撰文／林守德

開採資訊油田

巨量資料蘊含豐富價值，但就像開採原油般，需要複雜的探勘提煉技術。

Google前總裁施密特曾說，在2011年每兩天產生的資料量，等於人類有文明至2003年之前所產生全部的資料量，大約是1.8ZB，用容量4.7GB的光碟儲存，需要4500兆片。這些資料來自四面八方：例如單是2009年美國的製造業就產生了約一億GB的資料，存放這些資料的硬碟如果鋪在一個足球場上，高度超過101大樓；或是美國麻省理工學院教授羅伊將自己小孩三年的成長記錄下來，成為200TB的個人資料。

然而，巨量資料之所以巨大，不只是在「量」，還有驚人的產生速度。Facebook每天會增加上億張相片、一台噴射機上的感測器每小時記錄20TB的資料，更挑戰目前的資料處理技術。巨量資料最後一個重要的特性，是資料的「多樣性」很高，除了最原始的文字以外，也可能是照片、影音、GPS地理軌跡資訊、感測器偵測到的溫度、濕度等。多樣性雖然提升了資料的豐富程度，卻也增加了處理的困難度。

快速大量集結的高多樣性資料，成為巨量資料最重要的特性。那麼，這些巨量資料的意義跟價值何在？麥肯錫顧問公司分析，企業如果能善用巨量資料，將會以2~20倍的速度成長。這份分析更表示，未來10年，與個人地理定位相關的資料，商業價值高達8000億美元；而巨量資料的分析也可以幫助美國健保公司每年省下3000億美元。

巨量資料看似萬靈丹，但是就像所有藥物一樣，使用正確才有效。如何正確使用巨量資料，為資料分析與處理帶來了許多前所未有的挑戰。

首先，純人工分析巨量資料幾乎是不可能的任務，即使利用電腦輔助，也有很多需要克服的難題。例如，電腦在運算時，通常需要把資料存在記憶體中，但是目前一般的運算單位，記憶體最多也是幾百GB，所以在處理巨量資料的時候，通常需要把資料在硬碟跟記憶體之間往返傳送，更有甚者，還需要先藉由網路從雲端儲存設備傳到硬碟。這樣的傳輸相當耗費頻寬與時間，很多時候，整體計算效率低落不是因為運算太慢，而是資料傳輸的時間太長。

**巨量資料的多樣性
雖然提升了資料的豐富程度，
卻也增加了處理的困難度。**

除了資料傳輸的問題以外，因應資料快速產生的特性，分析方法也要有所修正。因為儲存容量有限，資料在生成一段時間後就會被遺棄或是覆蓋，分析巨量資料通常必須把資料當成「串流」，在很短的時間內盡快對其分析應用。在這樣的條件之下，資料使用以及分析的效率就變得很重要，如果處理的速度低於資料產生的速度，整個系統最後就會因為資料堆積太多而崩解。

另一個伴隨巨量資料的問題，是資料的完整性與可信度。大量蒐集的資料就像是大鍋炒的菜一樣，多少會有不理想之處。例如資料遺失不完整、外界干擾造成的錯誤、資料來源的不確定性造成可信度下降等，仍需資訊科學家努力找出完善的解決方法。

隨著巨量資料分析技術的成熟，我們可以期待未來隨身攜帶的電子裝置，會針對個人狀況，隨時提供各種量身打造的建議或是推薦。例如適合的職業、投資組合、旅遊景點，甚至朋友伴侶等。或是更準確的環境預測系統，讓我們可以更精確掌握天災、犯罪、氣象、天然資源蘊藏等資訊。

《經濟學人》指出，在巨量資料的年代，「資料」在企業上扮演的角色，將與「勞力」與「資本」這兩項眾所皆知的因素鼎足三分。巨量資料就像油田，蘊含了極豐富的價值，但是就如原油的開採需要探勘提煉的工具，巨量資料的分析也需要大量複雜的技術。以現下資訊科學進步的速度，相信巨量資料很快就可以發揮效能，為人類所用。 SA

林守德是台灣大學資訊工程系副教授。

影像來源：林守德