撰文／林守德

重賞之下必有奇謀

越來越多的資料探勘競賽，邀請全球專家一起解決問題！

在網路普及的今天，短時間內獲得大量資料已非難事，如何利用這些龐大的資料，也成為重要議題。應運而生的研究領域「知識發現與資料探勘」（knowledge discovery and data mining, KDD），目的在於設計智慧型程式，自動從大量資料中找出有用的資訊與知識。近年來，資料探勘技術已廣泛應用在相關產業：例如推薦系統，能經由蒐集並分析使用者購買行為的資料，來推薦使用者有興趣的產品；又如歧異點探測技術，已被銀行用來偵測信用卡盜刷行為。

對於資料探勘這樣的應用學科而言，評斷某個技術成敗最直接的方式，就是測試它能否真的從資料中找出有用的知識。為了建立公平的評比機制，並吸引更多人從事相關研究，美國電腦學會（ACM）的資料探勘小組（SIGKDD）從1997年開始，規劃了一年一度的資料探勘大賽——KDD Cup。競賽的主題都是當下非常熱門的議題，吸引學界與業界數百甚至上千個團隊參加。KDD Cup不僅有學術上的挑戰性，背後更擁有龐大的商業應用價值。參賽者必須結合理論發展與程式撰寫，在約四個月的比賽期間內開發智慧型探勘技術與系統。

KDD Cup跟很多運動比賽一樣，每年都由不同單位競逐舉辦權，舉辦單位可根據自己的需求設計主題並提供資料。例如2008年由西門子公司獲得舉辦權，由於西門子設有醫療資訊研究部門，當年的主題即為醫療資料探勘，參賽者必須設計出能夠由胸腔影像資料判別患者是否可能罹患乳癌的方法。2009年由歐洲最大的電信公司Orange主辦，參賽隊伍要找出忠誠度低或是傾向選擇高價位服務的顧客。2010年的主軸是教育資料探勘，參賽者要利用線上學習的資料，判斷學生是否已經學會某種知識。2011年則由Yahoo!主辦，要利用Yahoo Music的資料進行音樂推薦。台灣大學資工系團隊在2008~11年間的KDD Cup比賽中獲得三次冠軍，已成為世界知名的資料探勘競賽團隊。

舉辦國際比賽，吸引全球專家參與解決難題，可能成為企業新一代的研發模式。

KDD Cup的初衷在於提供平台，讓不同的方法與技術一較長短。但是近年來許多相關企業發現，舉辦這類國際比賽有助於解決當前面臨的重要問題，而且遠比投入研究經費給相關學者來得有效益。舉辦比賽的獎金以及行政支出雖然不低，但相較於長期聘僱專業研究人員，仍屬九牛一毛；若與學界合作，因經費有限，只能讓少數幾個研究團隊參與，舉辦國際資料探勘大賽動輒吸引數百甚至數千隊伍參與，效益之懸殊不言可喻。

於是，在KDD Cup之外，近年來各式各樣的資料探勘競賽如雨後春筍般增加。最有名的當屬2006年美國著名的線上電影出租公司Netflix舉辦的「百萬美元電影推薦系統」比賽，目標是推薦消費者可能喜歡的電影。顯而易見的，這樣的系統準確度對於出租公司的市佔率影響很大，這個比賽歷時三年，吸引了全球上萬支隊伍報名參加。至今獎金最高的，則是加州的醫療服務公司Heritage Provider Network今年舉辦的比賽，獎金高達300萬美元。參賽者必須利用病患醫療相關資料，預測每個病患未來一年是否會住院以及住院的日數。如果能夠精確預測，該公司就能夠提早進行治療以及病房管理等規劃，不僅可以提升醫療品質，也能節省醫療成本。

資料探勘比賽已經蔚為風潮，越來越多公司希望藉由這樣的比賽讓全球專家幫忙解決問題。根據專門協助公司舉辦資料探勘比賽的團隊Kaggle表示，目前已有上千家公司向他們表示希望舉辦比賽，而資料探勘競賽將會如網球或高爾夫等體育競賽一般蓬勃，甚至可能出現世界排名以及邀請賽等賽制。資料探勘領域利用競賽加速其貢獻的時代，於焉展開。 SA

林守德是台灣大學資訊工程系副教授。

影像來源：林守德