論述 》專論 · 評述

漫談大數據分析的是是非非

「大數據」到底是什麼？它到底可以給些什麼？傳統資訊系統技術與平台是否不再重要？現行資料蒐集與抽樣方法是否都應被淘汰？統計等量化分析方法是否已經落伍？... 等等問題，總會不斷地被提出。本文將針對在執行大數據專案時的實際狀況及其衍生的問題，由統計分析人的角度切入來聊聊「大數據」的平行運算與無抽樣推論的是是與非非。

梁德馨（輔仁大學統計資訊學系教授、台灣析數資訊股份有限公司整合策略長）

壹、前言

「大數據」（Big Data）是現今全球最熱門的科技主題之一。近二年來，政府機關、私人企業、研究或學術組織在其主要的策略或發展計畫中，若不能與「大數據」連上一些關係，似乎就無法與時俱進。

何謂「大數據」？麥塔集團（META Group；現爲高德納，The Gartner Group）分析員道格·萊尼（Doug Laney）在 2001 年指出，資料增長的挑戰和機遇有三個方向：量（Volume，資料大小）、速（Velocity，資料輸入輸出的速度）與多變（Variety，多樣性），合稱「3V」或「3Vs」。隨著資訊科技不斷地往前推進，資料量的複雜程度愈來愈高，「3Vs」已經不足以形容新時代的大數據，因此在 2012 年時，科技大廠 IBM、國際調查機構高德納（The Gartner Group）、IDC 等也紛紛在 3Vs 之外，再增加「準確性」（Veracity）的特色，而形成 4Vs。近 2 年來，「可視性」（Visualization）與「合法性」（Validity）等又被提出而形成「大數據」的 6Vs。此外，「大數據」最重要的是在這 6V 之上的第 7 個 V～價值（Value），代表著前述「大數據」的 6V，都必項以策略來指引分析，並以創造價值爲依歸。

「大數據」在上述 7 個 V 的基本觀念下，「大數據」到底是什麼？它到底可以給些什麼？傳統的資訊系統技術與平台是否不再重要？現行的資料蒐集（例如：電話訪問）與抽樣方法是否都應被淘汰？統計

及作業研究等量化分析方法是否已經落伍？運用組織內部資料庫做決策是否不如採用外部資料來得有價值？...等等問題，總會不斷地被提出。似乎在計畫內容中若無一套與過去不同的創新平台、技術與方法，就不足以稱其爲「大數據」的專案。本文將針對政府或企業在執行大數據專案時的實際狀況及其衍生的問題，由統計分析人的角度切入來聊聊「大數據」的平行運算與無抽樣推論的是是與非非。

貳、大數據的是是非非

一、量即是美

「數大便是美，...數大了似乎按照著一種自然律，自然的會有一種特別的排列，一種特別的節奏，一種特殊的式樣，激動我們審美的本能，激發我們審美的情緒。...西湖的蘆荻，與花塢的竹林，也無非是一種數大的美，不是智力可以分析的，至少不是我的智力所能分析。」早在90年前，徐志摩就以感性的眼看到這個世界的「大數據」之美了，但也感嘆人的智力對「大數據」分析之侷限。「大數據」不是現在才產生的或存在，但在資訊科技跳躍進步下，加速了各種型式資料的產生，也活化了被儲存而無法被運用的資訊。

無量不大，「量大」（Big Volume）似乎是「大數據」必要的條件。根據維基百科的定義，「大數據」的資料量從幾TB（Terabyte）到幾PB（Petabyte）不等，到目前爲止，沒有一個準確的標準來界定大數據的大小。依這樣的定義，若僅是依資料「靜態」的儲存量來衡量，那麼，以台灣的人口規模，絕大部分的政府機關、企業組織、學術單位所處理及應用的資料量都稱不上「大」。因此，在實務上，常會看到當長官交辦要建置大數據平台，而承辦同仁苦思不得「足量夠大」資料來展現「大數據」平台平行運算的效率之美。

