資料間隱含關係的挖掘與展望

沈清正、陳仕昇、高鴻斌、張元哲、陳家仁、黃琮盛、陳彥良
中央大學資訊管理學系

摘要

資料挖掘，指由大量資料中擷取出有價值之知識，亦即將資料轉換成知識的行爲。這些資料包括各型態的資料，如一般的交易資料與多媒體資料，而知識則是資料間隱含關係的具體表達與呈現。因爲資料挖掘能協助企業從資料中取得知識並創造競爭優勢，故引起廣大的重視，也促進了許多新的研究方法與系統的發展，而成爲一個快速成長的領域。對於目前現有的資料挖掘方法和資料挖掘系統，本文根據〝資料間隱含關係〞的不同，提出了九種不同的類別，分別是資料關聯性、順序性、結構性、週期性、類似性、有趣性、個人性、合用性、歸納性，對每一種資料關係，我們將介紹其定義、應用狀況、研究現況和其研究展望。本文除了可幫助讀者了解資料挖掘領域的現況外，也提供了有用的資料挖掘分類方法並且介紹了資料挖掘的比較性研究。

關鍵字：資料挖掘，知識，資料間隱含關係

An Overview on Mining Implicit Data Relation

Ching-Cheng Shen, Shih-Sheng Chen, Hong-Bin Gao, Yuan-Zhe Chang,
Jia-Ren Chen, Cong-Sheng Huang and Yen-iang Chen
Department of Information Management
National Central University

ABSTRACT

Data mining is an extraction of useful knowledge from a huge amount of data. The data can be of a variety of types, such as transaction data, relational data and multimedia data, whereas knowledge is an explicit expression and representation of implicit data relation. Since that data mining can assist business to get knowledge and create competitive advantage, it is not surprising that a great number of researches have been done in this field. Because of its fast-growing development and abundant results, it is difficult to provide a complete survey to cover all the issues in a single paper. Therefore, this paper only provides a reasonably comprehensive report for the recent development of data mining technology. As to the present data mining methods and systems, this paper suggests 9 distinct categories according to their implicit data relation. These relations include association, sequence, structure, periodicity, similarity, interestingness, personalization, suitability and generalization. For each of them, we will discuss its definition, applications, algorithms and future research directions. The contributions of this paper include (1) a classification based on the implicit data relation is proposed, (2) a comparative study between these categories has been done, and (3) The state of the art for each category is described.

Key words: Data mining, knowledge, Implicit Data Relation

壹、 序論

近來，資料產生和資料收集技術有非常快速的進展，政府和許多企業的營運皆已電腦化，條碼廣泛使用在許多商業活動中，電腦已成爲資料收集的主要工具，資料也因此大量的產生。同時，數以百萬計的資料庫正被企業、政府、科學和工程上使用在管理、決策、設計和許多其它的應用上。再加上資料庫能力的提升，讓資料產生爆發性的成長，資料量早遠超過人類能直接分析的能力，因此如何能有智慧且自動的將資料轉換成爲有用的資訊及知識，便成爲資料庫應用的前瞻目標，所以資料挖掘逐漸地成爲一個重要的研究領域。

資料挖掘指由大量資料中擷取出有價值之知識，亦即將資料轉換成知識的行爲。這些資料包括一般交易資料或多媒體資料，而知識則是資料間隱含關係的具體表達與呈現。對於目前現有的資料挖掘方法和資料挖掘系統，本文根據〝資料間隱含關係〞的不同，提出了九種不同的資料關係類別，分別是資料關聯性、順序性、結構性、週期性、類似性、有趣性、歸納性、個人性、合用性，對每一種資料關係，我們會先介紹其定義，接著再談到研究現況，最後談到研究展望。另外有兩種衆所週知的資料關係：群聚性（ clustering ）與分類性（ classification ），因這兩方面的研究歷史較爲久遠，本文因此割愛，有興趣的讀者可以參考文獻 [47] 。

貳、 資料關聯性

近來資料挖掘（ Data Mining ）相當熱門的研究議題，而挖掘關聯規則則是最重要的資料挖掘問題之一，所謂關聯規則就是某些項目會引發其他項目出現的規則，例如消費者購買雜誌後，通常有 75 % 的人會同時買咖啡，此時「買雜誌（買咖啡」便是一個關聯規則。而關聯規則的挖掘，就是在龐大的資料中，把一些資料項目的相關性找出來的方法。以交易資料庫爲例，每天均有大量的交易發生，經年累月累積下來的龐大資訊，無法由人進行分析來找出商品的關連性，然而這些交易記錄事實上隱含了許多有用的資訊（如消費者購買習慣）在裡面，如果能運用適當的方法將它找出來，便可能發現商機，創造利潤，挖掘關聯規則的演算法也就是在這種需求驅使下所產生出來的方法。

在探討關聯規則的挖掘之前，我們必須先了解最小支持度 (support) 和最小信賴度（ confidence) 的概念，最小的支持度界定一個規則必須涵蓋的最少資料數目，最小的信賴度則代表這個規則的預測強度。規則的支持度和信賴度可以評估規則是否有趣（ interesting），當挖掘演算法所找出的規則滿足使用者訂定的最小支持度和信賴度的門檻時，這個規則才算成立。

在挖掘關聯規則的領域中，其方法主要可以分成兩大類：(1)利用 Apriori-like 的方法產生 candidate set ，並找出符合最小支持度的大項目集合 (large itemsets)，再依據大項目集合產生關聯規則；(2) 使用 Non Apriori-like 的方法，找出大項目集合。

第(1)類的方法中是以 Apriori 演算法 [5] 爲基礎，它們共同的特點是第一次的 candidate set （以 C_1 表示）是直接利用掃描資料庫一次直接得到，其他的 C_k（ k>1 ）產生方式都包含了兩個主要步驟，第一個是合併產生 candidate set ，第二個則是將這些項目集合中，含有不是前一次作業的大項目集合者去除，然後針對這些留下來的 candidateset ，以掃描資料庫的方式獲取其支持度，再將未滿足最小

支持度要求的項目集合去除掉，即得到所謂的大項目集合。 ⇒

由於 Apriori 反覆處理 candidate itemsets 要花費很多時間，之後陸續有一些改良 Apriori 演算法的方式提出來，DHP [72] 利用 hash function 建立 hash table 來達到減少 candidate itemsets 數量的目的；Partition[84] 將交易分割成一些沒有重疊的 partitions 在主記憶體運作以增快速度；DIC[11]（Dynamic Itemset Counting）把項目集合存入 lattice 中並分割 transaction sequence 來達到儘早決定大項目集合的目的；Random sampling [95] 則以抽樣的方式找出合適大小的樣本，再針對這些樣本迅速找出大項目集合。

由於第 (1) 類的方法在檢驗 candidate set 是否爲 large 時，都需要重新掃描資料庫一次，需要大量 I/O 的時間，即使做了一些改良，所能提昇的效率仍然有限，所以最近的研究都不利用這種方式，改以直接快速的方法來得到大項目集合，我們將它們都歸類爲第 (2) 類的方法。第 (2) 類的方法不利用 Apriori 的原理來找出大項目集合，Close Algorithm[76] 利用 closed itemset lattice 遠小於 itemset lattice 的特性，以它取代 itemset lattice 進行 pruning，可以減少 database 存取次數和 CPU 的負擔；TreeProjection[1] 建構一個 lexicographical tree 做 candidates 的 counting 並提供快速且有彈性的挑選策略；FP-growth[46] 掃描資料庫兩次建立 FP-tree 後，從 FP-tree 可快速找出大項目集合。在第 (2) 類方法中，TreeProjection[1] 和 FP-growth[11] 是目前我們看到的文獻中最快的兩種方法，其中又以 FP-tree 的結構結合 FP-growth 演算法的作法效率最高，所以 FP-growth 可說是目前最有效率的方法，DBMINER 便是實際運用 FP-growth 的產品。

關聯規則有許多種類，不過我們大體上可以將它分成以下三類 [47]：

一、以屬性值的型態爲基礎：

如果我們所關注的只是項目是否出現，這種便稱爲布林值的關聯規則 (Boolean association rule)，例如「牛奶 ⇒麵包 (support=2%,confidence=60%)」即屬於這類關聯規則。如果我們也一併關注項目的購買單位數，這種便稱爲有重複項目的關聯規則 (association rule with repeated items)，例如「2 單位牛奶 3 單位麵包 (support=2%, confidence=60%)」即屬於這類關聯規則。如果我們所要描述的規則其項目或屬性是一個數值，這種就稱爲數量關聯規則 (quantitative association rule)。但因爲數量關聯規則的可能性太多，所以我們必須把數量值切割成不同的區間 (可以事先就切好，或根據資料分布情況來切割，或根據語意、模糊函數、資訊含量等不同方式切割)，才有辦法產生關聯規則。如下面的例子，X 是代表消費者的一個變數。

年齡 (X,"40...45")^ 收入 (X,"7 萬 ...8 萬 ")⇒購買 (X, 海外基金)

二、以規則中所涵蓋的資料維度爲基礎：

如果在關聯規則中的項目或屬性僅參照單一的維度時，我們稱之爲單一維度關聯規則 (single dimensional association rule)，例如我們將「牛奶 麵包」的關聯規則寫成「購買 (X," 牛奶 ") 購買 (X," 麵包 ")」，則其著眼的是「購買」這個維度。反之，如果關聯規則中的項目或屬性參照兩個以上維度時，便稱爲複合維度關聯規則 (multidimensional association rule)，例如上述定量的關聯規則中的例子，便包含了「年齡」、「收入」以及「購買」等三個維度。

三、以規則中所涵蓋的抽象層級爲基礎：

如果在關聯規則中的項目或屬性可以屬於不同的概念層級，例如「年齡(X,"中年")⇒購買(X,"味全果汁牛奶")」("中年"對於年齡而言屬於較高層級概念，但"味全果汁牛奶"對於購買項目而言屬於較低層級概念)，則稱這類規則爲跨層級關聯規則(multilevel association rule)。反之，如果沒有參照到不同層級的項目或屬性規則，則稱爲單一層級關聯規則(single-level association rule)。

