Data Mining 的企業應用範疇與方法論---SQL Server 2005

謝邦昌

輔仁大學統計資訊學系、中華資料採礦協會

Abstract

在資訊科技發展日進千里的今天，資料處理與儲存管理的問題，在軟體技術與速度不斷的改良，以及硬體設備的購置成本大幅降低之下，都變得簡單了，也因此間接帶動了企業在與營運相關的資料庫的建置與投資。

而在所謂「知識經濟」時代來臨，企業間的競爭模式，從傳統的「紅海策略」，採壓低成本與價格的殺價流血競爭，到近來倡導以「創新」為核心競爭力的「藍海策略」，不論哪一種策略模式，都是不斷在技術研發、製造生產、行銷販售、客戶服務，或資源配置等營運的相關問題上，尋求問題的發生原因，並嘗試找出解決方案。而在整個不同營運階段中，陸續累積的龐大資料，往往就是答案的隱身之所。因此，如何善用資料數據，從營運歷史的紀錄裡，採礦出深藏其中的寶貴經驗(金礦)，就是「資料採礦」(Data Mining)的目的。

企業在嘗試分析其資料時都面臨若干問題。一般而言，並不缺乏資料。事實上，很多企業感覺到他們被資料淹沒了；他們沒有辦法完全利用所有的資料，將其變成有用的訊息，尤其是當資料從不同的作業系統湧入時，如何得到一致性的資訊，是一直困擾企業營運的問題。為了處理這方面的問題，開發了資料倉儲(Data Warehousing)技術，以讓企業從源於各不同作業系統間的資料，加以並將其變成有用的資訊。

一個適當運作的資料倉儲是具有驚人強大功能的解決方案。公司可以對資訊進行分析，並將其加以利用，以進行明智的決策。透過使用資料倉儲，可以為您提供以下問題的答案：

- 哪些產品最受 15-20 歲的女性歡迎？
- 特定消費者的訂單前置時間和按時交付的百分比與所有消費者的平均值相比如何？
- 病房花在每個病患身上的成本和時間是多少？
- 在簽約階段停滯時間超過十天的項目所占的百分比為多少？
- 如果某個特定的實驗室在某類特定的藥品上投入了較多的資金，臨床試驗結果是否顯示病人健康狀況好於其他實驗室？

除了這些通常可透過使用分析應用程式得出答案的問題之外，資料倉儲還支援各種資料交換格式。分析應用程式設計為供分析人員使用，分析人員會對資料

進行分類，研究有助管理與決策的分析結果；報表應用程式會產出書面報表或線上報表，這些報表供功能要求略低的用戶使用，提供靜態內容，或提供有限的深入採礦功能；另對於業務決策者而言，計分卡是非常強大的功能，可以提供公司關鍵性能指標(Key Performance Indicator, KPI)的概況，使決策者知道其身處何處。

儘管資料倉儲功能強大而實用，但其自身有一個侷限；它實質上反映的是過去的歷史。由於資料倉儲經常在特定週期或時點進行資料載入和處理，因此它只是表示一個時點上的快照(Snapshot)。即使是建構了即時(real-time)或近似即時(near real-time)的資料倉儲，但其資料仍然只表示當前和歷史的資料，無法達到「預測」的需要，因此為了發現資料的因果關係，資料倉儲需要利用其他科學方法，進行定量的分析。

與傳統的統計分析方法不同的地方，「資料採礦」不是讓人提出假設，然後據此去找相關資料，而是讓資料倉儲確定資料關聯性，並允許採用以往不同的模式對資料進行分析。透過資料採礦，可以得出諸如以下這樣的問題的答案：

- 客戶將購買什麼產品？哪些產品將一起銷售？
- 公司如何預測哪些消費者可能會流失？
- 市場狀況如何，將會如何發展？
- 企業如何對其網站使用模式進行最佳的分析？
- 組織如何確定營銷活動是否成功？
- 什麼是分析非結構化資料（如無格式文件）的最好技術？

微軟公司(Microsoft)的Microsoft® SQL Server™ 2005是一個完整的商業智慧(Business Intelligence, BI) 平臺，為用戶提供了可用於構建典型和創新的分析應用程式所需的各種特性、工具和功能。其中引入了大量新的資料採礦功能，允許企業給出這些問題和其他問題的答案。本課程將討論資料採礦可以解決的各種問題，並介紹SQL Server 2005 處理這些問題的模式。