊密壓縮，使得它們更容易解讀與使用。

### 6.5.1 藉由樣式分群來探勘壓縮樣式

樣式壓縮可以藉由樣式分群來達成，分群技術將在課本第 9 與 10 章詳細介紹，在此節中，你不需要知道分群技術的詳細內容，相對地，你僅需

要知道如何套用分群的概念來壓縮樣式。分群 (clustering) 是將相似物件群組在一起的自動程序，所以在同一群集 (cluster) 內的物件是彼此相似的，而在不同群集間的物件彼此不相似。在現在的情況下，頻繁樣式就是物件，頻繁樣式集合透過 $\delta$-群集 ($\delta$-cluster) 緊密度量測來進行分群，並對每一個群集挑選一個代表該群集的樣式，因此，提供頻繁樣式集合的壓縮版本。

在我們開始前，先讓我們回顧一些定義，在資料集 $D$ 中，項目集 $X$ 是封閉的 (closed) 頻繁項目集，如果 $X$ 是頻繁的，而且不存在 $X$ 的超樣式集 (super-itemset) $Y$ 使得 $Y$ 與 $X$ 在資料集 $D$ 中具有相同的支持計數。項目集 $X$ 是最大的 (maximal) 頻繁項目集，如果 $X$ 是頻繁的，而且不存在超樣式集 $Y$，使得 $X \subset Y$，而且 $Y$ 在資料集 $D$ 上是頻繁的。僅使用這些概念並不足夠得到資料集很好的壓縮表示法，讓我們來看下面的範例 6.12。

**範例 6.12 使用封閉與最大樣式來壓縮的缺點**

表格 6.3 顯示大型資料集上的頻繁項目集的子集合，其中 a, b, c, d, e, f 代表個別的項目，此處並沒有封閉項目集，所以我們無法使用封閉頻繁項目集來壓縮資料，唯一的最大項目集為 $P_3$，但是，我們觀察到 $P_2$、$P_3$ 與 $P_4$ 這些項目集，在支持計數方面有顯著的差異。如果我們使用 $P_3$ 來代表資料的壓縮版本，我們將整個失去支持計數所攜帶的資訊。從視覺化的觀點，考慮底下兩對樣式 $(P_1, P_2)$ 與 $(P_4, P_5)$，在每對中的樣式，其在支持度與表達式方面都非常類似。因此，直覺地，可以挑選 $P_2$、$P_3$ 與 $P_4$ 做為資料集更佳的壓縮版本。

所以，讓我們來看能否找到對頻繁樣式分群的方法，以做為得到它們壓縮表示的一種手段。我們將需要定義一個良好的相似度量測，以根據此量測來對樣式分群，並接著選取與輸出每一群集的代表性樣式。由於封閉頻繁樣式集合是原始頻繁樣式集的無失真壓縮，所以在封閉頻繁樣式集合中發掘代表性樣式，是一個很好的想法。

**表 6.3** 頻繁項目集的子集合

| ID | 項目集 | 支持計數 |
|---|---|---|
| $P_1$ | {b,c,d,e} | 205,277 |
| $P_2$ | {b,c,d,e,f} | 205,211 |
| $P_3$ | {a,b,c,d,e,f} | 101,758 |
| $P_4$ | {a,c,d,e,f} | 161,563 |
| $P_5$ | {a,c,d,e} | 161,576 |

我們可使用以下對封閉樣式之間的距離量測，令 $P_1$ 與 $P_2$ 為兩個封閉樣式，它們的支持交易集分別為 $T(P_1)$ 與 $T(P_2)$，則 $P_1$ 與 $P_2$ 的樣式距離 (pattern distance) $Pat\_Dist(P_1, P_2)$ 定義為

$$Pat\_Dist(P_1,\ P_2) = 1 - \frac{|T(P_1) \cap T(P_2)|}{|T(P_1) \cup T(P_2)|} \qquad \textbf{(6.14)}$$