挖掘關聯規則的研究至今已算相當完整，然而不論是哪一種方式，在尋找大項目集合的過程中，所花費的時間成本均相當可觀，因此針對一個經常異動的資料庫作關聯規則的維護，是頗爲重要的問題，也就是如何以增量(Incremental)方式來局部調整經常發生項目集合(frequent itemsets)，使關聯規則保持正確性，並避免重新進行整個挖掘程序所需花費的成本。以 Apriroi 演算法爲基礎的增量維護技術已經有研究提出來[21]，然而其他方式的增量模式則仍有許多研究空間可以發揮。另一方面，如何在線上迅速獲得關聯規則也是一個重要的議題，因爲現存的大項目集合計算演算法常以離線或批次的方法進行，它給定一個使用者指定的支持度門檻，之後資料庫必須再次讀取才產生所有的大項目集合。然而，一般的使用者都無法事先知道該如何選擇合適的支持度門檻，如果選擇了一個不合適的支持度門檻，往往造成最後產生的關聯規則沒有用。

參、資料順序性

最常見的資料順序性研究的問題是要從交易記錄中尋找有趣的循序樣式(Sequential Pattern)，循序樣式的特點在於樣式中每個項目之間是有順序性的，因此在尋找循序樣式時，我們會有一個用以決定項目先後次序的衡量方式(如時間)，所有的項目或項目集合依據該衡量方式在一維的方向上呈現順序排列，而尋找循序樣式就是要在這些循序排列的資料中找到有趣的規則；以顧客的購買順序爲例，若我們發現有許多人在購買 A 物品後，會再購買 D 物品，這就是一種循序樣式。

順序性研究是一個很有價值的研究方向，因爲資料庫中的交易資料通常是有時間上的順序性，如在零售交易資料庫中交易發生的時間、時間序列資料庫中事件發生的時間、以及網站日誌中請求/回覆發生的時間等，如果在對這些資料進行挖掘時，能把順序性的考量納入，一方面可以在產生挖掘結果時去蕪存菁，另一方面其所呈現的資訊將會更具有意義。

目前順序性資料挖掘的研究與應用主要可概分爲四大類，第一類是在銷售記錄資料庫的挖掘中，將交易發生時間的順序列入考量，以期得到跨交易的顧客購買模式，上文所舉的例子就是這一類的挖掘應用，這類的研究主要在挖掘方法的設計與改良[6,7, 107]，其中[7]可以避免反覆的讀取資料庫，[107]則強調處理較長的循序樣式的能力，另外[60]可以對所得的循序樣式進行漸進式的更新維護，[73]則改善尋找循序樣式時進行序列比對的成本，[33,57,108]則是結合本法與其他領域的應用，例如應用於資料挖掘系統、計畫管理、資料庫系統等。

順序性資料挖掘的第二類是在時間序列資料庫中尋找相似的循序樣式，或是於時間序列資料庫中進行相似性的查詢，因爲時間序列資料庫的應用十分廣泛，所以順序性挖掘也被大量的應用，例如在股價歷史資料庫中挖掘股價變動的相似樣式、在氣象資料中尋找符合某相似(循序)樣

式的記錄、電信網路的警報分析 [51] 、在疾病資料中挖掘時間序列樣式等，目前這一類的順序性挖掘研究包括了一般化的時間序列樣式的挖掘演算法 [3,31,65] 、關鍵技術的改良 [103] 、特別化的時間序列挖掘與應用 [4,10,51] 。

順序性資料挖掘的第三類是於 WWW 的環境中尋找使用者的路徑尋訪樣式，我們可以將使用存取網頁的日誌合併或拆解成許多的路徑序列，然後從中挖掘相似的路徑尋訪樣式，目前這類的挖掘研究包括了一般化的 WWW 路徑尋訪樣式挖掘 [20,25, 92, 106] ，其中部分研究將挖掘的範圍擴大到一個提供多項服務的環境 [92] ，另外因為 WWW 的日誌資料異於一般挖掘所處理的交易資料，因此本類研究還包括了挖掘程序中的前置處理以及挖掘系統的架構 [26,27,34] 。

除了上述的三類，因為文字挖掘 (Text Mining) 也是處理循序文字資料，並應用所得的循序樣式，因此我們將之視為順序性資料挖掘的第四類，本類主要是於文字資料庫中挖掘文字序列樣式，研究的方向包括了一般化的挖掘演算法 [55] 、特殊化的挖掘系統與挖掘效率的改良 [54,102] 、文字序列樣式的應用 [2,16,82] 。

我們將資料順序性研究的領域劃分為四個議題，這四個議題從一般化的挖掘演算法、特殊化的挖掘演算法、效率相關的關鍵計算技術，到特殊的應用範例都已經有了一個完整的研究脈絡，然而相較於尋找大項目集合的交易資料挖掘，順序性資料挖掘在尋找大項序列時，會需要更多的計算成本與空間成本，因此發展更有效率的挖掘演算法是上述四個議題共同的未來發展方向，效率的改善一方面可以減少所需處理的候選樣式，另一方面則可以降低過濾候選樣式的計算成本，由此而言，目前的演算法尚有改進的空間；另外由於序列的特性，循序樣式有趣性的衡量，以及在預防資訊過量的考量下，挖掘系統與使用者的互動，也都是值得注意的發展方向。在四個議題之中，第三類使用者尋訪樣式挖掘必須有資料準備的前置處理動作，以將使用者存取日誌中的資料轉換成進行挖掘的記錄，因此前置處理與整個挖掘架構的有效性也是可能的研究方向。

肆、資料結構性

資料結構性的研究可視為資料順序性研究的延伸，在資料的性質上，順序性研究所要挖掘的資料具有序列的結構，也就是記錄與記錄之間可以用序列的結構加以組織，而在所欲求得的樣式上，順序性研究希望發掘可以反映這類資料特性的循序樣式，當我們將順序性研究的範圍加以擴大，不再將資料的結構限定於單純的序列時，便可算是資料結構性研究的範圍，也因為結構性研究所要處理的資料，其記錄間所形成的結構已不限於序列，所以本類研究所要挖掘的樣式更為多樣化，所需要的挖掘成本與挖掘的困難度都將會較循序樣式的挖掘來的高。

在許多情況下，交易資料庫中的記錄與記錄間是具有某些關係的，例如人口普查資料庫中，人與人之間會因為血緣、地域、社交等各種因素而互相牽連，我們可以根據這些關係將各筆記錄組織起來形成具有複雜結構的資料，而結構性研究的目的就是要從這種呈現複雜結構的資料中，挖掘常見的次結構；如果我們更廣義的解釋結構性研究的範圍，而不將組成樣式的最小單位限制於有良好定義、格式的交易資料庫記錄，則凡是資料的本身具有結構的特性（如由網頁構成的 WWW 分散式資料提供環境、由染色體構成的基因等），或資料之間具有關聯（如呈現樹狀的階層式文件集合、具有地理相關性的氣象資料等），我們都可以研究存在於這些資料之

中的子結構，也就是挖掘存在於結構化資料中的結構樣式 (Structural Pattern ，或拓撲樣式， Topological Pattern [99]) ，藉由結構樣式，我們可以對資料的特性進行更有效的分析或對現象進行更準確的預測。

結構化研究的應用並沒有一個很明確的範圍限制，只要語意上符合在結構化的資料中挖掘結構樣式者，就可以算是本類研究的應用，結構化的研究可以從兩個方向來討論，一種是一般化結構樣式的（通用）挖掘方法的研究，如 [24,99,100] ，一種是於特殊（結構的）資料上進行挖掘的研究，如 [18,61,81,100] 。

在通用型的挖掘方法研究中我們討論三種方式， 1. 爲先從結構化資料中抽取一部份的樣本，並從中尋找結構樣式，然後再用全部的資料去評估樣式的優劣 [99] 。 2. 是要從一群半結構化物件 (semi-structured object) 中進行結構樣式的挖掘，本文所指的半結構化物件每個都是一個圖型結構，而所謂的半結構指的是每個物件的圖型結構並沒有一定的輪廓，挖掘的目的就是從這些圖型結構中找出發生次數超出使用者限定的最小值的子結構（結構樣式）[100] 。 3. 提出一個在具有圖型結構的資料庫上進行結構樣式挖掘的系統，在這個系統中，結構樣式就是整個圖型結構資料中共同的子結構，而樣式的評估是依據其能「壓縮」原有圖型結構資料的程度 [24] 。

而在特殊結構的挖掘研究中我們討論四種方式， 1. 將每個使用者於存取日誌中留下的記錄轉換成較不會失去資訊的樹狀結構，然後在這些樹狀結構中，挖掘常出現的子結構樣式 (tree-like topology pattern) 以了解使用者的尋訪模式 [61] 。 2. 先從相關的網頁資料中萃取出代表有用的資訊、但型式並不明確的半結構資料，然後再對這些半結構資料進行挖掘以了解網站的資訊結構並幫助資訊的擷取 [18] 。 3. 是要從多個關聯表格中挖掘關聯樣式 (relational pattern) ，這些關聯表格彼此的參考關係必須形成一個單一路徑的樹，同時因爲這樣的參考關係，這些表格中的記錄會組成多個樹狀結構，而所要挖掘的關聯樣式就是常出現於這些樹狀結構中的子結構 [81] 。 4. 雖然並非尋找結構樣式，但其內容爲結構樣式的應用，該篇論文試圖以圖型結構表現軟體的原始碼與資料，和使用者所設定的模組內部的限制和模組間的限制，然後使用 Apriori 演算法去進行圖型之間的結構樣式的比對，以期在模組內高內聚、模組間低耦合的原則下還原軟體的架構 [83] 。

