輔仁管理評論 (專刊)
中華民國 91 年 9 月，第九卷第三期，11-34

商業智慧的工具－資料採礦

鄭宇庭*・蘇志雄**

*政治大學統計學系
**致理技術學院會計系

(收稿日期：91 年 3 月 14 日；第一次修正：91 年 5 月 8 日；
接受刊登日期：91 年 7 月 3 日)

摘要

本文介紹資料採礦 (Data Mining) 及資料倉儲的意義、比較資料採礦和統計分析，並進一步說明資料倉儲、KDD 與資料採礦的關係。同時也說明資料採礦的功能、應用、進行步驟、工具。最後並對資料倉儲與資料採礦在顧客關係管理上的運用作一說明。

關鍵詞彙：資料採礦，資料倉儲，商業智慧，顧客關係管理

壹・前言

資料採礦的工作 (Data Mining) 是近年來資料庫應用領域中, 相當熱門的議題。它是個神奇又時髦的技術，但卻也不是什麼新東西，因為 Data Mining 使用的分析方法，如預測模型 (迴歸、時間數列)、資料庫分割 (Database Segmentation)、連接分析 (Link Analysis)、偏差偵測 (Deviation Detection) 等；美國政府從第二次世界大戰前, 就在人口普查以及軍事方面使用這些技術, 但是資訊科技的進展超乎想像, 新工具的出現, 例如關連式資料庫、物件導向資料庫、柔性計算理論 (包括 Neural network、Fuzzy theory、Genetic Algorithms、Rough Set 等)、人工智慧的應用 (如知識工程、專家系統), 以及網路通訊技術的發展, 使從資料堆中挖掘寶藏, 常常能超越歸納範圍的關係；使 Data Mining 成為企業智慧的一部份。

Data Mining 是一個浮現中的新領域。在範圍和定義上、推理和期望上有一些不同。挖掘的資訊和知識從巨大的資料庫而來, 它被許多研究者在資料庫系統和機器學習當作關鍵研究議題，而且也被企業體當作主要利基的重要所在。有許多不同領域的專家，對 Data Mining 展現出極大興趣，例如在資訊服務業中，浮現一些應用，如在 Internet 之資料倉儲和線上服務，並且增加企業的許多生機。

我們對於這種 Data Mining 的產品應該有一個正確的認知，就是它不是一個無所不能的魔法。它不是在那邊監視你的資料的狀況，然後告訴你說你的資料庫裡發生了某種特別的現象。也不是說有了 Data Mining 的工具，就連不瞭解業務、不瞭解資料所代表的意義、或是不瞭解統計原理的人也可以做 Data Mining。Data Mining 所挖掘出來的資訊，也不是你可以不經確認，就可以照單全收應用到業務上的。事實上，Data Mining 工具是用來幫助業務分析策畫人員從資料中發掘出各種假設 (Hypothesis)，但是它並不幫你查證 (Verify) 這些假設，也不幫你判斷這些假設對你的價值。

貳．何謂資料採礦？

資料採礦 (Data Mining) 是指找尋隱藏在資料中的訊息，如趨勢 (Trend)、特徵 (Pattern) 及相關性 (Relationship) 的過程，也就是從資料中發掘資訊或知識 (有人稱為 Knowledge Discovery in Databases, KDD)，也有人稱為「資料考古學」(Data Archaeology)、「資料樣型分析」(Data Pattern Analysis) 或「功能相依分析」(Functional Dependency Analysis)，目前已被許多研究人員視為結合資料庫系統與機器學習技術的重要領域，許多產業界人士也認為此領域是一項增加各企業潛能的重要指標。此領域蓬勃發展的原因：現代的企業體經常蒐集了大量資料，包括市場、客戶、供應商、競爭對手以及未來趨勢等重要資訊，但是資訊超載與無結構化，使得企業決策單位無法有效利用現存的資訊，甚至使決策行為產生混亂與誤用。如果能透過資料發掘技術，從巨量的資料庫中，發掘出不同的資訊與知識出來，作為決策支援之用，必能產生企業的競爭優勢。

Data Mining 可說會合了以下六種領域：(1)Database systems, Data Warehouses, OLAP (2)Machine learning (3)Statistical and data analysis methods (4)Visualization (5)Mathematical programming (6)High performance computing

Data Mining 應用的行業包括了金融業、電信業、零售商、直效行銷、製造業、醫療保健及製藥業等等，應用領域如下表：

表一 Applications of Data Mining

Customer-focused	Operations-focused	Research-focused
Life-time Value	Profitability Analysis	Combinatorial Chemistry
Market-Basket Analysis	Pricing	Genetic Research
Profiling & Segmentation	Fraud Detection	Epidemiology
Retention	Risk Assessment	
Target Market	Portfolio Management	
Acquisition	Employee Turnover	
Knowledge Portal	Cash Management	
Cross-Selling	Production Efficiency	
Campaign Management	Network Performance	
E-Commerce	Network Performance	
	Manufacturing Processes	

現今電腦運算能力的躍進，以及資料儲存技術的進步，資料倉儲的廣泛建置，加上企業行銷策略轉為針對單一消費者個人行銷，更突顯 Data Mining 對於企業的迫切性。

參．資料倉儲

一、資料倉儲 (data warehouse)

所謂資料倉儲是具有主題導向 (subject-oriented)、整合性 (integrated)、長期性 (time invariant) 與少變性 (nonvolatile) 的資料群組，是經過處理整合，且容量特別大的關聯式資料庫，用以儲存決策支援系統 (Design Support System) 所需的資料，供決策支援或資料分析使用。

(一)主題導向 (Subject Orient)

1.Organized around major subjects, e.g., customer, supplier, product, sales

2.Not on day-to-day transaction, point of sale

3.Focus on modeling and analysis of data for decision making

4.Typically provides a concise view around particular subject issues...

(二)具整合性 (Integrated)

1.Data warehouse usually constructed by integrating multiple heterogeneous sources, such as: relational DB, flat files, on-line transaction records…

2.Data cleaning techniques are applied to ensure consistency of naming conventions, encoding structures, attribute measures, …

(三)具長期性 (Time Variance)

1.Data are stored to provide information from a historical perspective, every key structure in the data warehouse contains an elementary of time.

2.日常性作業的資訊系統，受限於軟硬體設備的容量及回應時間等因素，常無法保留太長期間的資訊 (約 60-90 天)。而資料倉儲系統，為了執行趨勢的分析，常須保留 1-10 年的歷史資料。而每一筆資料均會含有一個時間的標籤，用以區別資料的時點，以利執行特定期間的分析作業。

(四)具少變性 (Non-Volatile)

1.Data warehouse is always a physically separate store of data, transformed from the application data in the operational environment

2.It usually requires 2 operations for data accessing: initial loading and access of data.

