輔仁管理評論 (專刊)
中華民國 91 年 9 月，第九卷第三期，1-10

資料開門－商業智慧：讓資料為您開啟智識大門

張金榕*．蔣以仁**

*國禾科技公司
**台北醫學院醫學資訊研究所／國禾科技

(收稿日期：91 年 4 月 1 日；第一次修正：91 年 6 月 15 日；
接受刊登日期：91 年 7 月 24 日)

摘要

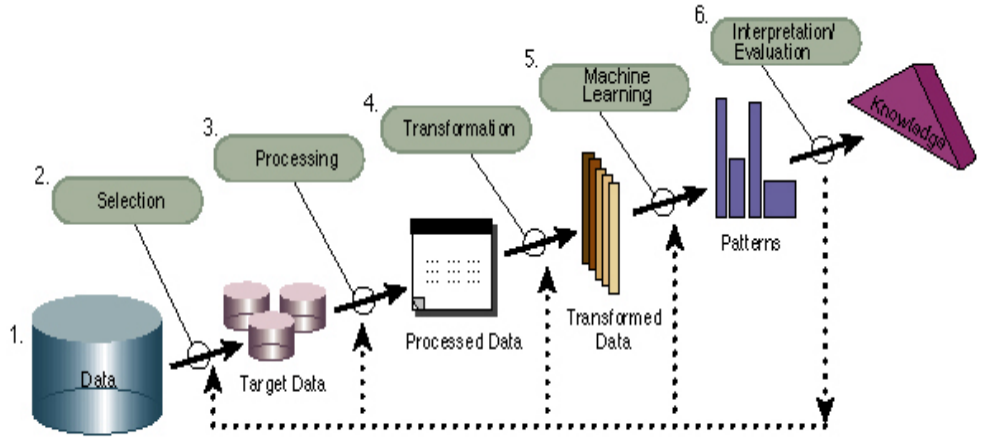
本文將從資料的角度，嘗試一窺資料探勘的風貌，並簡介資料探勘的正確做法，與評估方式。

關鍵詞彙：資料探勘，商業智慧

壹．介紹

「資料 (Data) 」一直是企業所擁有最大的無形資產，它散佈在各個地方，以有形和無形的方式在出現，然而在企業內部架設一個智識管理平台 (Knowledge Management Platform)，也是現今很多企業的當務之急，畢竟，它可以提昇企業之核心競爭優勢，更能讓企業行之多年的寶貴經驗得以薪火相傳，其重要性也可想而知。然而在近幾年來不斷的努力之下，隨著科技不斷的進步，技術也不斷的精益求精，但所謂「智識 (Knowledge) 」的呈現，卻也是讓 IT 部門最頭疼的問題，而如何把存在各地的資料整合起來進而產生「智識」，這整個過程是非常重要的，這也是今天筆者要跟大家報告的主要課題。

如何從企業中的原始資料 (Data) 轉換成可用智識 (Knowledge)，圖一即是說明這個重要的過程，一般稱為 KDD (Knowledge Discovery and Data Mining)，現在就來跟大家報告這整個過程，希望對有志從事資料探勘的朋友能有些幫助。



圖一 An Overview of the Steps That Compose the KDD Process

正本清源－從資料看起：「資料」是所有資訊的源頭，也是企業最寶貴的資產，但是，如果您仔細的觀察您的原始資料，相信大都會大失所望，因為您會看到一堆缺漏值 (Missing Value)，甚至還有很多資料矛盾的地方等等問題，然後，您期望從這些原始資料中可以得到好的結果，這幾乎是不可能的任務。是故，原始資料的處理就變得格外重要！

KDD 第一個步驟是選取資料，因為原始資料太多太大，不可能把所有的資料丟進去做分析，所以必須很清楚的知道要做什麼，根據您想做的題目，再仔細思考應該選取什麼樣相對應的資料，也就是從原始資料庫中拉出一個小的 data mart，以這個 data mart 為基礎來分析，而這個 data mart 也必須有足夠的代表性。

