



MIT, Boston, USA, by Frank S.C. Tseng
(<http://www2.nkfust.edu.tw/~imfrank>)

第十六章 資料庫系統研究方向



本章內容

- 16.1 資料模式的研究
- 16.2 使用者介面
- 16.3 資料庫管理系統的核心技術
- 16.4 資料儲存結構與存取方法
- 16.5 分散式技術
- 16.6 物件導向技術
- 16.7 從資料中採擷知識
- 16.8 資料庫系統的硬體架構
- 16.9 架於資料庫系統上的各種輔助系統
- 16.10 資料保密與安全性方面的議題
- 16.11 管理方面的議題
- 16.12 結語



簡介

- 為何要研究？

- 人會害怕，是因為「無知」。
- 做研究才能開啟我們了解當前未知的領域

Marie Curie (居禮夫人)：

*“Nothing in life is to be feared,
it is only to be understood.*

*Now is the time to understand more,
so that we may fear less.”*



資料庫相關的研究主題

- 資料模式
- 使用者介面與各類查詢語言的訂定
- 資料庫管理系統的核心技術
- 內部的資料儲存結構與存取方法
- 分散式技術
- 物件導向技術
- 資料倉儲 (Data Warehouse) 與資料發掘 (Data Mining) 、
- 資料庫系統的硬體架構
- 架於資料庫系統上的各種輔助系統



資料模式的研究

- 必須定義以下的內容才算完整：
 - 資料的表示法—資料結構，
 - 資料表示法上的限制條件—整合限制條件，
 - 資料表示法上的運算。
- 個體-關係模式 (E/R Model) [P.P.S. Chen (1976)]。
- 語意資料模式 (Semantic Data Model, SDM) [M.M. Hammer and D.J. McLeod (1978)]
- 函數式資料模式 (Functional Data Model) [D. Shipman (1981)]
- 物件導向式資料模式 (Objected-Oriented Data Model) — [W. Kim (1995)]



資料模式的研究

- 研究各種資料模式來反應真實世界裡的資料相關性
- 也可以是探討不同模式之間的關係或轉換關係。
- 在 D.C. Tsichritzis 與 F.H. Lochovsky 所合著的書 [D.C. Tsichritzis and F.H. Lochovsky (1982)] 中對各種傳統上由關聯式模式所延伸出來的模式有深入的剖析。
- 針對「空間資料庫系統」(Spatial Database Systems)、
「時間資料庫系統」(Temporal Database Systems) 與
傳統資料庫系統的結合而發展出來



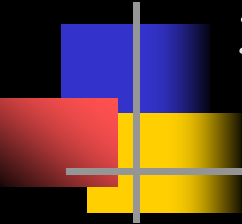
使用者介面

- 查詢語言的規格訂定 (Specification of Query Language) :
 - Ingres 的 QUEL。
 - SQL 的被標準化
 - 以 SQL 為主體再做進一步的擴充：Transact-SQL、UniSQL/X、用於整合異質性資料庫系統的 MSQL (Multi-Database Query Language)、XML 的 XQL



使用者介面

- 親和式使用者介面 (User-Friendly Interfaces)：自然語言、語音查詢
- Query-By-Example [M.M. Zloof (1975)]
- 將整個資料庫看成是單一個關聯表的 Universal Relation 概念 [J.D. Ullman]
- 「概念式查詢結果」 (Conceptual Query Answering) [A. Motro (1988)]
- 「概略式查詢結果」 (Intensional Query Answering)
- 「近似查詢結果」 (Approximate Query Answering) 研究