3.日常性作業的資訊系統，其資料內容常被頻繁地存取及異動。而資料倉儲系統，當資料從日常性作業的資訊系統中轉入後，主要用於大量資料查詢及分析；事實上，從忠於原始資料來源的角度來看，異動資料倉儲內的資料，是不合理且不道德的作法。

二、資料倉儲的規劃與建置的方式：

資料倉儲的規劃與建置的方式，將決定其效益。以規劃為例，理想的方式應採用集中式的方法而非分散式。

(一)分散式的架構 (稱為Independent Data Marts獨立式的資料超市，或稱為Departmental Data Warehouses部門級的資料倉儲)

資料超市是企業級資料倉儲的子集，建置的目的是為了企業中個別的部門或單位。與企業級資料倉儲不同的是，資料超市通常只為了特定的決策支援應用程式或使用群組，通常是由下到上 (bottom up) 利用部門的資源來建置。資料超市通常只有特定主題的彙總或詳細資料，而資料超市中的資料可以是企業級資料倉儲的一個子集 (獨立的資料超市) 或者可能直接使用來自運作中的資料來源 (相依資料超市，independent data mart)。無論是資料倉儲或資料超市，其組成與維護的程序是相同的，使用的技術元件也都類似。

資料超市雖然建置較為容易，卻無法達成企業對資訊有一致性的觀點，特定的 Data Mart 僅可滿足特定使用者族群的應用。當有跨 Data Marts 的應用時，必須再經由一次的資料轉換作業，故使用時極為不便。

(二)集中式的架構 (稱為Enterprise Data Warehouse企業級的資料倉儲)

企業級的資料倉儲所包含的是全企業的資訊，這些資訊是為了整體的資料分析而整合至多個運作系統的資料來源。一般而言，是由數個主題領域所組成。例如客戶、產品加上業務等，可用於戰術 (tactical) 與策略 (strategic) 上的決策支援。企業級資料倉儲的資訊包括即時的詳細資訊，也有彙總的資訊，資料庫大小的範圍可能從 50 gigabytes (GB) 到 1 terabyte (TB)。企業級的資料倉儲的建置與管理往往非常昂貴且耗時；建立的方法通常是從上到下 (top down) 由統籌的資訊服務單位主導。

在此集中式架構下，亦有二種建置模式：

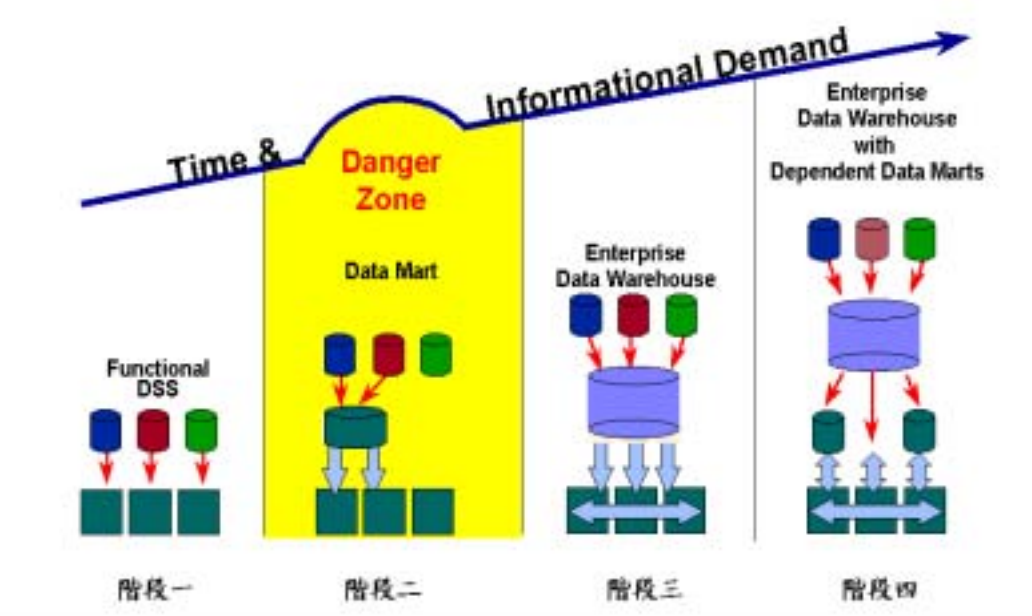
1. Data Mart Centric: 其做法雖然可將企業的資料統整在一起，卻不提供對所有明細資料的查詢，使用者僅可接觸部份的資料而已。此種架構並無法滿足各類的應用需求。

2. Enterprise Data Warehouse with Dependent Data Marts:即可滿足所有的應用需求。使用者可在 Dependent Data Marts 中快速地查詢大分類的資訊，亦可依需求至 Enterprise Data Warehouse 內查詢明細的資料。

建置資料倉儲，在規劃時必須包含整體的架構，而實際建置時是採由小而大逐步漸進的方式，先建置最重要的主題，而後慢慢的延伸下去。企業初期常購入一些簡單的建置初期的型態會以資料集市 (Data Mart) 形式存在，在資

料量與分析需求增加後遂漸擴展為資料倉儲 (Enterprise Data Warehouse)，而當考慮系統的回應時間及資料應用分析的習慣時，便會建立一個完整型態 (Full Scale) 的 Enterprise Data Warehouse with Dependent Data Marts。

故企業在建構其決策支援系統時，其資訊系統架構的演進將如圖一的四個階段，其中在階段二完成後，若資訊系統無法承受大量的資料與應用，而被迫朝向分散式的 Independent Data Marts 而非集中式的 Enterprise Data Warehouse 時，將註定走向失敗的命運 (故圖示中稱第二階段為危險區域 Danger Zone)：



圖一　資訊系統架構的演進

許多企業選擇實施較小型的部門或工作群組的 Data Warehouse 或稱 Data Mart，以兼顧低成本與快速上線能力。如日後有需要，該 Data Mart 可以予以擴充，在資料量或 Data Mart 數目擴充後，就可以成為整個企業的 Data Warehouse 解決方案。一般而言，資料倉儲必須能從多種異質 (Heterogeneous) 資料源中擷取作業資料，並將這些資料轉換成 Information Data 後儲存在 Data Mart 中。

目前「資料倉儲」技術，又以建立客戶關係管理 (Customer Relationship Management) 系統為主。「資料倉儲」對於企業的貢獻在於「效果」(Effectiveness)，能適時地提供高階主管最需要的決策支援資訊，作到 Deliver The Right Thing To The Right People At The Right Time。簡單地說，就是運用

資訊科技將寶貴的營運資料，建立成為協助主管作出各種管理決策的一個整合性「智庫」，利用這個「智庫」，企業可以靈活地分析所有細緻深入的客戶資料，以建立強大的「客戶關係管理」優勢。

