巨量資料的關鍵挑戰

結合資料與知識 巨量資料的創新價值

近年來透過資料探勘（Data Mining）技術，巨量資料（Big Data）已廣泛運用到各領域中，人類生活及思考模式有了全新的改變，資料不再是靜止的，而是能夠告訴人們「正是如此」的事實。但巨量資料應用上仍有許多關鍵挑戰，能掌握關鍵就能取得解決問題的先機。

口述／成功大學資訊工程學系教授曾新穆　整理／張勵德　攝影／蔡世豪　圖片來源／法新社

這些年巨量資料（Big Data）已經變成一種顯學，實際上也的確產生非常多的影響，今天利用這個機會和大家針對巨量資料分析及核心挑戰進行交流。

過去我個人和研究團隊做資料探勘（Data Mining）研究超過 15 年的時間，期間很幸運參與過許多政府與產學相關研究，今天提出許多現代正夯的領域與趨勢，包括健康醫療、社群網路等有關的案例和大家分享。

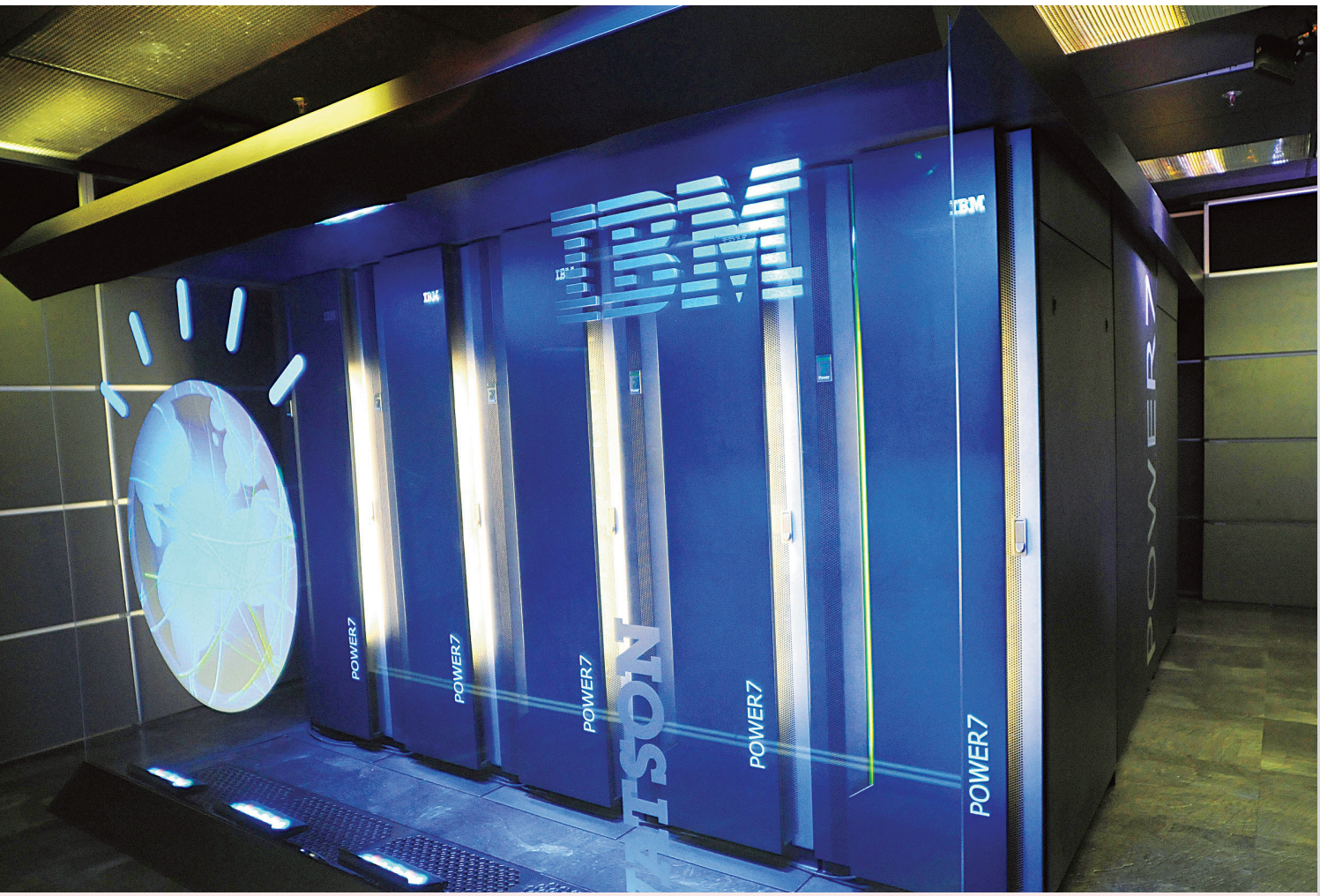
一般談到資料探勘分析都會提到一個耳熟能詳的故事，就是超級市場業者利用消費者購買商品資料進行分析，找出買尿布的消費者有很高比例也會順便買些啤酒，如果可以用更多、更好的資料分析找出存在裡面的使用者習慣與規則，進而提供促銷，就能增加購買力。