在許多資訊業者的強力宣傳研討活動中，誇大了Hadoop等平行運算在某些「大數據」案例中的成功事蹟，在這些熱門名詞及漂亮少數成功案例的重覆催眠下，許多組織做了不見得有其必要性的投入而建置了「大數據」平台。其實許多統計演算方法是不適合以分散式的平行運算方法來處理的，而絕大部分的組織雖擁有不小的資料量，但仍不足以大到需採用這些「大數據」的資訊平台。IBM華生研究中心System G團隊中的首席研究員林清詠就曾提出，「根據他帶領的IBM研究團隊採用Hadoop的經驗，最後常常因爲運算效能不夠好而浪費時間，由於Hadoop架構需要3倍的儲存空間，企業在採用時，常常會提出硬碟成本太高的問題。即使在美國，目前已經擁有大量資料的企業其實不多，但很多企業要導入大資料專案時，會盲目的採用如Hadoop這樣的大資料平臺。有些企業會在多臺機器上部署Hadoop，但可能每一臺機器都只使用了百分之二十的效能，

有高達 9 成 5 的企業，在採用 Hadoop 之後發現，其實根本使用不到。」[1]

諸如 Hadoop 此類平行運算設計，在分散分派資料時是需要有處理的工，在資料量不是極大的狀況下，分散進行演算再合併，最後整體的效率常不如一般的伺服器平台。本人與研究團隊曾針對 Hadoop 與一般 Server 進行效率測試比較，在資料筆數爲數千萬筆的資料排序演算中，發現二者間幾乎無差異，若將資料擴增到數十億筆資料時，才可測出 Hadoop 其較優化運算效能，但亦僅在 3 至 5 分鐘之間[2]。然而，Hadoop 架構需要 3 倍的儲存空間，組織在採用時，常常會提出硬碟成本太高的問題。

二、雜及速才是困

「大數據」的量大不見得是以儲存量來衡量，由資料分析的觀點來看，即使靜態的資料量並不算極大，但若其分析方法涉及複雜的反覆演算，那麼在進行數據分析時，那麼也必須要有足夠的儲存空間及隨機存取記憶體（Random Access Memory，RAM）才能解決演算所需大量記憶體及儲存空間的問題。由加州大學柏克萊分校 AMP Lab 所開發的 Apache Spark 開源叢集運算框架，採用了記憶體內運算技術（In-memory），由於可以用較少的節點數量，達到比 Hadoop 的 MapReduce 還高的執行效能。Spark 允許用戶將資料加載至叢集記憶體，並多次對其進行查詢，非常適合用於機器學習演算法。在這一、兩年內快速竄起，變得非常受歡迎，也較適合用於統計分析及圖像解析等運算。

其實，只要是傳統方式無法處理資料型態（多樣性，Variety），都可視之爲「大數據」。舉例來說，某金融機構客訴專線的文字或語音記錄，其每月或每年累積的資料筆數並不見得極多，但因資料形態並非結構化，如果沒有處理文字或語音的相關技術，就必須以人工方式逐筆來剖析處理，而這些處理的工就形成極大的工作「量」。如果演算的方法常會有大量的存取或者面對的是資料多樣性，那麼，在大資料架構與擴充性問題上，必須要考慮的 2 大問題，分別爲「水平擴充」（Scale out）和「垂直擴充」（Scale up），在「水平擴充」是運用大量資源及平行運算來處理資料，它在重覆存取的交易或演算中，有時反而會發生更高的資料延遲；而「垂直擴充」則是要讓同一個機器的運算效果提升，發揮最大價值。但同樣地，若所處理的資料量或資料複雜度不是極大，較優等硬體規格的 CPU 或 GPU 都能達到極佳的演算效能。

「殺雞用牛刀」的思維，只會徒增設備成本，而不見得有其生產力，但「宰牛用小刀」則必然會累死屠夫。何種情境是數據分析上的「殺雞」與「宰牛」？其考量重點及難處不僅在於量大，對處理回應的即時需求性（Velocity，速度）也是重要考量之一。而對於統計分析人而言，資料處理演算的複

