資料分析技術與情報應用之探討

王朝煌（Jau-Hwang Wang）
中央警察大學資訊管理系教授

摘要：資料探勘的目的就是從大型資料庫當中挖掘出這些潛藏的特徵與知識，提供管理者作爲決策的重要依據。資料探勘是一種自動化的模式建立，這些模式代表隱藏在大型資料庫或其他大量資訊儲存體中的知識。它涉及多學科領域，包括資料庫技術、人工智慧、機器學習、類神經網絡、統計學、模式識別、知識庫系統、資訊檢索和資料視覺化等技術。近年來已被廣爲應用於工商企業營運資料與文字資料之分析，以萃取決策所需之參考資訊。在資訊時代如何善用資料探勘技術，從浩瀚的數位資料發覺犯罪訊息，防患於未然，將成爲現代警察艱鉅挑戰之一。本文探討資料探勘的概念及其技術，並介紹最廣爲應用的關聯規則分析技術與文件分析技術之原理及其應用方法。

關鍵詞：資料探勘，關連規則分析，文字探勘，犯罪情報。
Keywords: Data Mining, Association Rule Analysis, Text Mining, Crime Intelligence.

綱目：

一、緒論
二、結論與建議
三、文字探勘技術與應用
四、結論與建議

一、緒論

我國警察機關自民國六十年起開始使用電腦處理資料[廖有錄 1994]，目前累積有大量的犯罪資料。研擬如何善用資訊科技，開發犯罪資料分析技術，將刑案資料作進一步分析統計獲取隱含資訊，作爲預測犯罪情勢之參考，以釐訂適當之刑事政策，實爲迫切需要之課題。此外，由於社會的進步與國民知識水準的提升，傳統佈建所能獲取的犯罪線索勢將越來越少，犯罪偵查工作所面臨的挑戰，不可言喻。此外隨著資訊科技的進步及其廣爲應用，以及資訊與通訊技術的結合，保密電話的發明，使得傳統以監聽獲取犯罪線索的方法，將漸被淘汰。因此如何利用科技開發新的犯罪線索蒐集方法爲刻不容緩之事。近年來資訊及網路科技的進展，人類無紙化的資訊網路空間(cyber-space)已然形成。廣大的網路空間爲未來破案線索之所寄，如何從浩繁的網路空間及無數稍縱即逝的網路訊息，萃取破案資訊，實爲迫切之研究課題[王朝煌 1998a&b]。

犯罪資訊分析[林燦璋 1993]是一套系統化的程序，特別是用來提供描述性與統計性的資料，專門協助警方，在策略研擬、偵辦步調及資源配置等過程上，做出迅速而正確的決策。近年來由於資訊科技的進步，資料的自動化分析技術有已

長足的進展，如資料探勘技術，已日趨成熟，且有相關產品問世。爲提昇資料分析效率，實有必要將資料探勘技術引進應用於犯罪資料之分析或犯罪線索之獲取。資料探勘技術乃整合資料庫技術、人工智慧技術、統計技術、資料檢索技術、及資料萃取技術等，以自動或半自動的方式分析、探索大量資料，以尋找資料之特殊型式(patterns)或規則(rules)[Han2001]。近年來已被廣爲應用於工商企業營運資料與文字資料之分析，以萃取決策所需之參考資訊。

資料探勘技術主要分爲：結構化資料探勘(data base mining)與非結構化或半結構化資料的文字探勘(text mining)兩大範疇，前者以分析較具結構的資料庫紀錄爲主，後者主要用以分析文件資料，如文章、網頁、新聞、及電子郵件訊息等結構較爲鬆散的資料。本文探討資料探勘的概念及其技術，並介紹最廣爲應用的關聯規則分析技術與文件分析技術之原理，並探討其應用方法。

二、資料探技術與應用

(一)、資料探勘技術

資料探勘的目的就是從大型資料庫當中挖掘出這些潛藏的特徵與知識，提供管理者作爲決策的重要依據。資料探勘是一種自動化的模式建立，這些模式代表隱藏在大型資料庫或其他大量資訊儲存體中的知識。它涉及多學科領域，包括資料庫技術、人工智慧、機器學習、神經網絡、統計學、模式識別、知識庫系統、資訊檢索和資料視覺化等技術。其功能有概念描述（concept description）、關聯分析（association analysis）、分類（classification）與預測（prediction）、群集分析（cluster analysis）等[Han2001]。

