巨量資料在學術研究上的革命

文．圖／林守德

無庸置疑的，在機器革命、數位化革命、及網路革命之後，巨量資料（或大數據）革命（Big Data Revolution）的時代已然來臨。近年來不管是政府機關，高科技產業，甚至傳統產業，無不努力找出自己與巨量資料的連結，把它當成可以通往無限可能的一扇門。巨量資料相關的研究造成了一股風潮，成為了炙手可熱的顯學，通稱「資料科學」（Data Science）。

在巨量資料對於工商業，民生經濟，政府政策的影響已被廣泛的討論與肯定下，本文希望從另一個角度來看巨量資料所注入的新活力——探討巨量資料對學術研究的影響及幫助。

臺大是臺灣最重要的學術研究重鎮。不管在自然科學，人文社會科學，或是應用科學領域，多年來皆有許多傑出的研究成果。身為一個研究者或許想知道：從學術研究的角度，巨量資料與資料科學扮演了什麼樣的角色？能夠帶來什麼樣的幫助？它如何改變研究的質與量？

在探討這個問題之前，先重述巨量資料的定義：巨量資料基本上是指資料的四大特性（4 v's）：數量大（volume）、變化的速度很快（velocity）、多樣性高（variety）、以及可能帶有雜訊與錯誤（veracity）。這四大特性只要有部分符合，就可以稱之為巨量資料。例如，飛機或其他交通工具上的引擎感應器數據，也許量並非巨大，但是卻是以極高的速度蒐集，也是屬於巨量資料。

資料科學對於研究上的影響或幫助，以下將由三個不同面向探討：

1. 巨量資料的產生

2. 應用巨量資料的研究

3. 巨量資料的方法學研究

巨量資料的產生

從事研究往往需要蒐集足夠的資料。例如研究某個古老部落快絕跡的方言，語言學家可能需要親自在田野調查中去跟老一輩的族人蒐集不同的語音資料；研究營養劑的濃度對於植物生長的影響，就需要從事植株摘種實驗蒐集數據；研究共乘服務對於交通阻塞的影響，就需要蒐集共乘及相關交通流量的資料；探討稅率對於生育率的影響，則需要蒐集多年在不同稅率之下各個地區生育率的資料。這些資料的量也許不一定很大，但是往往會符合「多樣性」（variety）

以及「不完全精準」（veracity）的條件。資料科學從數學、統計以及網路計算等角度，針對資料蒐集的議題有許多討論以及解決方案，闡述如下：

如何更快地蒐集到更多的資料？

在網際網路盛行的時代，資訊科學發展出許多「網路爬蟲」的模組，可以幫忙研究者自動在網路上抓取需要的訊息。例如，自動從網路上下載不同的國家省分的稅率與生育率。此外，人類計算（Human Computing）的領域，也著重在如何在網路服務平台上面設計遊戲或是問卷，幫助研究者蒐集標記的資料。

如何蒐集到最沒有偏差（bias）的資料？

研究者最擔心的就是蒐集的資料有偏差，使得分析的結果產生錯誤。例如，僅在中國大陸蒐集稅率與生育率的資料，可能會導致研究者誤判兩者之間沒有關係，殊不知這是國家「一胎化」政策造成的必然結論。所以，在資料科學中，有許多統計取樣的策略可以幫助減少資料中的偏差。相對應的，在分析的方法上，也存在像是「貝式網路」等模型可以避免一些不正確結論的產生。

要蒐集到多少的資料才足夠？

蒐集的資料量不足常會讓系統進入「過度解釋（overfitting）」的狀態，影響結論的可靠性。然而，資料是否足夠端視於要解決的問題的複雜度，並沒有一個絕對的答案多少的量才是足夠（要估計某群人平均的身高，可能幾千筆資料就已經足夠；要做出一套像是IBM華生一樣自動回答問題的系統，往往需要數以億計的文本）。從統計以及資訊科學實證的角度，已經存在一些指標可以回答這樣的問題。例如，做出「資料學習曲線」（learning curve）把資料量對於目標值的關係找出來，如果發現學習曲線已經平緩，就可以推論資料量對於精確度的影響已經不大。

如何有效率的儲存與分享蒐集的資訊？

巨量資料的量往往大到單一硬碟無法儲存（例如每個人的基因定序資料可以高達數十Giga Bytes）。資訊科學家提出了雲端分散式儲存的方式，並以快速搜尋（indexing）的技術，用以解決巨量資料在儲存與分享的挑戰。

如何將數據轉化成高品質的資訊？

巨量資料的蒐集難免會有雜訊以及不完整的資料（veracity）。資料清理（Data Cleaning）相關技術的目的，就是針對這樣的資料做除噪以及填補遺失的動作。此外，有些資料含有個資，所以巨量資料去機敏的技術，也是幫助研究者不會因為接觸機敏資料而觸犯個資法。