選取好原始資料後，第二個步驟是處理資料 (Processing data)，處理資料是一個非常繁瑣、卻常被忽略掉的重要工作，正確而乾淨的資料是得到正確資訊／智識的重要源頭，正所謂「Garbage in, garbage out」，處理資料的時候，要非常的小心，當有缺漏值、異常值、或矛盾的地方時，都要謹慎的處理，以缺漏值為例，一般處理方式可能是直接忽略掉，但是若資料樣本數不夠大時，若直接忽略會面臨實際的困難，故做法上可以給缺漏值一個「特殊值」做代表，以解決類似的問題；另外，資料的矛盾性也常常存在於資料中，只有定期檢視您的資料，去除資料中的矛盾，才能事半功倍。

拿到處理過的資料後，還不能直接丟入分析工具中去做分析，為什麼呢？因為分析工具中實際要用的變數，可能在您的資料中並不存在，是故要做資料

的轉換 (transformation)，舉例來說，一般資料中會存在與客戶往來的起始日期 (Date)，但分析時的變數可能是須要期間 (Duration)，這時候就要經過所謂的資料轉換，事實上資料轉換有很多技巧，有時還牽扯到一些複雜的數學運算，如 SIN、COSIN 等的轉換，這個步驟是分析前必要的工作，經過轉換的資料，成為可分析的變數，這在考驗您對企業本身及方法論的了解，因為惟有清楚企業的目的之後，才可以清楚的知道您要的是什麼變數，將資料做正確的轉換之後，才有可能得到最後的結果。

貳．資料處理

到目前為止，本文所談的都還沒有真正進入到方法論 (Methodology) 的領域，事實上，以上的資料準備會花您大部份的時間，而且大部份的事情也不能假手他人；等一切準備妥當之後，這時候再選用方法論，來產生模型 (Pattern)，不過，要注意的是，您的資料須分為二份，一份稱為 training data，它是為建立模型用的，一份稱為 testing data，它是用來測試所建立之模型的結果，無論如何，方法論的選擇可是一門學問，每個方法論中都有其代表的演算法 (Algorithm)，在使用方法論的同時，您必須小心幾件事：(1)其演算法之基本假設 (Hypothesis) (2)每個方法論中所隱藏的偏差 (Bias) (3)資料變數之間的關連性 (Dependency / Independency) (4)資料分佈狀況 (Distribution) (5)資料型態及其所代表的含意。

在使用方法論之後，當然就產生所謂的「結果」，這時候請捫心自問一下：什麼才叫做好的結果？而判定「好的結果」的條件又是什麼？

當完成以上所有步驟之後，這時候就剩臨門一腳了：您要如何解釋所產生出的「結果」呢？事實上從方法論所產出的「結果」，是一堆數字、編碼的組合，如何把它翻譯成企業所須的資訊 (或智識)，又是一個考驗主管 (或專業諮詢人員) 的地方，而這時候的工作不只是解譯所出現的結果，更要去評估 (Evaluation) 這個結果，以行銷活動為例，當您要預知一個行銷活動是否值得投資時，而利用資料探勘這個過程來協助您，一旦得到結果經評估後發現不符經濟效益時，就可以得到很清楚的建議，這也是所謂的智識平台，提供支援給決策者去做決定，累積企業的商業智慧，表現企業的核心競爭力…等等，這就是資料探勘的精神所在。

