資料採礦及其在國防應用之探討

溫志皓

提　要

資料採礦(data mining)是資訊科技應用的一項新技術，發展至今不過僅僅數十年，但是直至近十餘年來技術才較爲成熟，而且目前已經普遍應用在各種行業與範圍中，協助吾人能在短時間內發現值得注意的特殊事件，例如信用卡盜刷、行爲偏差預測、例外事件預防、太空探測等。在九一一事件後，美國對於反恐及防恐無所不用其極，運用各種資訊科技與情、監、偵系統，希望能夠防患於未然，並且提供有效情資給各國做預防工作。目前所倚賴的技術之一，就是資料採礦的科技，藉由資料採礦的龐大運算能力，將各種相關的旅遊與銀行交易資料分析比對，並鎖定特定對象提高注意力，以有效阻絕發生恐怖攻擊的發生。由於這一套分析技術與軟體已經能夠有效協助美國在九一一事件後提供防恐及反恐情資，澳洲已於2007年7月10日宣布將引進此分析流程與技術，協助擔任反恐最重要的資料分析工作。

關鍵詞：資訊管理、資訊安全、資料採礦、知識本體、恐怖攻擊

壹、緒　　論

資訊管理的範圍相當寬廣，近年來在這其中十分熱門且運用日漸廣泛的一支即爲資料採礦的技術。

資料採礦是近年來極爲熱門應用的技術之一，但是其熱絡的發展僅是近十餘年左右的事而已。雖然發展與應用的時間僅有十餘年，但是其背後所使用的演算法（例如：類神經網路演算法、基因演算法或關聯規則演算法等）或統計技術（如統計學、因素分析、羅吉斯回歸、時間序列或時序分析等），都已經發展了數十年甚至數百年之久，才能奠定資料採礦在這短時間內迅速蓬

勃發展的能量與能力，進而提供我們許多過去無法發現的特殊類型(pattern)，而進一步分析與使用。若是提到資料採礦的開端，仍應追溯至Usama Fayyad博士。在Fayyad於1987年就讀密西根大學參加GM的暑期工作時，爲了能自成千上萬的維修記錄中發掘特定規則(rule)與類型(pattern)，並能夠協助相關的維修記錄人員迅速的發現、解決問題。Fayyad所發展的Pattern辨識演算法，不但成了他1991年博士論文的主題，亦衍生出後來資料採礦的發展[註一]。之後，Fayyad加入美國太空總署(NASA)的噴射推進實驗室，其所發展的演算法在諸多領域如太空探測、生物基因與遺傳科技等均展現了相當驚人的結果。目前美國軍方也著手開始應用此技術來增強雷達解讀與辨識資料的能力[註二]。

爲有效維護國內治安及打擊犯罪，警政署刑事局資訊室自2000年起開始研發「刑案知識庫」，並於在2003年1月完成，也讓警方偵辦刑案再添一樁利器。這項「刑案知識庫」是應用最新資訊技術，整合司法院、法務部及警政署等機關之判決、執行、起訴及移送等刑案資料、前科相片、在監在所、同囚會客、通緝、流氓、幫派、典當、出入境及車籍等總計約5億筆的資料，提供警方在刑案發生後，僅掌握部分線索，如：地緣關係、犯罪手法、嫌疑犯年齡、性別等，即可利用資料採礦(data mining)、全文檢索(full text information retrieval)及跨部門資訊整合等先進科技，立即分析過去發生的刑案資料，迅速將相關案件、可疑人犯、相片及其共犯結構，在第一時間內，提供給偵辦刑案員警參考，成爲警方打擊犯罪最有效益的輔助工具，可媲美美國FBI所使用的電腦系統，也是亞洲第一個完成開發的國家。「刑案知識庫」另成立偵查專卷區，系統主動運用資料採礦的聚類分析，將刑案內含犯罪手法的同質資料如：「侵入住宅竊盜」、「破壞車窗竊盜」及「超商搶案」等以專卷方式全面提供員警參考。而爲了預防少數不肖員警不當使用犯罪資料，「刑案知識庫」採取創新作法，設立查詢紀錄預警稽核機制[註三]。

目前的資料採礦技術應用範圍相當廣泛，諸如：科學、行銷、工業、商業、體育、金融、財務、銀行、通訊、電信業、網路、零售商、製造業、醫療保健、製藥業、健康照護、教育與警政…等等，但在國內軍事範疇的應用，諸如國家安全、例外偵測、危險評估、故障檢核等仍是待開發的研究領域。

貳、從資料到知識

一、資料、資訊與知識

總的來說，資料、資訊與知識這三種截然不同的型態，經常讓人造成誤解，會認爲吾人已經在資料庫中記錄了許多的訓練資料、成績或是交易記錄等，是否可以馬上利用資料採礦的技術來挖掘出知識呢？首先來說明資料與知識的差異。

資料的型態可概分爲[註四]：

㈠非結構化資料(unstructured data)：所謂的非結構化毫無組織的資料或毫無結構可循的文件，稱爲完全非結構化文件。例如，缺乏段落結構的全文資料、遙測資料、監控

註一 「資料採礦簡介」，資料來源：http://www/teaching/stat_data_mining/introduction.htm

註二 參見網站http://is.arc.nasa.gov/IDU/DM.html

註三 謝明俊，「刑案知識庫，情資e點靈」，中國時報（臺北市：民92年2月16日），社會綜合版。

註四 DHS, "Early Attention to Privacy in Developing a Key DHS Program Could Reduce Risk", DHS Privacy Office Response to House Report (2007, Feb), p.13.