樣式距離是定義在交易集合上的正當距離尺度，注意，它包含了支持度的資訊，正如我們之前所期望的。

**範例 6.13** 樣式距離

令 $P_1$ 與 $P_2$ 為兩個樣式，使得 $T(P_1) = \{t_1, t_2, t_3, t_4, t_5\}$ 與 $T(P_2) = \{t_1, t_2, t_3, t_4, t_6\}$，其中 $t_i$ 為資料庫中的交易，則 $P_1$ 與 $P_2$ 之間的距離為 $Pat\_Dist(P_1, P_2) = 1 - 4/6 = 1/3$

現在，讓我們考慮樣式的表達式，給定兩個樣式 $A$ 與 $B$，我們說 $B$ 可以被 $A$ 表達，如果 $O(B) \subset O(A)$，其中 $O(A)$ 是樣式 $A$ 對應的項目集。根據此定義，假設樣式 $P_1, P_2, \ldots, P_k$ 在同一個群集中，此群集的代表性樣式 $P_r$ 應該要能夠表達群集中的其它樣式，顯然，我們有 $\bigcup_{i=1}^{k} O(P_i) \subseteq O(P_r)$。

使用距離量測，我們可以簡單地在頻繁樣式集合上套用分群演算法 ( 例如 10.2 節的 $k$-means 分群演算法 )，然而，這衍伸出兩個問題，第一個是，群集的品質沒有辦法保證，第二個是，可能沒辦法為每一個群集找

出代表性樣式 ( 即，樣式 $P_r$ 可能不屬於同一個群集 )，為了克服這些困難，所以有了 $\delta$-**群集** ($\delta$-cluster) 的概念，其中 $\delta(0 \le \delta \le 1)$ 量測群集的緊密程度。

樣式 $P$ 是被另一個樣式 $P'$ 給 $\delta$-**涵蓋**(δ-coverd)，如果 $O(P) \subseteq O(P')$ 而且 $Pat\_Dist(P, P') \le \delta$。一組樣式集合組成 $\delta$- 群集，如果存在一個代表性樣式 $P_r$，使得在此集合中的每一個樣式 $P$ 皆是被 $P_r$ 給 $\delta$- 涵蓋。

注意，根據 $\delta$- 群集的定義，一個樣式可能屬於多個群集，而且，使用 $\delta$- 群集，我們只需要計算每一個樣式與群集的代表樣式之間的距離，由於只有當 $O(P) \subseteq O(P_r)$ 時，樣式 $P$ 是被 $P_r$ 給 $\delta$- 涵蓋，所以我們可以簡化距離計算公式，只考慮樣式的支持計數：

$$Pat\_Dist(P, P_r) = 1 - \frac{|T(P) \cap T(P_r)|}{|T(P) \cup T(P_r)|} = 1 - \frac{|T(P_r)|}{|T(P)|} \qquad \textbf{(6.15)}$$

如果我們限制代表性樣式必須是頻繁的，則代表性樣式的數目 ( 即，群集的數目 ) 是不少於最大頻繁樣式的數目，這是因為最大頻繁樣式只能被自己給涵蓋。為了能夠更簡潔的壓縮，我們可以放寬對代表性樣式的限制，也就是說，代表性樣式的支持計數可以稍微地小於 min_sup。

對於每一個代表性樣式 $P_r$，假設他的支持計數為 $k$，由於它涵蓋至少一個頻繁樣式 ( 即 $P$ )，且此樣式的支持計數不少於 min_sup，所以我們有

$$\delta \ge Pat\_Dist(P, P_r) = 1 - \frac{|T(P_r)|}{|T(P)|} \ge 1 - \frac{k}{\min\_sup} \qquad \textbf{(6.16)}$$

也就是說，$k \ge (1-\delta) \times \min\_sup$，這是代表性樣式的最小支持度，記作 $\min\_sup_r$。

基於以上的討論，樣式壓縮問題可以定義如下：給定一個交易資料庫、最小支持度門檻值 min_sup 與群集品質量測 $\delta$，樣式壓縮問題是要找一組代表性的樣式集合 $R$，使得對每一個頻繁樣式 $P$ ( 根據 min_sup )，存在一個代表性樣式 $P_r \in R$ ( 根據 $\min\_sup_r$ ) 它能涵蓋 $P$，而且能最小化 $|R|$ 值。