巨量資料分析找出事實

去年還有一個很有意思的例子，美國著名的零售商 Target，有一天一位父親氣憤地跑到門市客訴，質疑為何把孕婦用品的 DM 寄給還是大學生的女兒。一個禮拜後，店經理再次去電這位客訴的爸爸致歉，但這個爸爸不好意思地說：「很抱歉，其實我女兒已經懷孕了，連我都不知道！」

Target 為何比父親還早知道女兒懷孕？就是利用公司內部所有的消費者購買資料來分析，找出孕婦懷孕初

> 巨量資料要產生價值，
> 最核心的就是資料探勘、
> 分析的技術。



IBM 超級電腦「華生」具備巨量資料運算力、記憶力、反應力及語言能力，並且透過學習，進一步把知識轉化為價值。

期、中期大概會買的物品，當這位消費者購買了某一項產品時，系統就會自動開始運作，提供消費者更多的產品訊息，從最早的消費習慣研究，到後來已經變成消費者行為預測。

另一個更有名的例子就是 Google 在 2009 年曾在《自然》（Natural）期刊發表了一篇論文，利用網友的搜尋習慣來預測流感的爆發。利用的技術很簡單，但預測出來的趨勢非常精準，與美國疾病預防管制中心 CDC 預測非常雷同，可以看出巨量資料的威力。

巨量資料要產生價值，最核心的就是資料探勘、分析的技術，目前巨量資料的增加速度、可靠度等方面提升很多，許多公司包括 IBM 等，都提供很多很好的架構，努力去開發巨量資料，未來發展性很快。

無論使用任何的方式進行資料探勘，最後都回歸實際的市場或服務上，可以整合支援客戶服務，在實際應用上產生價值。

巨量資料應用的關鍵挑戰

從資料探勘到現在巨量資料應用，其中有許多關鍵性的挑戰，首先在資料預先處理方面（Data Preprocessing），最重要就是資料的品質問題，包括資料是否有雜訊、是否完整，尤其在巨量資料環境下，影響最大的問題就是資料稀疏性（sparsity）。

舉例來說，美國火紅的線上影音租賃公司 Netflix，該公司最著名的就是有個人化的推薦系統，當消費者進入系統後，就會根據消費者過去看過的片子、評價進行分析後，提供推薦影片給消費者。全世界的影片就是一個巨量資料的代表，一般的推薦系統消費者使用過幾次後就會覺得平淡無奇，原因就是大部分推薦都是熱門、最新的影片，但 Netflix 最厲害的地方就是會推薦給消費

者比較冷門的影片，消費者試過之後發現非常好看、有興趣，之後就會對系統更有興趣，而系統也可以藉由顧客反應更有效地推薦影片，長此下來就有更多的資料可分析。

Netflix 目前有超過 250 萬個訂戶，每天累積近 400 萬次評價與 300 萬次搜尋，影音串流量從 2011 年第 4 季的 2,000 萬個小時，到 2012 年 1 月單月就有 1,000 萬個小時，成長非常快速，這些也是典型的巨量資料。2006 年 Netflix 曾舉辦過一個百萬美元獎金的比賽（Netflix Prize），提供一筆資料邀請各方好手來進行評價的預測，條件是比該公司目前的推薦系統精準度要提高 10%，就可以拿到大獎，最後花了 3 年的時間才終於把這個獎發出去，可以看出有多麼的困難。

這個問題難在哪裡呢？由於 Netflix 提供的資料沒有使用者背景，傳統上最簡單的方法，是把過去與目標使用者的評價有相似的資料拿出來，取平均數來預測；但在這個個案中，除了資料稀疏性相當大，各位可以想像一下，在上千萬名的使用者中，一個人的評價只占其中的極少數，用上述的概念去做恐怕很難做出來。最後冠軍的團隊跳脫窠臼，不用標準的方法來看，反而是用當時的社會環境與使用者心理現象去分析，才能脫穎而出。從這個例子發現，要解決資料稀疏、有效性等問題，首先要掌握資料的特性才行。

找到關鍵特徵　資料探勘事半功倍

其次是資料的理解與歸納（Data Understanding and Induction），常常我們對資料了解不夠清楚，其實很難做出很好的應用。其中有二個最重要的因素，一是能否找到資料中的關鍵特徵（Key Feature），當面對龐雜的資料時，最後的預測結果是否精準，可能就要靠關鍵特徵來決定。

例如諾基亞（NOKIA）在 2012 年曾經舉辦過一個競賽（Mobile Data Challenge），贈送手機及獎金給數百位使用者，但條件是必須把使用手機的一舉一動都記錄下來，希望藉此找出有用的使用者習慣應用。參賽團隊必須從資料中預測出使用者性別、所在地等資訊。

