簡介巨量資料探勘

文．圖／陳銘憲

近年來，人類社會已由「全球瘋雲」進入到「巨資時代」，巨量資料及其衍生之議題在未來將占據關鍵地位。因此如何搭上這班通往嶄新機會的列車，是大家所關切的課題。一般認為，大數據具有數量龐大(volume)、累積速度迅速(velocity)、以及資料多樣性(variety)的特性。如何從具有這些特性的數據當中發掘洞見，並藉以做出更好的決策以及最佳化整個過程，已非僅憑傳統之關聯式資料庫可以處理，而需要嶄新的資料處理方法。尤其近幾年在物聯網以及社群網路如Twitter、Facebook的推波助瀾下，全世界資料量的產生數量、速度和多樣性達到前所未有的高峰。若能妥善使用大數據，我們不僅能了解多數使用者的行為，還能藉此洞悉未來各種趨勢。要處理巨量資料，除了領域之知識外，受重視之技術包括機器學習、資料管理、人機介面和自然語言處理。目前，已有多家公司投入大數據的領域，例如處理資料分析的PARACCEL、多重資料庫的Rocket、資料探勘的KEEL、多模型的ArangoDB等。

然而，對不同領域之巨量資料進行正確、有效率的分析存在各種理論與現實上的挑戰。而面對資料領域特性及多元應用之眾多組合，要挑選正確、有效之分析方法與工具對資料分析者更不容易。一般而言，多數人容易忽略資料準備(data cleaning)之重要性及挑戰，同時也未能善用已有之工具。凡事從零開始自是事倍功半。另外，亦應瞭解資料探勘需有足夠領域知識之特性。以下介紹現行常用於數據分析的開源軟體（Open-Source Software），這些開源軟體分別隸屬大數據處理平台、大數據資料探勘與大數據圖形資料探勘。值得一提的是，巨量資料處理是個持續發展的領域，在此所列舉的工具僅是例子，新的軟體套件是會持續產生。

資料分析開源軟體介紹

巨量資料處理平台：許多大公司如Facebook、Yahoo、Twitter與LinkedIn等在處理大數據資料時，皆受益於開放原始碼的大數據處理平台與架構。當然，開源軟體亦因獲得使用而更受重視。舉例來說，Apache Hadoop是一款處理平行化應用程式的軟體，它以MapReduce模型與分散式檔案系統為基礎。基本上，Hadoop讓應用程式可以在擁有很多計算節點的叢集上快速地平行處理大量資料。而隨著Hadoop對大量資料平行處理的能力，許多相關的開源軟體亦被開發出來，例如Apache Pig，Apache Hive，Apache HBase，Apache ZooKeeper等。

資料探勘軟體：在資料探勘領域有許多開源軟體可幫助資料分析[註]。Apache Mahout是一個以Hadoop為基礎的機器學習和資料探勘開源軟體。Apache Mahout包含了很多機器學習及資料

探勘演算法，如分群（clustering）、分類（classification）、協同過濾（collaborative filtering）和頻繁特徵探勘（frequent pattern mining）等。透過個別設計，輔以相關的領域知識，這些演算法可以很有效率地對資料進行分析，協助使用者從資料中獲取相關資訊與知識。此外，開源程式語言R亦在大數據資料分析上日益普及。R可進行統計計算，是視覺化的開源程式語言，可用來統計分析大型的資料集。Python亦有許多程式庫可供資料分析之用。而MOA則能夠進行即時資料串流探勘，包含相關性（association）、分類、回歸（regression）、分群等演算法及頻繁的圖形探勘。目前在資料科學或相關之課程多有教授學生R及Python等軟體套件。

圖形資料探勘：在巨量資料分析的挑戰中，尤以圖形資料為最，因圖形資料可含各式訊息，目前網際網路圖（Internet Web graph）、社群網路（social network）、生物網路（biological network）等皆以圖形的方式存在。因此如何有效率地分析與探勘這些圖形十分重要。目前有一些針對大型圖形探勘的開源工具，如PEGASUS與GraphLab。PEGASUS是在Hadoop平台上開發的分散式平行化圖形資料探勘系統。它可以從生活中的大型圖形結構中找到特徵和異常，如分析病毒式行銷、疾病的傳播特徵或是網路資料傳遞的異常偵測等。另外，對於如何在巨量的圖形資料中擷取有用的資訊，所探討之技術有sampling（取樣），summarization（資訊彙整而使取樣有代表性）及extraction（針對應用作資料擷取）等。

