網路蜘蛛與資訊安全

【研究紀要】

【文／教育資源及出版中心副研究員　朱麒華】

網路就像浩瀚無邊的寶藏，大家都在搜尋自己想要的訊息。有的人想掌握即時新聞、有的人想要追尋過去的歷史，也有的人只想知道有趣好玩的地方。這些不同的需求都可以在網路搜尋引擎上，找到一些答案，得到某種程度的滿足。網站越來越多，訊息的累積量，早已超出一般人的想像。如何在這麼多的網路訊息中，迅速地、正確地找到使用者的答案，成為現代人日常生活的需求。

為了在網站中搜尋資訊，電腦工程師開發了網路蜘蛛這一類的網路程式。網路蜘蛛來自英文 web spider、web crawler 及 web robot 的通稱，也有人稱為網路爬蟲或網路機器人。網路蜘蛛的目的就是走訪整個網站，記錄資訊並製作索引（web indexing），以利日後檢索的需求。它的行為就像蜘蛛一樣，爬滿整個網路，而且對於重要的網站常常到訪，整日編製索引，傳回資訊。這樣的行為能讓搜尋引擎提供快速的搜尋服務，但另一方面卻有資訊安全上的隱憂。

在現實生活中，許多網站提供短暫的服務，服務功能結束後，網站也關閉了。教育網站常見的「研習報名」及「考試放榜」就是其中的例子。在活動期間，網站顯示已報名研習的教師資訊或是榜單上的考生資料。雖然活動結束後，這些網站功能也已經關閉，但原來的資訊卻可以網路上搜尋的到，這就是網路蜘蛛的副作用。有些資料原來並沒有提供線上瀏覽，只提供下載使用，如：EXCEL 檔、WORD 檔。雖然當時的網站已不再線上提供這些檔案，可是搜尋引擎中卻有備份資料，這也是網路蜘蛛的功能之一。

網路蜘蛛對網站有優點也有缺點：優點是可以提高網站的能見度，全世界的網路使用者可以藉著搜索找到網站提供的資料，這是網路蜘蛛在中間扮演了重要的角色；另一方面，若是網站有些敏感的資料，縱使網站不再提供這些資料，但仍能在搜尋網站中找到副本。而且網路蜘蛛在網站中爬行，增加網路流量及網站負荷，影響了網站正常服務的效能。許多網站的流量統計常常與實際觀感不同，就是因為到訪者不是一般民眾，而是自動化的網路蜘蛛程式。

谷哥（Google）與雅虎（Yahoo）是世界上主要的網路公司，也是利用網路蜘蛛程式搜集資料的主要來源，但並非只有網路搜尋公司會使用網路蜘蛛，一些網路研究者及電腦駭客也會撰寫類似的程式，挖掘自己想要的資訊。為了瞭解網站執行的成效，良好的網站管理者需要常常檢視網站瀏覽記錄（web log），作為網站效能改善的依據。凡是網路蜘蛛爬行過的網站，都會在瀏覽記錄中留下痕跡。由於網站瀏覽記錄中，特定欄位--”使用者代理人”（user-agent）會紀錄網路蜘蛛的行為，因此只要觀察網站日誌檔，就可以知道哪些網路蜘蛛曾經到訪。以下四行資料摘自「愛學網」的日誌檔

◎157.55.33.77 stv.moe.edu.tw - [30/Dec/2012:05:04:53 +0800] "GET /?p=109829 HTTP/1.1" 200 20601 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)"

◎123.125.71.82 stv.moe.edu.tw - [30/Dec/2012:05:04:53 +0800] "GET /?p=211439 HTTP/1.1" 200 22769 "-" "Mozilla/5.0 (compatible; Baiduspider/2.0; +http://www.baidu.com/search/spider.html)"

◎66.249.77.97 stv.moe.edu.tw - [30/Dec/2012:05:14:30 +0800] "GET /?s=%E5%B0%8F%E5%AD%B8&from_cat=62839&paged=61 HTTP/1.1" 200 45970 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"

◎72.30.198.98 stv.moe.edu.tw - [30/Dec/2012:08:39:53 +0800] "GET /wp-content/plugins/wp-pagenavi/pagenavi-css.css?ver=2.70 HTTP/1.1" 200 374 "http://stv.moe.edu.tw/" "Mozilla/5.0 (compatible; Yahoo! Slurp/3.0; http://help.yahoo.com/help/us/ysearch/slurp) NOT Firefox/3.5"

我們可以發現世界上知名的搜尋引擎都派有網路蜘蛛程式到愛學網來，而且都是在去年的 12 月 30 日凌晨時間到訪。他們分別是雅虎的 Slurp、谷哥的 Googlebot、大陸最大的網路搜尋引擎--百度的 Baiduspider 及微軟的 bingbot。一方面顯示愛學網的資訊受到國際網路公司的重視，一方面也影響了網站正常的流量。

由於國際間已有上百種以上的網路蜘蛛程式，過多的程式將大大影響網站的效能。為了在搜尋便利性及資訊安全取得平衡，利用 robots.txt 檔案及<meta>標籤來阻擋或引導網路蜘蛛程式成為一種國際規範。但是這些規範也僅對「正派」的網路公司有用，對於企圖挖掘敏感資料為樂的駭客而言，任何標註不允許的地方反而成為另一種重要的指標。因此，網路管理者關切的不應只是網路的順暢，更應該關注服務的對象到底是「誰」。