正如電腦科學領域中一句有名的諺語「Garbage in, garbage out.（垃圾進，垃圾出）」，如果把錯誤或無意義的資料輸入系統，那麼系統輸出的也會是錯誤或無意義的結果，因此，經由資料預處理來確保資料的品質是非常重要的一個步驟。

3 資料轉換（data transformation）

因爲要對不同類型的資料集進行不同的探索，而且現今的資料集可能都相當大，所以必須將資料轉換成適合接下來資料挖掘演算法處理的格式，如此也會有助於分析處理效率的提高。

4 資料挖掘（data mining）

這個步驟是進行知識挖掘的核心，主要是從資料庫中探勘出我們有興趣的隱藏樣式，或是建立可預測描述未來可能發生事件的模型，這些樣式或模型都可以視爲我們所要挖掘出的知識。

5 知識評估（knowledge evaluation）

經由前面步驟所挖掘出來的知識必須經過領域專家的評估，以確定其正確性與可用性。如果發現其正確性有誤，表示先前所進行的步驟有所缺失，就必須回溯到前面相關的步驟進行補強。例如，可能在資料預處理步驟中對於缺失值的處理不夠完整，就必須如圖 7-2 中的虛線所示，回到有缺失的步驟重新處理。或者，如果所建立出的預測模型之準確度不夠高的話，表示先前所採用的資料挖掘演算法不夠強健，就必須回到該步驟改變演算法。像這樣，把知識重複進行評估程序，直到通過領域專家的鑑定後，我們所挖掘出的就是如鑽石般發亮珍貴有用的知識了！

7-2 關聯規則與序列樣式

關聯規則探勘

關聯規則（association rules）探勘可用來找出資料庫中頻繁出現的項目組合，也就是具相關性的項目。舉例來說，一家美國的大型連鎖超市分析顧客的購買行爲後發現，每逢星期五晚上，如果顧客購買尿布，經常也會同時購買啤酒。因爲這些顧客是年輕的爸爸，在星期五下班之後會到超市幫家裡的小嬰兒買尿布，也會順便買啤酒，以便在週末看球賽時喝。這家超市挖掘出這個現象以後，便將一些商品和啤酒、尿布一起搭售，而獲得更高的收益。

假設項目 A 是資料庫中某事件（例如一筆交易）的項目之一，若項目 B 出現在該事件中的機率爲 P，且 A 和 B 同時發生在資料庫的頻率超過某一門檻值，則「A → B 爲一條關聯規則」。以 DVD 出租店爲例，假設表 7-1 是這家店的顧客租片資料，我們可以觀察到，所有租借《哈利波特》這部 DVD 的交易紀錄（即編號第 1、6、8 筆）中也都有《魔戒》，所以我們可以知道，當顧客租《哈利波特》時也會同時租《魔戒》，這就是一條關聯規則，我們可以把它記爲「哈利波特 → 魔戒」。有了這樣的規則，便可以應用在推薦系統上，例如，我們根據「哈利波特 → 魔戒」這條關聯規則，就可以推薦所有租看《哈利波特》的人也租看《魔戒》。

表 7-1　DVD 出租紀錄

紀錄編號	出租紀錄
1	移動迷宮、哈利波特、魔戒、鋼鐵人、美國隊長
2	異形、終極戰士、惡靈古堡、與神同行、飢餓遊戲
3	X 戰警、蜘蛛人、復仇者聯盟、蟻人、玩命關頭、變形金剛
4	移動迷宮、魔戒、美國隊長、X 戰警、蜘蛛人、不可能的任務
5	異形、惡靈古堡、飢餓遊戲、復仇者聯盟、蟻人
6	哈利波特、魔戒、鋼鐵人、玩命關頭、變形金剛
7	終極戰士、與神同行、飢餓遊戲、X 戰警、蜘蛛人、復仇者聯盟
8	哈利波特、魔戒、美國隊長、蟻人、玩命關頭、變形金剛