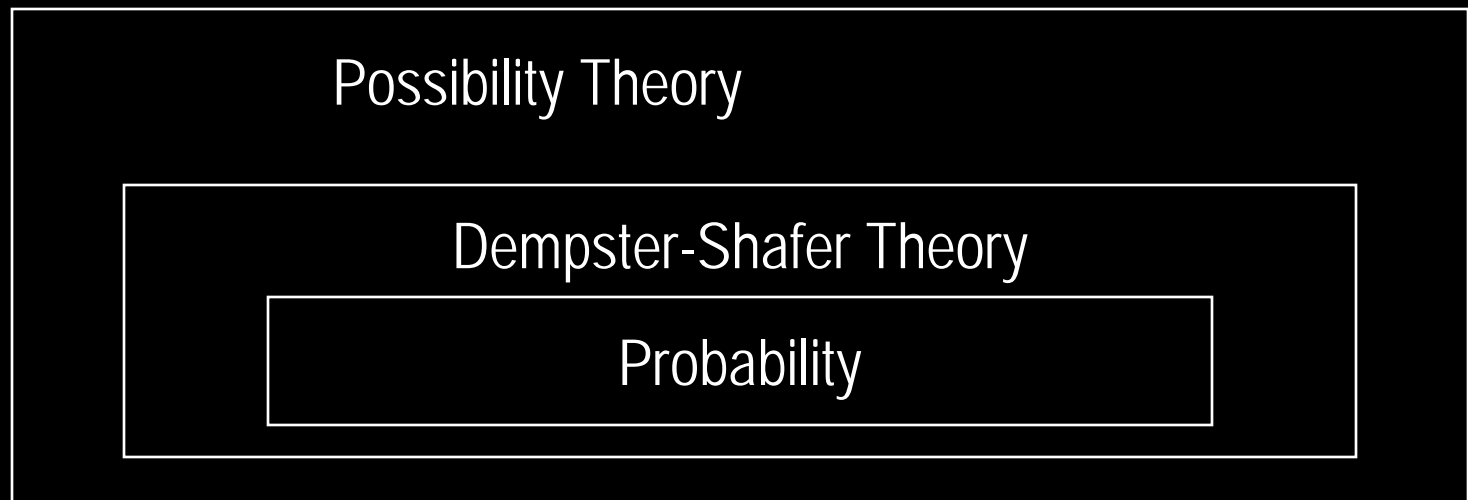
資料庫管理系統的核心技術

- 查詢處理與查詢的最佳化 (Query Processing and Optimization)
- 異動管理 (Transaction Management)
- 資料的整合與安全性 (Integrity and Security)
- 邏輯資料庫的設計理論 (Logical Database Design Theory)
- 不確定資料的處理 (Missing Information Handling)
 - 「不精確的資料」 (Imprecise Data) 或稱「不完整的資料」 (Incomplete Data)，以及「不確定的資料」 (Uncertain Data)



Theory for Uncertainty

- Probability Theory
- Dempster-Shafer Theory
- Possibility Theory (Fuzzy Set Theory)





資料庫管理系統的核心技術

- 資料的壓縮與快速索引(Data Compression and Fast Indexing) — 針對大量本文資料或影像的壓縮，同時還要考慮快速還原或存取。
 - 「赫夫曼編碼」 (Huffman Encoding)
 - 「簽名檔」 (Signature File)
 - 配合使用搜尋過濾器 (Search Filter)
 - 數位浮水印 (Stagnography)
- Real-Time Database Systems



資料儲存結構與存取方法

- 「搜尋樹狀索引結構」 (Search Tree Index Structure) 與 「雜湊技巧」 (Hashing)
 - 以儲存本文資料為主的結構，如： B-Tree, B+-Tree 等
 - Main-memory Database 的 T-tree
 - 以儲存影像資料為主的結構，早期所提出來的各種結構如： R-Tree, R⁺-Tree, 以及各種類型的 Quadtree
 - Cartesian product file



資料儲存結構與存取方法

- 靜態的雜湊技巧 (Static Hashing) — 雜湊表 (Hash Table) 不會隨著資料的增加而擴張。
- 動態的雜湊技巧 (Dynamic Hashing) — 雜湊表 (Hash Table) 會隨著資料的增加而擴張。如：Virtual Hashing, Extendible Hashing, Linear Hashing, Trie Hashing [W. Litwin (1981)] 等。