數據、聲音、影像資料…等，這一類型的文件在儲存時較爲簡單，然而由於欠缺文件結構資訊，使得檢索技術相對困難[註五]。如交易記錄、晤談記錄、裝備故障維修記錄或青年日報的新聞標題等。

㈡結構化資料(structured data)：所謂的結構化資料是依據資料綱要，分門別類、按部就班被建置於資料庫，透過資料庫內建的檢索功能，使用者可以檢索相關的資料[註六]。如產品資料、廠商資料、人員基本資料、軍中的飛行序列表、每日操課課表或是個人電子兵籍資料等。

無論結構化或是非結構化的資料，在一組織或單位中，日以繼夜的紀錄並運用各種不同的媒體予以儲存，例如紙張、數據、聲音、影像等等，資料膨脹與成長的速度在現今資訊爆炸的時代，更是驚人。在組織或單位中，如何從原始資料(raw data)轉換成可用知識(knowledge)的重要過程，一般稱知識發現和資料採礦(knowledge discovery and data mining, KDD)，如圖一所示。

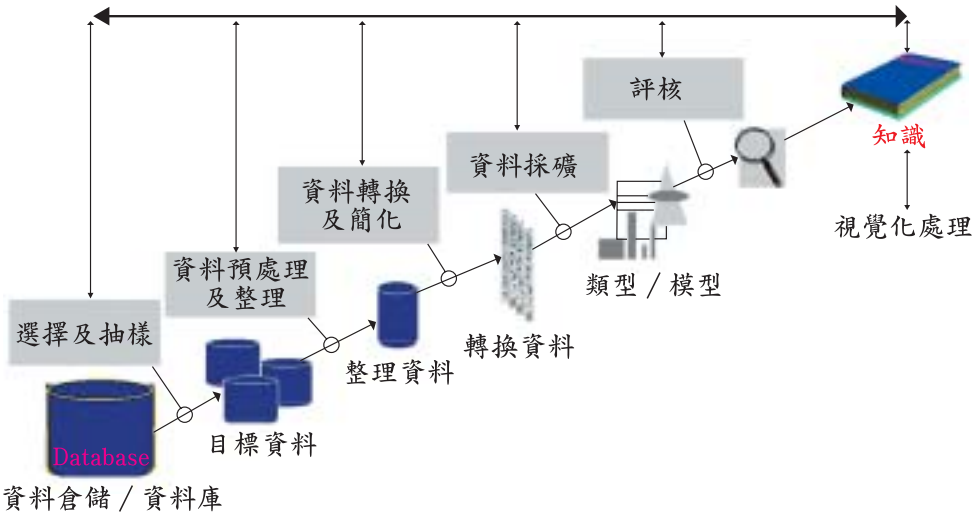
Fayyad & Stolorz (1997)認爲資料採礦程序包含六個步驟（如圖一）[註七]，以下則依此探勘步驟將之前美軍應用的實例逐步說明：

1.資料選擇與抽樣

資料庫中所累積的儲存資料極爲龐大且繁雜，必須要從巨量的資料中選擇需要的目標資料，且若是資料量的處理能力超過系統負荷，或是爲了節約時間與資源，抽樣是一項必要的工作。例如美軍之前在沙漠的某空軍基地中，曾經發生過數起F-16戰機的意外事件，當時美軍則以此事件爲分析標的，希望找出相關的因素，所以選擇包含了維修記錄、週期檢查、定期檢查等等的資料作爲分析的內容，這也是資料採礦的第一步。

2.資料預處理

資料庫中儲存的資料是日積月累所加總的結果，但是在這個過程中，常常會因爲資料輸入者的不經意或刻意而造成資料的錯誤或遺漏，產生了空白(blank)或空值(null)這種的遺漏值，故需要先將資料作預處理的工作。美軍在此步驟則先由人工方式



圖一　知識發現之流程

資料來源：Fayyad, U.M., Piatetsky - Shapiro, G., Smyth, P., & Uthurusamy, R.(Eds.), Advances in knowledge discovery and data mining(Cambridge, MA: The MIT press, 1996), p.13.

註五 林信成，「智慧型文件與智慧型系統整合之研究」，教育資料與圖書館學，第40卷，第4期，民92年6月，頁483。

註六 陳光華、呂明香(2003)，「知識探索及其於政府資訊之應用」，檔案季刊（臺北），第2卷第2期，2003年6月，頁4。

註七 Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R.(Eds.), Advances in knowledge discovery and data mining(Cambridge, MA: The MIT press, 1996), p.13.

預檢記錄中的電子記錄，並且委由工程師將資料會出成分析用檔案，檢視資料中的資料品質，剔除或修正記錄中缺漏或錯誤的部分。

3. 資料轉換

資料庫的格式是以儲存資料符合使用需求所設計，其未必符合資料採礦分析時的資料格式，所以適切的轉換能夠讓後續處理程序更順暢。如同一般企業的，對於美軍來說，資料轉換也是一項即爲頭痛的工作，因爲理行得保養、維修及檢查記錄，並不能夠符合電腦演算法分析的資料格式，必須先將資料轉換成適切的分析格式。

4. 資料採礦

依據我們要分析的目標，選擇適當的演算法以及來源資料，建立資料的模型，讓這個模型能夠代表所要尋找的目標或問題。在這個案例中，美軍利用了決策樹演算法以及類神經網路演算法等作爲分析的模型。分析時，先將資料分爲訓練組(training sets)及測試組(testing sets)。由訓練組中發展出模型，並在測試組中驗證模型效益。