在上述的研究中，有些結論 [18,100] 會因爲所處理的結構化資料在結構上會有不一致，或是所得的結構樣式本身會包括了多種不同的拓撲結構，而用「半結構化 (semi-structured) 」來說明其研究的對象，然而在此我們一律從廣義的結構化來看待之，不過必須了解的是，因爲半結構化會有結構不規則的性質，因此無論在挖掘方法或挖掘的效率都比較容易面臨挑戰。

從上段的文獻討論上可以發現目前資料結構性的研究尚處於發展的初期，相較於資料挖掘的其他領域而言，非序列性之結構性研究的相關著作與探討爲數相對較少；在一般化的結構挖掘方法研究上，也可以發現應用於大量資料上時，會有效率的問題，而在特殊結構資料的挖掘上，目前的研究相較於結構資料的多樣化則可以提醒我們還有很多發揮的空間，因此挖掘方法效率的改良，和結構性研究的多樣化應用（如生物資訊等）都是未來值得努力的方向；另外由於結構的多樣化，所得到的樣式通常會比其他領域來的大量，因此樣式有趣性的衡量以及挖掘系統的客製化也都是尚待努力的議題。

伍、 資料週期性

什麼是週期性分析？週期性分析是找出週期性樣式（ periodic patterns ）的分析方法，也就是由時間資料庫中，挖掘出循環性（ recurring patterns ）時間樣式。

挖掘週期性樣式的議題，我們大致可分爲三類：

一、完全週期性樣式

完全週期性樣式，是指週期中的每一時間點都會具有週期性行爲。例如：每一年內的情人節，玫瑰花的銷售會增加。

二、部份週期性樣式

部份週期性樣式，是指週期中允許僅有部份的時間點具有週期性行爲；部份週期性相對於完全週期性是較寬鬆的，但它卻是眞實世界中，更常發生的。例如：珍在每個工作日（星期一 ～ 星期五）早上7:00~7:30 通常閱讀「時代雜誌」。此處並不保證她一定會閱讀「時代雜誌」，但她通常會如此做。

三、週期性關聯規則

「週期性關聯規則」就是在規律的時間區間中，達到限定的最小支持度和最小信賴度的關聯規則，這種關聯規則並不一定在整個時間中都成立，而是可以在特定週期、特定時間區間內才會成立，例如：「若週末的下午茶在下午 3:00 - 5:00 時段銷售良好的話，則晚餐也將在下午 7:00 - 9:00 消費良好」。

完全週期性分析的技術，已被使用在信號分析與統計的研究上，其中最著名的方法是 FFT （ Fast Fourier Transformation ），其將時間資料轉換爲頻率資料，以方便提供分析使用。

然而目前用來挖掘關聯規則的演算法，並沒有辦法直接用以挖掘週期性關聯規則，爲了利用目前現有的演算法，其中的一個方法是擴充項目集合，將之加入時間屬性，並將資料庫中的交易資料，依相同的時間屬性加以分類成相同的區塊(Segment)，但是這並不是好的方法，因爲這有可能會找出非週期性關聯規則，而且無法找出任意週期長度的關聯規則。

大多數用以挖掘部份週期性樣式的研究和週期性關聯規則的研究，採用類似 Apriori 演算法的方法，如：用 Apriori-like 的演算法，挖掘連續性樣式(Sequence patterns)[6]；然而限制式方法(Constraints Method)，也被提出在用以處理連續性樣式與部份週期性樣式的處理過程上；也有人提出了結合兩種演算法用以解決週期性關聯規則的方式 [70]。

目前在週期性方面的研究仍嫌不足，特別是這些演算法大多根據陳舊的 Apriori 演算法修改而來的，所以一個可能的研究方向就是我們如何根據最近的 non-Apriori 演算法來發展較有效率的週期性演算法。此外，因爲不一定只有時間才會有週期，有可能其他的屬性也會有週期性的行爲，例如空間，所以如何在非時間屬性上挖掘週期性行爲應該是值得深入探究。

陸、 資料類似性

資料挖掘的技術中，有一種是以一個樣式 (pattern) 爲基準，去找出與它相似的資料。通常使用者要預定義搜尋目標序列 (target sequence) 和一個允許差異度，之後再找出資料庫中跟目標序列相似度在允許差異度範圍之間的序列。利用找尋相似的樣式可以應用在商業上，例如以股市的交易價格而言，可用各種的財務指標及走勢圖找出相似的樣式，擬定投資策略；以便利商店地點的設置而言，可以找出相

似條件的地理區域來設置據點；以個人化一對一行銷技術而言，由消費者曾經購買的商品，我們可以把類似商品的資料推薦給消費者，以促進銷售。

在時間或空間資料庫上找尋相似的樣式的操作，可將其分爲兩類 [19] ：

1. 物件相關相似查詢 (object-relative similarity query) ，使用者需先指定目標物件和允許差異距離，然後找出所有與目標物件的距離在範圍內的物件。
2. 全部相似查詢 (all-pair similarity query) ，使用者需先指定允許差異距離，然後找出所有兩兩物件間的距離落在範圍內的物件對 (pair of objects) 。

一般度量相似的方式主要用統計上的歐幾里德距離 (Euclidean distance) 及相關 (correlation) ，兩個序列的歐幾里德距離的定義如下：設 $\{ x_i \}$ 是目標序列，$\{ y_i \}$ 是資料庫中的序列， n 是 x_i, y_i 的長度，則 $\{ x_i \}$ 和 $\{ y_i \}$ 的相似性定義成

$$\min \sum_{i=1}^{n} (x_i - y_i)^2$$

兩序列間的相似性定義如下：設 $\{x_i\}$ 是目標序列，$\{y_i\}$ 是資料庫中的序列， n 是 x_i, y_i 的長度， i=1,...,N+n-1 。

$$c_i = \frac{\sum_{j=1}^{n} x_j y_{i+j}}{\sqrt{\sum_{j=1}^{n} x_j^2} \sqrt{\sum_{j=1}^{n} y_j^2}}$$

在類似物件的搜尋技術上，大多根據 Parseval 定理，使用 Discrete Fourier Transform (DFT) 將時間序列轉換成頻率序列，這樣的轉換不但不會改變彼此間的距離，且因爲頻率序列的前幾個頻率通常已足以代表整個頻率序列，所以可以只針對這幾個頻率作索引來找尋時間序列中的相似樣式 [3] 。對於轉換後的頻率序列，我們可將每一個序列對應到特徵空間上多維矩形的集合，這些多維矩形的集合以傳統空間儲取方式如 R*-tree 來索引，並使用 sliding window 用於序列上萃取它的特性，將序列作相似性的比對，這種方式可以比循序掃描節省序列比對的時間，而且不會漏掉任何子序列的比對並可減少空間使用的 overhead[31] 。找尋相似樣式的方法可以應用到多媒體資料上，例如 [32] 它將一個物件使用 k feature-extraction 函數到某些 k 維的空間，再使用空間資料結構及搜尋方法，找出相似的樣式。此外，對於時間序列資料，我們也可以找尋類似性樣式，例如 [9] 利用 dynamic time warping 的方式來找尋在時間序列資料中的相似樣式。而在 [4] 中只要兩時間序列的子序列不重疊且依時間順序排列，就可以作相似性的比對，也就是說這種作法，可以將兩時間序列中其中之一的振幅 (amplitude) 作適當的比例的放大或縮小，並將它的 offset 作調整來找尋其相似處，此外子序列在比對時不需用時間軸來排列。

樣式相似性就像是專家系統的案例推論 (case-based reasoning) ，可以提供許多的商業用途，讓使用者找尋到相似的樣式，作爲制定某些決策的參考。從先前的研究來看，許多的研究讓樣式相似性的比對更有效率，但我們可以發現過去比較著重於線性且連續性資料的探討，所以未來可行的研究議題包括 1. 研究平面及 3D 的樣式相似性比對，因爲現在的科技環境，有許多的資料都是由影像、聲音合成，如果能從中挖掘出知識，將比過去發現的更豐富； 2. 由於平面及 3D 的資料比較複雜，所以有必要研究提升在平面及 3D 的樣式相似性比對的效率； 3. 研究解決在實際生活上資料有許多都不能量化的問題，將物件中非數值性資料屬性加以量化，以便作爲相似性的比較； 4. 最近因電子商務的流行，爲解決其中對於個人化服務的迫切需求，有許多的相似性問題都急需釐清，包括哪些的網頁彼此相似？哪些的消費者有相似的購物傾向？哪些的網

路用戶有相似的瀏覽行爲？哪些的商品彼此相似？在這些個人化服務中，我們必須先判別其相似性，再據此提供可能的瀏覽建議、相關商品資訊給消費者，甚至可以預測使用者路徑、調整網站架構或提供個人化網站。

柒、資料有趣性

當利用資料挖掘的技術從資料庫中挖掘出許多的知識和規則時，由於產生的知識、規則的數量很多，但其中有些知識、規則對使用者而言是重覆性的、直覺的或無意義的，因此必須制定某些度量的標準去刪除這些不需要的規則，這個度量的標準稱爲有趣性（interesting）。

利用規則的有趣性，可以針對使用者的需要，找出使用者眞正有興趣的規則，避免使用者在衆多的規則中不知道那些是對他是有意義的規則，節省使用者分析規則的時間。以關聯規則爲例，如在有30000筆的人口普查資料中，產生的關聯規則超過20000條，如此多的規則，若沒有用有趣性來篩選規則，這些找出來的規則將沒有任何的用途[11,71]。