概念描述是要用簡單且容易瞭解的方式，來描述資料庫中的複雜狀況。一個正確的描述，可能啓發更多對該狀態的解釋，概念描述包含對資料特性的描述（characterization）與資料辨別（discrimination）描述等。

關聯分析的目的在於發現交易資料庫中項目之間的關係，也就是說哪些事件會同時發生，找出其中的關聯規則（association rule）。關聯分析經常被運用在購物籃（market basket）或交易資料分析。例如：分析在超級市場中，哪些商品常會一起被購買，藉此來瞭解顧客的消費習慣。衡量關連規則的兩大指標爲最小支持率(minimum support)與最小信賴度(minimum conference)。最小支持率指是否有足夠的資料支撐規則的效用，最小信賴度乃規則之前提(antecedent)與其結果(consequence)間信度。

資料分類與預測在機器學習、專家系統、統計學領域上是一種常見的技術，其功能在於找出能夠描述、區分資料類別的模型，並利用這個模型來預測物件或資料的類別歸屬。它是屬於一種監督式的學習（supervised learning model），資料分類有兩個步驟，首先使用包含各種特徵或屬性的訓練資料（training data），而訓練資料中的每一筆記錄必須事先標註其類別（class），利用這些訓練資料的各種特徵和屬性進行分析，產生學習模型，從而建構一個分類器。第二個步驟是使用與訓練資料特徵屬性相同，但內容不同的測試資料（testing data）對此模型進行測試，

判斷是否與測試資料所給定的類別相同，並評估模型的預測準確率，若準確率可以被使用者接受，往後即可利用此模型來預測資料的分類。目前已有許多分類的演算法被提出，包括決策樹歸納（Decision Tree Induction）、貝氏分類（Bayesian Classification）及類神經網路（Neural Network）等[Han2001]。

資料分群可以幫助簡化資料、建立資料的分類規則。所謂的分群（clustering）即將一群異質的群體區隔爲同質性較高的群集或是子群。簡單的說，資料分群即是將 n 筆資料自動分成 k 個群（groups）。除此之外，在某些應用領域上，對於資料分群集之後，落在群集以外的資料物件（outliers）會成爲分析研究的目標。分群演算法發展至今以有多年的歷史，目前已經有許多的分群演算法可以用來做資料分析，但不同的演算法適合不同的資料型態。選擇一個適合的演算法將是分類結果是否適當的關鍵因數之一。分群演算法依其分群方法可細分爲分割式（partitioning）、階層式（hierarchical）、密度基底式（density-based）、網格基底式（grid-based）與模型基底式（model-based）等不同種類。

(二)、關聯規則分析

關聯規則分析爲從大量的資料中，找出共同出現的特殊關聯規則，常被運用在大量的交易資料中找尋交易項目共同出現的型態，或稱爲購物藍分析(Market Basket Analysis)。一般而言交易紀錄 T 包含若干交易項目、數量、金額、及合計金額等資料，其型態如下：

I_1	Q_1	$Amount_1$
I_2	Q_2	$Amount_2$
I_3	Q_3	$Amount_3$
I_n	Q_n	$Amount_n$
	Total	

假如我們只對交易項目共同出現的型態感興趣，則交易資料 T 可簡化爲：

$T = <I_1, I_2, I_3, \cdots.., I_n>$。將所有交易資料紀錄簡化後運用關聯規則加以分析，即可找尋交易項目共同出現之型態，進而據以推論顧客購買行爲之共通型式(common pattern)，作爲支援行銷或產品陳設等決策的參考。共通型式必須滿足兩個條件才能作爲決策的參考；首先共通型式必須爲大量交易(例如所有交易的 30%或 40%)的共同購買行爲，稱爲滿足最小支持率(minimum support)，在少數交易的共通型式，一般較不具商業價值(在其他情報應用是否亦如此？)。此外這些共同出現的交易項目間必須有高度的正相關，例如顧客在購買產品 A 時，其亦購買產品 B2 的機率高於一定的門檻值(例如 75%)，稱爲滿足最小信賴度(minimum confidence)，否則較不具決策參考價值。