找出這樣一個最小集合是 NP-hard 問題，然而，已經開發出一些有效的方法，相較於原始的封閉樣式集合，它能夠讓產生的封閉頻繁樣式數量減少數個量級，這些方法能成功地找出樣式集合的高品質壓縮結果。

## 6.5.2 萃取冗餘感知 top-*k* 樣式

在探勘過程中，探勘 top-*k* 樣式是能縮減回傳樣式數目的策略。然而，在許多情況下，頻繁樣式並非彼此獨立，而是經常群集在一個小區域內。這有點像找出全世界人口聚集的前 20 個中心城市，所找出來的城市可能會群聚在少數某些國家中，而不是均勻的分佈在全球各地。相對地，大部分使用者更有意願找出前 $k$ 個最有趣的樣式，不僅是重要，而且還相互獨立，含有很少的冗餘性。這 $k$ 個具代表性的小型樣式集合稱為**冗餘感知 top-k 樣式** (redundancy-aware top-k pattern)，它們不僅高度重要，而且冗餘性很低。

**範例 6.14** 冗餘感知 top-*k* 策略與其他 top-*k* 策略

圖 6.11 闡述隱含在冗餘感知 top-*k* 樣式、傳統 top-*k* 樣式與 *k*-概括樣式背後的直覺概觀，假設我們有圖 6.11(a) 所示的頻繁樣式集合，其中每一個圓形代表一個樣式，而其灰階的顏色代表它的重要性，兩個圓形之間的距離代表兩個對應的樣式之間的冗餘性，兩個圓形的距離越接近，代表對應的樣式對另一個就越冗餘的。假設我們想要選取 3 個最能代表此集合的樣式，即 $k=3$，我們應該選哪 3 個？

我們使用箭頭來指示使用冗餘感知 top-*k* 樣式 ( 圖 6.11(b) )、傳統 top-*k* 樣式 ( 圖 6.11(c) ) 與 *k*-概括樣式 ( 圖 6.11(d) ) 方法所選取的樣式，在圖 6.11(c) 中，傳統 top-*k* 策略僅依靠顯著性，它選取最顯著的三個樣式來代表該集合。在圖 6.11(d) 中，*k*-概括樣式策略僅依靠非冗餘性來選取樣式，並且找出每個群集的“最中心”樣式做為代表性樣式，以它們來代表該集合。所選取的樣式視為“概括樣式”，因為其意義

上，是代表它所處的該群集，或為該群集提供“概括摘要資訊”。

相較之下，圖 6.11(b) 的冗餘感知 top-$k$ 樣式在顯著性與冗餘性之間取捨以獲得均衡，舉例來說，請觀察到有兩個高顯著性的樣式是彼此接近的 ( 根據它們的冗餘性 )，冗餘感知 top-$k$ 策略僅選取其中一個，因為考慮到兩個都選，則另一個會是冗餘的。為了能夠形式化定義冗餘感知 top-$k$ 樣式，我們需要定義顯著性與冗餘性的概念。

顯著性量測 $S$ 是將樣式 $p \in P$ 映射至實數值 $S(p)$ 的函數，使得 $S(p)$ 能代表樣式 $p$ 的有趣性 ( 實用性 ) 程度。一般而言，顯著性量測可以是主觀的或客觀的，客觀量測僅取決於給定樣式的結構與挖掘過程所使用的資料集，常用的客觀量測包含支持度、信賴度、相關性與 tf-idf，其中後者常用於資訊檢索 (information retrieval)。主觀量測是根據使用者對資料的