在規則有趣性的研究方面，規則的有趣性分成objective和subjective兩類[86]。objective是以在處理資料的過程中規則的結構及基本的資料爲基礎，根據支持度或信賴度等方式來測量有趣性，例如定義一個RI(rule-interest)函數規則的有趣性的度量方法[79]，在有A和B兩個條件中，利用RI函數可以判定A、B間的相關性是正相關、負相關或無關。爲了讓找出的有趣規則有更好的品質，針對RI函數有研究又提出了用1.disjunct size、2.the imbalance of the class distribution、3.attribute cost、4.misclassification costs、5.asymmetry in classification rules等五個因素來彌補RI函數在測量有趣性的規則時所產生的偏差[35]。所謂的disjunct size指的是符合某一規則的前提(antecedent)的資料筆數，the imbalance of the class指的是在度量規則時，對不同類的資料數量造成的誤差，attribute cost指的是考慮每個資料屬性時所需的成本，misclassification costs指的是對規則的有趣性分類錯誤時所需的成本，asymmetry in classification rules指的是對規則的前題和結論(consequent)對稱性的考慮。而利用使用者的認知建立信條(belief)並針對已經挖掘到的出乎意料（unexpectedness)的規則作知識的修正，再來找出乎意料的資料樣式，它的演算法產生的關聯規則比Apriori演算法的關聯規則少很多，而且也避免產生如Apriori產生的不相關或顯而易見的規則[71]。還有些研究提出一個概念利用partial order解決了最佳化規則的問題，找出最有趣的規則(most interesting rule）這種規則涵蓋了由support、confidence、gain、laplace value、conviction、lift、entropy gain、gini index和chi-squared value等方式找出的規則[11]，在實務上，使用者可以輕易的看出這個被找出數量不多的規則中最佳的規則所構成的集合，不會迷失在數量龐大的規則中。Subjective是以處理資料的過程中規則的結構及資料爲基礎，並且根據使用者指定的"有趣規則樣式"(interesting pattern)來測量規則的有趣性，Subjective有趣性的測量又分爲actionable和unexpected[87]，actionable指的是能讓使用者採取某些行動的規則，例如平均每個子公司的獲利率是30%，但有一個子公司C的獲利率是10%，則就管理者而言，子公司C就是會引起管理者有所行動去提醒、鼓勵子公司C。利用actionable的衡量方式，可以讓使用者就目前的現況，看出那些地方可以再改善。unexpected指的是利用信條能產生讓使用者覺得驚訝的規則，例如每個子

工廠的產品良率一般都是98%到99

% 之間，卻有一家子工廠 D 的產品良率只有 50% ，那子工廠 D 就是超乎管理者預期，令管理者訝異的子工廠。以關聯規則爲例，並不是所有有高信度及高支持度的規則就是有趣的，因爲有些規則的意義可能重覆或是意料中，有些沒有意義規則是由不相關的屬性所構成，因此讓使用者建立一些有趣及無趣的關聯規則，再去產生樣式表示式當作 template ，找出有趣的關聯規則 [52] 。也可讓使用者依他個人過去的知識及感覺輸入他期望的規則，再加上 fuzzy 的技術找出有趣的規則後，依使用者定義的有趣性大小作排序，此外並可讓使用者定義意外的規則樣式，分析其差異 [63] 。

規則有趣性的研究目的在提供更具商業價值的知識給決策者，並可以讓決策支援系統的功能更加完備。先前有許多的研究針對不同的情況，針對規則的有趣性作定義，並找出有趣的規則，未來的發展包含 *1.* 根據不同的領域定義各種不同的有趣性規則，讓使用者針對他的需要去挑選他自己感興趣的規則，包括主觀、客觀的，讓他覺得驚訝、會採取某些行動自然的規則。 *2.* 將已發現的規則，利用如專家系統的規則基 (rule base) ，將已發現的規則儲存起來，並可以用已發現的規則來作推理，供使用者作查詢。 *3.* 提升現有各種發現有趣性規則的演算法的效率，更迅速的提供給使用者想獲得的知識。

捌、資料個人性

資料個人性的研究是因應資訊的爆炸，爲幫助使用者過濾資訊，及網路的興起，使電子商務的一對一行銷取代傳統大量行銷變成可能，而在一對一的行銷時，因爲每個人需求的不同而需要個人化的服務。

資料個人性之研究的意義在於經過對使用者的瞭解、辨識後，能從大量的資訊中，以適當的方式提供與使用者最相關，或使用者有興趣的資訊給使用者，避免資訊過量，讓使用者被不相干的資訊淹沒，個人化系統的主要元件如圖 1[89] 。提供個人化資訊的方法有兩大類 [105] ，一爲 content based approach ，它是以該使用者過去的行爲爲基礎，所提供的資訊與使用者過去的喜好相類似，例如使用者過去曾經租過恐怖片，則提供他有關恐怖片的新片資訊給他。另一種爲 collaborative filtering approach ，此方法是先識別過去其他的使用者與使用者之間是否有相似的偏好，若有相似，再提供此類使用者過去的偏好給使用者，例如從使用者的行爲中發現他爲高收入族群，則提供使用者高收入族群較偏好的商品資訊。也有的作法是將兩者結合使用 [89] 。

data mining 的許多技術都可運用在提供個人化資訊，例如 clustering 可以幫助使用者及內容 (content) 的辨識，關聯規則可以找出內容間的關聯等等。

上述的兩類方法都有其缺點， collaborative filtering 的缺點是開始時需要大量的使用者參與及只能用於同質性高的商

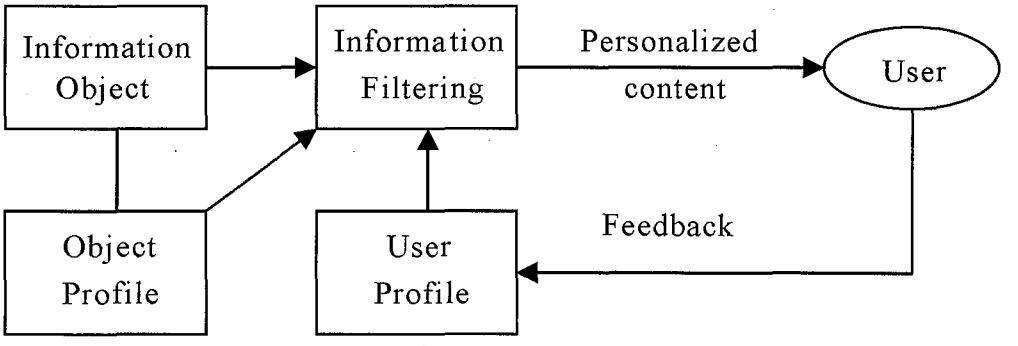

圖1：個人化系統的架構

品，而 content based approach 則爲建議的能力較差，因爲受限於過去使用者的資料，兩者以 collaborative filtering 較常被採用。

目前有關資料個人性的應用與研究大多是建立在 WWW 的環境上，對 web log 配合著使用者相關知識進行挖掘，例如運用在遠距教學上，可以依據學生的能力，提供符合學生程度的資訊 [94]。挖掘出使用者的行爲模式，使網站能客製化，針對不同使用者動態產生不同的連結 [30]。

在電子商務的應用方面，使用 rough set 的技術運用在客戶關係管理 (CRM) 上，可以進行一對一的服務，快速回應顧客不同的需求 [23]。結合 neural network 及 cluster 的技術，可以幫助企業對不同的顧客群，有效的進行目標行銷及直效行銷 [75]。

另外也有以使用者的相關資料、過去的操作紀錄，配合問答式的使用者介面，提供使用者更有興趣的資訊 [93]。GRAS 則結合 collaborative filtering 與 content based，可以不限於文字，在多媒體資料庫上提供個人化資訊 [89]。

資料個人性之研究未來的展望可分爲三個方向，第一是效率的問題，因爲個人化的服務必須是即時的服務，而且系統同時可能有多人使用，加上資料量十分龐大，所以如何即時、有效率的提供資訊是最基本的要求。

第二是延展性 (scalability)，在 collaborative filtering 中一個重要的議題是如何動態且即時的決定使用者所相對應的族群，因爲同一個人在不同的時間的行爲模式可能不同，例如某個使用者平常只買技術性書籍，而當他要買小說送朋友時，應該將使用者與喜愛小說的族群相對應，而非買技術性書籍的族群，並以小說族群的慣用行爲模式提出建議，不用重新比對使用者的檔案、興趣，再產生對應的族群來提供建議。

第三是品質，衡量我們給使用者的個人化資訊的品質，深受我們如何決定使用者是與哪一個族群相似所影響，因此如何建立一個好的類似函數或衡量標準，以判定兩個使用者的偏好相近到可以歸入同一個族群，及兩個 item 是否有足夠相似的特性能被歸類到相同的類別，是改善個人化資訊品質最基本的方法。

另外在多篇相關研究中都有參考一些 AI 的技術，所以結合 AI 的相關技術可能也是未來的研究方向之一。

玖、資料合用性

通常在資料庫中挖掘規則或樣式時，所得出的結果不是不夠而是太多，以致於使用者很難負荷，而且其中大多數結果對使用者而言，並不具有趣性，使用者必須進一步的篩選，才能找到他所要的資訊。因此若能在挖掘前先指定所要產生規則的型式，則可以大幅度減少所產生的規則或樣式，使用者也比較有可能去作進一步的分析，也因爲有效的限制範圍，進而提昇挖掘的效能，因此這一類型的研究在於如何制約制約規則的型式及過濾方法，使其達成資料挖掘步驟更有效率與產生的結果更有用的目標。

本類型的研究是以限制爲基礎的資料挖掘 (Constraint-based data mining) 方法，強調使用者的介入挖掘過程是必要的步驟，唯有使用者才知道他所想獲得的資訊，研究學者希望提供使用者在挖掘步驟之前、之間，能有效且正確的限制各種條件與範圍的方法，使其快速、正確取得想要的知識。

就本類型研究 " 限制 " 一詞的定義，其中最主要的精神是以各種方法有效的減少搜尋範圍，卻不影響結果的正確性，換言之，就是提供使用者對挖掘正確的結果

具有聚焦的能力，而這種能力建立在使用者現有的知識基礎上，因此研究學者提出強化傳統資料挖掘方法的方式，讓使用者可以將其現有的知識加入挖掘的過程中，去協助發覺所需的知識，這種方式事實上與人類一般學習過程極為相似，那就是在已知的知識基礎上探討未知的知識，有效的知識的建立方法是建在已有的知識上，並不應是每次都由零開始的挖掘過程，所以"限制"就是以已知的知識限定並控制挖掘的過程的方法。