資料儲存結構與存取方法

- 提昇「部份吻合查詢」(Partial Match Query) 與「正交區間查詢」(Orthogonal Range Query) 此兩種查詢類型的處理速度
 - K-D Tree、BD Tree、G-K-D Tree、Modified K-D Tree，與 Modified G-K-D Tree
 - Multi-key Hashing、String Homomorphism Hashing、Multi-Dimensional Directory、Optimal Cartesian Product File，以及 Greedy File



分散式技術

- 如何以最佳的方式將資料分散到各個資料站 (Data Distribution, Distributed Database Design) 。
- 考慮的方向有：如何對關聯表做水平切割與垂直切割？
- 如何充分讓各資料站同時做平行處理 (Maximize Parallelism of Query Execution) ？
 - 在查詢處理上儘量區域化 (Localized) 以減少網路上的傳輸 (Minimize Inter-Site Traffic)



分散式技術

- D.A. Bell [D.A. Bell (1984)] 曾經證明了 File Placement and Allocation Problem 是屬於 NP-Complete 的問題，所以一般都是以 Heuristic 的方法找出近似最佳的解法來解決
- 最近也有人嘗試以遺傳演算法 (Genetic Algorithms) 來解決此種問題 [A. Kumar, R.M. Pathak, and Y.P. Gupta (1995)]



分散式技術

- 分散式查詢做最佳化處理 (Optimization of Distributed Queries)
 - 「以查詢的特性來做最佳化」 (Qualitative Optimization)
 - 「以表格所含的數量來做最佳化」 (Quantitative Optimization)
- 網路錯誤時的回復 (Recovery From Network Failure)
 - Quasi-Copy, 2-phase-commit,...



分散式技術

- 死結的偵測與處理 (Deadlock Detection and Handling)
- 效能評估 (Performance Evaluation)— 透過測試標記 (Benchmark) 評估資料庫效能
- 關聯式查詢測試標記 Wisconsin
- 還有複雜的關聯式查詢測試標記 AS³AP
- 異動處理效能評議協會 (Transaction Processing Performance Council, 簡稱 TPC) 所訂定的 TPC-A, TPC-B, TPC-C, TPC-D, TPC-E 與 TPC-CS (Client/Server)。



分散式技術

- 專門針對物件導向資料庫系統的測試標記，如：
OO1 測試標記、HyperModel 測試標記 (The HyperModel Benchmark)，與 OO7 測試標記 (The OO7 Benchmark)
- 行動計算 (Mobile Computing) 所引發的問題
 - Optimal Stop Algorithms
 - ...



異質性的分散式環境

- 分散式環境下的整合架構 (Integration Architecture) —
 - 聯邦式的整合方式 (Federated Schema Integration)
 - 整體式的整合方式 (Global Schema Integration)。
 - 以 MSOL 則整合介面
- 不同資料庫在綱要上的語意衝突 (Semantic Discrepancies)



Advanced Transaction Model

- Nested Transactions 、
- Long Transactions
- Cooperative Transaction Model,
- Open Nested Transaction Model,
- Multilevel Transaction Model,
- Saga,
- Polytransactions, 等



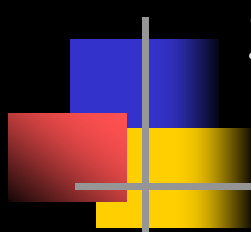
物件導向技術

- 物件導向式分析
 - Grady Booch 的物件導向式分析設計方法論
 - James Rumbaugh 物件模型技術 OMT Ivar Jacobson的使用個案方法論 (Use Case Driven Approach)
 - 統合為一，稱為UML
- 物件導向式資料庫綱要設計
- 物件導向式資料模式上的問題探討
- 物件導向式模式與 XML, HTML 的關係