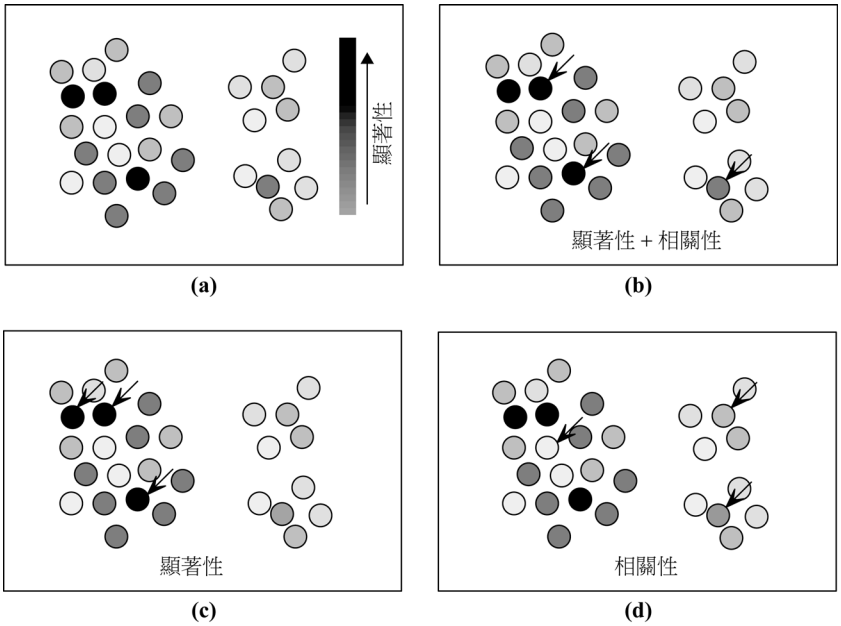


**圖 6.11** 比較 top-$k$ 方法的概念觀點，其中每一個圓形代表一個樣式，其灰階的顏色代表它的重要性，兩個樣式的距離越接近，代表該樣式對另一個就越冗餘：(a) 原始樣式 (b) 冗餘感知 top-$k$ 樣式 (c) 傳統 top-$k$ 樣式 (d) $k$-概括樣式。

信念，因此它們取決於檢驗樣式的使用者，主觀量測通常是根據使用者的先驗知識或背景模式的相對分數，它通常根據樣式與背景模式的偏移程度來量測樣式的非預期性。令 $S(p,q)$ 為樣式 $p$ 與 $q$ 的**聯合顯著性** (combined significance)，而 $S(p|q)=S(p,q)-S(q)$ 為給定 $q$ 下之 $p$ 的**相對顯著性** (negatively significance)，請注意，$S(p,q)$ 量測樣式 $p$ 與 $q$ 的共同顯著性，而不是單個超模式 $p \cup q$ 的顯著性。

給定顯著性量測 $S$，兩個樣式 $p$ 與 $q$ 之間的**冗餘性**(redundancy) $R$ 定義為 $R(p,q)=S(p)+S(q)-S(p,q)$，因此，我們有 $S(p|q)=S(p)-R(p,q)$。

我們假設兩個樣式的聯合顯著性不少於任何單一樣式的顯著性 ( 因為它是兩個樣式的共同顯著性 )，並且不超過兩個樣式的顯著性的總和 ( 因為它們之間存在冗餘性 )，也就是說，兩個樣式之間的冗餘性應當滿足

$$0 \le R(p,q) \le \min(S(p),S(q)) \qquad \textbf{(6.17)}$$

理想的冗餘量測 $R(p,q)$ 通常是難以得到，然而，我們可以使用兩個樣式之間的距離 ( 諸如 6.5.1 節定義的樣式距離 ) 來近似冗餘度。

冗餘感知 top-$k$ 樣式的問題，可以轉換成找出具有最大邊際顯著性 (marginal significance) 的 $k$ 個樣式，這問題已經在資訊檢索領域中研究透測了，在該領域中，如果一份文件與查詢語相關，而且它與先前選取的文件具有很小的邊際相似度 ( 其中邊際相似性是由選取最相關的文件來計算 )，則此文件具有高度邊際相關性。實驗證實此方法能夠有效的找出高顯著性與低冗餘性的 top-$k$ 樣式。

