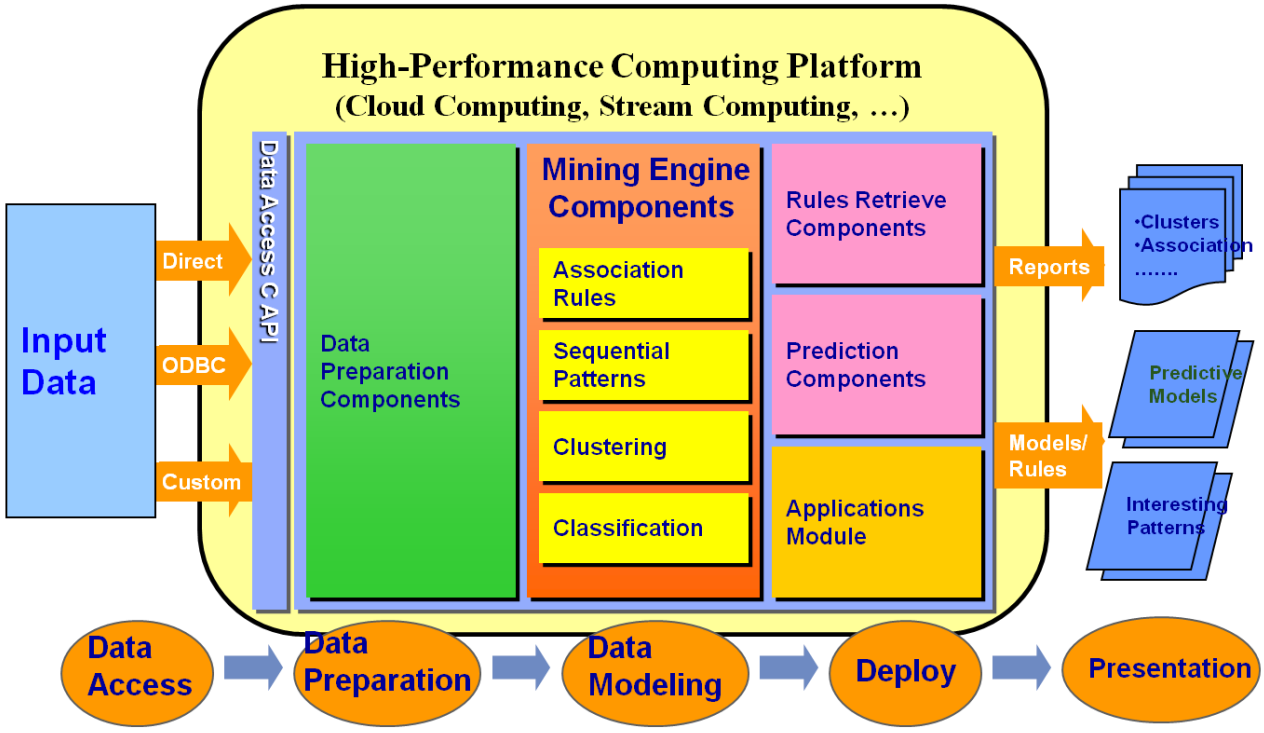
雲報專欄：巨量資料分析前處理程序之關鍵議題及挑戰

—國立交通大學資訊工程系曾新穆教授/技術專家委員會委員

巨量資料(Big Data)具有 3V's 的特性，亦即在 Volume, Variety, Velocity 等面向上均極為巨大，而更重要的是要能從中產生 4th V，亦即 Value (價值)，而此即有賴於巨量資料分析(Big Data Analytics)技術。巨量資料分析之程序涵括前處理、特徵篩選、學習及模型化，以至後處理等，近年來雖已有許多巨量資料分析工具被提出，多數企業或研究者欲導入巨量資料專案時多將注意力集中於各種分析技術，但對於分析程序中的前處理部分常忽視了其重要性。本專欄探討巨量資料分析中前處理程序中的幾項關鍵議題及挑戰，提供作為在現今巨量資料時代下要由巨量資料中挖掘出金礦並產生高度產業價值之參考。

巨量資料分析(Big Data Analytics)為巨量資料應用中之關鍵環節，打通此環節方能由巨大之資料中挖掘出有價值之金礦。巨量資料分析之程序如圖一所示，包含了對於輸入資料(包含結構化及非結構化資料)之前處理(Pre-processing)、特徵篩選(Feature Selection)、學習及模型化(Learning & Modeling)以至後處理(Post-processing)等，並加入雲端計算(Cloud Computing)及串流運算(Stream Computing)等高效能計算技術，以達成能處理巨量資料之大量、高變項度、高流量等複雜特性。