三、資料倉儲的運作方式：

1.Top-down：

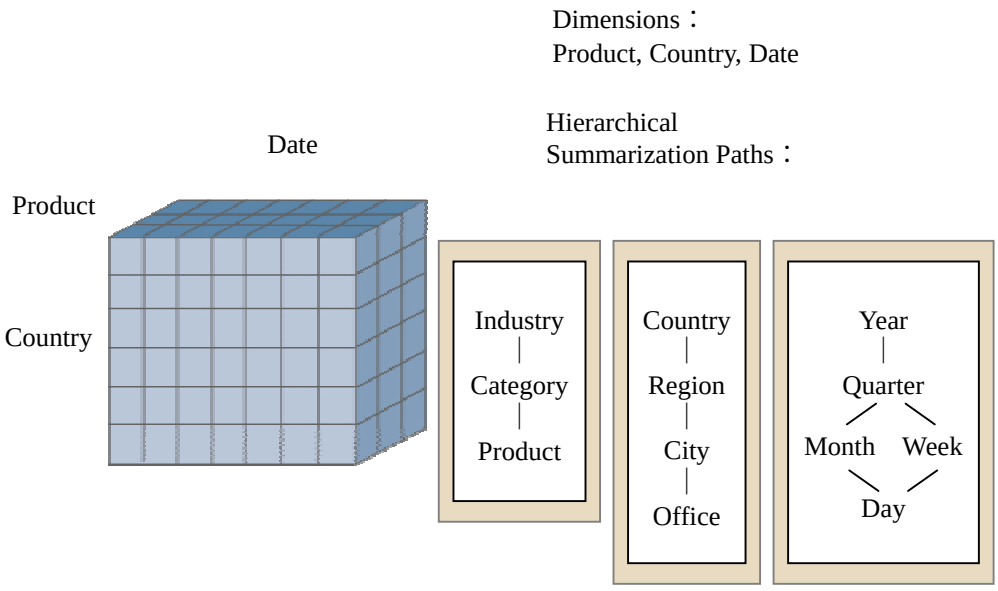
Starts with overall design and planning (mature). ― 整體規劃，但耗時。(Water fall：structured and systematic analysis at each step before proceeding to the next.)

2.Bottom-up：

Starts with experiments and prototypes (rapid). ― 從有需要的地分先做，有實際的效果，且速度快。(Spiral：rapid generation of increasingly functional system, short turn around time, quick turn around.)

A data warehouse is based on a multi dimensional data model which views data in the form of a data cube.

Cube: A lattice of cuboids.(立方體)



圖二　多向度資料模式

四、Modeling data warehouses：dimensions & measures

1.Star Schema：a single object (fact table)

2.Snow Flake Schema (雪花)：從模型的右邊再分 dimensions 出去

3.Fact Constellation (星座、銀河)：有很多主題

五、資料倉儲與運作系統 (Data Warehouse VS. Operational System)

資料倉儲與傳統運作系統在資料庫的設計上以及對系統預期的作用是不同的。運作系統組織架構的設計是為了支援線上交易處理 (On Line Transactional Processing, OLTP)，例如訂單輸入、銀行的存取款等作業，需具備快速的交易處理需求。而資料倉儲的組織架構設計，則是依據特定的主題 (subject) 而定，例如行銷、產品等。

表二　傳統運作系統與資料倉儲的比較表

	OLTP	Data Warehouse Systems
組織架構	依執行的交易而定 (如訂單處理、進銷存處理等)	依主題而定 (如產品、顧客等)
使用者的數量	大量的同時上線使用者	經常的使用者人數可能只有數百人或更少 (如分析、業務或行銷資料)
交易的數量／處理時間	交易時間通常很短，只存取數筆資料；交易處理在數秒間完成	處理查詢的交易時間可能長達數分鐘到數小時；查詢次數少，但很複雜，通常是檢索許多筆資料，並需要結合 (join) 多個資料庫的表格
資料庫的大小	OLTP 的資料庫通常小於資料倉儲，因為歷史資料會移出，不在線上	資料倉儲的資料量通常遠大於 OLTP 資料庫，因為其中整合了數種資料來源 (包括 OLTP)，也包括歷史資料
結構	高度正規化的資料結構，由許多表格組成，每一個表格設計是越少欄位越好	表格數量較少 (因為有特定的主題)，但欄位的數量很多
歷史資料	線上只維持在較短期間的資料	維護的資料可能回溯到數年以前
更新處理	當新交易被處理或輸入時，持續性的更新 (近乎即時)	期間性的更新，通常以批次的方式 (例如為了反應 OLTP 資料的變更)

六、Typical OLAP Operations

1.Roll up (drill-up): summarize data by climbing up hierarchy or by dimension reduction

2.Drill down (roll down): reverse of roll-up from higher level summary to lower level summary or detailed data, or introducing new dimensions

3.Slice and dice: 切片和切塊 (個別找出內容來看)

4.Pivot (rotate): 用不同的旋轉角度看

七、OLAP Server Architecture (儲存資料方式)

1.Relational OLAP (ROLAP):

(1)資料儲存方式與傳統方式一樣，用表格型式 (二維)

(2)耗時

(3)容易擴充 (因為是用一般方式)

2.Multidimensional OLAP (MOLAP):

(1)Array-based (陣列方式) multidimensional storage engine—以多維角度分析資料

(2)速度快

(3)耗空間 (因為要以 Cube 的方式呈現)

3.Hybrid OLAP (HOLAP):

User flexibility, e.g., low level：relational, high level：array

八、Data Mining VS. OLAP

所謂 OLAP (Online Analytical Process) 意指由資料庫所連結出來的線上查詢分析程序。它是由使用者所主導，使用者先有一些假設，然後利用 OLAP 來查證假設是否成立。

Data Mining 不需要假設或期待可能的結果，透過 Mining 技術可找出存在於資料中的潛在規則，於是我們可能得到例如尿布和啤酒常被同時購買的意料外之發現，這是 OLAP 所做不到的。再者，Data Mining 常能挖掘出超越歸納範圍的關係，但 OLAP 僅能利用人工查詢及視覺化的報表來確認某些關係。

OLAP 可以和 Data Mining 互補，但這項特性是 Data Mining 無法被 OLAP 取代的。

九、Data Mining VS. Data Warehousing

所謂資料倉儲是具有主題導向 (subject-oriented)、整合性 (integrated)、長期性 (time invariant) 與少變性 (nonvolatile) 的資料群組，是經過處理整合，且容量特別大的關聯式資料庫，用以儲存決策支援系統 (Design Support System) 所需的資料，供決策支援或資料分析使用。