雜度、資料型態的多元性等才是在建置「大數據」分析平台上，首要考量的重點。這二點的效能估算常需要透過實測才能得知。

三、社群及輿情的危險推論

在「大數據」資料型態多元性的議題上，「社群媒體與輿情分析」是最近這二年來最熱門的話題之一。在幾次選戰中候選人成功地運用「大數據」協助掌握民意後，許多人又將「大數據」與「社群媒體與輿情分析」畫上等號。在「大數據」新潮流下，民意的掌握、施政的成效、消費者的意見及行爲觀察 ... 等，似乎都應改用網路爬文的推論來進行，過去電話調查或傳統抽樣所運用的統計推論好像不再那麼重要。「大數據」分析的推論常強調著運用全部母體資料納入分析，所以分析的結果就代表著母體的實際情況，無需再運用到任何抽樣方法。而這樣的論點，似乎解放了由樣本來推論母體的這一套傳統理論的束縛，但其實卻產生了「母體的代表性不足」、「樣本的隨機性破壞」及「結論來自於過度配適模型」等三大推論上的危機。

在「社群媒體與輿情分析」上，首先分析者必須選定資料源，衆所周知的資料源包含網路新聞，論壇、Facebook、Twitter、LinkedIn、各類網址 ... 等。而這些被選定的資料源母體是否能代表分析所要推論的母體呢？再者，在這些被選定的母體中，是否眞能針對所有的元素（各類發文）進行普查呢？以 Facebook 的發文爲例，對於不公開的 ID 或貼文，非被授權者是無法具「合法性」（Validity）的取得。依據本人針對某議題的研究，在 Facebook 中以單一關鍵字進行所需 ID 的搜尋，其中公開的 ID 僅占 20%[3]。而在公開的 ID 之中，亦只能針對公開的發文進行爬取。合法搜尋海搜只能取得「主動」提供訊息者的資料，對於「不喜歡」或「不習慣」提供訊息者的意見是忽略及漠視的。在調查實務及消費者行爲研究上，都曾有研究提出「主動」與「被動」表達意見者，其人格特質、意見傾向、消費態度及行爲上都會有極大的不同。若全部以「主動」發言者的意見來做爲決策之參考訊息，則會忽略了沈默大衆的心聲及需求。

運用網路爬文蒐集訊息依然會運用到抽樣的程序。此作法與「集群抽樣」（Clustering Sampling Method）的步驟較相似。亦即，可將各個資料源視爲「群」，抽出群中再針對群中所有的元素進行普查或再抽樣。但此方法卻無法吻合集群抽樣中「隨機選群」以及在被抽到的「群中全部資料普查」或「再隨機抽樣」的理論原則。在樣本選取無法隨機的狀況下，「寫手灌文」或「反輿情」等手法就很容易介入操作，而造成其推論失去「準確性」（Veracity）。此外，不同的資料源其發文的傾向意有極大差異。根據本人針對食安、服貿等議題的追蹤研究中得知，Facebook 及 YouTube 的意見通常較爲負向，新聞報導的意見

較為中立，而新聞論壇則較理性及正向，是以，這些看似是「群」的不同資料源，其實反而較像「分層抽樣」（Stratified Sampling）中的「層」（Stratum），而各「層」的母體元素總數，則是推論時計算分層加權統計量的不可或缺資訊。然而，在不同資料源或不同發文者其發文數量的多寡要如何計算其權重呢？則又無法釐清而很難有定論。

上述種種問題呈現出「母體的代表性不足」及「樣本的隨機性破壞」二大推論危機。統計的推論是信賴度及推論誤差必須建立在樣本的隨機性。所謂樣本的隨機性代表著在母體中的每一個元素都必須有機會被抽取，且被抽取的機率是可知的。「大數據」的抽樣方法其實大多為「立意抽樣」（Judgement Sampling），而其推論則較接近質化調查研究中的「文本分析 」（Content Analysis）。採用此二項研究方法進行研究，研究者必須極具實務上專業經驗的判斷力，其研究過程的管控及結論才會具有一定的信度及效度。