關聯分析技術主要分爲兩個步驟：1. 尋找共同交易項目，2.關聯規則推導。本文例子說明關連規則分析技術。

1.尋找共同交易項目

假設所有交易項目的集合爲：<A, B, C, D, E>，且所有的交易紀錄如下：

交易	交易項目

T1	A, C, E
T2	A, C, D
T3	C, D
T4	B, D, E
T5	C, D, E

假設最小支持率設定爲 40%，關聯規則分析首先統計所有交易項目之交易筆數及其支持率，本例的統計結果爲：

交易項目	交易別	交易筆數	支持率
A	T1, T2	2	40%
B	T4	1	20%
C	T1, T2, T3, T5	4	80%
D	T2, T3, T4, T5	4	80%
E	T1, T4, T5	3	60%

其中交易項目 B 僅出現在交易紀錄 T4，支持率爲 1/5 = 20%，未達最小支持率，其資訊不具商業決策參考價值。因此滿足最小支持率之組合個數爲一的共同交易項目爲：

交易項目	支持率
A	40%
C	80%
D	80%
E	60%

組合個數爲二的共同交易項目其任意子集合亦必須滿足最小支持率，因此滿足最小支持率之組合個數爲一的共同交易項目可以作爲計算組合個數爲二的共同交易項目之基礎。本例中因爲交易項目 B 之支持率僅 20%，交易項目 B 不可能與其他的項目共同出現且具有超過 20%的支持率，因此在計算組合個數爲二的共同交易項目時，可以忽略交易項目 B 而不影響分析結果。因此在計算組合個數爲二的共同交易項目時，僅需考慮下列組合：

交易項目組合	交易別	交易筆數	支持率
A, C	T1, T2	2	40%
A, D	T2	1	20%
A, E	T1	1	20%
C, D	T2, T3, T5	3	60%
C, E	T1, T5	2	40%
D, E	T4, T5	2	40%

其中交易項目組合{A, D}及{A, E}之支持率僅 20%可以略去，因此滿足最小支持率、組合個數爲二的共同交易項目爲：

交易項目	支持率
A, C	40%
C, D	60%
C, E	40%
D, E	40%

同理，滿足最小支持率之組合個數爲三的共同交易項目，其任一子集合亦必須滿足最小支持率，因此我們可以從滿足最小支持率之組合個數爲二的共同交易項目計算組合個數爲三的共同交易項目，可能的組合如下：

交易項目組合	交易別	交易筆數	支持率
A, C, D	T2	1	20%
A, C, E	T1	1	20%
A, D, E		0	0%
C, D, E	T5	1	20%

以上交易項目組合{A, C, D}, {A, C, E}, {A, D, E}, 及{C, D, E}皆未達最小支持率，因此滿足最小支持率之組合個數爲三的共同交易項目爲空集合。同理，可推知滿足最小支持率之組合個數爲四以上的共同交易項目亦爲空集合，因此無須繼續找尋其他共同交易項目組合。

2.關聯規則推導

本例中滿足最小支持率之共同交易項目組合之個數最大者爲二，因此我們僅須從下表中推導關聯規則：

交易項目組合	支持率
A, C	40%
C, D	60%
C, E	40%
D, E	40%

關聯規則推導的邏輯爲：

- 尋找每一個交易項目組合 L 之非空正子集合(nonempty proper subset)
- 每一非空正子集合 s, 如果 $\frac{L\text{的的支持}}{s\text{的的支持}}$ 大於最小信賴度，則推導出規則 " s => (L – s)" 。
- 計算每一規則 s => (L – s)之增益(*improvement* or **lift**)[Berry 1997]：$\frac{P(L)}{P(L-s)\quad P(s)}$，剔除增益小於一之規則。

假設最小信賴度設定爲 75%，則本例可推導出下列規則及其增益如下：

規則編號	規則	信賴度	增益
1	A => C	100%	P(C/A)/P(C) = 1/(4/5) = 5/4
2	C => A	50%	P(A/C)/P(A)= (1/2) / (2/5) = 5/4
3	C => D	75%	P(D/C)/P(D) = (3/4)/(4/5) = 15/16
4	D => C	75%	P(C/D)/P(C) = (3/4)/(4/5) = 15/16
5	C => E	50%	P(E/C)/P(E)=(1/2)/(3/5)=5/6
6	E => C	66.7%	P(C/E)/P(C)= (2/3)/(4/5)=5/6
7	D => E	50%	P(E/D)/P(E)=(1/2)/(3/5)=5/6
8	E => D	66.7%	P(D/E)/P(D)=(2/3)/(4/5)=5/6