## 6.6 總結

- 頻繁樣式探勘的研究範圍，已遠遠超過第 5 章所介紹探勘頻繁項目集與關聯規則的基本概念與方法。本章節呈現該領域的路觀圖，其主題是根據探勘樣式與規則的類型、探勘方法與應用問題來組織。
- 除了探勘基本頻繁項目集與關聯規則，也可以探勘以下的**進階樣式形式**，諸如多層級關聯規則與多維度關聯規則、量化關聯規則、罕見規則

與負向規則，我們也可以發掘出高維度樣式、壓縮或近似樣式。

- **多層級關聯規則** (multilevel association) 涉及在多層抽象層級 ( 例如，“購買電腦”或“購買膝上型電腦”) 上的資料，它們可以使用多個最小支持度門檻值來探勘。**多維度關聯規則** (multidimensional association) 包含超過一個維度，探勘這些關聯規則的技術依據處理重複謂詞的方法而有所不同。**量化關聯規則** (quantitative association rule) 涉及數量化屬性，離散化、分群法與統計分析等能夠揭露異常行為的方法，可以整合至樣式探勘程序。
- **罕見樣式** (rare pattern) 是指那些很少出現，但是特別有趣的樣式。**負向樣式** (negative pattern) 的成員顯示出負相關行為，定義負向樣式應該小心謹慎，並考慮 null-invariance 的性質。罕見樣式與負向樣式可能突顯出資料中異常行為，這可能是很有趣的。
- **限制式探勘** (constraint-based mining) 策略能幫助指引探勘程序，前往符合使用者直覺或滿足特定限制的樣式，許多使用者指定的限制，可以嵌入探勘程序當中，限制式可以區分為**樣式修剪** (pattern-pruning) 與**資料修剪** (data-pruning) 限制，限制式的性質包含單調性、反單調性、資料反單調性與簡潔性，具有這些性質的限制式，可以適當地融合入有效的樣式探勘程序中。
- 已開發出在高維度空間中探勘樣式的方法，這包含以列枚舉為基礎的樣式增長方法，來探勘具有大量維度與少數值組的資料集 ( 例如，微陣列資料 )，還有藉由樣式融合方法來探勘**巨型樣式** (colossal patterns) ( 即，非常長的樣式 )。
- 為了縮減探勘程序回傳的樣式數目，我們可以探勘壓縮樣式 (compressed pattern) 或近似樣式 (approximation pattern)。壓縮樣式可以透過以分群概念為基礎所定義的代表性樣式來探勘得到，而近似樣式可以藉由萃取**冗餘感知 top-k 樣式** (redundancy-aware top-k pattern) 來探勘得到，即具有 $k$ 個代表性樣式的小集合，它們不僅具有高度顯著性，而且彼此之間冗餘性很低。


**6.1** 請提出一個層級共享探勘方法的綱要，來探勘多層級關聯規則，其中每一個項目由其層級位置來編碼，請妥善的設計它，使得他能夠在初始掃描資料庫時，蒐集每一個項目在每一個概念層級的計數值，並且判別出頻繁與子頻繁項目。並就處理成本方面，對於探勘多層級關聯規則與單層關聯規則的比較做出評論。

**6.2** 假設你身為連鎖商店的管理者，你想要使用交易銷售資料，來分析你的商店的廣告的效果。特別是，你想要研究哪些特定的因素會影響特定商品項目促銷廣告的效果，想要研究的因素包含顧客居住的區域、廣告撥出的頻率 ( 一周幾天與一天幾次 )。請探討如何設計一個有效的演算法來探勘交易資料集，並解釋多維度與多層級探勘方法如何幫助你推導出好的解答。

**6.3** 量化關聯規則可以發掘出資料集中異常的行為，其中“異常”可以根據統計理論來定義，舉例來說，6.2.3 節顯示下述關聯規則暗示一個異常樣式：

性別 = 女性 ⇒ 平均工資 = $6.90/hr ( 整體平均工資 = $9.02/hr )