應用巨量資料的研究
（任務型的研究 vs. 分析型的研究）

「使用巨量資料」達成某項研究目的也是許多學術研究必經的過程。基本上又可以細分為兩大類，第一類是「任務型的研究」，第二類是「分析型的研究」。所謂任務型的研究，就是研究者心中已經存在有一個要解決的任務，而希望利用蒐集到的巨量資料來做出一個系統達成這個任務。例如，財經相關研究可能會希望利用長年各國股匯市的資料，預測未來市場走向，甚是下次金融風暴會發生的時間；交通運輸研究希望利用高速公路車流的資料，來判斷在哪裡設置新的交流道最有機會抒解阻塞；天文學研究則可以利用太空望遠鏡的影像資料，預測未知天體的存在與方位；計算語言學希望能夠利用多國語言文本資料，做出能夠達成自動翻譯的系統。任務型的研究通常比較偏應用科學，而所採用的方法基本上是屬於資料科學中的主流方法，如機器學習中的分類演算法（classification）、分群演算法（clustering）、深度學習（deep learning）、貝式模型（Bayesian）等。其他方法如最佳化分析（optimization）與數值方法、搜尋與規劃（search and planning）等也常常被使用。

而所謂分析型的研究，往往研究者或使用者並沒有預設一個需要達成的目標，而是純粹對於所持巨量資料作一些分析，希望發現一些前所未知的知識與現象。例如運輸相關研究可以利用乘客上下計程車資料，探討出在不同時間地域人們對計程車的需求；政治學的研究可以利用投票的資料，分析出不同族群對於某個政策支持度的相關性；醫學研究可以利用健保資料，找出用藥與併發症的關係。這些分析的挑戰在於如何從這麼大量而且多樣性的數據中，找出真正有用而且沒有偏差的知識。目前主要是利用數學與統計的方法，在資料間找尋因果或是相依的關係；或是利用機器學習的「結構學習」（structure leaning）方法找出資料的具體結構。其他有用的技術還有資料視覺化模型、資料摘要技術等，讓研究者更容易從視覺化或摘要過後的結果中發現新奇的現象。

巨量資料的方法學研究

除了前兩類型的研究，另有專門開發能夠處理分析巨量資料的技術的研究。這些技術通常沒有特定針對某類的數據，而是希望能夠發展更好更快的演算法，讓前兩類的應用者可以使用。機器學習堪稱是巨量資料分析最普遍也最有用的技術，但仍有許多需要更進一步研究的挑戰。例如：如何能在分散式的環境下學習；如何利用沒有標記的資料增進學習的效能；如何不讓資料的雜訊影響學習結果；如何處理資料量大於記憶體或硬碟空間的情況等等。其他研究方向如知識發現、資料結構、數據倉儲等，也都因應巨量資料時代的來臨，發展出各自相應的新技術。

巨量資料除了帶給學術研究新的方法之外，自己本身也是個研究議題：因為巨量資料讓人工智慧變成了可能，從哲學與社會學的角度也衍生出相關議題如「電腦是否能夠有人類的智慧」、「人工智慧機器對社會的責任與義務」；從法律面衍生的議題如「巨量資料與個人隱私的衝突」、「個資的價值判定」；甚至從文學及語言學習的角度討論「電腦自動作詩」、「電腦批改作文」的成熟度。

巨量資料之所以強大，是因為這些資料不是一人的貢獻，而是人類整體集體創造的價值。如果牛頓還在世，也許會考慮把他的名言改成：「如果說我可以看得比別人遠一些，那是因為我站在巨量資料的肩膀上」。（本專題策畫／電資學院陳銘憲院長）

參考文獻：

[1] “Data, data everywhere” . The Economist. 25 February 2010. Retrieved 9 December 2012.

[2] Sharma, Sugam; Tim, Udoyara S; Wong, Johnny; Gadia, Shashi; Sharma, Subhash (2014). “A BRIEF REVIEW ON LEADING BIG DATA MODELS.” Data Science Journal

[3] Kalil, Tom. “Big Data is a Big Deal” . White House. Retrieved 26 September 2012.

[4] Executive Office of the President (March 2012). “Big Data Across the Federal Government” (PDF). White House. Retrieved 26 September 2012.

[5] E. Sejdić. “Adapt current tools for use with big data.” Nature, vol. vol. 507, no. 7492, pp. 306, Mar. 2014.

[6] Rui Yan, Han Jiang, Mirella Lapata, Shou-De Lin, Xueqiang Lv, Xiaoming Li. “i, Poet: Automatic Chinese Poetry Composition through a Generative Summarization Framework under Constrained Optimization.” IJCAI 2013.

林守德小檔案

臺大電機系畢業，美國南加州大學資訊工程博士，現任臺大資工系。研究著重於機器學習與巨量資料的分析方法與應用，利用電腦從大量的資料中做出有用的預測判斷及分析。特別熱衷於開發新的演算法讓電腦可以從事「發明與發現」的動作。

資訊科學是應用科學，所以林教授非常重視研究成果在應用上的價值。團隊參加國際競賽 ACM KDDCUP 獲 5 次冠軍，成功解決業界提問的各種挑戰，包括醫學影像辨識（西門子），音樂推薦（Yahoo!）和作者姓名判別（微軟），在國際資料探勘領域中得到廣泛肯定。個人亦獲多項學術獎項如國際學術會議最佳論文獎，傑出人才基金會年輕學者創新獎，吳大猷獎等。此外，林教授也非常重視研究成果對業界在實務上的價值。近年分獲 Google, INTEL, Microsoft，IBM 等 IT 公司獎助，以及美國空軍研究獎（AOARD）5 年獎助。與微軟合作「空氣品質自動偵測」系統，已是線上使用的服務。