其中規則 1 之信賴度爲 100%，增益爲 1.25 ，符合最小信賴度設定爲 75%及增益大於一之標準，較具參考價值。其他規則 2(信賴度：50%)，規則 5(信賴度：50%)，規則 6(信賴度：66.7%)，規則 7(信賴度：50%)，及規則 8(信賴度：66.7%)等之信賴度皆小於最小信賴度，規則 3(增益：15/16)及規則 4(增益：15/16)之增益皆小於 1，則較不具決策參考價值。

(三)、關連規則分析於電腦鑑識之應用

關聯規則最常見的應用爲購物藍分析，以分析歸納一般顧客共同的購物行爲型態，如 75%購買乳製品的顧客亦會購買麵包。將關連規則分析適度的調整，則可廣泛地應用於其他歸類分析的應用，如以內容爲基礎之推薦系統(content-based filtering recommendation systems)及協同推薦系統(collaborative recommendation systems)、顧客剖繪(customer profiling)、及網頁瀏覽分析(web browsing activities)[Abraham2002]。推薦技術其主要立論乃顧客在購物(如買書)時對類似於之前曾經買過的較感興趣，以及同類型之顧客其購物傾向亦類似的原理，分析歸納顧客之購物行爲，作爲推薦產品之依據。顧客剖繪乃將顧客的基本資料及購物行爲加以分析以了解顧客的特質，已漸廣用於電子商務系統。網頁瀏覽分析則包含分析網友對網頁之喜好程度、瀏覽路徑之偏好等瀏覽行爲，作爲調整個人化網頁、網頁內容與架構之參考。

在電腦鑑識的應用方面，可以將計算機系統的日誌、網路連線日誌、資料庫紀錄、及使用者之個人資料加以分析，例如系統的使用者依其功能分別隸屬於資訊部門(系統管理者，程式設計師，系統操作員)，財務部門，會計部門，人力資源部門等等。運用關聯規則分析之後，可以歸納每一類使用者之使用習慣，推導系統使用者之慣性，例如：

規則一：

(使用者類別爲系統操作員) 且 (日期周五且爲下班時間) ➔ (則系統資料備份爲合理之操作)

規則二：

(使用者類別爲系統管理員) 且 (日期爲星期二) → (則系統使用者帳戶更新維護爲合理之操作)

規則三：

(使用者類別爲人事管理員) 且 (日期爲月底) → (則存取人事薪資系統爲合理之操作)

…

假如系統偵知使用者的行爲偏離正常使用之慣性，則可送出警訊給安全人員作深入之調查。

三、文字探勘技術與應用

(一)、文字探勘技術

資訊科技的進步及全球資訊網的廣爲應用，使得電子化文件急速增加。因而如何從電子化文件篩選過濾有用資訊訊及萃取知識的文字探勘技術(text mining)，便成爲資料探勘另一重要課題。文字探勘技術主要應用於搜尋文件內的特殊內涵，以提升文件管理及資料檢索效能，例如文件的自動分類、分群、查詢增強、及事件追蹤[張毓修 2001]。

文件分類乃依文件的內涵加以分析，並將文件加以歸納爲已知的類別，以有效組織管理文件。常見的方法爲將文件篩選分析，設定代表文件之關鍵詞集合，再利用決策樹、類神經網路(Neural Network)、貝氏網路(Bayesian Network)、最近 K 鄰近演算法，加以分類。

文件分群主要依據文件與文件之間的相似程度，將文件分成數個群集，使在同一群集的文件之間的相似度很高，而不同群集的文件間相似度很低。亦即將文件依其領域的不同分成不同的群集，使得相同領域的文件分在同一群集之中。分群（clustering）與分類（classification）的主要差異在於分群前並不知物件可以被分爲哪幾種類別，或是不想限定分爲幾種類別，可以依照資料的組成型態，做適當的分類。同時，利用分群的方法也不需以訓練資料庫產生分類規則。