此規則指明女性的平均薪資只有每小時 $6.90，這明顯的低於整體的平均工資每小時 $9.02，請討論如何系統化與有效地從擁有量化屬性的大型資料集中發掘量化規則。

**6.4** 在多維度資料分析中，萃取資料方塊中與量測顯著改變相關聯的一對相似的單元，是很有趣的一件事情，其中單元被視為相似的，如果它們在上捲 ( 即，祖先 )、下鑽 ( 即，後代 ) 或 1D 突變 ( 即，堂兄弟 ) 等操作上是相關的，此分析稱為資料方塊梯度分析 (cube-gradient analysis)。

假設方塊中的量測為 average，使用者提出一組探測單元，並想要找出它們對應的梯度單元，它們滿足特定的梯度門檻值。舉例來說，找出平均銷售價格高於給定探測單元 20% 的對應梯度單元的集合，請開發一個能在大型資料方塊中有效

探勘限制式梯度單元集合的演算法。

**6.5** 6.2.4 節呈現多種方法來定義負相關樣式，考慮定義 6.3：“如果項目集 $X$ 與 $Y$ 皆是頻繁的，也就是說，$\sup(X) \geq \min\_sup$ 與 $\sup(Y) \geq \min\_sup$，其中 min_sup 是最小支持度門檻值，如果 $(P(X|Y)+P(Y|X))/2 < \epsilon$，其中 $\epsilon$ 是負向樣式門檻值，則樣式 $X \cup Y$ 是負相關樣式”。請設計一個有效的樣式增長演算法來探勘負相關樣式。

**6.6** 請證明在頻繁樣式探勘中，下述表格中每一個規則限制所對應的特徵是正確無誤的。

| | 限制 | 反單調的 | 單調的 | 簡潔的 |
|---|---|---|---|---|
| (a) | $v \in S$ | 否 | 是 | 是 |
| (b) | $S \subseteq V$ | 是 | 否 | 是 |
| (c) | $min(S) \leq v$ | 否 | 是 | 是 |
| (d) | $range(S) \leq v$ | 是 | 否 | 否 |
| (e) | $variance(S) \leq v$ | 可轉換的 | 可轉換的 | 否 |

**6.7** 商店中每個商品項目的價格都是非負的，商店經理只對特定形式的規則有興趣，那些規則須滿足 (a) ~ (d) 的限制。請對每一種情況，判別它們限制的類型，並簡略的探討如何使用限制式樣式探勘，來挖掘出這樣的關聯規則。

(a) 包含至少一個藍光 DVD 電影。

(b) 包含商品的總價格是低於 $150。

(c) 包含一個免費商品，而且其它商品的總價格至少為 $200。

(d) 所有商品的平均價格是介於 $100 與 $500 之間。

**6.8** 課本 6.4.1 節介紹使用樣式融合方法來探勘高維度資料，請解釋，為何如果資料集中存在巨型樣式時，它們較容易被此方法發掘出。

**6.9** 課本 6.5.1 節定義封閉樣式 $P_1$ 與 $P_2$ 之間的樣式距離 (pattern distance) 量測為

$$Pat\_Dist(P_1, P_2) = 1 - \frac{|T(P_1) \cap T(P_2)|}{|T(P_1) \cup T(P_2)|}$$

其中 $T(P_1)$ 與 $T(P_2)$ 分別為 $P_1$ 與 $P_2$ 的支持交易集，他是否為有效的距離量測尺

度？請顯示推導過程來支持你的答案。

**6.10** 關聯規則探勘通常會產生大量的規則，其中有許多規則是相似的，所以沒有包含太多新穎的資訊，請設計一個有效的演算法來將大量的規則，壓縮到一個小型的緊密集合。並探討在使用不同樣式相似度定義下，你的方法是否為強健的？

**6.11** 頻繁樣式探勘可能產生大量的樣式，因此，開發能探勘壓縮樣式的方法是很重要的一件事情，假設你只想要得到 $k$ 個樣式（$k$ 是一個小整數），請規劃一個有效的方法來產生 $k$ 個最具有代表性的樣式，其中傾向選取越不相同的樣式，而不偏好選取相似的樣式。使用一個小型資料集來闡述你的方法的有效性。