資料倉儲系統

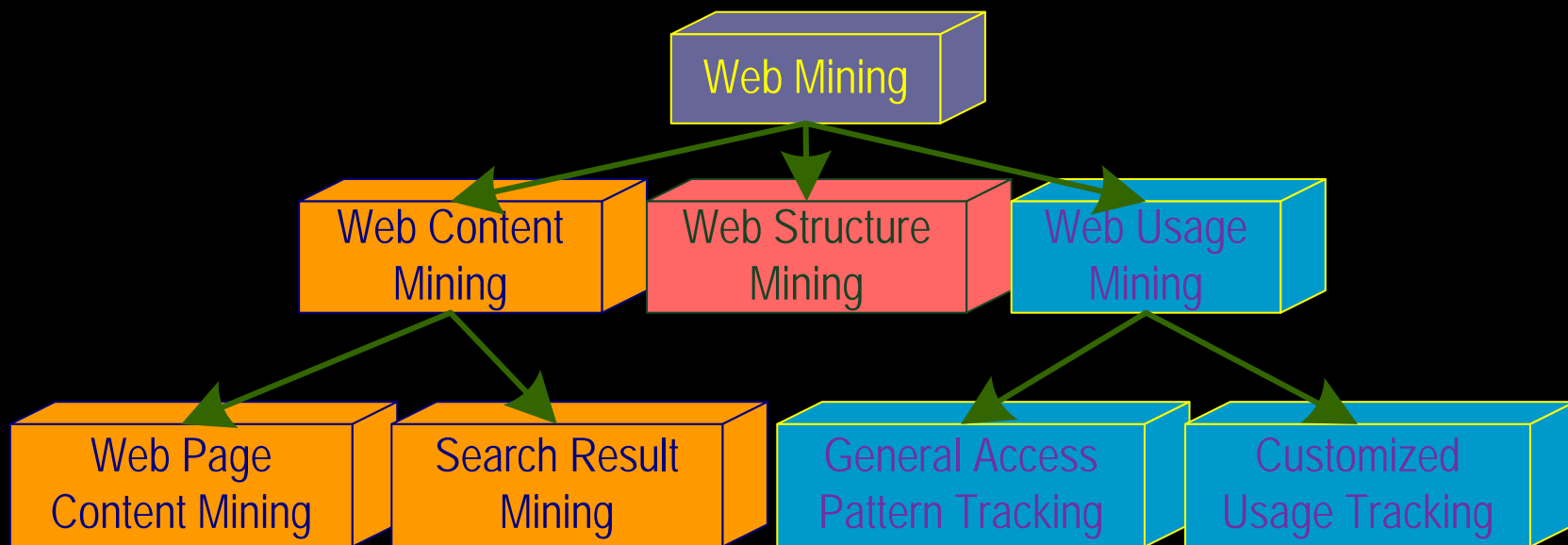
- 決策品質完全是取決於「資料品質」(Quality of Data) 的好壞。
- 「垃圾進，垃圾出」(Garbage In, Garbage Out)。
- 它的建置過程要求內存的資料必須要經過品管，也就是要經過淨化、修補與增強後才能進到資料倉儲中，需配合「異質性資料庫整合」，讓資料倉儲的內容能夠在整合多方來源之後變得更充實
- 整合後的資料須經過淨化 (Cleaning)、修補與強化，才能讓產出結果形成有意義的決策。
- 以多維度的結果呈現



資料採礦 (Data Mining)

- 從雜亂的資料當中整理出某些頭緒，以得到以往所無法觀察得知的現象
- 結合了資料庫系統、人工智慧、統計學、高效能運算架構與視覺化等技術所產生的新興領域？
- Web Mining 利用此一技術來了解
 - 網站內容的特徵 (Web Content Mining)
 - 網站結構的特徵 (Web Structure Mining)
 - 使用者的行為模式 (Web Usage Mining)

Web Mining 的分類





Data Warehousing/Data Mining

- OLAP Structure
- 找出分類的規則 (Classification Rules)
- 找出關聯性規則 (Association Rules)
- 找出順序規則 (Sequence Rules)
- 找出同質時間序列 (Similar Time Series)
- 找出群集規則 (Clustering Rules)
- ...