查詢擴充主要運用字詞關聯及文件的關聯分析(association analysis)結果，改寫查詢以增加查詢的檢出率(recall)。運用字詞關聯的查詢擴充，例如使用者以關鍵字 computer 查詢進行資料檢索，如 computer 與 information, data processing 等字有密切關聯，可將查詢擴充爲 computer, information, 及 data processing等三字之組合，將內涵 computer, information, 或 data processing 之所有文件檢出。運用文件關聯的查詢擴充，例如以關鍵字 computer 檢索出文件資料 A, B, C，其中 A 及 B 經使用者確認爲相關文件，文件 A 及 B 內涵關鍵詞 information system, data processing，可據以將查詢擴充爲 computer, information system, 及 data processing等三字之組合，重新進行資料檢索以檢索出更多之相關資料。

事件的追蹤主要用於分析文件資料，例如公司說明文件及例行記者會說明文件，分析出事件的發展趨勢。常見的分析方法爲將說明文件進行分析，賦予代表文件之關鍵詞及權重，如果文件間之相似度極高，則將文件歸類爲描述同一事件之說明文件，再由說明文件關鍵詞的變化，分析事件的演進情形。

(二)、概念萃取與斷詞

以上這些應用皆須將文件進行前置之概念萃取或斷詞處理，以便萃取出代表文件之關鍵詞及計算其權重。目前中文文件的斷詞，主要分爲詞庫斷詞法、統計斷詞法、及混合式斷詞法[賴錦慧 2004]。

詞庫斷詞法：爲最普遍的斷詞方式，利用詞庫和文件中的句子做字詞比對，以找出文件內之字詞。爲保持詞庫斷詞之正確性，詞庫內容必須時常維護與更新。我國教育部提供有國語辭典列表，內涵近十六萬筆各類中文詞。

統計斷詞法：從大量領域相關的文件資料庫中分析統計找出鄰近字元共同出現的頻率及前後字之分佈情形作爲斷詞的依據。例如某一系列的字 ABCD 共同出現的頻率頗高，且 A 之前及 D 之後出現的字則較爲多樣化，則 ABCD 爲一詞的可能性較高。

混合式斷詞法：利用詞庫斷出不同的組合字詞，再利用字詞的統計資訊，找出最佳的斷詞組合。

關鍵詞權重設定之原理乃利用字(詞)頻統計技術[Jones1972, Luhn1957, Salton1973, Salton1975]，首先對一大量文件資料之每一字(詞)被使用頻率作統計，再對個別文件中之字(詞)使用頻率作一統計。然後再根據統計資料，找出在個別文件中使用次數較多，但在一般文件使用頻率較少之字(詞)，兩項之比值較大者，其代表性越強。因此可由比值之大小來決定某一字(詞)是否適合成爲文件之代表(亦即可爲該文件之關鍵詞)。關鍵詞權重設定公式如下：

$$w_{ij} = tf_{ij} * \log \frac{N}{Cf_j} \text{.....(1)}$$

在(1)式中，w_{ij} 代表第 i 個文件的第 j 個關鍵詞之權重，tf_{ij} 表示第 i 個文件第 j 個關鍵詞在第 i 個文件出現之頻率，N 代表文件資料庫之文件數，而 Cf_j 則表示含第 j 個關鍵詞之文件數。

(三)、文字分析技術於犯罪模式之應用

「犯罪模式」一詞，原出於拉丁文的 Modus Operandi，英文稱其爲 Methods of Operation，日本譯爲「犯罪手口」，我國則有人譯稱爲「犯罪方式」或謂「犯罪手法」，係指慣行犯於每次作案時所習用的犯罪手段、方法或形式，即犯罪定型或類型的意思。犯罪者依其生長環境之差異，而養成獨特之性格與行爲特質，而這些特質也常反應於其犯罪行爲上。一般犯罪行爲之實施，大都先有犯罪動機、意念，再產生計畫、準備，而後實施完成犯罪行爲;因此當犯罪者決定犯罪時（除偶發犯外），常經一番考慮、評估，尤其在財產性犯罪，由於投資報酬率高，故再度犯下類似案件之機率非常高，且其犯罪手法皆有重復出現之現象，進而形成所謂的「犯罪模式」。因此在十九世紀末，奧地利刑事法學者漢斯克魯茲將犯罪模式之特性加以研究，付之理論形式，並提出其具有必存性、反覆性(慣行性)、及固定性之特性