其實限制的想法早就存在開始資料挖掘的方法中，如門檻值（threshold）的限制應用於「經常發生樣式」(frequent pattern) 的發覺過程中，用其過濾不明顯的樣式，以減少後續的搜尋空間與時間的花費，以利有效樣式的取得。

但是，傳統資料挖掘的方法中只提供使用者少許限制能力，這些方法就像是黑箱作業一樣 [68,69]，就如在關聯法則的挖掘時，使用者只能在開始提供門檻值，以後就看資料挖掘的方法了，其過程使用者完全使不上力，雖然如此，傳統資料挖掘方法的門檻值就是一種限制的觀念，我們可以將這類的方法視為本類型研究的第一階段，之後有學者發現這些問題，開始有些研究 [29,62,87,91] 使用已知樣式與述詞 (predicate) 加入資料挖掘方法中作為篩選的指標，可有效限制挖掘過程資料範圍與挖掘方向，以提昇挖掘效率與結果的有用性，這一類的方法較第一階段明顯改進，因為研究者發現以已知的知識限定挖掘過程，可明顯獲取所需新的知識，因此我們將這類的方法視為本類型研究的第二階段。

如果細心觀察，可以發現所有的資料挖掘方法就是運用各種限制在巨量的資料中取得知識的方法，然而部分的研究 [68,69] 提出現行的資料挖掘模式常有的三大問題：

一、缺乏使用者探究與控制的機能

應將資料挖掘方法的黑箱打開，並提供使用者回饋的機能，使系統可以結合使用者回饋，作進一步的修正。

二、缺乏聚焦的能力

使用者在其腦中對要挖掘現象可能已有定見，系統應提供多元的機會讓使用者有聚焦的能力，將資源可以聚焦的導入挖掘步驟中使其能快速且正確的取得想要的知識。

三、對關係的概念過於僵化

類似關聯法則門檻值應不限於支持度與信賴度的使用，如相關度（correlation）也可以應用，而在挖掘步驟不同階段門檻值應有調整的空間，並可以混用各種衡量關係的尺規，使挖掘過程更有彈性，這類的減少僵化研究已有相當多的研究報告 [44,56,69,90]。

明顯的，現在的資料挖掘觀念已不只是知識的取得而已，應是提供使用者與系統互動並介入控制挖掘的過程的管道，也應提供問題聚焦的能力與提供使用不同的關係概念尺規的能力，藉以快速且正確的縮小搜尋範圍，同時快速取得有用的知識。

事實上近來的本類型研究方向，就是以上述的問題作為研究的主題進行探討，我們將其視為本類型研究的第三階段，而為了近一步對研究方向與定位做說明，研究學者就限制 [45] 而言，將其區分為五大項：

1. 知識型態限制 (Knowledge type Constraints)：
   限制所挖掘知識的類型，如關聯性規則、資料分類性規則或資料群聚性規則。

2.資料限制 (Data Constraints) ：
資料庫與其資料來源的選擇或篩選。

3. 維度/層度限制 (Dimension/level Constraints) ：
指定挖掘過程中要使用哪些的資料維度 (dimension) 或指定概念階層 (concept hierarchy) 的層度 (level) 。

4.有趣性限制 (Interestingness Constraints) ：
如支持度與信賴度的限制就是有趣性限制。

5.規則限制 (Rule Constraints) ：
規範所要挖掘規則的型式，這類限制的例子有 metarule 或指定規則中最多包含的項目數。

其中以第 3,4,5 項限制有較多的研究文獻，主要的原因是 1,2 項是在挖掘步驟之前運用，第 3,4,5 項是在挖掘步驟之間運用，也就是原始黑箱部分，情況較爲渾沌也因此比較有研究價值，如第 3 項相關的有資料項限制的研究文獻 [91] ，又如第 4 項相關的有支持度與相關度限制的研究文獻 [8,91,101] ，最後第五項規則限制，又可以區分出幾項特徵，所以有較多的研究文獻進行探討，下一段爲其內容明細說明。

在規則限制的格式中，運用的方式可區分爲兩種類型 [45] ：

1.規則型式限制 (rule form constraint):
如 P(x,y) ∧ Q(x,w) --> Takes(x," database system")

2.規則內容限制 (rule content constraint):
如 sum(price) > 1000 ∧ sum(price) <= avg(price)
規則內容限制可分爲單邊受限型，如上面例子的第一項，雙邊受限型，如上面例子的第二項，因爲規則內容限制加入聚合函數 (aggregation) 與集合關係，在不同的聚合函數與集合關係、不同資料的內容與不同邏輯比較式的組合下會有不同的規則內涵變化，因此就單邊受限型的研究就是在釐清上述組合的變化所導出規則內涵變化結果，其研究討論在下段說明，而雙邊受限型而言，兩邊都是聚合函數會因資料的變化發生交互影響，因而發生內涵的變化，已有學者加以研究 [53] 。

文獻中將規則內容限制的涵義細分爲下列限制類型：

1.非單調 (antimonotone) 限制 [45,53,68,69]
如果一項目集合 (itemset) 不滿足此限制，則其超集合 (superset) 也不會滿足。例如 count(I) ≤ 10 或 min(J.price) ≥ 500 爲非單調。

2.單調 (monotone) 限制 [37]
如果一項目集合滿足此限制，則其超集合也會滿足。爲非單調限制的相反，如 max(J.price) ≥ 50 。

3.簡潔的 (succinct) 限制 [45,53,68,69]
不需要計算 support 就可以將滿足此一限制的所有項目集合列出來，如 max(J.price) ≥ 50 。

4.可變換的 (convertible) 限制 [77,78]
某些的限制雖然不屬前三者，但若我們將項目集合中的項目按某種次序加以排序，它便可以變成非單調限制或單調限制，如 avg(J.price) ≤ 50 ，我們若把項目按照價格由小到大排序，則這個限制便成爲非單調限制，因爲若 J 不滿足此一限制，則我們再加入一價格更貴的項目，一定也不會滿足。

5.不可變換 (inconvertible) 限制
以上皆不滿足者，如 sum(s.price) ≥ 50 。

因爲規則內容限制運用了聚合函數與集合關係，在實際上可讓使用者能進一步有效限制過濾資料，上述的研究明顯想在以前較棘手的 (tough) 較灰色地帶，且不易釐清的部分中分離出明顯可用的理論思維。

進一步的可能研究方向有，1.爲對規則內容限制進一步的分離出明顯可用的理

論思維，*2.*爲結合不同的限制類型進行整合，如有趣性限制與規則限制整合的研究[37]，*3.*將資料合用性之研究的思維運用到不同的知識型態限制上，如群聚性規則上[98]，*4.*提出新的或結合現有的方法，使其能提供使用者與系統互動並介入控制挖掘的過程的管道，同時能提供問題聚焦的能力與提供使用不同的關係概念尺規的能力，藉以快速且正確的縮小搜尋範圍，同時快速取得使用者合用的知識的方式，都是本類研究可發展的方向。

拾、資料歸納性

一般而言，在資料庫中存放的資料和物件通常是大量的原始明細資料，例如：在銷售資料庫中存放的明細資訊，如姓名、地址、項目名稱、銷售量、日期、價格等。但是通常使用者想看到的卻是一種具簡明性、總結性的歸納描述性資料，所以我們會希望用某些方法將這一大串明細資料概括起來，用一個較高的概念層級(concept level)來表示它們的內涵；再者，使用者喜歡在不同層度的顆粒度與不同的角度下，用簡單與彈性的方式去操控描述性資料。這就需要運用在資料挖掘中一個很重要的方法：資料歸納(data generalization)。而所謂資料歸納，即是在資料庫中，從一個具有較低概念層級資料集中，摘要歸納相關的資料成爲另一個較高的概念層級的資料集過程稱之。而如何有效率及彈性地歸納這些資料的方法主要有二種：1.資料方塊法 DCA(Data Cube Approach)或稱爲 OLAP(OnLine Analytical Processing)；2.屬性導向歸納法 AOI(Attribute-Oriented Induction Approach)。此一部份的研究在實務上引起廣大的迴響，因爲除了可以應用於資料倉儲中挖掘知識外，也提供決策者不同概念層級多層次的知識，其中第一種方法在資料倉儲的領域已有相當多的研究文獻，本文將不深入討論，而將重心放在第二種方法屬性導向歸納法上。

屬性導向歸納法，從資料分析的角度看是一種概念化描述(concept descriptive)的資料挖掘[47]，概念化描述的方式是以簡明的、總結的方法描述資料集並以歸納且有趣的屬性表現資料，這些資料可能提供資料集全面性的描繪，或是區分出相對的資料集，而概念化的動作可視爲資料聚集的行爲如：研究生、中學生等概念化描述，概念化描述並非只是對資料進行計數加總，它是對資料的特徵(Characterization)與相對性(Comparison)進行描述，概念化描述其具體方法就是屬性導向歸納法，對存放巨大量資料的資料庫，它提供有效的多層次抽象化(Abstraction)的能力，以協助使用者調查資料集所具有的一般化行爲，它並不對個別的客戶進行調查而是針對客戶群進行資料特性調查分析，如銷售經理希望看到某一地區範圍或是某一收入範圍的客戶群的特性等。

而概念化描述與傳統 OLAP 的差異，可用下列 3 項加以說明，*1.*處理資料型態的差異，概念化描述具有處理複雜的資料型態，如數值、文字、空間與影像等的資料型態，*2.*聚集處理能力的差異，概念化描述具有處理非數值聚集的能力，如聚合非數值性資料、合併空間範圍、壓縮影像、整合文字與將物件點群體化等能力，並可運用在不同型態的架構上，而 OLAP 是架在資料倉儲的架構上，以數值性資料爲基礎進行數值聚集的處理。*3.*使用者自行控制與自動化的差異，一般而言 OLAP 的使用是使用者介入進行上捲與下展等功能，由使用者自行決定抽象化的維度與層級，而概念化描述提供較多自動化機制，協助解決使用者決定抽象化的維度與層級的問題。雖然兩者有相當的差異，但抽象概念化描述整體資料的精神是一致