若將 Data Warehousing (資料倉儲) 比喻作礦坑，Data Mining 就是深入礦坑採礦的工作。

資料倉儲應先行建立完成，Data mining 才能有效率的進行，因為資料倉儲本身所含資料是乾淨 (不會有錯誤的資料參雜其中)、完備，且經過整合的。

因此兩者關係或許可解讀為「Data Mining 是從巨大資料倉儲中找出有用資訊的一種過程與技術」。

肆．資料採礦和統計分析之比較

硬要去區分 Data Mining 和 Statistics 的差異其實是沒有太大意義的。一般將之定義為 Data Mining 技術的 CART、CHAID 或模糊計算等等理論方法，也都是由統計學者根據統計理論所發展衍生，換另一個角度看，Data Mining 有相當大的比重是由高等統計學中的多變量分析所支撐。但是為什麼 Data Mining 的出現會引發各領域的廣泛注意呢？主要原因在相較於傳統統計分析而言，Data Mining 有下列幾項特性：

1.處理大量實際資料更強勢，且無須太專業的統計背景去使用 Data Mining 的工具；

2.資料分析趨勢為從大型資料庫抓取所需資料並使用專屬電腦分析軟體，Data Mining 的工具更符合企業需求；

3.純就理論的基礎點來看，Data Mining 和統計分析有應用上的差別，畢竟 Data Mining 目的是方便企業末端用者使用而非給統計學家檢測用的。

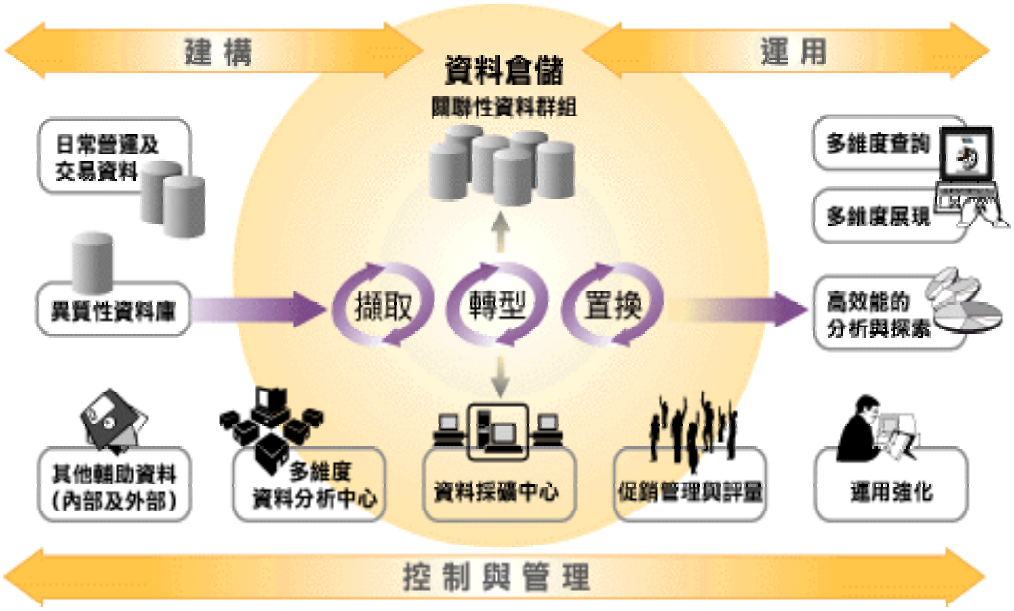
伍．資料倉儲、KDD與資料採礦的關係

許多人對於資料倉儲 (Data Warehouse) 和資料採礦 (Data Mining) 時常混淆，不知如何分辨。其實，資料倉儲是資料庫技術的一個新主題，在資料科技日漸普及下，利用電腦系統幫助我們操作、計算和思考，讓作業方式改變，決策方式也跟著改變。另外，決策支援系統和主管資訊系統也日漸普遍，它們操作資料的方式不盡相同，因而有必要把作業性資料庫和資料倉儲分隔開來，利用不同資料庫系統與技術操作，才能達系統最佳化。由於關聯式資料庫、平行處理及分散式資料庫技術的進步，不論是主從式架構或主機型架構的資料庫系統，資料倉儲技術皆可以利用原有作業中或已有的 (Legacy) 系統，進而提供一個穩固的基礎以支援全公司的決策支援系統 (DSS)。

資料倉儲本身是一個非常大的資料庫，它儲存著由組織作業資料庫中整合而來的資料，特別是指從線上處理系統 (OLTP) 所得來的資料。將這些整合過的資料置放於資料倉儲中，而公司的決策者則利用這些資料作決策；但是，這個轉換及整合資料的過程，是建立一個資料倉儲最大的挑戰。因為將作業中的資料轉換成有用的的策略性資訊是整個資料倉儲的重點。也就是，資料倉儲應該具有這樣的資料：整合性資料 (integrated data)、詳細和彙總性的資料 (detailed and summarized data)、歷史資料、解釋資料的資料 (Metadata)。如果資料倉儲集合具有成功有效率地探測資料的世界，則挖掘出決策有用的資料與知識，是建立資料倉儲與使用 Data Mining 的最大目的。而從資料倉儲挖掘有用的資料，則是 Data Mining 的研究重點，兩者的本質與過程是兩碼事。換句話說，資料倉儲應先行建立完成，Data mining 才能有效率的進行，因為資料倉儲本身所含資料是「乾淨」(不會有錯誤的資料參雜其中)、完整的，而且是整合在一起的。因此，或許可說 Data Mining 是從巨大資料倉儲找出有用資訊之一種過程與技術。

KDD (Knowledge Discovery in Database) 和 Data Mining 的關係也是需要釐清的，根據 Fayyad 等人對 KDD 的定義：「The nontrivial Process of identifying valid、novel、potentially useful, and ultimately understandable pattern in data」，其流程步驟是：先理解要應用的領域、熟悉相關知識，接著建立目標資料集，並專注所選擇 (Selection) 之資料子集；再從目的資料中作前置處理 (Pre-processing)，去除錯誤或不一致的資料；然後作資料簡化與轉換工作 (Transformation)；在經由「Data Mining」的技術程序成為樣型 (Patterns)、做迴歸分析或找出分類型態；最後經過「Interpretation/Evaluation」成為有用的

知識。這些程序是一個循環的關係，一直重複的步驟，最後才得到一些有用的知識。所以，KDD 是一連串的程序，Data Mining 是其中的一個步驟而已。



圖三　資料倉儲與資料採礦的關係

總而言之，Data Mining，Data warehouse，KDD 三者的關係可以如此釐清，即 Data warehouse 是一個經過處理、整合之資料庫，而 KDD 是一種知識發現的一連串程序，Data Mining 只是 KDD 的一個重要程序。它們最終目的，乃為組織取得決策支援所需的資訊，這個資訊是突破盲點、見人所未見的知識和訊息，能替組織取得競爭優勢。