[守屋恆浩]。犯罪模式分析主要運用於累犯的行爲特徵分析，蓋初犯的手段和方法，尚無一定之習慣性，不能稱之爲模式，但若其再犯，便可從其作案的手段、方法或形式分析歸納建檔[倪秋煌 1987]。由於犯罪者考量安全、容易得手及實質利益等因素，往往會反覆運用自己最得意的手段與方法從事犯罪行爲，因此在連續犯、常業犯、或職業犯的犯罪行爲往往會重複出現同一犯罪模式。在犯罪發生之後，雖然犯罪者已經離去，但根據「痕跡論」與「無痕跡論」，犯罪行爲雖具隱密性與消失性，但在犯罪現場必留下跡證或犯罪徵候，而這些跡證或犯罪徵候即爲犯罪者獨特性格與行爲特質之表徵。因此如能將每一犯罪者之獨特性格與行爲特質加以分析，歸納其犯罪模式並予存檔，當犯罪發生時，即可比對犯罪現場跡證之犯罪模式與犯罪模式資料庫，找出可能之涉嫌者。

綜合上述犯罪模式的主要內涵犯罪手法特徵及其代表性。刑事偵查人員對於轄區的犯罪疑慮人口資料，雖可作適當的管理與運用。然而犯罪的特徵資料非常廣泛，以強盜案件爲例，犯罪特徵即達十大類共三百八十七項之多[謝慶欽 1994]，然而一般人腦只能同時處理五到九塊不同的資訊[Miller1956]，以人工分析犯罪模式，恐難克服人類先天的限制。因此犯罪模式的推導除了經由刑事偵查人員的經驗累積之外，在數位化時代可以透過資料分析技術偵對犯罪人資料，犯罪報案資料，犯罪偵查報告等等，進行探勘及推倒犯罪模式。例如運用特徵頻率統計法，將經常出現在某一嫌犯之犯行中，卻鮮少在其他的嫌犯中發現的獨特特徵，賦予較高的權重。亦即權重應是每一特徵在各偵查報告與其他地方所出現之頻率的一個函數。基此，可以下列公式作爲設定權重之標準[王朝煌 1997]：

$$w_{ij} = tf_{ij} * \log \frac{N}{Cf_j} \text{.....(2)}$$

在(2)式中，w_{ij} 代表第 i 個嫌犯的第 j 個特徵之權重，tf_{ij} 表示第 i 個嫌犯的第 j 個特徵在與該嫌犯有關之報告中出現之頻率，N 代表資料庫內所有之嫌犯數，而 Cf_j 則表示出現過第 j 個特徵之嫌犯數。再將代表性不足之特徵略去，即可歸納推導出犯罪模式。並據以進行犯罪模式比對、分類、分群、及關聯規則分析。例如在犯罪特徵與權重計算出來之後，犯罪案件或者犯罪者之犯罪模式可以轉化爲向量空間之向量如下：

$$A_i = (a_{i1}, a_{i2}, \ldots, a_{it}) \text{.....(3)}$$

兩向量 A_r 及 Q_s 間之相似程度(similarity)，可以向量內積計算而得：

$$\text{similarity}(A_r, Q_s) = \sum_{i=1}^{t} a_{ri} * q_{si} \text{.....(4)}$$

公式(4)可作爲計算犯罪案件間、犯罪者與犯罪案件間、以及犯罪者間之相似程度，提供偵查方向研判之參考資訊。

四、結論與建議

資料探勘的三大支柱爲資料探勘方法、資料、及模型建立技巧[吳旭志 2001]。資料探勘技術日漸成熟，市面上已漸有套裝軟體出現，在關聯分析方面，市面上已有商業軟體可供使用(如 I2)，其主要功能爲從所蒐集整理的資料中，分析資料間之關係。這些應用均以龐大的資料爲基礎，因此我國將來如欲發展類似的計畫，應先著手規劃完整的資料蒐集機制。此外資料探勘技術成功的重要因素，除了統計及分析技術之外，通常亦須融入相當的應用領域知識(Domain Knowledge)以建立分析模型及精確判讀資訊，才能獲得較佳之效果，例如資料探勘建立預測模型的主要假設之一爲”過去的紀錄適合用來預測未來的”，然而有些案例指出過去的紀錄，由於外部事件的影響，而不適於用來預測未來，決策者仍必須必須具備應用領域知識，才能有效運用探勘之資訊。

