林守德
台灣大學資訊工程系教授。

微軟學術圖的雄心與挑戰

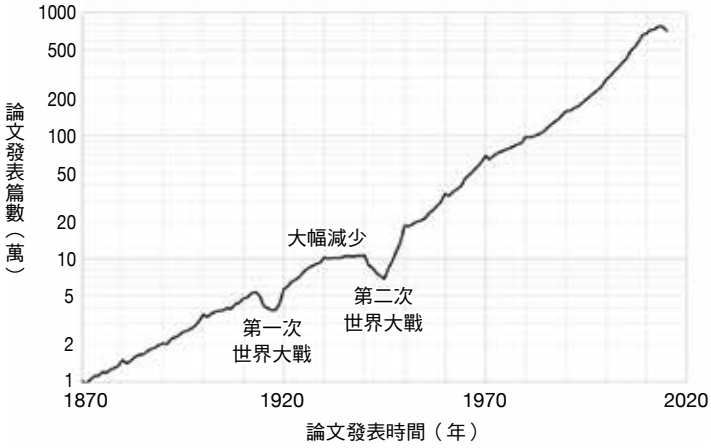
蒐羅、分析150年的科學發展趨勢，讓未來研究者在巨人的肩膀上站得更穩。

撰文／林守德

「知識就是力量」是我們從小就耳熟能詳的諺語，那如果把全世界成億上兆的知識集合起來，會帶來多大的力量？

微軟研究院研究員王冠三帶領的團隊正逐步探索這個問題。微軟在2015年公開釋出微軟學術圖（Microsoft Academic Graph）的資料，蒐集了1870年之後所有數位化的論文資料，包含作者、出處、引用、分類、關鍵字和摘要等。

這些資料構成一個很大的學術網絡：其中作者會跟其出版論文相連、論文與會議期刊相連、論文間又可以利用引用與被引用的關係相連；加上論文的關鍵字以及分類都有索引，同類的論文又可以在這個網絡中彼此連結。微軟學術圖至今有大約1億5000萬篇文章、30多億個連結、超過4000萬名作者，其中每三年至少發表一篇論文的作者數超過1000萬人。



分析這麼巨量的資料，可以看到科學發展的趨勢。王冠三就指出，從1870年起，研究論文篇數一直以指數成長，大約每隔10年就成長一倍；這代表科學知識可以自我繁殖，後人可以站在知識這個巨人的肩膀上，更快速累積新知與發明。唯二例外是兩次世界大戰之際，論文的產出量呈現負成長。微軟研究院以這個學術圖資料庫為基礎，推出「微軟學術」搜尋引擎服務，不僅能夠按照關鍵字搜尋，還能用更複雜的條件找到使用者想要的論文，例如可以搜尋某個作者在特定時間內發表在某個期刊上的論文。

除此以外，研究者也開始思考如何利用巨量的學術論文資料提供更有價值的服務，「學術聲望排名」即為其中之一。在2016年的「網路搜尋與資料探勘」（WSDM）國際會議舉辦了一場比賽，希望參賽隊伍能夠提出利用巨量學術資料來判斷論文影響力高低的演算法。如果能夠自動評估論文影響力，對於搜尋的排序就會有貢獻。這個比賽的冠軍隊伍來自台灣，他們發展出在「論文、作者、會議期刊」這些異質節點中遞迴影響的模型，來評估論文或是作者的影響力。所謂遞迴影響，舉例而言，就是「有影響力的作者發表有影響力的論文，而有影響力的論文發表於有影響力的期刊，有影響力的期刊出版有影響力作者的論文」這樣循環加乘的概念。

另一個實用的服務是學術論文推薦，學術搜尋網站可以利用使用者搜尋的記錄，預測使用者可能有興趣的論文。未來這類服務甚至可以擔任「圖書館員」的角色，推薦最符合使用者要求的文獻。

當然，這樣巨大的學術論文資料也為資料探勘帶來了挑戰。舉例來說，這些學術資料中人名、會議名、期刊名等專有名詞的寫法往往不盡相同（例如英文名，有時姓在名之前，有時則倒過來），如何在眾多的資料中找到對應，就是很大的挑戰。所幸可利用的資料很多，例如可借助論文中名詞的重複性、領域重合度等訊息來增加判斷的速度。

雖然人類產出的新知是以指數成長，但是並不代表我們「利用知識」產生力量的能力也以同樣的速度增長。如何使用這樣龐大的學術論文資源，創造能夠改變人類社會的價值，是未來值得我們關注的課題。SA

影像來源：林守德（上）、王冠三