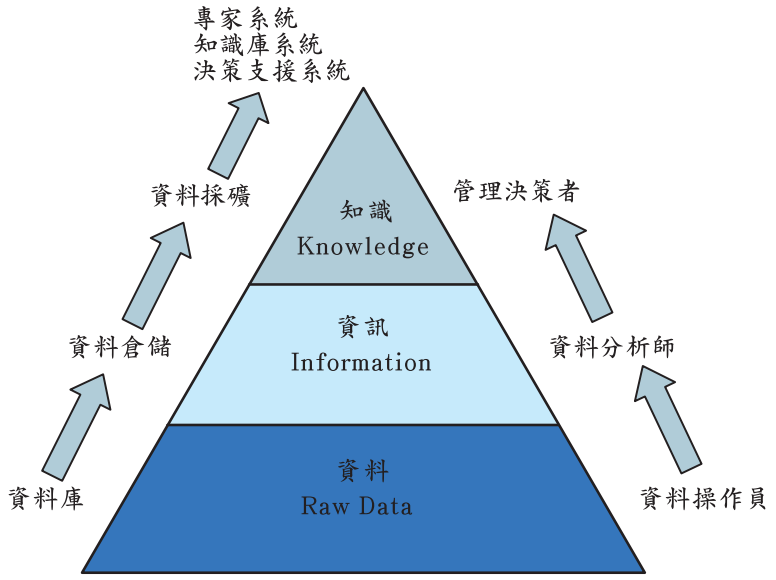
5. 評估效益

建立了模型後，必須要評估模型的效益，大於所設定標準值或門檻值的模型才算是一個有效的模型，否則必須另外再建立適當的模型。美軍藉由兩組資料的驗證，建立一個符合期望值的有效演算模型。

6. 結果解釋與應用

既然已經確認模型的效益符合需求，那麼就必須將模型予以實例化，並且把模型傳達予使用者理解，再將其應用到實際狀況中，讓資料轉化成資訊，並從中抽取關鍵的知識來運用。此一案例中，美軍透過模型的建立，找出F-16意外頻傳的因素肇因於所申請潤滑油的係數錯誤而造成，而更正後狀況不再出現，讓知識能夠有效被利用，發揮資料採礦的價值與經濟效益。

此外，資料庫系統中存放的是未經整理的原始資料，資料倉儲系統內則是經過整理大量現況與歷史資料所得的資訊，而資料採礦系統所儲存的則是經過整理大量資訊所得到的知識。資料、資訊和知識這三者間的關係如圖二這個金字塔所示[註八]。在資料管理這個階層，負責的人員是低階的資料操作員，使用的工具是資料庫系統；在資訊管理這個階層，負責的人員是中階的資料分析師，使用的工具是資料倉儲系統；而在知識管理這個階層，負責的人員是高階的管理決策者，所使用的工具是資料採礦系統。由金字塔的結構可以看出，越往頂端，管理的資



圖二　資訊管理的金字塔

資料來源：曾憲雄、蔡秀滿、蘇東興、曾秋蓉、王慶堯，資料探勘（臺北市：旗標出版社，2005年），頁1-8。

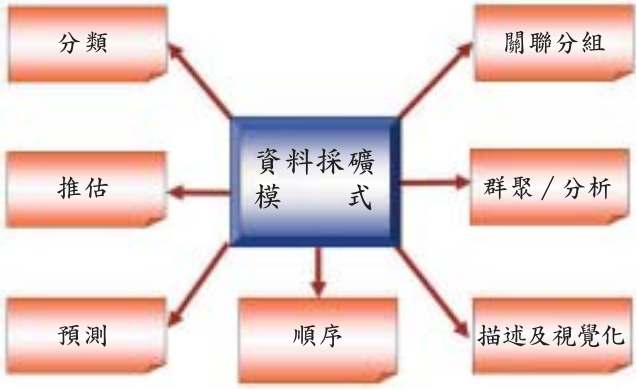
註八 曾憲雄、蔡秀滿、蘇東興、曾秋蓉、王慶堯，資料探勘（臺北市：旗標出版社，2005年），頁1-8。

料量越少，負責的人員也越少，與一般企業體的人事結構有相似之處[註九]。

二、資料採礦的涵意及模式

在資料採礦的範疇中，Liao & Chen (2004)說明了資料採礦乃是爲近期遽速成長的新研究領域，此一範疇，乃是爲了解決自巨量的資料中，分析、淬取其中的知識，有些甚至還能夠提供視覺化的決策功能以協助使用者[註十]。MIT於2002出版的「Technology Review」中提到改變未來的十項新興趨勢中，第三名即爲「資料採礦」的技術[註十一]；因此，在近幾年，知識發現與資料採礦已經成爲一個快速成長的跨學科領域，並結合如資料庫、統計方法、機械學習及相關的領域，以從大量的資料中挖掘出有用的知識(Kawano, Huynh, Ryoke & Nakamori, 2005)[註十二]。Fayyad & Stolorz(1997)等將KDD定義爲從資料中找出有效的(valid)、新穎的(novel)、潛藏有用的(potentially useful)以及最終能被了解的(understandable)類型，爲一連串重要(nontrivial)的程序[註十三]。Keim, Pansea, Sipsa & Northb(2004)認爲資料採礦即爲從所觀察的資料中，淬取吾人感興趣的類型 (pattern)或模型(model)[註十四]。

在資料採礦的領域中，包含了許多的模式(model)（如分類、推估、預測、群聚/分析、同質分組或關聯規則、描述及視覺化、順序等七種）(如圖三)及應用的方法(method)(如關聯性法則、時間序列分析、序列類型、群組式法則、分類式法則、機率經驗分析等6種）[註十五]。



圖三　資料採礦的模式

資料來源：本研究整理。

㈠分類(classification)

根據不同團體的物件特性建立屬性變數，當新物件進來時，可以前述的屬性加以判定並分類。例如當不明飛行物體出現於雷達螢幕時，即可利用已建立之屬性加以判定並分類爲轟炸機或是運輸機。常使用的技巧包括有決策樹(decision tree)或類神經網路(neural network)等。

㈡推估(estimation)

分類出來的結果是不連續的，而推估所得的結果則是連續性的數值。例如以一個家庭擁有之汽車款式來推估該家庭的年收

註九　曾憲雄、蔡秀滿、蘇東興、曾秋蓉、王慶堯，資料探勘（臺北市：旗標出版社，2005年），頁1-9。

註十　Liao, S. H. and Chen, Y. J., "Mining customer knowledge for electronic catalog marketing." Expert Systems with Applications, Vol.27(2004), pp.521~532.

