撰文／林守德

從嬰兒身上學到的事

迎接巨量資料紀元的來臨！

「巨量資料」（big data）是近年崛起的新名詞，泛指因為網際網路、社群網站，以及許多網路上多媒體服務的興起，所迅速產生與累積的資料。這些資料不僅量大、形式互異（如文字、圖像、影片），更以驚人的速度產生。如何善用這些巨量資料來從事預測與分析，成為資訊科學家非常關心的議題。

美國麻省理工學院多媒體實驗室的教授羅伊（Deb Roy）自2005年起開始了「人類語音之家」（human speechome）計畫，利用錄影及錄音的方式，記錄自己小孩出生之後三年的活動，進而用這些資料研究人類如何學習語言。他在家中佈建了10幾台全方位錄影機以及10多個收音麥克風，希望能夠把父母、嬰兒以及保姆的行動與對話都錄下來。數年內共累積了12萬小時的聲音及9萬小時的影像，其中涵蓋了小嬰兒70%清醒時間的活動記錄。

這些資料堪稱有史以來對單一個體最完整的記錄，總共佔了250 TB的硬碟空間。在開始分析之前，羅伊的團隊還需要幫這些資料加上標記。例如需要辨識小嬰兒聽到跟講出的聲音，並產生文字檔。然而，現今的語音辨識技術還無法準確辨識嬰兒的語音，對於辨識離麥克風較遠的成人聲音精確度也有待加強。所以他們團隊製作了一個「半自動」的語音標記模組，利用電腦過濾雜訊以及非語音訊號，並把音波自動接成一個一個單元，以增加人工辨識的速度與精確度。最後他們總共辨識了800萬個字。此外，他們還利用視訊資料標記了這個小嬰兒每段時間的位置以及他是否醒著，這樣的資訊，有助於之後判斷他有沒有聽到大人的某段對話。

一開始，羅伊團隊希望能夠從這些資料中，了解嬰兒如何從與環境的互動中逐漸培養出字彙以及學習語言。例如，從語音資料中可以擷取所有小嬰兒表達「水」這個意念的詞，然後再串接起來，就可以形成一連串從一開始gaga這樣的狀聲詞一直演化到後來water這個正確讀法的過程。同時，羅伊也想利用這些資料來得知小嬰兒會先學到什麼樣的詞彙，以及其背後的原因。於是他們利用機器學習方法，把蒐集到的資料自動分類，這些分類通常可以對應到某種情境或動作（如用餐、換尿布），然後利用訊息原理中「熵」（entropy）的概念，去計算每一個字彙在每一種情境中出現的分佈是否平衡。例如有些字詞對於情境比較不敏感（譬如「要」跟「來」等），反之也有一些字詞對情境比較敏感（譬如「吃」跟「再見」），只會在特定的情境出現。最近，他們從蒐集來的巨量資料發現了一個有趣的結果：小嬰兒會較快學到對於情境敏感的詞，對於情境比較不敏感的詞，學習速度就比較緩慢。

善加應用網路及電腦
產生的巨量資料，可幫我們
做出更正確的決定。

這樣的巨量資料，還可以讓學者更進一步了解語言學習背後的一些現象，例如分析在哪段時間或是哪個情境，小嬰兒的語言學習能力比較強；探討字彙的學習是同時認知某一主題的許多詞彙，還是交錯學習不同主題的詞彙；甚至可以研究成人言談中的情緒會不會影響小孩的語言學習（例如激動時講的字眼，是否比較容易被學到）。

網路及電腦所產生的巨量資料，可以幫我們回答許多以前無法回答的問題，或是做出比較正確的決定。羅伊的研究就是一個最好的例子，他首先仔細定義出想要回答的問題，然後利用高科技設備蒐集巨量資料，對原始資料做適度的整理以及標記後，即可利用資訊科學上發展出來處理大量資料的方法深入解讀資料，進而找出問題的答案甚至發現之前未曾發現的現象。

在巨量資料充斥的時代，只要能夠學會傾聽資料、理解它們背後的含意，即使像是嬰兒這樣無法充份表達思想的個體，都有機會傳達給我們一些過去所不知道的知識。SA

林守德是台灣大學資訊工程系副教授。

影像來源：林守德