雲報專欄：巨量資料(Big Data)的技術發展與挑戰—中研院資訊科技創新研究中心陳銘憲主任/雲端技術專家委員會委員、林與絜博士

資訊產業發展已由「全球瘋雲」進入「巨資時代」。根據國際研究暨顧問機構 Gartner 的定義[1]，巨量資料為「大量、快速累積、具有多樣性的資訊資產，需要新的處理技術以提升決策品質、發掘問題、最佳化流程」。本文將簡介巨量資料的成因、因應巨量資料所需發展的技術及挑戰。

簡介 Introduction

近年來由於行動裝置的普及、社群網站的盛行以及雲端服務的進步，大量資料以驚人的速度產生。比方物聯網(Internet of Thing)普及之後，各種裝置(device)上都配備了感測器(sensor)，因而能夠廣泛而大量地收集各式各樣的資料。又比方智慧型手機裡也裝置了各式各樣的感測器，可以感應並收集 GPS 定位、移動速度、環境亮度等各種資料，這些資料在使用各種應用服務的過程中被記錄下來。此外，也有許多資料是由使用者主動分享給彼此：照片、影片、具名或不具名的個人意見等等。Big data 累積的速度非常驚人，在短短的一分鐘之內，上傳至 YouTube 的影片總長度為 30 小時，發佈在 Twitter 上的新訊息有十萬則，Google 搜尋被利用了兩百萬次，Facebook 上的訊息則被瀏覽了六百萬次[2]。

大量的資料隱藏著巨大的價值。以網路服務為例，當使用者在使用服務的時候，服務提供者也同時蒐集了使用者的資料，包含使用者主動輸入的資料以及使用服務過程中產生的行為資料。當服務提供者擁有越多和特定使用者相關的資料，就能分析此使用者的行為，進而提供此使用者更高品質、個人化的服務。假如某位 Google search 使用者同時也使用了 Google map、Google+等服務，則在使用搜尋功能的時候，Google 可以利用此使用者的地點資訊（從 Google map 使用紀錄得知）及朋友資訊（從 Google+個人資料得知）將搜尋結果個人化，甚至可以進一步利用使用者的朋友或同地點的人之歷史資料，來分析此使用者的需求，並提供更貼近此需求的搜尋結果。在此例中，使用者提供了資料讓自己能得到更好的服務品質，也提升了其他人得到的服務品質（例：此使用者的朋友、和此使

用者在同樣地點使用搜尋的人、和此使用者行為相似的人)，而提升的服務品質自然也提高了服務提供者的價值。

定義 Definition

Big data 具有三項主要特性：大量(Volume)、快速(Velocity)、多樣性(Variety)；同時，big data 的真實性(Veracity)也需要被檢驗。由於傳統的關聯式資料庫系統及在此系統上運作的資訊探勘、統計分析等各種技術，無法直接因應上述幾項特性，是以需要設計能夠大量儲存並實行平行運算的各種新方法來處理、利用 big data，以期找出 big data 中蘊含的巨大價值(Value)。

挑戰 Challenges

以下簡單描述上述 big data 之各項特性所帶來的挑戰並舉例說明。

Volume：大量是 big data 最基本的特性。如前所述，物聯網的感測資料及使用者主動、被動分享的資料以前所未有的速度累積成非常大量的資料。所謂的大量往往不只需要分散放在數台機器上，而需要上百台、上千台機器來存放。如何將大量資料分散儲存在雲端之大量機器上從而能提供有效率的平行存取？原本在一台機器上執行的運算，要如何統整儲存在大量機器上的所有資料進行分散式或平行式之運算？這些當然需要我們對現行之演算法作重新評估或設計。

一般而言，資訊探勘的應用與領域專業知識是高度相關(domain dependent)。根據不同應用，需要設計不同方法來收集大量有用的資料。以醫療照護為例，要使用什麼樣的感測器、將感測器裝置在哪些地方，才能有效地量測到需要的資料、確保資料正確性、且不至於增加使用者太大負擔，這都需要針對應用的領域來規劃。

Velocity：由於資料流量太大，先存起來再慢慢分析並不足以因應作即時處理之需求。如何在大量資料流進來的時候直接執行過濾、或輔以適度的後設資料(metadata)以加快處理速度是個技術挑戰。尤其在 big data 環境中，收集到的資料很多，但資訊的密度卻往往很低，絕大部份資料可能是沒有用處的。若能夠過濾掉大量沒有用的資料、保留真正有價值的部分，即可提升資料的價值、方便

後續分析使用。此外，若系統處理資料的速度趕不上資料流進的速度則無法做到即時反應，對於某些需要即時反應的應用（例：偵測攻擊行動），就必須針對其需求設計出能夠即時處理大量流入資料、做出反應的方法。

Variety：非結構(unstructured)是 big data 的重要特性，也是在關聯式資料庫上廣泛被使用的資訊探勘及統計分析等方法無法直接套用在 big data 上的主要原因。關聯式資料庫的資料具有固定格式，big data 則包含各種不同格式的資料。在全球資訊網上(World Wide Web)以文字資料(text data)為主，圖片、影片資料為輔；社群網路服務(social network service)上的資料則包含了社群網路本身的資訊(social network graph)以及使用者彼此之間的互動。各種資料都具有不同特性，所以需要用不同的方法來處理（例：NoSQL）。文字資料在過去十年已累積大量處理經驗，而社群網路服務資料的處理則是目前熱門研究領域。

異質性與多元性也是 big data 的重要特色。當 big data 由不同型式的資料集合而成，如何利用不同型式的資料來提供整合性的服務？而當 big data 由不同資料來源的資料集合而成，如何整合不同資料結構、不同的語意(semantics)，亦是重要課題。

Veracity：由於 big data 的資料可能來自不同來源、甚至來自廣大群眾，在利用資料的時候，不可避免地需要考慮資料的真實性。舉例來說，若有惡意的使用者利用社群網路散播假消息、或是不公正的使用者刻意於評價網站上詆毀特定賣方，都會影響到資料的真實性。

在 big data 環境中，隱私也是非常重要的問題。如同傳統資料庫系統，資料收集者應該要誠實完整揭露其收集及使用資料的範圍。此外，由於 big data 往往會整合不同資料來源的資料，不同資料收集者間的資料分享比以前更加頻繁而複雜，因此，需要設計新的規範以作為資料收集者分享資料的依據，也需要設計新的方法以在複雜的分享過程中確保使用者的資料安全。

結論 Conclusion

使用 big data 應該使得資料收集者和資料提供者達到雙贏的局面。除了發展各種技術來處理並利用 big data 之外，也需要根據不同需求整合相關知識(例：

社群網路科學、醫療照護、環境監測）來設計各種應用，更需要根據社會科學和法律專業來設計相關規範以確保資料提供者和資料收集者之間的權利義務。我們相信隨著多媒體技術之發展、雲端平台的建立、社群網路與行動應用的迅速普及，巨量資料的時代已然來臨，而這也帶來了資訊領域研究人員前所未有的機會與挑戰。

[1] *"Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."*

[2] http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html