近年來已有許多巨量資料技術工具及平台被發展出，多數企業或研究者欲導入巨量資料專案時多將注意力集中於各種分析技術之運用，但對於前處理程序常忽略了其中的許多關鍵要項，以致未能達成預期成效。事實上，所謂 Garbage in, garbage out，巨量資料之前處理至關重要，巨量資料由於其複雜之特性，前處理程序較一般資料更要困難許多，除了習知之資料遺漏(Missing Value)等資料品質問題外，還多出了許多新議題及挑戰。在此探討剖析幾項關鍵性之議題及挑戰：



圖一、巨量資料分析架構

一、資料稀疏性問題：由於巨量資料之極大量性(High Volume)及高變項度(High Variety)等特性，資料中常存有嚴重之稀疏性問題，為資料分析上之一大挑戰。舉例而言，美國之串流媒體服務巨擘 Netflix 公司在多年前即致力於運用巨量資料探勘技術發展個人化推薦服務，並舉辦獎金達 100 萬美元之 Netflix Prize 競賽。基本上，其欲達成有效之個人化推薦之關鍵在於如何利用其客戶對影片之訂閱瀏覽及評等(rating)等大量記錄來學習建立出有效之模型，以精準預測客戶對影片之喜好度。事實上，此種推薦應用在學界及業界已研究多時，並已發展出類如協同過濾(Collaborative Filtering)等有用之技術。然在類如 Netflix 所具之巨量資料環境下，其客戶數高達數千萬人，而影片數亦達上百萬部，以致其有效可運用之客戶-影片相關聯之訂閱瀏覽及評等資料變得非常稀疏。因此，若套用一般之協同過濾方法將無法產生有效之模型，而必須針對此種資料稀疏特性先加以處理進而設計合適之分析方法。

二、資料真實性問題：巨量資料中因混雜著各式各樣的結構性與非結構性資料，常存有資料真實性(Veracity)之問題。舉例而言，Google 在 2006 於 Nature

上發表了利用分析其龐大的 User Search Log 所發展之 Flu Trend 服務，可準確預測流感之爆發，並且能較美國之 CDC(疾病管制局)提早一周以前即預測出。然而，此預測功能後來被學者發現在其發表數年之後已不復準確，甚至有高度之誤差。分析其原因為原本此流感預測模型之原理乃基於人們於罹患流感之不同階段會查詢不同之流感相關資訊(如症狀、治療方式等)之現象；然而在 Google 推出此功能後引起許多研究者之好奇，紛而於 Google 之搜尋引擎中輸入與流感相關之關鍵字欲查詢了解此 Flu Trend 功能，結果這些查詢字詞混雜於所有之 Search Log 中，被錯認為流感病人之查詢，而導致錯誤預測流感之發生趨勢。此種資料偏差問題在許多巨量資料應用中極為常見，由此可見於資料前處理時即需鑑別資料真實性之重要。

三、關鍵特徵挖掘之挑戰：特徵篩選(Feature Selection)向為資料探勘分析之極重要之一環，而巨量資料因其高變項度(High Variety)特性而更顯重要。而為能達成資料分析及應用之即時性，更需挖掘出其關鍵特徵(Key Features)。2012 年由 Nokia 所舉辦之全球性 Mobile Data Challenge 競賽中，蒐集了數百位歐洲地區智慧型手機使用者長達一年以上之手機使用紀錄，經匿名化之後以預測使用者之個人特質及行為作為競賽主題。筆者所率領之研究團隊有幸於本競賽中榮獲第二名之佳績，以此競賽中其中一項主題—預測使用者之性別為例，此資料所包含之使用者特徵多達數萬項，而我們經過完整之資料探勘程序建立出之模型其預測準確度可高達 95%。此結果看似相當好，應具有可運用到類如個人化行動商務等應用之高度潛力。然而，若我們要將此種模型嵌入智慧型手機之應用中，則需同時考慮到行動載具之記憶體、耗電量等限制，以及運算之即時性等問題，因此過多之特徵計算將不可行，而必須能挖掘出關鍵特徵加以利用。而此議題中我們最後即由上萬種特徵中找出可判定性別之關鍵特徵—加速計(accelerometer)(其主要原因為多數男女生在行走時置放手機之位置有明顯之差異，因而加速計之震動型態呈現對應明顯不同之型態)，如此將更能符合實際應用情境之需求。

綜合上述，資料前處理實為巨量資料分析專案成敗之最重要關鍵之一，現今各企業單位於導入巨量資料專案時實應慎重待之。基本上在此部分除運用資料品質分析工具做基本檢視外，可由資料科學家與領域專家就領域特性參考上述於資料稀疏性、資料真實性、關鍵特徵發掘等面向逐一檢視，以完成深度之資料前處理程序，將更能確保巨量資料分析專案之成功度。