雖然「大數據」的「社群媒體與輿情分析」有前述母體涵蓋度及抽樣代表性等種種問題，但其所使用的「文字探勘」技術仍具有其應用的價值。透過「斷字」、「斷詞」及「語意分析」... 等方法，針對網路媒體與輿情以及組織內部資料（例如政府單位的 1999 記錄、企業組織的客服進線服務記錄 ... 等）進行議題導向的剖析，探討人、事、時、地、物及態度意見彼此間的關聯，將焦點著眼在即時掌握訊息及監控危機上，不再以推論為其終極目的，便能產生極有價值（Value）運用。

另外，在處理組織內部的「大數據」資料時，也會在不同的情境下採用抽樣技術。當面對極大資料量的狀況下進行資料整理時，常先抽出部分樣本進行剖析，了解資料的問題及找出最佳的清理及整合方法，經過數次的反覆抽樣及確認，再將這些清理及整併變數的方法套入整個大數據資料庫去進行處理，如此才可省去不必要時間浪費，並提升資料清理的效能。當採用全部的過去歷史資料做為決策參考依據時，「未來的一切等於現在的種種」的假設就需被設立及接受。因此，在運用組織內部資料進行建立相關的預測模型時，仍應採用多次抽樣的結果來建模，每一個樣本都像是對未來資料的不同組合模擬，運用資料採礦的重覆抽樣（Resampling）再整合推論（Ensemble）所有模型，減少「結論來自於過度配適模型」的現象，對於未來的預測及推論反而會較具可使用性。

四、視覺分析多樣化以準確為底線

在「大數據」的「可視性」（Visualization）上。各類視覺

化軟體在親和度及易用性上已極佳，一般使用者可自行使用各類視覺化軟體展現出商業智慧（Business Intelligence）訊息。透過這些工具將各類統計及分析訊息以地理資訊圖及各類有創意的圖表來展示，讓資訊使用者可以很容易的掌握重點。然而，值得提醒及注意的是，在選擇各類五顏六色且立體變化的圖形時，最重要的仍在於統計分析訊息表達的「準確性」（Veracity）。就以圓餅圖爲例，若採用立體設計，立體端的占比在視覺上會產生放大效果；若選擇彩色浮凸的光面設計，那麼顏色較亮的浮凸區塊會產生放大，而黑或深色就會有縮小的視覺誤判。因此，在選擇視覺化圖形時，最好仍以平面及不具浮凸效果的爲佳，以免造成視覺決策上的誤判。

參、結論

總而言之，「大數據」的計畫推動需要有策略來引導方向，並在以創造「價值」爲目的下，進行一系列資訊整合與分析推論的流程及活動。由資訊系統平台層面來看，「大數據」是一種提升儲存、萃取、載入、轉換、快速演算環境、視覺展現的技術組合，透過高效能的硬體、作業系統及各類軟體的整合服務，達到大量儲存、快速平行運算及展現的功效。以資料匯流層面來看，「大數據」著重在築建可容納百川的渠道，引流匯整來自多元管道的各類異質資料。再由方法及應用層面來看，「大數據」是人工智慧、軟體工程、統計分析、作業研究、各應用領域專業知識及方法（例如：財務金融、生產製造、顧客行銷 ... 等各類領域）的整合應用。站在統計分析的角度來看，「大數據分析」的蹲馬步基本功在於如何整合及清理這些數據，唯有好的「數據品質」才能有「準確性」（Veracity）的推論。而統計人想要成爲「大數據新貴」，必須具備有「資料匯整力」、「數據分析力」以及「領域策略力」。然而，要找到三者能力兼具的人才極爲不易，所以，針對「大數據」專案的推動，應組成包含資訊、統計、專業領域等三種能力成員的小組，才能避開「大數據之非」所造成的危險推論，讓「大數據之是」得以彰顯其美。

註釋

1. 由 iTHome, 2015；http://www.ithome.com.tw/news/98285 中截錄。
2. 伺服器或平行運算架構都會因設備的規格等級及系統調優有不同的演算效能，本測試之結論僅爲單一個案測試結果。在配置伺服器時，僅以 Hadoop 一半成本來配置最佳伺服器架構。
3. 不同議題不同關鍵字結果會有不同，此爲單一案例，不足做爲所有公開 ID 搜尋占比的結論。❖