撰文／林守德

保護隱私的資料探勘法

模糊資料的精確度、隨機修改、分散儲存……，智慧型資料探勘正在崛起。

今年初，網路巨擘Google宣告新的隱私權政策，將整合旗下各式各樣的服務，如Gmail、YouTube、行事曆中的個人資料。這個從3月1日起生效的隱私權政策，引發許多人權團體的疑慮，因為它代表未來Google可以輕而易舉知道每個人在何時何地使用何種服務，並藉此進一步分析、追蹤每個人的網路行為。

對於隱私權可能的侵犯，一直是一般大眾質疑資料探勘研究的主要原因之一。2003年，美國參議院通過一項法令，禁止國防部利用資料探勘方法蒐集、分析一般民眾的資料，就是擔心人民隱私被這項技術侵犯的最好例子。然而，相關技術的研究學者認為，資料探勘的主要目的，是希望從大量資料中找出有用的資訊與知識，而非利用資料分析特定個體，本質上與坊間所謂的「人肉搜索」不同。例如探勘的目的是希望從賣場資料中，找出消費者可能會同時購買的商品，然後將這些資訊應用在商品陳列或是促銷機制上，提升顧客的購買慾望，並非對於特定某位顧客的購買行為進行分析。

儘管如此，為了因應侵犯隱私權的疑慮，資料探勘界開始發展「注重隱私保護的資料探勘機制」（privacy preserving data mining），希望能夠在不侵犯隱私的前提下，對資料做有用的分析。保護隱私最直覺的方法，就是把比較敏感的資料移除，例如把患者姓名以及身分證字號從醫療資料裡面移除或取代。但是這樣做不保證能夠完全保障個人隱私，因為有心人士還是可以藉由對照其他資料（如年齡、地址、學歷）與一些公開資訊來推測某些人的身分。為了解決這類問題，學者提出可以模糊敏感資料的精確度，例如將身高體重四捨五

資料探勘的主要目的，是希望從大量資料中找出有用的資訊與知識，而非利用資料分析特定個體。

入，或是用區間來表示（如160~170公分），甚至捨棄顯示過大或過小等特別有代表性的數字。此外，為了不讓其他人猜出某些資料欄位的意義，甚至可以將存在區間明顯的資料（例如生日皆為1~31），放大或縮小（如同步除以二）避免他人猜出含意。這類的方法統稱「資料抑制」（suppression）。當然，基於保護目的而修改過資料，有時會影響到之後探勘結果的精確度與實用性。

「資料提供者」希望提供「資料探勘者」已經隱蔽隱私的資料，但是卻希望這樣的處理不會妨礙找出有用知識的目的。於是一些「資料隨機化」（randomization）的方法應運而生，這類方法是利用特殊的方式修改資料，去除隱私資訊，卻盡量不妨礙探勘的結果。假設資料中「年齡」為隱私資料，我們可以先隨機產生一組平均為零的隨機數，然後將這個干擾訊號加到年齡這個項目上面，於是每個人的年齡都會被這個隨機數干擾。但是因為這個隨機數平均為零，我們可以在探勘的過程中，利用「平均」的概念將之消除，如此便不會影響探勘的結果。

另一類方法是利用「分散式儲存」的方式探勘資料，也就是把資料分散在不同的伺服器上面處理。比如：把年齡資料放在某個伺服器、姓名資料在另一個伺服器，而薪水、病歷等資料放在其他的伺服器裡。各伺服器之間則利用某些加密傳輸協定互相交換加密後的資料，以利從事後續探勘的工作。如此將資料分散的做法，可防止任何一個從事探勘的個體（伺服器）窺探到整個資料的全貌，達到保護隱私的目的。

在社群崛起、資訊爆炸的年代，人們對於隱私的重視程度與日俱增。而隱私保密的智慧型資料探勘，不僅代表一個新的研究方向，更是科技與公民人權共生的一個正面嘗試。 SA

林守德是台灣大學資訊工程系與資訊網路多媒體研究所副教授。

影像來源：林守德