所以我們可以這麼說：資料探勘 (Data Mining) 的整個流程，首先須從了解您的業務 (Business) 開始，了解業務之後，再進一步知道您要準備什麼樣

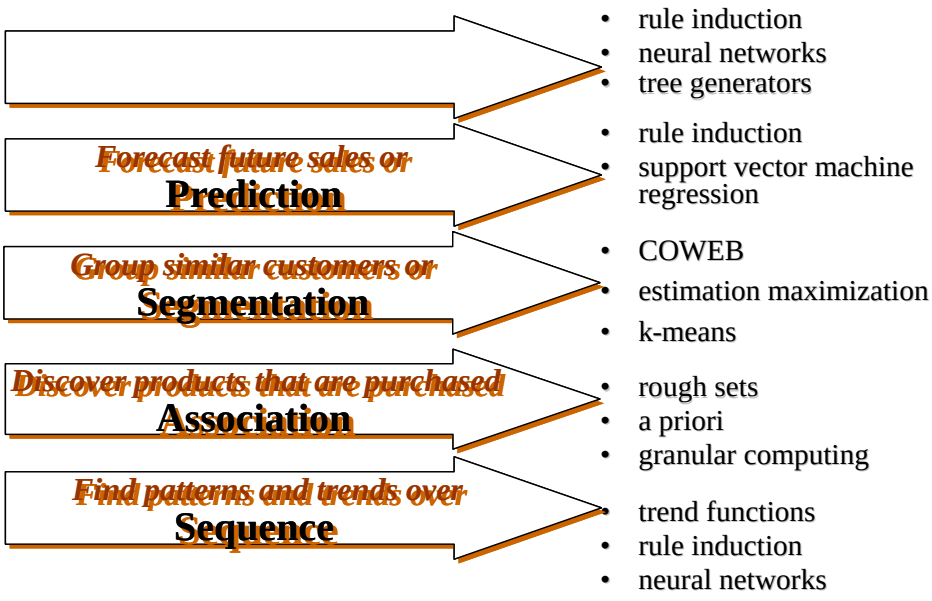
的資料，從了解業務到要準備什麼樣的資料，看似簡單，其實卻是資料探勘相當困難的部份，因為每一筆資料到底代表什麼樣的特性，都必須要能夠被清楚的解讀出來。以上例來說，當您做一個行銷活動 (Campaign) 的時候，要準備什麼資料，才能做好這個行銷活動呢？以台灣本土的特性來看，Demograph，地理環境或人口統計等等資料要放進哪些呢？若這些資料拿到美國來看的時候，其代表性又有什麼不同呢？舉例來說，每個地區都有所謂的區域號碼 (Zip Code)，但每個國家對 Zip Code 的處理方式又不盡相同，以美國來說，Zip Code 可以對應到當地人口的分佈，所以美國在 Zip Code 有幾個不同的 Code 來做相互對應；同樣的，對企業而言，是不是可以找到這樣的 Code？它可以代表這個區域人口的特殊屬性、特殊族群、甚至其學經歷等等，都有其相對應的 Code？當然在台灣目前是還沒有這樣的 Code，從這裡可以看到，台灣還有非常多的改善空間。

在資料倉儲／資料探勘技術不斷的被要求的同時，其實很多人都忽略掉了「資料」這個最原始的課題，今天，我不想長篇大論介紹世界上的技術有多新多強，反而是從問題的最根本來著手，也就是說：凡事正本清源，從資料看起！當您把資料準備好之後，接著可能要做 Retention Rate，或做出一個有效的 Campaign，但重點是 Retention Rate 或「有效」的 Campaign 又要如何去定義？又要經過什麼樣的 Transformation 才能達到您的要求？這樣的 Transformation 是非常重要的，而在 Data Mining 中，也是較困難的地方；這些地方可以用一些方法論來解決，在套用方法論部份或用電腦程式來跑，而其中大部份都有現成的工具可直接來執行，還不算太困難，最困難的地方是您要準備什麼樣的 Data？及用什麼樣的 Transformation 來處理？

事實上，您還可以找到一些 Rules 來協助您，這些 Rules 大致可分為二大模組，一是預測模組 (Predictive Models)，另一為描述模組 (Description Models)，Predictive Models 得到的是一個量化的數值，Description Models 大部份得到的是一個 Rule，大致上我們可以將此二大模組做這樣的區分；至於做一個 Campaign，須要什麼資料的佐證，使得這個活動比較有效，並評估其投入／產出之投資報酬率，是企業最關心的課題，在對方法論有進一步了解之後，再去計算整個投資報酬率，亦即評估 (Evaluation) 效益，下次有機會再跟大家介紹。

參．方法概述

在對整個資料探勘 (Data Mining) 過程有些了解之後，一般在資料探勘會常用的 Models 如圖二所示：



圖二　The Models of Data Mining

圖二左半邊是企業想利用 Data Mining 來完成的任務 (Task)，而右半邊則是完成任務之相對應的方法論 (Methodology)，在這裡一一為大家做一個簡單的介紹如下：

1.Classification

Classification 是運用已知的結果，結合其相關之屬性，來推導出在資料中存在什麼樣的規則 (Rules) 及事實 (Fact)，而這裡一般會用到的方法像是類神經網路 (Neural Network)、決策樹 (Decision Tree) ...等等方法，其主要對資料做分類之外，它還能找出離群值 (Outlet)。