註十一　李佳珍，「關聯性法則應用於品牌聯盟與異業結盟之研究」，淡江大學管理科學研究所碩士論文（臺北），民國94年，頁13。

註十二　Kawano, S., Huynh, V.N., Ryoke, M. & Nakamori, Y., "A context-dependent knowledge model for evaluation of regional environment," Environmental Modeling & Software, Vol.20 (2005), pp.343~352.

註十三　Fayyad, U. M., Piatetsky, S. G. and Smyth, P., "The KDD Process for Extracting Useful Knowledge from Volumes of Data." Communications of the ACM, Vol.39, No.11 (1996), pp.1~29.

註十四　Keim, D. A., Pansea, C., Sipsa, M. & Northb, S. C., "Pixel based visual data mining of geo-spatial data," Computers & Graphics, Vol.28(2004), pp.327~344.

註十五　廖述賢，資訊管理（臺北市：雙葉書廊，2007年），頁254。

入。

㈢預測(prediction)

利用一或多種獨立變數來找出某個標準(criterion)或因變數的值就叫預測。使用的相關技術包括迴歸分析、時間序列分析(time series analysis)、類神經網路及案例庫推理(case-based reasoning)等。

㈣關聯分組(affinity grouping or association rule)

用以辨識資料間的關聯，而這些關聯通常以規則來表示。例如含有項目A和B的紀錄中，有60%也含有C和D。事件發生的百分比是關聯的支持度及可靠度，較常使用Apriori演算法來辨識事件間的關聯。例如美軍曾利用此一技術，發掘出因爲在F-16戰機上使用錯誤號數之潤滑油品，導致失事率遽增的始末。

㈤群聚/分析(clustering/segmentation)

將異質母體中區隔爲較具同質性之群組(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及agglomeration。

㈥描述及視覺化(description and visualization)

利用視覺化的方式，將分析以及淬取的結果呈現出來，以解釋複雜或繁瑣的內容。使用圖像的表達方式，對於使用者來說能夠更容易解釋及接受。

㈦順序(sequential modeling)

例如，假設在研究的結果中發現三五快砲實彈射擊時，未能換用適當之瓦斯噴嘴，則有25%的機率將會在500小時後影響瓦斯活塞桿的運動，同時有53%的機會將會造成送彈機構的故障，這樣的分析就是「順序」的研究結果。

不論哪一種資料採礦的模式，都是爲了讓知識能夠有效且完整的轉移，減少在世代遞嬗中因人爲或無意造成的損耗。

參、資料採礦與國家安全

恐怖主義長久以來一直是國際安全的潛在威脅，依據美國國務院(Department of State)的估計大約有十萬的恐怖份子或與恐怖主義接觸的名單中，蓋達組織訓練過約有7萬名的恐怖份子[註六]。恐怖團體通常應用航空器或機動車輛犯罪(motor vehicle violation)、移民詐欺(immigration fraud)、非法製造武器或爆裂物、武裝搶劫銀行或偷竊、走私大規模毀滅性武器(weapons of mass destruction)、跨國組織犯罪等[註七]。

根據美國國務院公布的報告，全球恐怖攻擊事件去年(2006)增加了25%，因恐怖攻擊不幸喪生的人數更飆高40%；主要原因是極端主義份子在伊拉克以化學武器和自殺炸彈攻擊民眾聚集的地點所造成。這份報告指出，伊朗仍是最大的支持恐怖主義國家，伊朗政府在背後支持中東地區各個組織，尤其是在伊拉克，他們對什葉派叛軍組織提供實質支援和指導方針，用來攻擊遜尼派武力、美軍以及伊拉克部隊。國務院指出，2006年大約發生1萬4,000起恐怖攻擊事件，主要在伊拉克與阿富汗，這些攻擊造成2萬多人喪生，其中三分之二喪命於伊拉克，和2005年相比，攻擊事件多了3,000多起，死亡人數多了5,800多人。總計全球因手段愈加致命的恐怖攻擊事件死亡人數，較前一年增加了40%。報告將傷亡增加的部分原因歸咎於以

註六 Paul Rosenzweig, "Civil Liberty and the Response to Terrorism", Duquesne University Law Review, Vol.42, Summer 2004, p676.

註七 莊坤龍，「印尼回教祈禱團恐怖組織之研析」，國防雜誌（桃園），第22卷第1期，民國96年2月，頁72。

大批聚集民眾爲目標的非車輛型自殺炸彈攻擊，這類型自殺攻擊事件大幅成長了25%，車輛自殺炸彈攻擊則下降了12%。報告並指出，2006年11月23日在伊拉克薩德市攻擊事件中，首次出現使用化學武器的狀況，顯示恐怖攻擊戰術向危險策略轉變。隨著死亡人數的攀升，2005年到2006年之間，在恐怖攻擊中受傷的人數也大幅了54%，在伊拉克受傷的人數更成長一倍。這些數字由美國國家反恐中心彙整，死傷者主要是「非戰鬥人員」，而被當做攻擊目標的孩童、教師和新聞工作者的死傷數目也明顯增加[註六]。

由於恐怖主義的蔓延，導致第三世界的回教國家醞釀已久的仇美情節不斷發酵，並在2001年9月11日的時候以恐怖行動的方式，向全世界展示其手段並宣告新的、更強烈的恐怖活動。2001年9月11日至今仍是個讓全世界都難以忘懷的慘痛經驗，其所帶來人員傷亡與情緒撼動至今仍難完全撫平。尤其是對於美國的政府與人民更刻骨銘心。遭受重大恐怖攻擊後，美國政府重新思考了「國土防衛」與「先發制人」等議題，爲了落實這些政策，除了大舉出兵企圖殲滅以賓拉登爲首的恐怖組織外，更積極的希望以資訊科技協助美國減少或是避免此一威脅。其中，資料採礦就是美國極度仰賴的技術之一。從九一一事件後，美國政府成立新的聯邦部門──美國國土安全部(US Department of Homeland Security, DHS)，其主要目的是負責美國國土的安全以及防範各項恐怖活動的發生。DHS整合了聯邦政府內22個與國防和情報相關的機構，人數達17萬餘人，每年花費國家預算超過400億美元，是一個相當龐大又重要的一個組織[註九]。