陸．資料採礦的功能

一般而言，Data Mining 功能可包含下列五項功能：(1)分類 (classification) (2)推估 (estimation) (3)預測 (prediction) (4)關聯分組 (affinity grouping) (5)同質分組 (clustering)。茲將這些功能的意義及可能使用的技巧簡述如下：

一、分類

按照分析對象的屬性分門別類加以定義，建立類組 (class)。例如，將信用申請者的風險屬性，區分為高度風險申請者，中度風險申請者及低度風險申

請者。使用的技巧有決策樹 (decision tree)，記憶基礎推理 (memory-based reasoning) 等。

二、推估

根據既有連續性數值之相關屬性資料，以獲致某一屬性未知之值。例如按照信用申請者之教育程度、行為別來推估其信用卡消費量。使用的技巧包括統計方法上之相關分析、迴歸分析及類神經網路方法。

三、預測

根據對象屬性之過去觀察值來推估該屬性未來之值。例如由顧客過去之刷卡消費量預測其未來之刷卡消費量。使用的技巧包括迴歸分析、時間數列分析及類神經網路方法。

四、關聯分組

從所有物件決定那些相關物件應該放在一起。例如超市中相關之盥洗用品 (牙刷、牙膏、牙線)，放在同一間貨架上。在客戶行銷系統上，此種功能係用來確認交叉銷售 (cross-selling) 的機會以設計出吸引人的產品群組。

五、同質分組

將異質母體中區隔為較具同質性之群組 (clusters)。同質分組相當於行銷術語中的區隔化 (segmentation)，但是，假定事先未對於區隔加以定義，而資料中自然產生區隔。使用的技巧包括 k-means 法及 agglomeration 法。

柒．資料採礦的應用

Data Mining 導入企業，其重點在於企業領域方面的知識，而它的 Domain-specific Tools 要結合企業中使用者的語言和分析過程，才能發揮工具的效能與增進企業的智慧。換句話說，就是要顛覆常規和超越平日的想像，展現企業目標與問題的知識，以支援解釋別人看不到、看不出的資訊來。企業必須能夠從巨大資料庫中挖掘到濃縮、先前不知、可理解的資訊，並從使用中獲利。例如，一個發行管理共同基金 (mutual funds) 的企業體要發掘潛在客戶，它要能整合客戶的帳戶、人口統計、生活型態等資料。也就是說要能把資料庫

中人口資料切分成為一些關鍵子集合：都市化情況、婚姻狀態、家庭所得、年齡、風險偏好、高淨值等。最後，依據資料挖寶分析結果，可區分集群和從事推廣促銷活動，成功的把共同基金推展至市場上。

目前企業界把 Data Mining 應用在許多領域。例如，行銷、財務、銀行、製造廠、通訊等。並且產學合作下，發展出許多實用的系統，例如 MDT、Coverstory and Spotlight、NichWork visualization system、LBS、FALCON、FAIS、NYNEX、TASA 等等。這些資料發掘的系統，應用非常廣泛，例如有一個應用在行銷領域的例子：經由記錄客戶的消費記錄與採購路線，超級市場可以設計出更吸引顧客購買的環境。根據資料採礦出特別的資訊來，因此現在超級市場的廚房用品，是按照女性的視線高度來擺放。根據研究指出：美國婦女的視線高度是 150 公分左右，男性是 163 公分左右，而最舒適的視線角度是視線高度以下 15 度左右，所以最好的貨品陳列位置是在 130 至 135 公分之間。

企業界實際發展 Data Mining 時，效能並不能預期，因為有許多因素影響著。例如，不充足的教育訓練、不適當的支援工具、資料的無效性、過於豐富的樣型 (patterns)、多變與具時間性的資料、空間導向資料 (spatially oriented data)、複雜的資料型態、資料的衡量性 (scalability)。這說明資料與知識的發掘是一項資訊豐富性的工作，面對易變的環境，沒有現成的 Model 馬上可用，也不要期望按照程序即能成功。因此，我們要體會一些潛在的因素，如資料取捨、實體關係性、數量多寡、複雜性、資料品質、可取得性、變遷、專家意見等因素，才能做好資料採礦工作。

Data Mining 對每個公司來說都是一種重要的策略性的的計畫，而將之列為高度機密，所以要調查各家公司到底用 Data Mining 來做什麼樣的事其實相當不容易。根據 Two Crows Corp.最近的調查顯示，Data Mining 主要的三個應用方式-如我們所預期的-都在市場推廣方面，分別是：Customer Profiling、Targeted Marketing、以及 Market-Basket Analysis。

在 Customer Profiling 方面，我們希望找出客戶的一些共同的特徵，希望能藉此預測哪些人可能成為我們的客戶，以幫助行銷人員找到正確的行銷對象。Data Mining 可以從現有客戶資料中找出他們的特徵，再利用這些特徵到潛在客戶資料庫裡去篩選出可能成為我們客戶的名單，作為行銷人員推銷的對象。行銷人員就可以只針對這些名單寄發廣告資料，以降低成本，也提高行銷的成功率。

Market-Basket Analysis 主要是用來幫助零售業者瞭解客戶的消費行為，譬如哪些產品客戶會一起購買，或是客戶在買了某一樣產品之後，在多久之內會

買另一樣產品等等。利用 Data Mining，零售業者可以更有效的決定進貨量或庫存量，或是在店裡要如何擺設貨品，同時也可以用來評估店裡的促銷活動的成效。

客戶關係的管理是 Data Mining 的另一個常見的應用方式。我們可以由一些原本是我們的客戶，後來卻轉而成為我們競爭對手的客戶群中，分析他們的特徵，再根據這些特徵到現有客戶資料中找出有可能轉向的客戶，然後公司必須設計一些方法將他們留住，因為畢竟找一個新客戶的成本要比留住一個原有客戶的成本要高出許多。

近來電話公司、信用卡公司、保險公司、股票交易商、以及政府單位對於詐欺行為的偵測 (Fraud Detection) 都很有興趣，這些行業每年因為詐欺行為而造成的損失都非常可觀。Data Mining 可以找出可能的詐欺交易，減少損失。財務金融業可以利用 Data Mining 來分析市場動向，並預測個別公司的營運以及股價走向。Data Mining 的另一個獨特的用法是在醫療業，用來預測手術、用藥、診斷、或是流程控制的效率。