知識管理系統

- 隱藏在人與人之間 (People-to-People) 的「隱含知識」 (Tacit Knowledge)
- 隱藏在人與文件資訊 (People-to-Information) 之間的「明確知識」 (Explicit Knowledge)
- 隱藏在人與資料 (People-to-Data) 之間的「潛在知識」 (Potential Knowledge)



文件倉儲系統

- Survey.com 的分析結果顯示：
 - 其實企業所需要的商業智慧大約只有 20% 是由存放在傳統關聯式資料庫中的結構化資料所推導出來的。
 - 其餘 80% 左右的商業智慧必須要到各式各樣的商業文件中去找尋
 - 與 Vilfredo Pareto 所提倡的 80:20 法則相符
- 例如：市場調查報告、專案進度報告、會議記錄、客戶的抱怨信件、專利申請書、競爭對手的廣告內容等，都是以文件形式儲存。



文件倉儲系統

- 自動萃取文件特徵內容 (Key Feature Extraction)、自動做文件分類 (Automatic Document Categorization)、自動做文件內容總結與歸納 (Automatic Text Summarization) 的操作環境
- 資料倉儲的建置僅能協助企業找出結構化資料中的商業智慧，並協助決策人員了解某些營運現象中所產生的 Who, What, When, Where, 以及 Which，
- 而文件倉儲的建置目標是協助使用者了解 Why？



文件採擷 (Text Mining)

- 以文件關鍵字索引的建立、文件特徵的擷取、文件的分類、文件的總結及文件的分群等文字分析的技術來對以純文字構成的文件做分析，產生可結構化的資訊，
- 盡可能的以結構化的形態將已分析過的文件資料加以表示，
- 最終則是根據企業模式來建立文件倉儲
- 以前稱為 “Information Retrieval”



論文投稿信件應用範例

敬啟者：

欣逢第七屆國際資訊管理學術研討會在 貴校舉辦，謹寄上由本人所撰寫的論文投稿。我們的論文題目為：

從資料處理的演進過程看資料庫系統的發展走向

我們投稿的組別為 “資訊管理組”，可以歸類為下列幾個領域之一

1. Database Management

2. Data Engineering

期盼能參與盛會，共襄盛舉。敬祝

教 安

國立高雄第一科技大學

資訊管理系 曾守正 敬上

E-mail: imfrank@ccms.nkfust.edu.tw

URL: <http://www2.nkfust.edu.tw/~imfrank>



論文投稿信件轉入資料表格

屬性	內含值
研討會名稱	第七屆國際資訊管理學術研討會
作者	曾守正
論文中文題目	從資料處理的演進過程看資料庫系統的發展走向
論文組別	資訊管理組
歸屬領域	Database Management, Data Engineering
網址	http://www2.nkfust.edu.tw/~imfrank
電子郵件信箱	imfrank@ccms.nkfust.edu.tw

如何萃取這些資料並存入結構化的資料表是新興的研究領域!



資料倉儲 vs. 文件倉儲

	Document Warehouse	Data Warehouse
相同點	<ol style="list-style-type: none">1. 具有相同的建置步驟，但是處理的資料一為 Documents，一為 Formatted Data。2. 都要面對大量的商業資料，並產生有用的資訊。3. 使用者可以快速瀏覽這些有用的資訊，萃取所需的資訊並進行比較。4. 都必須要能吸納各種不同的異質性資訊來源。5. 兩者形成互補關係 (Data Warehouse 協助我們看出 Who, What, When, Where, Which ; Document Warehouse 則協助我們了解 Why?)。	