2.Prediction

Prediction 顧名思義就是做預測，其是運用歷史資料去預測未來會產生的變化，一般最常用的方法論是迴歸 (Regression)。

3.Segmentation

Segmentation 是使用 Clustering 方法去產生物以類聚的現像，很多人會把它跟 Classification 搞混，其最主要區別是 Clustering 並沒有在事前特別給資料的屬性，而是直接做分群，再做資料分析，不同於 Classification 的是其已先定義每群資料，對每群資料的特性事前就知道了，故二者是不同的。

4.Association

Association 是在一堆看似無關聯的資料中去找到一些關聯性，它一般被用來做貨品上架的分析，最有名的例子莫過於尿布與啤酒的案例，它是在 1990 年代的時候，Teradata 這家公司對全美國的超市做貨品分析，意外發現尿布與啤酒有高度相關，也就是會買尿布的人同時買啤酒的機率較大，後來有一超商試著把這二樣商品擺在一起販賣，結果果然大賣，事後解讀這個結果是：因為會買尿布的家庭，家中必然有小嬰孩，而母親通常會在家中照顧小孩，買尿布的工作當然也交給父親，而父親在買尿布的同時，順便幫自己帶一些啤酒，從這個案例中，可以看到在資料中隱藏著很多用人腦無法找到的關連，也因為這個案例讓 Data Mining 一炮而紅，Association 一般常用的方法論如 Rough Set、A Priori…等。

5.Sequence

Sequence 通常是用來指說這個事件的發生是否有連續性？舉例來說，在過去經驗中所做的 Campaign 會有一定的效果，然而過去的經驗，對我這次所要做的 Campaign 是否也會達到同樣的效果呢？在時間序列上，依照經驗法則是否可以看到某種趨勢呢？這裡用到的方法論像是 Neural Network…等等。

肆．結語

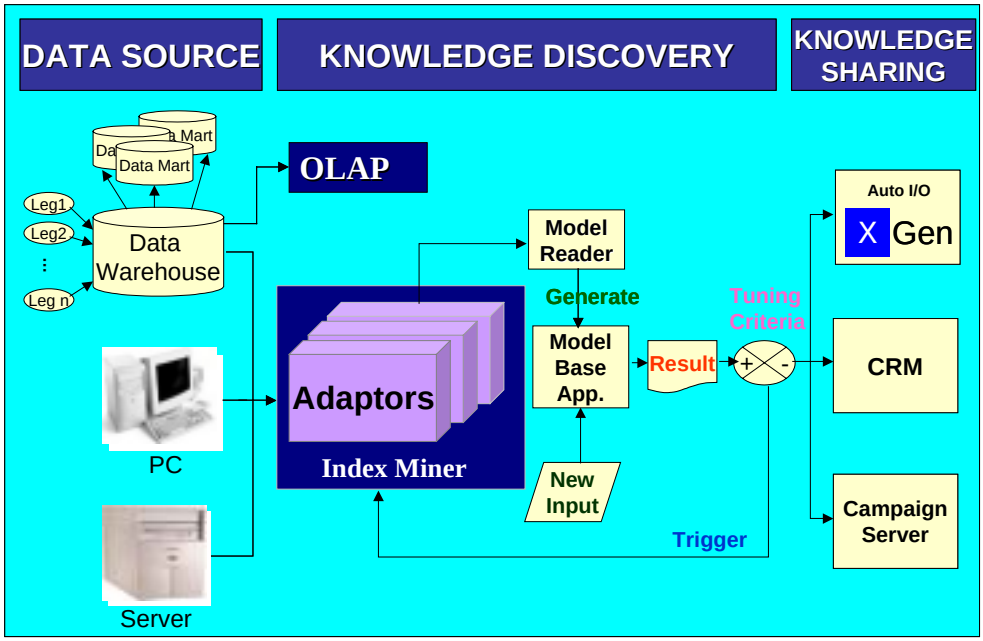
走筆至此，不得不為本文作一結尾，然本文之結語，係為資料探勘的楔子。談到這裡，如果您是剛踏入 Data Mining 這個領域的朋友，希望可以給你一些啟發，而上面所提及的專有名詞，相信未來您也會一直不斷的看到，這時候您可能會問：那 Data Mining 的定義是什麼呢？這就是筆者現在所要提的。

Data Mining 的定義：

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if

its performance at tasks in T, as measured by P, improves with experience E.

