篇名：

數據分析—大數據

作者：

趙漢彊老師　私立高英高級工商職業學校

大數據（Big Data）

一、前言

近年來討論度非常高的大數據（Big Data），它是數位時代的產物，由於行動裝置的普及，人們在網路上製造了大量的資訊，因此政府或企業便利用 Big Data 來預測未來趨勢和發展商機，而如何彙整及分析龐雜的巨量資料，已經成為 21 世紀最重要的課題。

二、正文

（一）什麼是大數據

大數據（Big Data）又被稱為巨量資料，亦即就是過去 10 年廣泛用於企業內部的資料分析、商業智慧和統計應用之集合。但大數據現在不只是資料處理工具，更是一種企業思維和商業模式，因為資料量急速增加、儲存設備成本下降、軟體技術進化和雲端環境成熟等種種客觀條件成立，使得資料分析從過去的洞悉過往進化到預測未來，以開創從所未見的商業模式。

（二）為什麼需要大數據

大數據已影響到每個人的生活以及每家企業。對企業而言，大數據可望提升服務品質、增加管理效率、協助決策和創造商業模式；而對一般民眾而言，大數據是另一個自我，它可能比本人更了解本人，為你預先解決每個未知，當一切都開始數據化，你能夠不需要數據嗎？

（三）大數據的特點

大數據有四個基本特徵：第一，數據體量巨大，根據百度資料提到，其新首頁導航每天需要提供的數據超過 1.5PB（1PB=1024TB），這些數據如果列印出來將超過 5 千億張 A4 紙。而到目前為止，人類生產的所有印刷材料的數據量僅為 200PB。第二，數據類型多樣，現在的數據類型不僅是文本形式，更多的是圖片、影音、地理位置信息等多類型的數據，個性化數據佔絕大多數。第三，處理速度快，數據處理遵循「1 秒定律」，可從各種類型的數據中快速獲得高價值的信息。第四，價值密度低，以視頻為例，一小時的視頻，在不間斷的監控過程中，可能有用的數據僅僅只有一兩秒。

（四）大數據的分析

對大數據進行分析，才能獲取更多智能的、深入的、有價值的信息。那麼越來越多的應用涉及到大數據，而這些大數據的屬性，包括數量，速度，多樣性等等都是呈現了大數據不斷增長的複雜性，所以大數據的分析方法在大數據領域就

顯得非常重要，可以說是決定最終信息是否有其價值的重要因素。而大數據分析普遍存在的方法理論，本文簡述五項分析法，如下：

1.可視化分析：大數據分析的使用者有大數據分析專家，同時還有普通用戶，但是他們二者對於大數據分析最基本的要求就是可視化分析，因為可視化分析能夠直觀的呈現大數據特點，同時能夠非常容易被讀者所接受，就如同看圖說話一樣簡單明瞭。

2.數據挖掘演算法：大數據分析的理論核心就是數據挖掘演算法，各種數據挖掘的演算法是基於不同的數據類型和格式，因此才能更加科學的呈現出數據本身具備的特點，也正因為這些被全世界統計學家所公認的各種統計方法，才能深入數據深層，挖掘出其公認的價值。另外，也是因為有這些數據挖掘的演算法才能更快速地處理大數據。

3.預測性分析：大數據分析最終要的應用領域之一就是預測性分析，從大數據中挖掘出特點，通過科學方法來建立模型，之後便可透過模型帶入新的數據，從而預測未來的數據。

4.語義引擎：非結構化數據的多元化給數據分析帶來新的挑戰，我們需要一套工具去系統地分析，提煉數據。設計語義引擎能有足夠的人工智慧可以從數據中主動地提取信息。

5.數據質量和數據管理：大數據分析離不開數據質量和數據管理，高質量的數據和有效的數據管理，無論是在學術研究還是在商業應用領域，都能夠帶來分析結果的真實性和價值性。

（五）大數據的常見誤解

1.數據不等於信息

經常有人把數據和信息當作同義詞來用。其實不然，數據指的是一個原始的數據點（無論是透過數字、文字、圖片或是影音等等），信息則是直接與內容有相關，需要具備資訊性（informative）。數據量越多，不一定代表信息就越多，更不能代表信息就會成比例的增加。例如：社交網站上的信息，隨著我們加入的社交網站越多，我們獲得的數據量就會成比例的增多，我們獲得的信息雖然也會增多，但卻不會成比例的增加。

2.信息不等於智慧

信息要能轉化成智慧，至少要滿足以下三個標準：

(1)可破譯性：越來越多的企業每天都會生產出大量的數據，卻還沒想好怎麼運用，因此，他們就將這些數據暫時非結構化（unstructured）的儲存起來，然而這些非結構化的數據卻不一定可以破譯。比如說，當我們記錄了某客戶在網站上三次翻頁的時間間隔為：3 秒、2 秒與 17 秒，但卻忘記標註這三個間隔時間到底代表了什麼，這些數據是信息（非重覆性），卻不可破譯，因此就無法成為可運用的智慧。

(2)關聯性：毫無相關的信息，至多只是個噪音或是可隨處丟棄的資料。

(3)新穎性：很多時候我們無法僅僅根據手中的數據和信息來進行判斷。舉個例子，某電子商務公司透過一組數據（信息），分析出了客戶願意為當天送貨的產品多支付 10 元，然後又透過另一組完全獨立的數據（信息）得到了同樣的內容，這樣的情況下，後者就不具備新穎性。不幸的是，很多時候我們只有在處理了大量的數據和信息之後，才能判斷它們的新穎性。

三、結語

我們對未來的認知，主要是基於常識和對未來的想像。根據統計，現在「紐約時報」一周的信息量比 18 世紀一個人一生所收到的資訊量更大，現在 18 個月產生的信息比過去 5000 年的總和更多。大數據的發展已成為一股無法阻擋的洪流，這項技術的發展是漸成的，並非像工業革命或電腦革命一樣讓世界轉瞬間有巨大變化；然而它仍將一步步改變你我的日常生活，舉凡金融、健康、醫療都將息息相關，就讓我們一起擁抱大數據的未來吧！

四、參考文獻

1.引用網路資料 http://wiki.mbalib.com/zh-tw/%E5%A4%A7%E6%95%B0%E6%8D%AE
2.周正、陳楓，「大數據時代」來了—專訪國防信息學院研究所所長孟寶巨集，解放軍報，2013.1.17.。
3.趙繼海，大數據時代圖書館面臨的挑戰機遇與對策，浙江大學寧波理工學院。
4.李欣宜，數位時代，第 251 期。