

Crawlzilla 安裝與 Demo 方法

一、Crawlzilla 安裝步驟：

【Step 1. 取得安裝檔】

- [取得 crawlzilla 最新安裝檔](#)
 - 目前最新：
<http://sourceforge.net/projects/crawlzilla/files/stable/Crawlzilla-0.2/Crawlzilla-0.2.2.tar.gz/download>

【Step 2. 解壓縮並執行安裝程式】

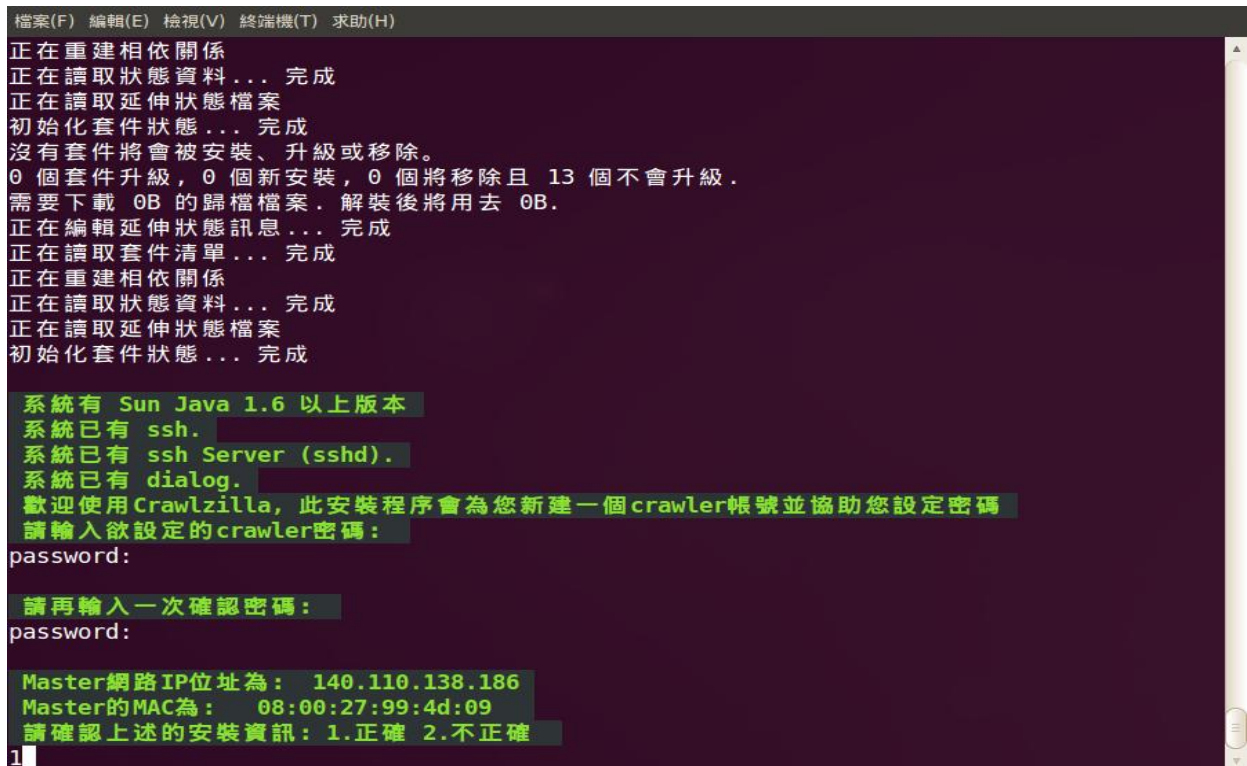
- 在下載的該資料夾，執行解壓縮與執行動作

```
tar zxvf Crawlzilla-0.2*.tar.gz
./Crawlzilla_Install/install
```

ps：此指令會切換成 sudoer，因此有可能會要您的 sudoer 密碼

【Step 3. 設定密碼及確認網路資訊】

- 設定密碼並確認網路狀態資訊後，等候完成安裝即可。畫面如下



```
檔案(F) 編輯(E) 檢視(V) 終端機(T) 求助(H)
正在重建相依關係
正在讀取狀態資料... 完成
正在讀取延伸狀態檔案
初始化套件狀態... 完成
沒有套件將會被安裝、升級或移除。
0 個套件升級，0 個新安裝，0 個將移除且 13 個不會升級。
需要下載 0B 的歸檔檔案。解裝後將用去 0B。
正在編輯延伸狀態訊息... 完成
正在讀取套件清單... 完成
正在重建相依關係
正在讀取狀態資料... 完成
正在讀取延伸狀態檔案
初始化套件狀態... 完成

系統有 Sun Java 1.6 以上版本
系統已有 ssh.
系統已有 ssh Server (sshd).
系統已有 dialog.
歡迎使用Crawlzilla, 此安裝程序會為您新建一個crawler帳號並協助您設定密碼
請輸入欲設定的crawler密碼:
password:

請再輸入一次確認密碼:
password:

Master網路IP位址為: 140.110.138.186
Master的MAC為: 08:00:27:99:4d:09
請確認上述的安裝資訊: 1.正確 2.不正確
1
```