傳統警察工作的主要重點之一在於發覺犯罪，防患於未然。在資訊時代，隨著電腦的普及應用，利用電腦程式的擴散與執行而達到犯罪目的之行爲將日益增加，如以電腦病毒從事犯罪之行爲。因此如何於資訊社會發覺犯罪，防患於未然，亦將成爲現代警察之艱鉅挑戰之一。美國在 911 事件之後，將傳統於犯罪事件發生之後再進行調查的偵查作爲，調整爲事前主動調查，以提早發現不法，瓦解犯罪圖謀，避免遭受犯罪之破壞。並將犯罪前的資料探勘(precrime data mining)列爲重點工作之一[Mena2003]。因應資訊社會犯罪偵查之需要，除了增設專責單位(如刑事局偵九隊)，負責資訊、網路犯罪之偵查外，研發資訊萃取及資訊分析過濾等技術，以提供執法人員偵測及搜集違法資訊之輔助工具，應爲現代化警政建設迫切之課題。

中文參考文獻

[王朝煌 1997] 王朝煌，林璨璋，謝慶欽，1997，"犯罪嫌疑犯資訊擷取系統"，*中央警察大學學報*，第 30 期，355 頁-380 頁。

[王朝煌 1998a**]** 王朝煌，”網路犯罪之防治—從網路對現行法律秩序的挑戰談起”，*曠野雜誌*，6 頁-12 頁，1998。

[王朝煌 1998b] 王朝煌，”資訊社會之犯罪偵查”，前瞻觀點，中央日報，1998 年六月十五日。

[守屋恆浩] 守屋恆浩，*犯罪手口之研究*，立花書房，昭和五十二年八月，第 3-12 頁。

[吳旭志 2001] 吳旭志，賴淑貞譯，資料採礦理論與實務：顧客關係管理的技巧與科學，Michael J. A. Berry & Gordon S. Linoff 原著，維科圖書，2001。

[林燦璋 1993] 林燦璋，系統化的犯罪分析:程序、方式、與自動化犯罪剖析之探討，警政學報第二十四期，中央警官學校，民國八十二年十一月，第 124 頁。

[倪秋煌 1987] 倪秋煌，*警察紀錄學*，中央警官學校，民國七十六年十月七版。

[張毓修 2001] 張毓修，*次序性文件探勘：事件演化關連之研究*，國立中山大學資訊管理研究所碩士論文，2001。

[廖有錄 1994] 廖有錄 編著，*電子計算機概論--含警政電腦化簡介*，中央警官學校印行，第十五章，臺灣，1994。

[賴錦慧 2004] 賴錦慧，*新聞事件之變化探勘以支援決策制定*，國立交通大學資訊管理研究所碩士論文，2004。

[謝慶欽 1994] 謝慶欽，*資訊擷取技術應用於犯罪偵查之研究—以強盜案件爲例*，中央警官學校警政研究所碩士論文，1994。

英文參考文獻

[Abraham 2002] Abraham, T. and Oliver de vel, “Investigative Profiling with Computer Forensic Log Data and Association Rules”, *Proceedings of the 2002 IEEE International Conference on Data Mining*, 2002.

[Berry 1997] Berry, M. J. and G. Linoff, *Data Mining Techniques for Marketing, Sales, and Customer Support*, John Wiley and Sons, 1997.

[Han 2001] Han, J. and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2001.

[Jones 1972] Jones, K.S., "A Statistical Interpretation of Term Specificity and Its Applications in Retrieval", J*ournal of Documentation*, Vol. 28, No. 1, 1972.

[Luhn 1957] Luhn, H.P., "A Statistical Approach to the Mechanized Encoding and Searching of Literary Information", *IBM Journal of Research and Development*, Vol. 1, No. 4, October 1957.

[Mena 2003] Mena, J., *Investigative Data Mining for Security and Criminal Detection*, Butterworth Heinemann, 2003.

[Miller1956] Miller, G. A., “The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capability for Processing Information.”, *Psychology Review* 63(2): 81-97, 1956.

[Salton 1973] Salton, G., and C.S.Yang, "On the Specification of Term Values in Automatic Indexing", *Journal of Documentation*, Vol. 29, No. 4, 1973.

[Salton1975] Salton, G., "A Theory of Indexing", *Regional Conference Series in Applied Mathematics, No. 18,* Society for Industrial and Applied Mathematics, Philadelphia, 1975.