恐怖攻擊的方式與範圍在「九一一」事件之後，已逐漸由美國爲中心向全球的各個地區不斷擴散，例如2003年10月12日，渡假天堂峇里島，被一輛載有100多公斤爆裂物的箱型車，在滿是外國遊客的夜總會前引爆，造成187人喪生及逾300人受傷[註十]。2005年7月7日，4名英國回教徒以自殺炸彈方式，攻擊了倫敦的三個地鐵列車以及一輛雙層公車巴士，炸死了自己也造成另外52人喪生及700多人受傷[註二]。2005年11月10日，於約旦首都安曼的飯店遭到當地有史以來最慘烈的恐怖炸彈攻擊後，總共造成57人喪生及115人受傷[註三]。這些重大傷亡的恐怖事件，所付出的代價何其沈重！

由於恐怖主義蔓延，對於諸多國家來說是揮之不去的夢魘，但是防範的工作除事倍功半、效果有限外，其所耗費之資源、人力、時間等均所費不貲，所以若是無法藉助科技的協助，可能在恐怖行動來臨之前，就已經讓美國因爲資源耗盡且人心惶惶而自行宣告失敗。

美國在911事件之後，爲了要避免類似案件再發生並能提早察覺可疑事件，所以大量使用資料採礦的相關技術可以提供早期預警。工作內容之一包括透過多種通訊及資料傳輸媒體，採擷、分析特定的關鍵字或主

註六 雅虎新聞網，「反恐戰爭進行多年，全球恐怖攻擊去年飆升」，2007年6月28日，資料來源：http://tw.news.yahoo.com/article/url/d/a/070501/19/dngb.html

註九 美國國土安全部組織架構，資料來源：http://www.dhs.gov/xabout/structure/

註十 雅虎新聞網，「印尼成功處理峇里島爆炸案 方法不同於美國」，2006年9月7日，資料來源：http://tw.news.yahoo.com/article/url/d/a/060907/19/3b5j.html

註二 雅虎新聞網，「倫敦2年前連環爆炸3男子意圖謀殺罪成立」，2007年7月10日，資料來源：http://tw.news.yahoo.com/article/url/d/a/070709/1/h0zn.html

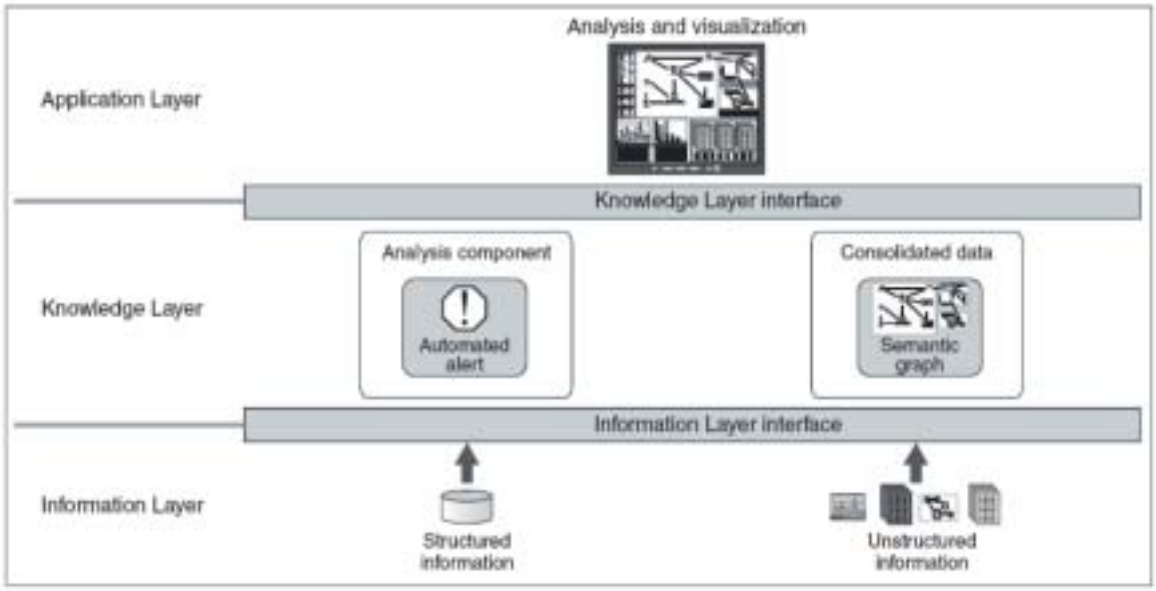
註三 雅虎新聞網，「約旦飯店爆炸增至23死學校政府機關明關閉」，2005年11月10日，資料來源：站 http://tw.news.yahoo.com/article/url/d/a/051109/19/3pp0.html

題，以從中發現相關可疑訊息。同時，在2002年的元月中旬正式成立「資料處理局(Information Awareness Office, IAO)」，其主要任務是利用如資料採礦之資訊科技來維護國家安全，並遏止在資訊不對稱(Information Asymmetry)[註三]下所可能發生的「不對稱威脅(Asymmetric Threats)」[註四]。此外並發展出一套資料採礦的工具──ADVISE──來監控或管制特定人員的各項流通資訊，包括電話、無線電、電子郵件、網路資料、傳輸數據、資金流動、銀行交易、行動等等的記錄資料，再將其分析、比對以供提出早期預警的警告訊息。而ADVISE表示美國目前運用資料採礦處理資料來源的過程，也就是分析、傳播、視覺化、洞察以及語意增強(Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement)[註五]。

現在由於對資料分析與知識萃取的日漸重視，因此美國國土安全部也在其報告中定義出其知識萃取與應用的流程（如圖四）。在分析的流程可以分爲三個層級，分別是資訊層(Information Layer)、知識層(Knowledge Layer)與應用層(Application Layer)等三級[註六]。

一、資訊層(Information Layer)

資訊層是將資料導入的第一步驟，將各式各樣不同種類的資料從四面八方匯集後進入分析工具之中。資訊層的階段把資料分爲



圖四　DHS（美國國土安全部）的知識萃取流程

資料來源：DHS,"Data Mining Report Psychology and Aging", DHS Privacy Office Response to House Report(2006, July), p.12.