- 待出現"恭喜您完成 Crawlzilla 安裝,按 Enter 鍵離開..."即表示單機環境已安裝完成！

【Step 4. 設定密碼及確認網路資訊】

安裝完成後開啓網頁將會顯示畫面如下：

在瀏覽器內輸入網址：<http://localhost:8080>

第一次需要設定網頁密碼



The screenshot shows the CrawlZilla web management interface in a browser window. The title bar says "CrawlZilla". The page has a blue header with the CrawlZilla logo and navigation tabs: "首頁", "爬網設定", "索引庫管理", "系統狀態", and "使用者設定". The main content area has a yellow background with a warning message: "這是你第一次登入安全考量，預設的密碼不該被使用". Below this is a form with three rows: "原密碼為" (Original password) with a masked input field, "新設定的密碼" (New password), and "確認新設定的密碼" (Confirm new password). At the bottom of the form are "送出" (Submit) and "重設" (Reset) buttons. On the right side, there is a sidebar with links: "搜尋引擎快速連結", "CrawlZilla 搜尋引擎範例", "相關資源", and "CrawlZilla@GoogleCode". The footer contains copyright information: "copyright © 2010 Free Software Lab@NCHC" and "Template provided by: DesignsByDarren.com". The browser status bar at the bottom shows "完成" (Done).

之後就可以用此網頁密碼，透過網頁介面使用 crawlzilla 的功能。

Ps :

這些步驟可以從 <http://code.google.com/p/crawlzilla/wiki/SystemInstall> 找到

安裝影片網址 http://code.google.com/p/crawlzilla/wiki/Install_video

二、Crawlzilla 使用步驟

網頁管理介面的使用方法：

<http://code.google.com/p/crawlzilla/wiki/WebManagement>

2.1 建立第一個搜尋引擎

【Step 1.至 Crawl 網頁中設定爬取項目】

- 依序填入：索引庫名稱，欲抓取的網址（可多行，如圖所示）及設定爬取深度即可送出

CrawlZilla 網頁管理介面

首頁 爬網設定 索引庫管理 系統狀態 使用者設定 登出系統

爬網設定

如何使用

索引庫名稱

索引庫名稱 tracCloud_and_nchcTW_3

輸入抓取的網址

http://trac.nchc.org.tw/cloud/
http://www.nchc.org.tw/tw/

輸入抓取的網址

抓取的深度設定

選擇抓取的深度 3

送出 重設

搜尋引擎快速連結

CrawlZilla 搜尋引擎範例

系統功能

修改管理者密碼

相關資源

CrawlZilla@GoogleCode

完成

CRAWLZILLA [HOME](#) [Crawl](#) [資料庫管理](#) [系統狀態](#) [登出系統](#)

Crawl Status


Crawl Setup Status

URL: <http://trac.nchc.org.tw/cloud/> <http://www.nchc.org.tw/tw/>
Depth: 3

Setup Success !!! But, it need time to crawl !!!

(ex. 4URLs with 1 depth -> 10~20 minute
4URLs with 2 depth -> 40~80 minute
100URLs with 10 depth -> very very long)

The Crawl speed is depend on your system performance, URLs number, and depth.



If you don't want to wait, click below link !!!
(1)[Crawl operation page](#)
(2)[Mian page](#)

完成

【Step 2.瀏覽網頁爬取進度】

- 透過系統狀態頁面，可即時了解網頁爬取進度

CrawlZilla 網頁管理介面

CRAWLZILLA [首頁](#) [爬網設定](#) [索引庫管理](#) [系統狀態](#) [使用者設定](#) [登出系統](#)

系統狀態

索引庫狀態

索引庫名稱	抓取狀態	刪除狀態
tracCloud_and_nchcTW_3	crawling	Delete

Jobtracker 工作排程器狀態 ([New Window](#))

Running Jobs

[Quick Links](#)

[none](#)

Completed Jobs

Jobid	Priority	User	Name	Map Con
job_201008181050_0001	NORMAL	crawler	inject tracCloud_and_nchcTW_3/urls	100