下面是一些 Data Mining 的在科學、行銷、工業、商業、體育…等各方面運用的類型：

1.在財務金融方面，預測市場動向，防範犯罪詐欺。

2.分析客戶的行為，可以讓您看出您的客戶是不是準備要轉向您的競爭對手。資料挖採中的前後行為分析 (Sequential Pattern Detection) 功能讓您分析那些已經轉向您的競爭對手的客戶在轉向期間的行為，如此您就可以在現有客戶中找到可能轉向的客戶，想辦法留住他們。

3.資料挖採可以幫您找出從前的一些信用不良的客戶的特徵，而從這些特徵您就可以從現有客戶中找出可能有不良信用的客戶，防止產生壞賬，也可以過濾這些人成為您的客戶。

4.資料挖採中的客戶分類 (Segmentation) 功能，可以讓您更瞭解您所服務的客戶，這樣您就可以設計更好的產品來滿足您的客戶的需求。

5.商業智慧所要解決的問題還包括如何減低詐欺或不實的申報 (Fraud)。利用資料挖採的技術，您可以在特定的客戶群中找出可能的詐欺行為，如此才能減少損失，增加利潤。

6.如果採用不同的價格策略，是否能增加市場佔有率？

7.什麼時候才是推出新產品的好時機？

8.我們與競爭對手的優劣勢如何？

9.讓我們獲利高的客戶們有什麼共同的特徵？

10.當我們的客戶要轉向我們的競爭對手之前，是否有何前兆？

11.如何認定客戶的信用風險狀況？

12.如何設計更好的保險產品來吸引客戶，讓客戶滿意？

13.一個經紀人在一個星期中應該可以賣出多少共同基金？

14.於銷售資料中，發掘顧客的消費習性。

15.根據以往審核的資料，找尋核發信用卡的規則。

16.在 NBA 球賽資料中，找出球員的強弱點。

17.從消費及繳費資料中，預警信用卡呆帳可能。

18.從通話記錄資料中，預警盜打電話可能。

19.從太空船拍攝的影像資料，找尋星球上的火山。

20.星際星體分類。

捌．資料採礦的進行步驟

資料採礦既然可以增加企業智慧，提昇企業競爭優勢，到底應該如何進行呢？根據 Glymour 等人的研究，提出一個參考的進行步驟如下：

1.理解資料與進行的工作

2.獲取相關知識與技術 (Acquisition)

3.融合與查核資料 (Integration and checking)

4.去除錯誤或不一致的資料 (Data cleaning)

5.發展模式與假設 (Model and hypothesis development)

6.實際資料採礦工作

7.測試與檢核所挖掘的資料 (Testing and verfication)

8.解釋與使用資料 (Interpretation and use)

從八個步驟來看，Data Mining 牽涉大量的規劃與準備，而從其他文獻得知，專家聲稱高達 80%的過程花在準備資料階段，這包括表格的 Join 以及可能相當大量的資料轉換。從這個角度看，Data Mining 只是知識發掘過程中的一個步驟而已，而達到這個步驟前還有許許多多的工作要完成。

玖．資料採礦的工具

Data Miming 的工具是利用資料來建立一些模擬真實世界的模式(Model)，利用這些模式來描述資料中的特徵 (Patterns) 以及關係(Relations)。這些模式有兩種用處，第一，瞭解資料的特徵與關係可以提供你做決策所需要的資訊，譬如 Association Model 可以幫助超級市場或百貨店規畫如何擺設貨品。第二，資料的特徵可以幫助你做預測，例如你可以從一份郵寄名單預測出哪些客戶最可能對你的推銷做回應，所以你可以只對特定的對象做郵購推銷，而不必浪費許多印刷費郵寄費而只得到很少的回應。

Data Mining 可以建立六種模式：Classification、Regression、Time Series、Clustering、Association、以及 Sequence。Classification 以及 Regression 主要是用來做預測，而 Association 與 Sequence 主要是用來描述行為 (例如消費行為)。Clustering 則是二者都可以用的上。

一、Classification

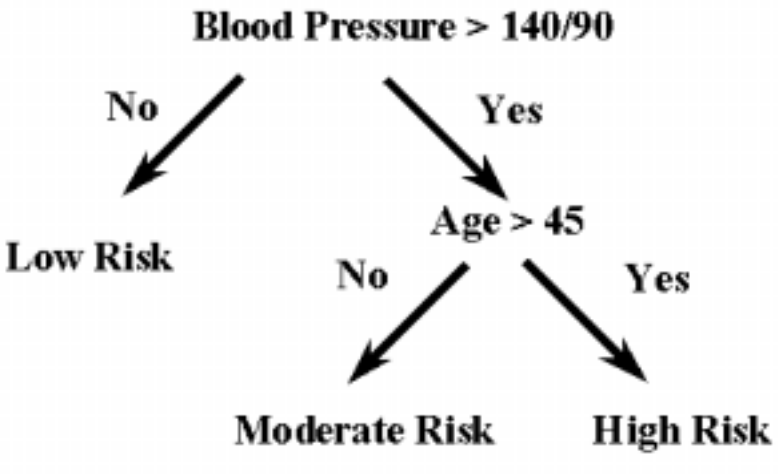
Classification 是根據一些變數的數值做計算，再依照結果作分類。(計算的結果最後會被分類為幾個少數的離散數值，例如將一組資料分為"可能會回應"或是"可能不會回應"兩類)。Classification 常常被用來處理如前面說到的郵寄對象篩選的問題。我們會用一些已經分類的資料來研究它們的特徵，然後再根據這些特徵對其他未經分類或是新的資料做預測。這些我們用來尋找特徵的已分類資料可能是來自我們的現有的歷史性資料，或是將一個完整資料庫做部份取樣，再經由實際的運作來測試；譬如利用一個大的郵寄對象資料庫的部份取樣來建立一個 Classification Model，以後再利用這個 Model 來對資料庫的其他資料或是新的資料作預測。

Classification 通常會牽涉到兩種統計方法：Logistic Regression 以及 Discriminant Analysis。然而因為 Data Mining 已漸普遍，所以 Neural Nets 以及 Decision Tree 也漸漸受到採用。雖然這些統計方法本身都十分複雜，但使用者並不會牽涉到這些繁雜的統計。

Neural Nets 使用許多參數 (每個參數代表 Net 上的一個 Node) 來建立一個模式, 這個模式接受一組輸入值來預測出一個連續值或分類值。每一個節點 (Node) 都是一個函數，這個函數是使用輸入該節點的相鄰節點值的加權總和 (Weighted Sum) 做運算。

在建立一個模式的過程中，我們要用一些資料來'餵'給這個網路，'訓練'它來找到一組能夠產生最佳輸出結果的加權值 (Weights)。有一種最常用的'訓練法'稱為 Back-Propagation，它是把輸出結果與一個已知的正確結果相比。每次相比之後就產生另一組調整過的 Weights，然後再產生一個新的輸出值再與該已知值相比。這個過程經過反覆的執行後，這個 Neural Net 就被'訓練'得能夠相當正確的做預測了。