註三　資訊不對稱下的誘因理論大致內容是指，資訊較少的一方如何設計一套誘因制度克服資訊較少的劣勢，誘使資訊較多的一方透露出其所擁有的資訊，或誘使資訊較多的一方行爲符合資訊較少一方的要求。維基百科，「資訊不對稱」，資料來源：http://wiki.planetoid.info/index.php/Information_Asymmetry

註四　九一一事件發生後，美國受到重創，恐怖組織以區區十餘人的攻擊行動，卻達到傳統戰爭難以企及的效果，使得美國在司法、情報、國防、經濟、財政、運輸、資訊等各部門遭到重大衝擊而必須進行組織改造與任務調整。這一切都說明了「不對稱威脅」不再是假設性的問題，而是已經存在的事實。「九一一事件美國聯邦政府反恐怖作爲」，資料來源：http://www.vghtpe.gov.tw/~ged/left4/left4_04_05.htm

註五　DHS, "Early Attention to Privacy in Developing a Key DHS Program Could Reduce Risk", DHS Privacy Office Response to House Report (2007, Feb), p.2.

註六　DHS, "Early Attention to Privacy in Developing a Key DHS Program Could Reduce Risk", DHS Privacy Office Response to House Report (2007, Feb), p.14.

結構化（例如在資料庫中的資訊或是電腦報表所列出的資料）與非結構化（例如e-mail的內容、文本報告或是新聞標題等等）等兩種。

結構化資料在軟體的應用上較爲簡易與便利，因爲已經藉由固定格式設定與存取，將資料由儲存媒體（如資料庫）中取出，交由分析軟體ADVISE進行解析，並在資訊層中將資料修正至格式一致，就如我們經常聽到的「知識本體(ontology)」。知識本體是一個正式明確的規格，旨在說明大家都能共同接受的概念，但是概念仍須澄清後才能成爲知識。因此，知識本體是特定領域中的概念描述，包含此特定領域的重要基本概念及彼此間的關係。知識本體不僅定義出特定領域中的重要概念，亦可呈現出概念之間的關係，包括垂直的階層關係、水平對等的關係、及群組的相依關係等[註七]。所以知識本體能夠協助我們定義實體（例如人、位置、組織或事件）、屬性（例如姓名或是住址），以及它們之間的關聯。

相對於結構化資料來說，非結構化的資料則是較難處理的區塊。因爲非結構化的資料，是以常態的原貌儲存於不同的媒體中，所以資料格式紊亂且不一致，因此在資料分析之前的資料洗刷或是資料過濾(data clean)的工程常常會耗費分析者80%的精力。所以對於非結構化的資料，ADVISE包含了許多工具以便萃取關於實體及屬性間的資訊。類似於結構化的資料，這些分析之後的結果是根據知識本體而產出固定格式與結構化的資料以利後續的分析及處理。ADVISE允許分析者以手動的方式去識別那些資料中的實體、屬性及其關聯。也就是說，分析人員並非僅僅是軟體的使用者而已，還必須包含許多專業的領域知識(domain know how)才能適時調整相關的決策變數與門檻值，讓重要的資訊浮出檯面。目前DHS仍在持續並積極的發展更有效率與效能的辨識方法，以輸入不同格式的非結構化資料[註八]。

二、知識層(Knowledge Layer)

知識層，是指接收大量來自資訊層的實體語意圖像(semantic graphing)，並藉由各種實體的線條來串連彼此，達到圖像與實體意義相符。圖四則表示一般典型的語意圖像實體連結圖，這一部分的處理是藉由本體論的組織而產生報告。此外，當新進入的資料符合特定的預設類型(patterns of interest)時，知識層亦能夠自動警告分析人員。此外知識層中來自資訊層的事實和關聯已經強化到大規模語義的圖和各種各樣的子圖形(sub graphs)。語意圖是一種資料建模的技術，使用節點來表示特定的實體，以及使用連接線代表其間的關係，因其能夠適當的表示資料間的關係、連結，圖四表示一種典型的語義圖形。

三、應用層(Application Layer)

應用層藉由知識層的介面，能夠運用知識層處理過的資料，並視需求自知識層中擷取數據資料交叉使用，再把萃取的知識轉換成圖像或是表格這種的格式，將知識分析化與視覺化，協助決策者能更易做出最佳判斷。應用層提供分析人員兩種彈性的資料查詢功能，分別是分析人員執行以圖形爲主或是以主題爲主的查詢，並且搜尋符合特定類型的資料後可以連結到特定的實體。例如，分析人員能尋找已經在選定的一段期間內旅行