的，所以就長遠發展，概念化描述與OLAP最後終將整合在一起[47]。

屬性導向歸納法是一種以關聯查詢導向，歸納爲基礎，並具有概念階層性的線上資料分析技術[14,39,40,41]，處理步驟爲1. 收集任務相關的資料集並區分出各屬性不同的內涵值個數，2. 以屬性移除(attribute removal)與屬性歸納(attribute generalization或是概念樹爬升(concept-tree climbing))進行歸納，再將歸納後相同的tuple加以合併，並聚集加總其產生的筆數，以減少歸納資料集的大小，再用屬性門檻值控制歸納程度，再決定是否要再歸納，3. 以不同的方式呈現特徵屬性的資料。以上三個步驟也道出資料歸納性的研究中所要討論的重心所在。

就本文所言的屬性而言，並不一定是指最低階的屬性，也可能是歸納後的屬性，事實上這與OLAP的維度有不同層級的觀念相似，只是OLAP用星狀綱目(star schema)，而屬性導向歸納法用包含(⊂)規則顯現其概念階層的歸納關係。在屬性導向歸納法中，概念階層是處理歸納的過程中所必備的背景知識，概念分類法具有主觀性，同時與資料內涵絕對相關。一個概念階層有「一般-至-特定」(general-to-specific)的順序性，最一般化的概念，是以"ANY"來表示之，最特定的概念，則對應到資料庫中某一特定的屬性值，下面爲一個典型的大學資料庫的概念階層的歸納關係。

{freshman, sophomre, junior, senior}
⊂ undergraduate
{M.S., M.A., Ph.D.} ⊂ graduate
{undergraduate, graduate}
⊂ ANY(status)

屬性導向歸納法的重點是在減少資料量卻不會減低知識內涵的正確性。就屬性移除的原則而言，如果在大量的資料歸納過程，*1.*屬性內中有許多不同的屬性值，且沒有較高的概念層級可以表示它的話，*2.*高層的概念可經由其他的屬性明白表示[47]，*3.*增加該屬性或移除該屬性並不影響整體歸納的結果[49]，也就是這屬性與資料集其他屬性群間呈現不相關或是弱相關的特徵(可用gain ratio、Gind index、χ^2等或其他方法加以衡量)時，則在歸納的過程，就必須將這些屬性移除。就概念樹的爬升而言，若某一屬性在概念階層中存在著一個更高層級的概念，則該屬性值就應以其更高層級的值來取代。屬性值向上爬升後，若產生相同值的tuple，則將相同值的tuple合併歸納爲一筆一般化tuple，並將相同值資料的筆數值累加到歸納後的tuple中vote欄位中。然而歸納要做到什麼地步才停止呢？我們利用門檻值來做控制。如果屬性中不同屬性內涵值的數目超過預先設定的門檻值，則必須再進一步針對這個屬性進行歸納。再者，一個歸納後的關聯表，其tuple數目超過預先設定的門檻值時，則必須做再進一步的歸納。最後我們再做規則的轉換，即將最終關聯表的tuple轉換成規則，或是將它作爲資料挖掘分類性的準則用以區分出相對的資料集。經過這些步驟之後，資料庫中原始概念層級的資料就可以被歸納成層級較高、較一般性的規則了。而爲何要用門檻值來做控制呢？，主要是要因防止過度歸納或是歸納不足，所造成歸納後的結果較不具有參考價值的現象後果。

經過上面的說明，我們將屬性導向歸納法研究現行與未來的趨勢區分爲五種，*1.* 提出改良或提昇效率的方法，如運用不同的資料結構提昇歸納與再歸納的能力[13,14,15]，運用rough set與統計方法以減少或區分參予歸納的屬性[17,49,85,96]，*2.*接受概念化爲主觀性與多樣性的事實，提出解決處理屬性具有多概念階層的方法[22,38,48,67]，*3.*將屬性導向歸納法的方法結合其他的方法進行研究[12,50,85]，*4.*

應用屬性導向歸納法到關聯資料庫以外的資料庫 [28,36,42,64,80] ，5.屬性導向歸納法的應用，特別是當資料庫的知識內涵適合概念歸納時 [66,96,97,104] 。

再者，第六種可能的趨勢，就是在前面我們曾經提到許多高層管理的本質就是歸納資料以找出簡明性、總結性的概念，而這也是另一個與屬性導向歸納法相似議題「資料倉儲與 OLAP 」爲何流行的原因之一。但相對於資料倉儲，本法還多出對非數值屬性歸納處理的能力，較 OLAP 更具有一般化的特性，但其目標與精神是一致的。所以就長遠發展，概念化描述的 AOI 與 OLAP 終將整合在一起，所以如何將兩者整合並應用於管理上將是值得探討的方向。

拾壹、結論

資料挖掘是一個快速成長的領域，最近有許多新的研究報告、新系統或雛形的發展。因此要在短短的文章中提供廣泛的資料挖掘方法的概論是一個極難的目標。這篇文章是從資料庫研究者的觀點，對於最近發展的資料挖掘技術提供一個合理廣泛的報告。事實上，在作者原先的規劃中，擬探討十一種的資料隱含關係，因爲群聚性、分類性研究，研究文獻相當多而且歷史較爲久遠，故未將其納入文章中。

由於資料挖掘方法的多樣性，最近有許多不同的資料挖掘系統和雛形被發展，當中有些是從大型資料庫中成功的挖掘知識，也有一些是由機器學習和統計學方面的研究者所完成的。此外，資料倉儲系統和資料挖掘工具的整合在這些系統中也普遍的出現，幾乎已成爲一個趨勢。除了資料庫研究者外，在許多其它的領域中，對於資料挖掘和發現知識上，也都有相當豐富的成果。例如：統計學方面的學者已經發展了許多有利於資料挖掘的技術。歸納邏輯方法在邏輯方法上是屬於快速成長的子領域，其與資料挖掘是緊密關聯的。也有許多研究在探討視覺化資料挖掘技術和如何把挖掘出來的結果予以視覺化與並提供互動的機能。

如同許多年輕和大有可爲的領域，資料挖掘仍然面對許多的挑戰和未解決的問題，這些問題可以產生新的研究議題，做更多的研究。除了可以針對不同種類的資料庫如：主動式資料庫、物件導向式資料庫、時間和空間資料庫、多媒體資料庫、生物資料庫進行更多的探討外，在網際網路資訊系統的資料挖掘；發現知識的應用；如何和專家系統或專家知識整合；資料挖掘中保證安全性和隱私權保護的方法，這些都是重要的研究。此外，目前資料挖掘的研究大多偏重技術面且評估的指標也都著重於方法的效率，非常缺少管理面的研究探討如資料挖掘關鍵成功要素、導入程序、對組織的衝擊與影響、對學習行爲與創新的影響等議題，這些議題亦值得進一步的研究。

參考文獻

1. Agarwal, R., Aggarwal C. and Prasad, V. V. V. "A tree projection algorithm for generation of frequent itemsets," In J. Parallel and Distributed Computing , (2000).
2. Agrawal, R., Bayardo Jr, R.J. and Srikant, R. "Athena: Mining-based Interactive Management of Text Databases," IBM Research Report RJ10153, July 1999.
3. Agrawal, R., Faloutsos, C. and Swami, A. "Efficient Similarity Search in Sequence Databases," in Lecture Notes in Computer Science 730, Springer Verlag, 1993, pp. 69-84.

4. Agrawal, R., Lin, K., Sawhney, H.S. and Shim, K. "Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases, " Proc. of the 21st Int'l Conference on Very Large Databases, Zurich, Switzerland, Sep. 1995.
5. Agrawal, R. and Srikant, R. "Fast Algorithms for Mining Association Rules, " Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, Sep. 1994.
6. Agrawal, R. and Srikant, R. "Mining Sequential Patterns," Proc. of the Int'l Conference on Data Engineering (ICDE) , Taipei, Taiwan, March 1995.
7. Bayardo Jr., R. J. "Efficiently Mining Long Patterns from Databases," In Proc. of the 1998 ACM-SIGMOD Int'l Conf. on Management of Data, 1998, 85-93.
8. Bayardo Jr., R.J., Agrawal, R. and Gunopulos, D. "Constraint-Based Rule Mining in Large, Dense Databases," Proc. of the 15th Int'l Conf. on Data Engineering, Sydney, Australia, March 1999.
9. Berndt, D.J. and Clifford, J. "Finding Patterns in time Series: A Dynamic Programming Approach," Advances in Knowledge Discovery 1996, AAAI MIT Press.
10. Bettini, C., Wang, X.S., Jajodia, S. and Lin, J-L. "Discovering Frequent Event Patterns With Multiple Granularities In Time Sequences," IEEE Transactions on Knowledge and Data Engineering, Vol. 10, No. 2, 1998, pp. 222-237.
11. Brin, S., Motwani, R., Ullman, J.D. and Tsur, S. "Dynamic Itemset Counting And Implication Rules For Market Basket Data," SIMOD, 1997, pp. 255-264.
12. Cai, Y., Cercone, N. and Han, J. "An attribute-oriented approach for learning classification rules from relational databases," Data Engineering, 1990. Proceedings. Sixth International Conference on , 1990, pp. 281 -288
13. Carter, C.L. and Hamilton, H.J. "Performance evaluation of attribute - oriented algorithms for knowledge discovery from databases, " Tools with Artificial Intelligence, 1995. Proceedings., Seventh International Conference on , 1995, pp. 486 -489
14. Carter, C. L. and Hamilton, H. J. "A Fast, On-Line Generalization Algorithm for Knowledge Discovery," Applied Mathematics Letters Volume: 8, Issue: 2, March, 1995, pp. 5-11
15. Carter, C.L. and Hamilton, H.J. "Efficient attribute-oriented generalization for knowledge discovery from large databases," Knowledge and Data Engineering , IEEE Transactions on , Volume: 10 Issue: 2 , March-April 1998, pp. 193 -208
16. Chakrabarti, S., Dom, B., Agrawal, R. and Raghavan, P. "Using Taxonomy, Discriminants, and Signatures for Navigating in Text Databases," Proc. of the 23rd Int'l Conference on Very Large Data Bases, Athens, Greece, August 1997.
17. Chan, C.-C. "A rough set approach to attribute generalization in data mining ," Information Sciences Vol: 107, Issue: 1-4, June, 1998, pp. 169-176
18. Chen, E. and Wang, X. "Semi-Struc-

tured Data Extraction And Schema Knowledge Mining," Proceedings. 25th EUROMICRO Conference, Volume: 2, 1999, pp. 310 -317.