但是，如何評估關聯規則的關聯度有多強，則是我們要思考的另外一個要點。根據關聯規則的定義，關聯規則的關聯度是採取一個名爲**信賴度**（confidence）的評估方式來表現，它的計算方式是：

$$\text{信賴度} = \frac{\text{規則中前項與後項共同出現的次數}}{\text{前項出現的次數}}$$

舉例來說，「哈利波特 → 魔戒」這條規則的前項爲《哈利波特》、後項爲《魔戒》，前項與後項共同出現在第 1、6、8 筆資料中，所以共同出現次數爲 3，而前項出現的資料也是第 1、6、8 筆，次數也爲 3，因此，「哈利波特 → 魔戒」這條規則的信賴度爲 3/3 = 100%。這裡必須注意到，當前後項互換時，信賴度會不同，例如，當關聯規則變成「魔戒 → 哈利波特」，即前項爲《魔戒》、後項爲《哈利波特》時，《魔戒》是出現在第 1、4、6、8 等共 4 筆資料中，因此「魔戒 → 哈利波特」這條規則的信賴度是 3/4 = 75%。根據這個結果，我們可以利用**最小信賴度**（minimum confidence）作爲門檻值，來篩選出在資料中可能存在的關聯規則。

根據上述範例，當我們把最小信賴度設定爲 80% 時，「哈利波特 → 魔戒」這條規則的信賴度是 100%，高於 80%，於是它就會被探勘出來；相反地，由於「魔戒 → 哈利波特」這條規則的信賴度是 75%，低於 80%，就不會被探勘出來。

然而，如果只是單純地觀察是否有共同出現的資料，似乎會落入小概率事件的盲點。所謂小概率事件，就是有些事情的發生機率非常低，而恰巧這些事情發生的時候伴隨著某件事情，我們就會以爲兩者相關聯。像古人觀察到彗星經過時剛好出現旱災，就認定一旦彗星出現勢必會有天災，是不祥之物，還給了它掃把星、妖星等稱呼。但事實上，彗星與旱災兩者之間是沒有關係的，只是恰巧同時發生罷了。

舉例來說，我們在表 7-1 的第 4 筆資料中觀察到《不可能的任務》這部 DVD，我們或許會得到一個可能有偏差的結論，就是當有人租《不可能的任務》這部 DVD 時，必然會租《移動迷宮、魔戒、美國隊長、X 戰警、蜘蛛人》這些片。因此，關聯規則的代表性強度是另一個我們需要關心的重點。評估關聯規則的代表性強度是用一種名爲**支持度**（support）的評估方式來表現，其計算方式是用前項與後項共同出現的次數在整個資料庫中所佔的比例來代表，例如，「哈利波特 → 魔戒」這條規則的前項與後項共同出現在第 1、6、8 筆資料中，所以出現次數（筆數）爲 3，而整個資料庫總共有 8 筆資料，故支持度爲 3/8 = 37.5%。

據此，我們在探勘關聯規則時，會先設置一個**最小支持度**（minimum support）的門檻值，並且先利用這個門檻值過濾出支持度夠高的項目集合。例如，如果我們將最小支持度設定爲 30%，那麼，{ 哈利波特，魔戒 } 這個項目集合就必須滿足這個最小支持度的要求，由這個項目集合所能組合出來的關聯規則「哈利波特 → 魔戒」或「魔戒 → 哈利波特」就是支持度夠高的。這裡必須注意一點，關聯規則不僅是要支持度高於最小支持度，也要信賴度高於最小信賴度，所以，一般而言，探勘關聯規則的方法是先用最小支持度過濾出所有支持度夠高的項目集合，再利用這些項目集合組合出信賴度高於最小信賴度的關聯規則。

頻繁項目集計算的挑戰

從前面描述的內容，我們可以知道，挖掘關聯規則的過程分爲兩個步驟：

1. 找到所有支持度夠高的頻繁項目集。
2. 利用這些頻繁項目集來生成信賴度夠高的關聯規則。