可是 Neural Net 有兩個問題。首先，Neural Net 最受質疑的是它的'曖昧不明'的特性，也就是它做的預測所根據的因素並不明確。第二，Neural Net 對測試資料可以做相當正確的預測，但是對真實資料預測的準確性則較差。但是現在已經有一些新的技術可以改正這個缺點。Decision Tree 則是利用一系列的規則來得到一個類別或數值。例如，你想把申請貸款的人歸類成'風險高'與'風險低'兩種。有了這個 Decision Tree，銀行的放款人員就可以審查申請人的條件，決定該人是屬於高風險或低風險群。例如'收入高於 40000'而且'高負債'的人會被歸為高風險之類，而'收入低於 40000'而且'工作超過 5 年'則會被歸為低風險之類。Decision Tree 現在相當普遍，因為它所做的預測相當正確，而且又比 Neural Net 容易瞭解。Decision Tree 與 Neural Net 也可以用來做 Regression，某些種類的 Neural Net 甚至可以用來做 Clustering。



圖四　Decision Trees

二、Regression

Regression 是使用一系列的現有數值來預測一個連續數值的可能值。

三、Time-Series Forcasting

Time-Series Forcasting 與 Regression 很像，只是它是用現有的數值來預測未來的數值。Time-Series Forcasting 的不同點在於它所分析的數值都與時間有關。Time-Series Forcasting 的工具可以處理有關時間的一些特性，譬如時間的階層性 (例如每個禮拜五個或六個工作天)、季節性、節日、以及其他的一些特別因素如過去與未來的關連性有多少。

四、Clustering

Clustering 是將資料分為幾組，其目的是要將組與組之間的差異找出來，同時也要將一個組之中的成員的相似性找出來。Clustering 與 Classification 不同的是，你不曉得它會以何種方式或根據什麼來分類。所以你必須要有一個分析師來解讀這些分類的意義。

五、Association

Association 是要找出在某一事件或是資料中會同時出現的東西。Association 主要是要找出下面這樣的資訊：如果 Item A 是某一事件的一部份，則 Item B 也出現在該事件中的機率有 X%。(例如：如果一個顧客買了低脂乳酪以及低脂優酪乳，那麼這個顧客同時也買低脂牛奶的機率是 85%。)

六、Sequence Discovery

Sequence Discovery 與 Association 關係很密切，所不同的是 Sequence Discovery 中相關的 Item 是以時間區分開來 (例如：如果做了 X 手術，則 Y 病菌在手術後感染的機率是 45%。又例如：如果 A 股票在某一天上漲 12%，而且當天股市加權指數下降，則 B 股票在兩天之內上漲的機率是 68%)。

有一點很重要的是，沒有一種 Data Mining 的工具可以應付所有的要求。對於某一種問題，資料本身的特性會影響你所選用的工具。所以你可能會需要用到許多不同的工具以及技術從資料中找到最佳的模式。在產學界合作下，近二年有驚人的發展，而各種工具只在某些領域下有特別的效能，也就是說尚無

適用所有業種 用途的工具問世。以下介紹一般常用的工具分類：(1)Case-Based Reasoning (2)Data Visualization (3)Fuzzy Query and Analysis (4)Knowledge Discovery (5)Neural Networks。列於表二：

從表二可以發現資料採礦技術的多樣化，從傳統分析工具，例如統計迴歸預測模型、資料庫分割、連接分析、偏差偵測等。但是，重要的是這些產品應用新的技術，如類神經網路、機器學習、專家系統等人工智慧的工具，使 AI 找到新的應用 Domain。但是，近年浮現的新技術：遺傳演算法 (Genetic algorithms)，卻無確切證據顯示在 Data Mining 工具產品中使用，本文認為遺傳演算法的特性，必然在 Data Mining 領域中有出色的演出。

表二　資料採礦分析工具

Data mining tools	定義	代表性產品
Case-based Reasoning	在關聯式資料庫中提供一個 Means 找出 record 以發現類似規範的記錄或一般記錄	1.CBR Express 2.Esteen 3.Kate-CBR 4.The Easy Reasoner
Data Visualization	其目標是從不同的角度，讓資訊以圖形方式呈現，讓使用者容易和快速的使用。這工具把不同資料次集合，或不同彙總性資料，讓使使用者快速的了解。	1.Alterian 2.AVS/Express 3.Visualization Edition 4.Axum 5.Discovery 6.SPSS Diamond 7.Visual Insight
Fuzzy Query and Analysis	模糊理論積極的承認人主觀性問題的存在，進而以模糊集合來處理不易量化問題，故能找出意想不到的資訊。模糊理論發展的工具能使使用者容易導入既定的標準中，而此種工具最大用途是，當使用者要查核多重標準，以及要改變每一種標準時。	1.CubiCalc 2.FuziCalc 3.Fuzzy TECH for business 4.Quest
Knowledge Discovery	這些工具特別設計以便確認那些已存在變數間的顯著關係，也就是當它們有可能多重關係時，特別有用。這些 data mining 工具能幫助指出巨量變數間的關係，發現盲點創造巨大的商機。	1.Aria 2.Answer tree 3.CART 4.DARWIN 5.Enterprise Miner 6.DataEngine
Neural Networks	類神經網路技術的目標是發現與預測資料的關係，它與傳統統計方法的區別是，它可以訓練學習發現的關係，並且可適用於線性與非線性的情況，並可以彌補資料品質較差的情況，而處理出品質不錯的資訊來。	1.BackPack 2.BrainMaker 3.Loadstone 4.NeuFrame/NeuroFuzzy 5.Neural network Browser 6.Neural connection 7.Neural network Utility 8.Neuralyst For Excel

遺傳演算法是一種全新的最佳化空間搜尋法，其最初概念是由 John Holland 於 1975 年提出，其主要目的如下：1.以嚴密而具象的科學方法解釋自然界中「物競天擇、適者生存」的演化過程。2.將生物界中基因演化重要機制以資訊科學軟體實作模擬。近年來，資訊科技的長足進步，在更快穩定的系統支援下，遺傳演算法被各領域廣泛應用。於是人工智慧領域中的自我學習機制、各類最佳化問題的快速求解，它提供了一種不同以往的思考模式，運用在 Data Mining 上，可以在巨量資料中快速搜尋、比對、演化出最佳點，並且具有學習機制，可在 Data Mining 領域綻放光芒。

