

D16：實作 Day：Wiki的爬蟲實作練習



簡報閱讀



範例與作業



問題討論

Wikipedia爬蟲練習 >

重要知識點 >

重要知識點複習 >

解題時間 >

Wikipedia爬蟲練習



維基百科
自由的百科全書

首頁

分類索引

特色內容

新聞動態

近期變更

隨機條目

說明

說明

維基社群

方針與指引

互助客棧

知識問答

字詞轉換

IRC即時聊天

聯絡我們

關於維基百科

資助維基百科

列印/匯出

下載為 PDF

可列印版

條目 討論 臺灣正體 漢

閱讀 編輯 檢視歷史

搜尋維基百科

網路爬蟲 編輯

維基百科，自由的百科全書

！ 本條目存在以下問題，請協助**改善本條目**或在**討論頁**針對議題發表看法。

展開

網路爬蟲（英語：web crawler），也叫**網路蜘蛛**（spider），是一種用來自動瀏覽**全球資訊網**的**網路機器人**。其目的一般為編纂**網路索引**。

網路搜尋引擎等站點通過爬蟲軟體更新自身的**網站內容**或其對其他網站的索引。網路爬蟲可以將自己所存取的頁面儲存下來，以便搜尋引擎事後生成**索引**供用戶搜尋。

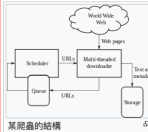
爬蟲存取網站的過程會消耗目標系統資源。不少網路系統並不允許爬蟲工作。因此在存取大量頁面時，爬蟲需要考慮到規劃、負載，這需要講「禮貌」。不願意被爬蟲存取、被爬蟲主人知曉的公開站點可以使用 robots.txt檔案之類的方法避免存取。這個檔案可以要求**機器人**只對網站的一部分進行索引，或完全不作處理。

網際網路上的頁面極多，即使是最大的爬蟲系統也無法做出完整的索引。因此在公元2000年之前的全球資訊網出現初期，搜尋引擎經常找不到多少相關結果。現在的搜尋引擎在這方面已經進步很多，能夠即刻給出高品質結果。

爬蟲還可以驗證**超連結**和HTML代碼，用於**網路抓取**（參見資料驅動編程）。

目錄 編輯

- 命名
- 概述
- 爬蟲策略
 - 選擇策略
 - 連結跟蹤限制
 - URL規格化



某爬蟲的結構

範例2：擷取文章中，延伸出的外部連結關鍵字。



維基百科
自由的百科全書

首頁

分類索引

特色內容

新聞動態

近期變更

隨機條目

說明

說明

維基社群

方針與指引

互助客棧

知識問答

字詞轉換

IRC即時聊天

聯絡我們

關於維基百科

資助維基百科

列印/匯出

下載為 PDF

可列印版

條目 討論 臺灣正體 漢

閱讀 編輯 檢視歷史

搜尋維基百科

網路爬蟲 編輯

維基百科，自由的百科全書

！ 本條目存在以下問題，請協助**改善本條目**或在**討論頁**針對**外部連結**發表看法。

展開

網路爬蟲（英語：web crawler），也叫**網路蜘蛛**（spider），是一種用來自動瀏覽**全球資訊網**的**網路機器人**。其目的一般為編纂**網路索引**。

網路搜尋引擎等站點通過爬蟲軟體更新自身的**網站內容**或其對其他網站的索引。網路爬蟲可以將自己所存取的頁面儲存下來，以便搜尋引擎事後生成**索引**供用戶搜尋。

爬蟲存取網站的過程會消耗目標系統資源。不少網路系統並不允許爬蟲工作。因此在存取大量頁面時，爬蟲需要考慮到規劃、負載，這需要講「禮貌」。不願意被爬蟲存取、被爬蟲主人知曉的公開站點可以使用 robots.txt檔案之類的方法避免存取。這個檔案可以要求**機器人**只對網站的一部分進行索引，或完全不作處理。

網際網路上的頁面極多，即使是最大的爬蟲系統也無法做出完整的索引。因此在公元2000年之前的全球資訊網出現初期，搜尋引擎經常找不到多少相關結果。現在的搜尋引擎在這方面已經進步很多，能夠即刻給出高品質結果。

爬蟲還可以驗證**超連結**和HTML代碼，用於**網路抓取**（參見資料驅動編程）。

目錄 編輯

- 命名
- 概述
- 爬蟲策略
 - 選擇策略
 - 連結跟蹤限制
 - URL規格化

外部連結



某爬蟲的結構

本文

重要知識點複習



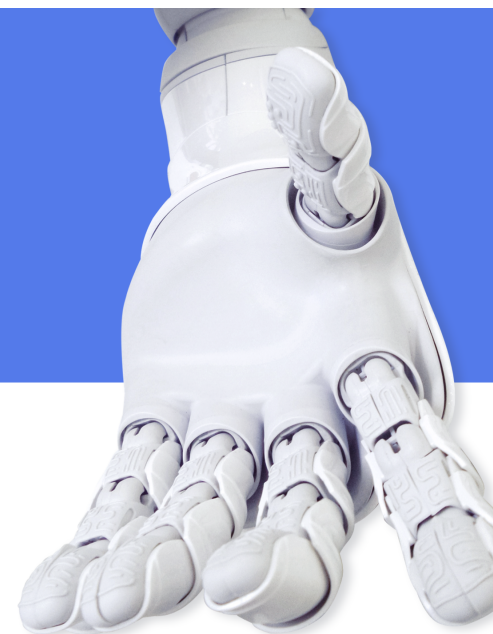
練習：定義一個爬蟲函數，重複前面兩個步驟的流程，可遞迴爬取更多關鍵字的解釋文章。其流程如下：

- 爬取當前關鍵字的解釋，並存入檔案(因為文章內容太多會佔滿整個頁面，所以存程檔案，方便後續檢視)。
- 萃取出當前關鍵字所引用的外部連結，當作新的查詢關鍵字。
- 把第(2)擷取到的關鍵字當作新的關鍵字，回到第(1)步，爬取新的關鍵字解釋。

解題時間 LET'S CRACK IT

Sample Code & 作業

開始解題



[下一步：閱讀範例與完成作業](#)