19. Chen, M-S., Han, J. and Yu, P.S. "Data Mining : An Overview from a Database Perspective," IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, 1996, pp. 866-883.
20. Chen, M-S., Park, J-S. and Yu, P.S. "Efficient Data Mining for Path Traversal Patterns," IEEE Trans. on Knowledge and Data Engineering, Vol. 10, No. 2, April 1998, pp. 209-221.
21. Cheung, D., Lee, S.D. and Kao, B. "A General Incremental Technique For Maintaining Discovered Association Rules," in the Proceedings of the Fifth International Conference On Database Systems For Advanced Applications (DASFAA '97), Melbourne, Australia, March 1997.
22. Cheung, D.W., Hwang, H.Y., Fu, A.W. and Han, J. "Efficient Rule-Based Attribute-Oriented Induction for Data Mining," Journal of Intelligent Information Systems v.15 n.2 pt.0 2000, pp. 175
23. Chiang, I-J, and Lin, T.Y. "Using Rough Sets to Build-up One to One Customer Services," The 24th Annual International Computer Software and Applications Conference, 2000, pp. 463-464.
24. Cook, D.J. and Holder, L.B. "Graph-Based Data Mining," IEEE Intelligent Systems, Vol. 15, No. 2, 2000, pp. 32-41.
25. Cooley, R., Mobasher, B. and Srivastava, J. "Web Mining: Information and Pattern Discovery on the World Wide Web," in Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), Nov. 1997.
26. Cooley, R., Mobasher, B. and Srivastava, J. "Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns," Proceedings of the 1997 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX-97), Nov. 1997.
27. Cooley, R., Mobasher, B. and Srivastava, J. "Data Preparation for Mining World Wide Web Browsing Patterns," Journal of Knowledge and Information Systems, Vol. 1, No. 1, 1999.
28. Dao, S. and Perry, B. "An overview of data mining in heterogeneous schema integration," WESCON/96 , 1996, pp. 478 -483
29. Dhar, V. and Tuzhulin, A. "Abstract-Driven Pattern Discovery In Databases," IEEE Transactions on Knowledge and Data Engineering, Vol. 5, No. 6, 1993, pp. 926-938
30. Dua, S., Cho, E., and Iyengar, S.S. "Discovery Of Web Frequent Patterns And User Characteristics From Web Access Logs: A Framework For Dynamic Web Personalization," Proceedings. 3rd IEEE Symposium on Application-Specific Systems and Software Engineering Technology, 2000, pp. 3-8.
31. Faloutsos, C., Ranganathan, M. and Manolopoulos, Y. "Fast Subsequence Matching in Time-Series Databases," SIGMOD Conference 1994, pp. 419-429.
32. Faloutsos, C. and Lin, K-I. "FastMap: A Fast Algorithm for Indexing, Data-

Mining and Visualization of Traditional and Multimedia Datasets, "SIGMOD Conference 1995, pp.163-174.

33. Feng, L., Lu, H. and Wong, A. "A Study Of Database Buffer Management Approaches: Towards The Development Of A Data Mining Based Strategy," 1998 IEEE International Conference on Systems, Man, and Cybernetics, Vol. 3, 1998 .

34. Feng, T. and Murtagh, K. "Towards Knowledge Discovery From WWW Log Data," Proc. International Conference on Information Technology: Coding and Computing, 2000, pp. 302-307.

35. Freitas, A.A. "On Rule Interestingness Measures," Knowledge-Based Systems, Vol 12, 1999, pp. 309-315.

36. Goh, C.-L., Tsukamoto, M. and Nishio, S. "Knowledge discovery in deductive databases with large deduction results: the first step," Knowledge and Data Engineering, IEEE Transactions on , Volume: 8 Issue: 6 , Dec. 1996, pp. 952 -956

37. Grahne, G., Lakshmanan, L.V.S. and Wang, X. "Efficient Mining Of Constrained Correlated Sets," Proc. Of the 16th International Conference on Data Engineering, 2000, pp. 512 -521.

38. Hamilton, H.J., Hilderman, R.J. and Cercone, N. "Attribute-Oriented Induction Using Domain Generalization Graphs," Tools with Artificial Intelligence, 1996., Proceedings Eighth IEEE International Conference on pp. 246 - 252

39. Han, J., Cai, Y. and Cercone, N. "Knowledge Discovery in Databases: An Attribute-Oriented Approach", Proc. of 1992 Int'l Conf. on Very Large Data Bases (VLDB'92), Vancouver, Canada, August 1992, pp. 547-559

40. Han, J., Cai, Y. and Cercone, N. "Data-driven discovery of quantitative rules in relational databases," Knowledge and Data Engineering, IEEE Transactions on , Volume: 5 Issue: 1 , Feb. 1993, pp. 29 -40

41. Han, J. and Fu, Y. "Exploration of the Power of Attribute-Oriented Induction in Data Mining," U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.), Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996, pp. 399-421

42. Han, J., Nishio, S., Kawano, H., and Wang, W. "Generalization-based data mining in object-oriented databases using an object-cube model," Data and Knowledge Engineering, 25 1998, pp. 55-97

43. Han, J., Dong, G. and Yin, Y. "Efficient Mining of Partial Periodic Patterns in Time Series Database," 15th International Conference on Data Engineering, 1999, pp. 106-115.

44. Han, J. and Fu, Y. "Mining Multiple-Level Association Rules In Large Databases," IEEE Transactions on Knowledge and Data Engineering, Vol. 11, 1999, pp. 798 -805.

45. Han, J., Lakshmanan, L.V.S. and Ng, R.T. "Constraint-Based, Multidimensional Data Mining," Computer, Vol. 32, 1999, pp. 46-50.

46. Han, J., Pei, J. and Yin, Y. "Mining Frequent Patterns without Candidate Generation," Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data

(SIGMOD'00), Dallas, TX, May 2000, pp. 1-12.

47. Han, J. and Kamber, M. Data mining: Concepts and Techniques, Academic Press, 2001.

48. Hilderman, R.J., Liangchun, L. and Hamilton, H.J. "Data visualization in the DB-Discover system," Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on , 1997, pp. 474 -477

49. Hu, X. and Cercone, N. "Mining knowledge rules from databases: a rough set approach," Data Engineering, 1996. Proceedings of the Twelfth International Conference on , 1996, pp. 96 - 105

50. Kamber, M., Winstone, L., Gong, W., Cheng S. and Han, J., "Generalization and decision tree induction: efficient classification in data mining," Research Issues in Data Engineering, 1997. Proceedings. Seventh International Workshop on , 1997, pp. 111 - 120

51. Klemettinen, M., Mannila, H. and Toivonen, H. "Interactive Exploration Of Interesting Findings In The Telecommunication Network Alarm Sequence Analyzer (TASA)," Information and Software Technology, Vol. 41, No. 9, June 1999, pp. 557-567.

52. Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H. and Verkamo, A.I. "Finding Interesting Rules for Large Sets of Discovered Association Rules," Proc. of the Third International Conference on Information and Knowledge Management, Gaithersburg, Maryland, 1994, pp. 401-407

53. Lakshmanan, L.V.S., Ng, R., Han, J. and Pang, A. "Optimization of Constrained Frequent Set Queries with 2-Variable Constraints," Proc. 1999 ACM-SIGMOD Conf. on Management of Data (SIGMOD'99), 1999, pp. 157-168.

54. Lee, J., Grossman, D., Frieder, O. and McCabe, M.C. "Integrating Structured Data And Text: A Multi-Dimensional Approach," Proc. International Conference on Information Technology: Coding and Computing, 2000, pp. 264-269.

55. Lent, B., Agrawal, R. and Srikant, R. "Discovering Trends in Text Databases, " Proc. of the 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, August 1997.

56. Lent, B., Swami, A. and Widom, J. "Clustering Association Rules," Proc. of the Thirteenth International Conference on Data Engineering, Birmingham, UK, April 1997, pp. 220-231.

57. Lesh, N., Zaki, M.J. and Oglhara, M. "Scalable Feature Mining For Sequential Data," IEEE Intelligent Systems, Vol. 15, No. 2, 2000, pp. 48-56.

58. Li, S., Shen, H. and Cheng, L. "New Algorithms For Efficient Mining Of Association Rules," Information Sciences, Vol. 118, No. 1-4, Sep. 1999, pp. 251-268.

59. Li, C-S., Yu, P.S. and Castelli, V. "HierarchyScan: A Hierarchical Similarity Search Algorithm for Databases of Long Sequences," Proc. of the Twelfth International Conference on Data Engineering, New Orleans,

Louisiana, 1996, pp. 546-553.

60. Lin, M-Y. and Lee, S-Y. "Incremental Update On Sequential Patterns In Large Databases," Proc. Of the Tenth IEEE International Conference on Tools with Artificial Intelligence, 1998, pp. 24-31.

61. Lin, X., Liu, C., Zhang, Y. and Zhou, X. "Efficiently Computing Frequent Tree-Like Topology Patterns In A Web Environment," Proc. Technology of Object-Oriented Languages and Systems, 1999.

62. Liu, B., Hsu, W. and Chen, S. "Using General Impressions to Analyze Discovered Classification Rules," Proc. of the Third International Conference on Knowledge Discovery and Data Mining, 1997, pp. 31-36.