在性別的預測方面，由香港科技大學的團隊奪冠，但進一步分析其操作方法，竟然是分析了 1 萬多個各種可能特徵再做出推測，最後準確度雖然高達 96%，但在實際應用上這樣的做法卻有很大的問題。後來再進一步研究，發現其實只要用「加速計感應器」這個關鍵特徵進行分析，就可做到 95% 的使用者性別預測，主要就是考量到男女使用者的習慣不同，男生會把手機放在褲子、衣服內，女生通常是把手機放在包包裡，也就是說，只要找到一個關鍵特徵，就能把預測做到最好。

其次是能夠看到資料的各種面貌，例如 Google 預測流感，就是利用全球最龐大的搜尋資料，依照使用者在感冒不同期間，搜尋的各種關鍵字進行分析，若某處越來越多類似資料出現，可能就是流感要爆發的跡象。有這樣的正確、全面且廣泛的資料庫，呈現資料的各式面貌，效果更甚於設計出各種演算法的推論。

另一個很重要的面向則是在於資料分析的即時性與設計架構的問題。由於巨量資料有流動的特性，要越快完成才有即時的價值，但因資料量非常大，分析必須要用到很多方法，除了掌握關鍵特徵，也經常會用到抽樣來試做，但樣本是否足以代表整體資料又會產生問題。例如亞利桑納州立大學做了一個研究，將「推特」（twitter）需收費的完整推文資料，與免費的、僅 1% 推文抽樣資料進行比對，發現雖然在熱門話題及使用者地理分布上有很高的相關度與代表性，但在標籤功能（Hashtags）等其他方面卻幾乎沒有明顯代表性與關聯性。

目前在巨量資料的環境下，公開資料研究也是一種趨勢，不過公開資料有許多也是抽樣資料。例如以往國內進行公共衛生、醫療等研究時，使用國衛院提供的 100 萬個國民抽樣數據，這份數據雖依照年齡、性別、每年出生人數分布等各方面因素決定抽樣，理論上是全民縮影，但進一步整合環境資料分析致病因子後，卻發現樣本得到的數據代表性嚴重不足，顯見抽樣樣本應用必須非常小心。

此外，當前巨量資料分析時多把注意力放在前端的資料蒐集上，但另一個很重要的面向是後端處理問題，若無法制定出規則或規則太多無法判讀，對資料探勘會有很大的影響。舉例來說，利用健保資料分析一些疾病

成功大學資訊工程學系教授曾新穆認為，巨量資料的運用還存在很多挑戰。

可否在 1 年前、10 年前就被發現，從先期出現那些症狀、後來會演變出哪些疾病的關聯進行分析，雖然表面上看起來有很明顯的規則及代表性出現，但最後研究經常卻是負相關。因此，為了減少偏差，研究者必須從文獻研究開始，篩選出一些症狀回過頭去進行資料分析。不過，巨量資料的分析在醫學研究上確實可以克服臨床時間太少的問題。

至於巨量資料在研究應用上也可能面臨到一些爭議問題，最明顯的例子就是隱私權的疑慮，雖然現在有很多技術可以遮蔽個人資料，但因為很多個人的特徵可以推論出來，仍可能造成個資外洩，因此未來從法律層面進行規範，應可解決此問題。

跨越資料與知識間鴻溝　產生價值

最後，在應用上還有一個很重要的問題，就是如何解決資料、知識與價值之間的鴻溝問題。有時候一份資料可能有一些知識存在，到最後是否可以產生價值？

2012 年 IBM 超級電腦「華生」，在全美的益智節目中挑戰並擊敗 2 位歷史紀錄保持人，華生雖具備巨量資料運算力、記憶力、反應力及語言能力，但很多資料並不是光靠記憶就可以使用，還必須透過學習才能了解背後的真正意涵。為此 IBM 特別找了很多文學、語言、歷史專家，來「教導」華生學習知識的深層意義，包含很多問題的隱喻、人類的常識等，才能把知識轉化為價值，絕非單靠資料就可以做到。

此外，巨量資料分析應用在製造業領域上，也是縮短知識與資料之間距離的例子。很多高科技產業的製程非常繁雜，有時要有上千種的程序，要找出良率的問題，很多情況已無法靠人力進行檢測，透過資料探勘雖可找出機台本身偏離正常等問題，但發掘出太多的問題點卻也讓效果大打折扣。

為了讓廠商能實際應用於解決問題，團隊必須和製程工程師從頭了解整個生產程序中的每個特性，幫助在資料探勘中找出最有可能性的問題，把知識納入資料探勘之中，兩者結合起來，最後得到的實際上的結果可找出 90% 以上的問題。

巨量資料分析研究運用還存在很多挑戰，但也有很多新的應用應運而生，未來還有很大的努力空間。