遺傳演算法是應用演算法的適應函數來決定搜尋的方向，再運用一些擬生物化的人工運算過程，例如選擇 (selection)、複製 (reproduction)、交配 (crossover) 和突變 (mutation) 等進行演化，週而復始地進行一代一代的演化，以求得一個最佳的結果。它具有強固性 (robustness) 與求值空間的獨立性 (domain independence)。強固性使問題的限制條件降到最低，並大幅提高系統的容錯能力；而求值空間的獨立性則使遺傳演算法的設計單一化，且適用於多種不同性質、領域的問題。因此，利用它於 Data Mining 領域中，可以發掘出不同的資訊、別人看不出的資訊，必然帶給企業體巨大的商機。遺傳演算法實際運作，非本文主題，然可斷定它必然成為 Data Mining 的分析利器。

拾．Data Mining軟體

1.MLC++ (pd)

2.MOBAL (pd)

3.MOBAL (pd)

4.Emerald (rp)

5.Kepler (rp)

6.Clementine (cp)

7.DataMind DataCruncher (cp)

8.Darwin (cp)

9.Intelligent Miner (cp)

10.INSPECT (cp)

11.NeoVista Solutions (cp)

12.Nuggets (cp)．Partek (cp)

13.Polyanalyst (cp)

14.SAS Data Mining (cp)

15.SGI MindSet (cp)

16.Knowledge Explorer (cp)

17.DataEngine (cp)

18.Delta Miner (cp)

19.S-PLUS (cp)

20.MATLAB (cp)

21.Mathematica (cp)

22.XGOBI (pd)

23.Crystal Vision neé ExplorN．sphinxVision

24.Graf-FX．IRIS

25.Spotfire

26.Netmap

27.Visible Decisions Inc.

28.Visual Mine

拾壹．資料倉儲與資料採礦在顧客關係管理上的運用

資料採礦 (Data Mining) 指收集和顧客有關的資料作分析，並把原始資料轉換成商機。從 CRM 的整體架構來說，資料採礦是整個 CRM 的核心精神，也是構成商業智慧的基礎。 完整的資料採礦不單可以做到準確的目標市場行銷，當分析的工具及技術成熟時，加上資料倉儲 (Data Warehousing) 提供大量儲存顧客資料能力，能讓資料採礦作到大量客製化 (Masscustomization)，做到準確的對個人顧客作行銷，也就是所謂的一對一行銷。

企業對所服務的顧客作利潤貢獻度的分析，將有助於了解服務顧客的成本及顧客利潤貢獻度的關係，而對不同的顧客提出不同的服務策略。企業利潤貢獻度高且服務成本低的顧客，這一群組的顧客應是對企業最有利的客群，維繫並發展這一群組的顧客應是經營上最主要的考量方向，企業內部的資源也應該主要投注在維繫且發展這一群組的顧客。

對於顧客充分的了解，才能有效的和顧客建立關係，進而有效的行銷出擊，創造訂單。資料採礦是 CRM 中商業智商 (Business Intelligence) 的基礎，透過資料採礦，有效的提供行銷上、銷售上、服務上的決策支援，讓作業人員可以得到充分的資訊來行動，達到在適當的時間、地點，提供顧客適量的產品及服務，大幅提高作業的效率，這也就是所謂的行銷智慧 (Marketing Intelligence)、銷售智慧 (Sales Intelligence) 及服務智慧 (Service Intelligence)。

一旦對於顧客的了解程度提高以後，針對目標市場行銷的準確度 (hit rate) 大幅提高，將直接影響到成交的比例，同樣數量的 DM 出去，成交比例從過去 10%可能提高到 80%，行銷的效益/成本比將大幅提高。因為了解顧客，所以可以有效的過濾無效的樣本，在未接觸顧客以先，就已經知道顧客是會成交的對象，因此以往浪費的行銷成本作散彈打鳥的方式，變成一擊必中 (One-Shot) 的方式。

企業在導入客戶關係管理前，必需先誠實地作一次全面體檢，瞭解自身的優勢與缺點，進而傾聽客戶的聲音，確實瞭解所有與客戶互動的管道，開始規劃整體的 CRM 架構，CRM 架構中最重要的是客戶互動資料庫的建立與全面客戶關係管理的心態，在建立完整的系統後，便可整合所有公司資源，開始將資料庫豐富化，並利用資料庫與客戶互動，進而透過分析工具和資料擷取，得到更有價值的企業營運資訊，在營運的過程中，必須適度地將資訊回饋至客戶互動資料庫，而成為一個良性的循環。

未來不僅是一個數位化的資訊時代，更是一個消費者導向的時代。如果一家企業無法好好掌控其主要消費顧客族群，那麼這家公司降尚失其競爭力和應有的優勢。唯有建立良好的顧客關係管理，並不時的監督和提出因應決策，方可挽留住消費者的心。要建立良好的顧客關係管理，並須依賴公司顧客資料的完善資料倉儲，只有完善的資料倉儲是不夠，資料倉儲並不會幫你分析並提供決策，所以還需要資料擷取和相關的專業人士的結合。

CRM 客戶關係管理在台灣正處於起跑點上，可預見的是客戶關係管理將是繼電子商務後下一個熱門的課題，在面對網際網路中數量驚人的潛在客戶，

企業營運者如果想要持續地擴充客戶、留住客戶進而提高利潤及營業額，請不要遲疑，現在就導入客戶關係管理，絕對讓你有出乎意料的效益。

參考文獻

謝邦昌、葉瑞鈴，「統計在資料探勘之應用」，主計月報，第 530 期，頁 67-84，2000 年。

Berry, M. J. A. and G. Linoff, "Data Mining Techniques: for Marketing, Sales, and Customer Support," New York: John Wiley & Sons, 1997.

Chen,M. S., J. Han and P.S. Yu, "Data Mining: An Overview from a Database Perspecitve," IEEE Trans. Knowledge and Data Engineering, 8:886-883, 1996.

Fayyad, U.M., G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence, pp.37-54, Fall, 1996.

Fayyad, U. M., G. Piatetsky-Shapiro, P.Smyth and R. Uthurusamy, "Advances in Knowledge Discovery and Data Mining," Cambridge, MA: AAAI/MIT Press, 1996.

Fayyad, U.M., "Mining Databases : Towards Algorithms for Knowledge Discovery", IEEE Computer Society Techinical Committee on Data Engineering, pp.1-10, 1998.

The Tool of Commercial Intelligence -Data Mining

YU-TING CHENG*, CHIH-HSIUNG SU**

**Department of Statistics, National Chengchi University*

***Department of Accounting, Chihlee Institute of Commerce*

ABSTRACT

The main purposes of the article are to descript the meaning of data mining and data warehouse, to compare the differences of data mining and statistics analysis, and to descript the relationships between data warehouse, KDD and data mining.

At the same time, we also descript the functions, applications, processes, and instruments of data mining.

Finally, we descript how to apply data warehouse and data mining in customs relationship management.

Keywords: data mining, data warehouse, commercial intelligence, customs relationship management