資料探勘的應用與挑戰

針對未來應用，許多行業都可由大數據分析後的結果獲益。對金融保險業，大數據分析可用於信用評等、客製化金融服務、授信、壞帳分析等；對零售業而言，大數據分析的結果可作為輔助購買與物流整合決策之依據；對觀光及連鎖業，大數據分析可作為展店址選擇、分店貨品品項選擇等決策輔助依據；而對電信業，可提供最佳化之網路配置與用戶行為分析等；對教育業而言，大數據分析可作為學生課程規劃與職涯規劃之依據；而針對廣告業，可提供廣告點閱來源分析、回應率分析、行銷策略提供等。尤其是行動裝置之螢幕大小有限，如何適時推出最合適之廣告就更是挑戰（如Real-Time Bidding 之研究）。這些應用在我們對資通訊技術使用日增的今日勢必日益重要。

未來在大數據資料管理和分析上仍有很多重大挑戰，源自於大數據的特性：數量龐大、速度迅速、以及多樣性。例如，在希望達到探勘之目的下，不犧牲對資訊安全和隱私權的保障。即使是做去識別化，如何在兩者之間取得平衡很重要。在資訊技術上，累進運算(incremental process)可避免資料因部分新增而需全部重新檢視，有發展潛力。巨量資料分析與Cognitive Computing，如何相輔相成，且適時與人互動，亦值得探討。此外，在分散式探勘的課題上，對龐大且分散之資料，需要很多理論分析與研究來提供增進效率或是進行平行化。由於資料可能會隨著時間持續不斷演變，所以分析資料串流也是一個十分重要的研究方向，必須能有效地處

理演變。針對特定應用作軟硬整合亦是重要方向。最後，資料視覺化（visualization）提供使用者友善方式來分析與理解資料。因人類可接受之資料維度有限，所以有效地視覺化大數據資料並不容易。

結語

由於雲端運算的典範轉移，行動裝置的普及，物聯網與社群網路應用正蓬勃發展，我們已進入巨量資料的時代，這亦使臺灣產業發展面對更巨大的挑戰，但未來也將因此而有更多的機會。而人們帶著接上雲端之行動裝置（如智慧眼鏡、手機、智慧手錶），就如同有個超級電腦在身邊，可使用世界各地之資料庫，可以是上知天文、下知地理的超人。這樣一來，人與人之間如何競爭與合作，便需有新的思維（在課堂上稱之為Human 2.0）。我們認為透過這新平台、新環境的創新將成為分別差異之處。天下大亂，形勢可能大好，這或許是臺灣資訊業轉型與提升之契機。我們正處於一個新資訊時代的開端，希望藉由本文的簡介讓讀者對巨量資料分析與探勘有基本了解並一同迎接巨量資料時代的來臨。（本專題策畫／電資學院陳銘憲院長）

獲教育部國家講座，總統授獎。

註：W. Fan, A. Bifet. Mining Big Data: Current Status, and Forecast to the Future. ACM SIGKDD Explorations, 14(2):1-5, 2012.

陳銘憲小檔案

在美國密西根大學獲電腦、通訊和控制工程學博士學位。現任臺灣大學電機工程學系特聘教授並擔任電機資訊學院院長。陳教授於 1988 至 1996 年間於美國紐約州之 IBM Thomas J. Watson Research Center 從事研究工作。2003 至 2006 年擔任臺大電信所所長，2007 至 2008 年擔任資訊工業策進會執行長，2008 至 2015 年間擔任中央研究院資訊科技創新研究中心特聘研究員兼主任。其主要研究領域為資料庫、資料探勘、社群網路及多媒體網路。

陳銘憲教授為教育部國家講座教授，獲有教育部學術獎、國科會傑出研究獎、有庠科技講座、潘文淵研究傑出獎、東元獎、及資訊榮譽獎章。陳教授於 IBM Research 之研究成果已為業界產品使用，曾獲 IBM 傑出創新獎 (IBM Outstanding Innovation Award)。陳教授亦為許多重要期刊之編輯或總編輯，其研究、教學等成就獲得許多重要獎項，並榮膺 ACM Fellow 及 IEEE Fellow。