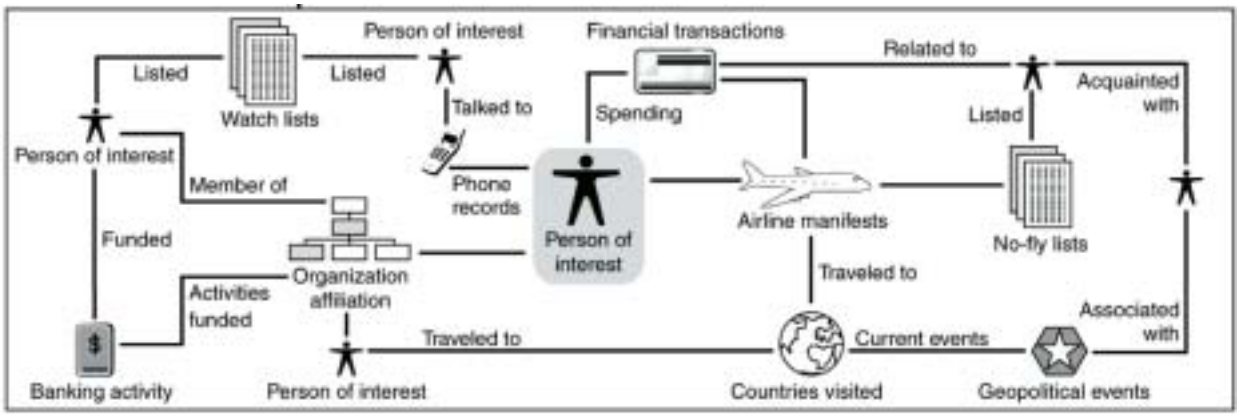
註七 黃保瑞，「符合SCORM之網頁教材設計──以電腦的資料處理爲例」，逢甲大學資訊工程學系碩士論文（臺中），民國92年6月，頁39。

註八 DHS, "Early Attention to Privacy in Developing a Key DHS Program Could Reduce Risk", DHS Privacy Office Response to House Report (2007, Feb), p.26.

到某站的所有個人(individual)，或者搜尋關於特定人員、位置或者機構的全部訊息。搜尋的結果可以藉由語意圖來呈現(如圖五)。

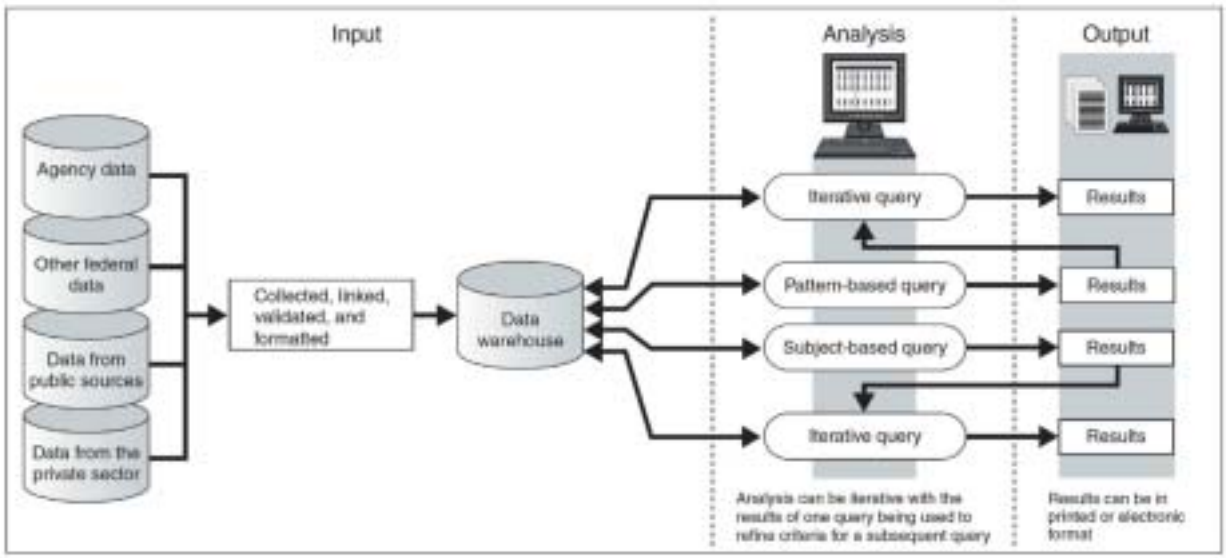
一個分析人員能夠從語義圖中準確判讀重要的特徵節點究竟是在哪一個環節，同時與這個節點相關的重要的附加訊息全部都可以導入ADVICE中分析處理，並由分析的過程中瞭解資料來源的可靠度、可信水準以及是否涉及美國人民的議題等等。而ADVISE應用層提供了分析人員能夠監控資料中特定類型圖案的相關能力。這些圖案一直持續在資料庫內被監控，並且於符合類型時主動提供警告。例如，一個成熟的分析人員能選擇適當的決策變數，同時定義特定的類型，例如「在過去6個月內從美國到中東去旅遊的所有個人(Individual)」，此外ADVICE還能夠在當這些特定類型出現時主動發出警告。

DHS的資料採礦流程很簡單的分爲「輸入」、「分析」與「輸出」等三個步驟（如圖六）。在「資料輸入」階段，資料儲存在資料倉儲的架構中，以利於資料採礦的分析與使用。在資料分析階段中，資料以查詢與



圖五　典型語意圖的格式

資料來源：DHS, "Early Attention to Privacy in Developing a Key DHS Program Could Reduce Risk", DHS Privacy Office Response to House Report (2007, Feb), p.2.



圖六　Vipin K. & Mohammed J. Z.的資料採礦流程

資料來源：DHS, "Early Attention to Privacy in Developing a Key DHS Program Could Reduce Risk", DHS Privacy Office Response to House Report (2007, Feb), p.5.

比較的方式，尋找我們所關注的項目。查詢的方式分爲兩種，分別是類型查詢(pattern-based queries)以及主題查詢(subject-based queries)[註元]。

類型查詢：搜尋的原則是依據預先設定的類型去比對符合的內容（例如在保險契約中特定而又稀少的事件）。

主題查詢：搜尋的原則是依據預先決定的主題，並且使用具體可識別的符號，去比對已獲得的任何資訊（如人名或是身分證號碼）或是一種對於某些社群特定的符號（如美國海軍使用資料採礦在其艦艇故障原因的鑑定）。