63. Liu, B., Hsu, W., Mun, L-F. and Lee, H-Y. "Finding Interesting Patterns Using User Expectations," IEEE Transactions on Knowledge and Data Engineering, Vol. 11, No. 6, 1999, pp. 817-832

64. Lu, W., Han, J. and Ooi, B.C. "Discovery of General Knowledge in Large Spatial Databases," Proc. of 1993 Far East Workshop on Geographic Information Systems (FEGIS'93), Singapore, June 1993, pp. 275-289

65. Mannila, H., Toivonen, H. and Verkamo, A.I. "Discovery Of Frequent Episodes In Event Sequences," Data Mining and Knowledge Discovery, No. 1, Nov. 1997, pp. 259-289.

66. McClean, S., Scotney, B. and Shapcott, M. "Using background knowledge with attribute-oriented data mining," Knowledge Discovery and Data Mining (Digest No. 1998/310), IEE Colloquium on , 1998, pp. 1/1 -1/4

67. McClean, S., Scotney, B. and Shapcott, M. "Incorporating Domain Knowledge into Attribute-Oriented Data Mining," International Journal of Intelligent Systems v.15 n.6 pt.0 2000, pp. 535-548

68. Ng, R., Lakshmanan, L.V.S., Han, J. and Pang, A. "Exploratory Mining and Pruning Optimizations of Constrained Associations Rules," Proc. of 1998 ACM-SIGMOD Conf. on Management of Data, 1998, pp. 13-24.

69. Ng, R., Lakshmanan, L.V.S., Han, J. and Mah, T. "Exploratory Mining via Constrained Frequent Set Queries," Proc. Of 1999 ACM-SIGMOD Conf. on Management of Data (SIGMOD'99), Philadelphia, PA, June 1999, pp. 556-558.

70. Ozden, B., Ramaswamy, S. and Silberschatz, A. "Cyclic Association Rules," International Conference on Data Engineering, April 1998.

71. Padmanabhan, B. and Tuzhilin, A. "Unexpectedness As A Measure Of Interestingness In Knowledge Discovery," Decision Support Systems, 1999, Vol. 27, pp. 303-318

72. Park, J-S., Chen, M-S. and Yu, P.S. "Using a Hash-Based Method with Transaction Trimming for Mining Association Rules," IEEE Trans. on Knowledge and Data Engineering, Vol. 9, No. 5, Oct. 1997, pp. 813-825.

73. Park, S., Lee, D. and Chu, W.W. "Fast Retrieval Of Similar Subsequences In Long Sequence Databases," Proc. 1999 Workshop on Knowledge and Data Engineering Exchange, 1999, pp. 60-67.

74. Park, S., Chu, W.W., Yoon, J. and Hsu, C. "Efficient Searches For Similar Subsequences Of Different Lengths In Sequence Databases," Proc. 16th International Conference on Data Engineering, 2000, pp. 23-32.
75. Park, S. "Neural Network and Customer Grouping in E-commerce: A Framework Using Fuzzy ART Research Challenges," Proc. Of Academia/Industry Working Conference, 2000, pp. 331-336.
76. Pasquier, N., Bastide, Y., Taouil, R. and Lakhal, L. "Efficient Mining Of Association Rules Using Closed Itemset Lattices," Information Systems, Vol . 24, No. 1, March 1999, pp. 25-46.
77. Pei, J., Han, J. and Lakshmanan, L.V.S. "Mining Frequent Itemsets with Convertible Constraints," Proc. 2001 Int. Conf. on Data Engineering (ICDE '01), Heidelberg, Germany, April 2001.
78. Pei, J. and Han, J. "Can We Push More Constraints into Frequent Pattern Mining?," Proc. 2000 Int. Conf. on Knowledge Discovery and Data Mining (KDD'00), Boston, MA, August 2000.
79. Piatetsky-Shapiro, G. "Discovery, Analysis, and Presentation of Strong Rules," Knowledge Discovery in Databases, AAAI/MIT Press, 1991.
80. Rainsford, C.P. and Roddick, J.F. "The Attribute-Orinted Induction of Rules from Temporal Interval Data," Data mining, Data Warehousing Client / Server Database proceeding of the 8th International Database Workshop Hong Kong, July, 1997, pp.29-31
81. Saar, T.M., Nava, P., Gadi, R. and Avi, P. "Mining Relational Patterns From Multiple Relational Tables," Decision Support Systems, Vol. 27, 1999, pp. 177-195.
82. Sadakane, K. and Imai, H. "Text Retrieval By Using K-Word Proximity Search, " 1999 International Symposium on Database Applications in Non-Traditional Environments, pp. 183-188.
83. Sartipi, K., Kontogiannis, K. and Mavaddat, F. "A Pattern Matching Framework For Software Architecture Recovery And Restructuring," 8th International Workshop on Program Comprehension, 2000, pp. 37-47.
84. Savasere, A., Omiecinski, E. and Navathe, S. "An Efficient Algorithm for Mining Association Rules in Large Databases," Proc. Int'l Conf. Very Large Data Bases, Zurich, Switzerland, Sep. 1995, pp. 432-444.
85. Silberschatz, A. and Tuzhilin, A. "On Subjective Measures of Interestingness in Knowledge Discovery," First International Conference on Knowledge Discovery and Data Mining, August 1995.
86. Shan, N., Hamilton, H.J. and Cercone, N. "GRG: knowledge discovery using information generalization, information reduction, and rule generation," Tools with Artificial Intelligence, 1995. Proceedings., Seventh International Conference on , 1995, pp. 372 -379
87. Silberschatz, A. and Tuzhilin, A. "What Makes Patterns Interesting in Knowledge Discovery Systems," IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, 1996, pp. 970-974.
88. Silverstein, C., Brin, S. and Motwani,

R. "Beyond Market Baskets: Generalizing Association Rules to Dependence Rules," Data Mining and Knowledge Discovery, Vol. 2, 1998, pp. 39-68.

89. Specht, G. and Kahabka, T. "Information Filtering and Personalization in Databases using Gaussian Curves," 2000 International Database Engineering and Applications Symposium, pp. 16-24.

90. Srikant, R., and Agrawal, R. "Mining Generalized Association Rules," Proc. of the 21st Int'l Conference on Very Large Databases, Zurich, Switzerland, Sep. 1995.

91. Srikant, R., Vu, Q. and Agrawal, R. "Mining Association Rules with Item Constraints," Proc. of the 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, August 1997.

92. Srinivasa, S. and Spiliopoulou, M. "Modeling Interactions Based On Consistent Patterns," International Conference on Cooperative Information Systems, 1999, pp. 92-101.

93. Sumi, K., Sumi, Y., Mase, K., Nakasuka, S-I. and Hori, K. "Takealook: Personalizing Information Presentation According to User's Interest Space," 1999 IEEE Conference on Systems, Man, and Cybernetics, pp. 354-359.

94. Tang, C., Lau, R.W.H., Qing, L., Huabei, Y., Tong, L. and Kilis, D. "Personalized Courseware Construction Based On Web Data Mining," Proceedings of the First International Conference on Web Information Systems Engineering , Vol. 2, 2000, pp. 204-211.

95. Toivonen, H. "Sampling Large Databases For Association Rules," The 22th International Conference on Very Large Databases (VLDB'96), Mumbay, India, Sep. 1996, pp. 134-145.

96. Tsumoto, S. "Knowledge discovery in medical databases based on rough sets and attribute-oriented generalization," Fuzzy Systems Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on , Volume: 2 , 1998, pp. 1296 -1301

97. Tsumoto, S. "Knowledge discovery in clinical databases and evaluation of discovered knowledge in outpatient clinic," Information Sciences v124 n1 2000, pp. 125-137

98. Tung, A.K., Han, H.J., Lakshmanan, L.V.S. and Ng, R.T. "Constraint-Based Clustering in Large Databases," Proc. 2001 Int. Conf. on Database Theory (ICDT'01), London, U.K., Jan. 2001.

99. Wang, J. T-L., Chirn, G..W., Marr, T., Shapiro,G., Shasha, B.D. and Zhang, K. "Combinatorial Pattern Discovery For Scientific Data: Some Preliminary Results," Proceedings of
100. ACM SIGMOD, 1994, pp. 115-125.
Wang, K. and Liu, H. "Discovering Structural Association Of Semistructured Data," IEEE Transactions on Knowledge and Data Engineering, Vol.
101. 12, No. 3, 2000, pp. 353-371.
Wang, K., He, Y. and Han, J. "Mining Frequent Itemsets Using Support Constraints," Proc. 2000 Int. Conf. on Very Large Data Bases, Cairo, Egypt,
102. Sept. 2000.
Weiss, S. M., Apte, C., Damerau, F.J., Johnson, D.E., Oles, F. J., Goetz, T.

and Hampp, T. "Maximizing Text-Mining Performance," IEEE Intelligent Systems, Vol. 14, No. 4, 1999, pp. 63-
103. 69.

Yi, B-K., Sidiropoulos, N.D., Johnson, T., Jagadish, H.V., Faloutsos, C. and Biliris, A. "Online Data Mining For Co-Evolving Time Sequences," Proc. 16th International Conference on Data Engi-
104. neering 2000, pp. 13-22.

Yoon, S.-C. and Park, E.K. "An approach to intensional query answering at multiple abstraction levels using data mining approaches," Systems Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International
105. Conference on , 1999, pp. 1-9

Yu, P.S. "Data Mining And Personalization Technologies," the 6th International Conference on Database Systems for Advanced Applications, 1999, pp. 6-13.

106. Zaiane, O.R., Xin, M. and Han, J. "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs," Proc. Advances in Digital Libraries Conf. (ADL'98), Santa Barbara, CA, April 1998, pp. 19-29.

107. Zaki, M.J. "Efficient Enumeration of Frequent Sequences," 7th International Conference on Information and Knowledge Management, Washington DC, Nov. 1998, pp 68-75.

107. Zaki, M.J., Lesh, N. and Ogihara, M. "PlanMine: Predicting Plan Failures using Sequence Mining," Artificial Intelligence Review, special issue on the Application of Data Mining, 1999.