資料倉儲 vs. 文件倉儲

	Document Warehouse	Data Warehouse
相 異 點	1. 希望得到 Text-oriented business intelligence。	1. 希望得到 Numeric-oriented business intelligence
	2. 資料來源為：企業內部文件檔案（或是文件庫）、會議記錄、研討會論文集、市場調查報告、市場研究報告、產業公報、政府機構發行之文件、E-mail、合約書、廣告信件。	2. 資料來源為：企業內部資料庫、POS (Point-of-Sale) 系統、ERP (Enterprise Resource Planning) 系統、財務或會計系統。
	3. 可以協助使用者快速過濾大量文件，但是難以精確分析出精準的結果，目標是萃取出某些問題的原因 (Why)。	3. 分析結果相當精準，並且可以依據人 (Who)、事 (What)、時 (When)、地 (Where)、物 (Which)，動態切換與檢視。
	4. 將大量文件的特徵與摘要萃取出來後予以分類歸納。	4. 將交易後的數字資料依據所需的維度加以統計、加總後呈現出來。
	5. 比較難以使用固定結構的 Relational Database 來儲存。或許可以利用 UML 設計，並配合 Native XML 資料庫管理系統，以物件導向的型式來儲存。	5. 可以使用固定結構的 Relational Database 來存放原始交易資料。
	6. 需配合 Text Mining 技術來做文件特徵的萃取。	6. 需配合 Data Mining 技術來做各種資料分佈與群聚特徵的萃取。



資料庫系統硬體架構

- Disk Array : RAID 0, 1, 3, 5
- 主記憶體資料庫系統 (Main Memory Database Systems) 最重要的是 Data Recovery Mechanism Design
- 硬體資料庫機器 (Database Machine)
- 平行計算的資料庫伺服器 (Parallel Data Servers)
- 大量平行計算 (Massively Parallel Computing)



資料庫系統上的各種輔助系統

- 主動式資料庫系統 (Active Database Systems)
- 時間資料庫系統 (Temporal Database Systems)
- 空間資料庫系統 (Spatial Database Systems)
- 地理資訊系統 (Geographical Information Systems, GIS)
- 知識庫系統 (Knowledge-Based Systems, KBS)
- 決策支援系統 (Decision Support Systems, DSS)
- 主管查詢系統 (Executive Information Systems, EIS)



資料庫系統上的各種輔助系統

- 多媒體資訊系統 (Multimedia Information Systems)
- 「工作流程軟體」 (Workflow Software)—：IBM Flowmark、HP WorkManager、Novell Groupwise、Wang OPEN/Workflow、Fujitsu 的 Regatta、Staffware Staffware、Action Technology 的 Action Workflow、Xerox 的 InConcert 或 Lotus Notes 4.5 (Domino)



Workflow System

- 將作業流程自動化，以形成橫向的整合，建構成更完整的資訊系統
- 考慮以下四種個體：使用者 (users)、活動 (activities)、程式 (programs)、以及資料 (data)。
 - 工作流程模式 (workflow model) 則是一個不含迴圈的有向圖型 (acyclic directed graph)，其中包含了節點 (nodes)，代表了執行的步驟；另外還有一些箭頭 (edges)，代表了不同步驟間的資料與控制流向。



資訊保密與安全性

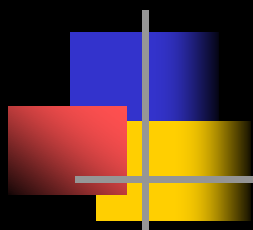
資訊安全的服務項目	所可能遭受的威脅	可採用的資料安全防護法
資料完整性 (Data Integrity)	遭人篡改、偽造、刪除	數位簽章、序碼、時間、資料辨識碼、安全雜湊函數。
資料來源認證 (Authentication)	遭人冒充傳送假資料	數位簽章、資料辨識碼
存證 (Non-repudiation)	遭對方否認已收到或送出資料	數位簽章
資料隱密性 (Confidentiality)	洩密或非法取得資料	使用保密系統
存取控制 (Access Control)	非法取得資料	防火牆系統、稽核追蹤、身份密碼等。



研究歷程

■ 如王國維的人間詞話所描述

- “昨夜西風凋碧樹，獨上高樓，望盡天涯路。”（第一境）
- “衣帶漸寬終不悔，為伊消得人憔悴。”（第二境）
- “眾裡尋他千百度，驀然回首，那人卻在，燈火闌珊處。”（第三境）



本章結束
The End.