最後在「知識輸出」階段能夠產生書面或是電子格式的檔案，並將這些報告讓他人共享。

肆、資料採礦的應用

由於資訊科技的進展快速，造成資料量的成長速度，已遠遠大幅超過往日我們所能夠處理與分析的速度及能力。相同的一分恐怖行動威脅警告，在攻擊發起前與攻擊發後起被解譯出來，產生的效果將大不相同。相對於我國，美國的公共事務部門（包含政治、軍事、經濟等等）爲了處理與分析大量的資料而使用此一技術，對於資料採礦的應用與發展遠遠超過民間的範圍與能力，尤其是面對來自第三世界回教國家的恐怖行動威脅。在DHS於2004年5月的報告中，已經有52個相關機構已經正在使用或已經計畫將資料採礦應用在199個專案中；其中有68個專案已經在規劃中，而有131個專案已經在使用[註千]。此外，美國在經歷了在2001年9月11日的恐怖攻擊之後，資料採礦技術的推廣更是不遺餘力，藉由蒐集與分析許多公開或是私人的資料來協助偵測恐怖威脅。這些包含追蹤恐怖活動，涵蓋了金錢的移轉、通訊，以及追蹤恐怖主義者的移民或旅遊紀錄。根據2006年8月DHS督察長室對資料採礦的初步研究報告，DHS已經使用或發展了12個資料採礦的計畫，其中的9個已經完全可以運轉並服役中，另外的3個計畫則還在繼續開發中[註三]。

相對於美國及澳大利亞等國家已經開始廣泛的運用資料採礦技術於國家安全的相關事務上，我國對於資料採礦仍尚處於發展中期，多應用於企業、商業或是工業等方面，而在國防方面的應用，目前多侷限於軍事預算[註三]、軍事訓練[註三]、後勤補保[註亖]、人力資源管理[註五]、軍事教育[註六]……等等的範圍。

註元 DHS, "Early Attention to Privacy in Developing a Key DHS Program Could Reduce Risk", DHS Privacy Office Response to House Report(2007, Feb), p.17.

註千 DHS, "Early Attention to Privacy in Developing a Key DHS Program Could Reduce Risk", DHS Privacy Office Response to House Report(2007, Feb), p.5.

註三 DHS, "Early Attention to Privacy in Developing a Key DHS Program Could Reduce Risk", DHS Privacy Office Response to House Report(2007, Feb), p.6.

註三 高克志，「資料探勘運用於國防預算規劃及績效衡量之研究」，國防大學管理學院資源管理研究所碩士論文（臺北），民國95年6月。

註三 陳仲禮，「資料探勘應用於空軍軍官訓練成效評估之研究」，國防大學管理學院資源管理研究所碩士論文（臺北），民國95年6月。

註亖 盧奎龍，「運用資料探勘技術建議後勤維保單位物料管理模式之研究」，私立大葉大學工業工程與科技管理學系碩士論文（彰化），民國96年6月。

註五 蔡志雄，「資料挖掘技術在國防人力資源管理之研究」，國防大學管理學院國防資訊研究所碩士論文（臺北），民國91年6月。

註六 陳慧生，「資料採礦模型應用於國防管理學院基礎教育評鑑之研究」，國防大學管理學院國防決策科學研究所碩士論文（臺北），民國93年6月。

以目前資料採礦發展成熟的幾個領域應用在國軍可帶來的效益，在下列簡述：

作戰情報：美軍現已利用資料採礦來提高雷達的辨識率，減少人員誤判或是漏失的狀況，這僅是資料採礦運用的一部分。此外，長期蒐集重要的情報資料，或是累積長期的雷達搜索記錄或是電信監聽記錄，透過語意探勘(Semantic mining)或是文本探勘(Text mining)，則可由其中挖掘出有價值的資訊。

資安防護：目前中共的「網軍」對我不論是政府機關或國防單位，均以各種不同的方式入侵，對我形成資安威脅甚至產生資訊缺口。利用資料採礦的技術，可於網軍入侵時，主動依入侵類型或方式，調節資訊系統資源，避免遭致過量資訊使系統資源耗盡而當機。亦可將入侵的紀錄予以聚類或分群，將知識以視覺化的方式提供予決策單位參考使用。

政戰心輔：我們可以預先將人員的資料及屬性建立資料庫，並且產生適切的模型，待新進人員到部時，即可依模型預判該員的生、心理狀況，及早並適時做好準備，除了可以提供人員適切的生理及心輔需求外，更能夠大幅降低新進人員不適應的情況。業界中普遍利用此技術於顧客關係管理(Customer Relationship Management, CRM)方面，目前發展已相當成熟。

財務支用：可以利用資料採礦來稽核單位金錢支用狀況，避免不當支用，同時主動提出警告。目前金融業利用資料採礦於信用卡盜刷的詐欺偵測(Fraud Detection)及預防已有顯著功效。

伍、結　　語

資料採礦能夠將資料轉換成資訊並且淬取成知識。知識永遠有助於我們戰備整備的遂行，了解敵人在何處，及如何遂行攻擊，方可決定是否或在何處接戰，以獲得最佳戰果[註一]。美國現正積極的發展並應用資料採礦技術於國土安全之上，而澳洲政府亦於日前採購並啓用與美國類似的軟體加入國家安全防護的一環[註二]。我國資料採礦在各行各業的應用仍在發展階段，身處在高科技時代的國軍，自然不能置身事外，善加利用科技的力量，可以達到事半功倍的效益。尤其是在這重要的時期，發揮資料採礦的技術並應用於國土防衛及國家安全等方向，將可對我國未來的反恐制變相關作爲，增加反應時間及緩衝，減少我國因面臨恐怖攻擊而造成的重大政治、經濟、軍事與心理等損失。

收件：96年08月27日
修正：96年09月10日
接受：96年09月14日

作者簡介

溫志皓上尉，空軍官校88年班、空軍作戰參謀軍官班96年班、國防管理學院資源管理研究所94年班；現任國軍防空砲兵訓練中心裁判官。

註一 馬丁李比奇原著，張天虹譯，掌握明日戰爭（臺北，國防部史政編譯局，民國90年2月），頁33。

註二 公共電視新聞網，澳洲反恐比對可疑人物資料，新系統九月上路」，2007年7月9日，資料來源：站 http://tw.pts.org.tw/php/news/new_