要做資料探勘時，首先須定義清楚您的任務 (task T) 是什麼？它是一個行銷活動呢？亦或是一個收入／利潤之預估？針對每個不同的任務，再來決定須要什麼樣的資料 (experience E；historical data)，再從這些資料中找出其所須的特徵 (feature)，這些特徵是可以區分出不同的類別且具代表性，而它所表現的地方就是您要選的屬性 (attribute)，最後則是觀察其效能 (performance P)，從花下的成本中能帶出多少的收入？或是在可預見的未來，是否可以產出更高的利潤？經過這樣一個可量測 (Measurable) 的過程，就可以決定這個任務 (Task) 是否是一個值得執行的投資。以上三者 (task T、experience E、performance P) 是構成 Data Mining 的要件，也就是說，在做 Data Mining 時，必須要有明確的目標或問題，再思考能提供什麼樣的資料屬性，最後得到的是一個可量測的效能，請注意：整個過程都必須是可以被量測的過程，才能稱得上是做資料探勘。



圖三　CIS 的架構圖

從了解您的問題 (如 Business Target)，並找到其適用的資料，再根據此資料去找出適用的方法，然而適用的方法所產生的模組 (Model)，最重要的是希

望可以跟企業的應用系統直接連結在一起，而變成一整套完整的系統，而不是產生模組 (Model) 後，再到資料庫去 Select 出資料來做獨立的系統，很不幸的是，目前大多數的企業都採用後者的方式來做，這也說明了為什麼很多企業投入了很多金錢及時間來做 Data Mining 或 Knowledge Management，卻看不到太大成效的原因，事實上，企業真正要關心的是如何賺更多的錢、節省更多的成本、提高更好的服務等跟企業本身有切身關係的課題，是故如何建立一套完整的協同式智慧型系統 (CIS，Collaborative Intelligent System) 才是提高競爭力的核心，圖三即是 CIS 的架構圖。

圖三可以很清楚的看到企業現今面臨的困境：

1.後端資料倉儲 (Data Warehouse) 的建立

這是多麼巨大的工作啊？更何況原本在各地的資料既然都已經存在了，我們還有必要去做一個集中式的超級大怪物嗎？是否有更好的解決方案呢？事實上是有的，我們可以用分散式的資料庫來解決這個問題，當然這個課題我們再找個時間另外談。

2.前端 CRM 系統的孤立無援

事實上，現在存在企業內部的 CRM 系統似乎還沒有真正發揮它強大的功能，想想看，您現在的 CRM 系統嚴格說起來，可能只是一個客服中心 (Call Center) 而已，如何跟後端資料庫所產生的智慧結合在一起運做，恐怕才是老闆們心中的期望。

3.建立前端與後端的橋樑

相信在每個 IT 部門主管的工作之一，就是如何把存在後端的資料庫中的資料，可以順利的支援到最前端的業務，當企業電腦化越深的同時，或許您享受到它帶來的便利，但同時也代表著有一定的包袱存在，在智識發掘的過程中，企業通常要的是結果，不管您過程做的多辛苦，老闆是不管的，它要的是結果，而且是「好的結果」，正所謂成敗論英雄，這也是為什麼您要一個 CIS 系統，限於篇幅的關係，筆者希望下次有機會再跟大家好好談論這個概念。

筆者行筆至此，其實是希望不要用太艱深的論調來談資料探勘，事實上，我相信有很多朋友對這個課題是心急如焚的，也希望這篇文章能對您有些小小的幫助，讓資料能打開智識之門，讓企業能運用其最寶貴的資產來創造更多的利潤、降低更多的成本、提供更多元化的服務等等，在這不景氣的大環境之下，能有些不同的作為。

參考文獻

Mitchell, T., "Machine Learning", McGraw Hill, 1997.

Data Mining-Commercial Intelligence: Let Data Opening Your Knowledge

JENNY CHANG*, I-JEN CHIANG **

**Index Software, Inc.*

***Taipei Medical University of Graduate Institute of Medical Informatics*

ABSTRACT

The purpose aim of this paper is to use concept of data, introduction how to use accuracy methods of data mining and evaluated rules of data mining.

Keywords: data mining, commercial intelligence