減

- 待出現"Finish"表示索引庫已建立，並可將此一訊息刪除



- 完成此步驟，第一個搜尋引擎已建置，右側快速連結中的"tracCloud_and_nchcTW_3"即為此次所建立的搜尋引擎。

【Step 3.測試搜尋引擎功能】

- 點選右側快速連結中的"tracCloud_and_nchcTW_3"進入搜尋引擎後，輸入一組關鍵字測試搜尋結果，下圖為輸入"nchc"為例：



->



2.2 使用索引庫

- 索引庫管理頁面中將會顯示目前已建立的所有索引庫，管理者可於此頁面進行瀏覽，刪除及提供網頁嵌入語法，如下圖所示：



【索引庫瀏覽】

進入索引庫管理頁面後，在欲瀏覽的索引庫欄位點選"preview"即可瀏覽此一索引庫的資訊，目前提供瀏覽的資訊包括：爬取網址、爬取文字數、爬取文件數、相關索引排名

由於有加入中文分詞功能，因此可以明顯看出索引庫的建立是以"中文字詞"作為基本單位

索引庫管理

索引庫名稱	建立時間	刪除索引庫	預覽統計資料	嵌入搜尋引擎到網頁的語法
nchc-en_3	2010-08-24 16:16:14	Delete	Preview	embed code
nchc-tw_3	2010-08-24 15:22:48	Delete	Preview	embed code

資料總覽

起始URL	http://www.nchc.org.tw/tw/		
本機索引路徑	/home/crawler/crawlzilla/archieve/nchc-tw_3/index		
總共文字數	37095	文件檔數量	1036
索引庫更新日期	Tue Aug 24 15:22:46 CST 2010	使用者名稱	crawler

被搜尋分析到的網址:

排序	內容	引用次數	排序	內容	引用次數
0	site:www.nchc.org.tw	336	1	site:pccluster.nchc.org.tw	87
2	site:bioinfo.nchc.org.tw	66	3	site:www.narl.org.tw	57
4	site:edu.nchc.org.tw	53	5	site:service.nchc.org.tw	35
6	site:accta.nchc.org.tw	28	7	site:colife.nchc.org.tw	14
8	site:wlanrc.nchc.org.tw	13	9	site:elib.nchc.org.tw	13
10	site:www.medicalgrid.org	13	11	site:volunteer.nchc.org.tw	9
12	site:www.stpi.org.tw	7	13	site:noc.twaren.net	7
14	site:ecogrid.nchc.org.tw	6	15	site:www.sipa.gov.tw	3
16	site:asp.104ehr.com.tw	3	17	site:viml.nchc.org.tw	3
18	site:www.ym.edu.tw	2	19	site:www.tnu.edu.tw	2
20	site:www.usc.edu.tw	2	21	site:www.ssvs.tp.edu.tw	2
22	site:www.smelearning.org.tw	2	23	site:ecocam.nchc.org.tw	2

完成

分析的文件型態:					
排序	內容	引用次數	排序	內容	引用次數
0	type:text/html	989	1	type:html	989
2	type:text	989	3	type:application	47
4	type:application/pdf	34	5	type:pdf	34
6	type:xml	10	7	type:application/xml	10
8	type:msword	3	9	type:application/msword	3
出現次數前五十分的字符:					
排序	內容	引用次數	排序	內容	引用次數
0	content:網	805	1	content:路	777
2	content:國	758	3	content:中心	750
4	content:計	744	5	content:資	742
6	content:與	740	7	content:訊	734
8	content:頁	712	9	content:電	705
10	content:學	698	11	content:算	696
12	content:家	692	13	content:的	684
14	content:關	676	15	content:議	674
16	content:統	666	17	content:1024	665
18	content:768	664	19	content:系	662
20	content:高速	648	21	content:一	643
22	content:號	636	23	content:區	635
24	content:站	633	25	content:導	632
26	content:解析	628	27	content:建	627
28	content:會	627	29	content:解析度	624
30	content:務	622	31	content:覽	618
32	content:請	614	33	content:發	614
34	content:體	608	35	content:上	607

完成

【索引庫刪除】

- 在欲刪除的索引庫中點選刪除，確認後即完成刪除索引庫

檔案 (F) 編輯 (E) 檢視 (V) 歷史 (S) 書籤 (B) 工具 (T) 說明 (H)

http://140.110.138.186:8080/crawlzilla/Statistics.do?fileName=r

Google 搜尋

CrawlZilla

CrawlZilla 網頁管理介面

CRAWLZILLA

首頁 爬網設定 索引庫管理 系統狀態 使用者設定 登出系統

索引庫管理

索引庫名稱	建立時間	刪除索引庫	預覽統計資料	嵌入搜尋引擎到網頁的語法
nchc-en_3	2010-08-24_16:16:14	Delete	Preview	embed code
nchc-tw_3	2010-08-24_15:22:48	Delete	Preview	embed code

資料總覽

起始URL	http://www.nchc.org.tw/tw/		
本機索引路徑	/home/crawler/crawlzilla/archieve/nchc-tw_3/index		
總共文字數	37095	文件檔數量	1036
索引庫更新日期	Tue Aug 24 15:22:46 CST 2010	使用者名稱	crawler

被搜尋分析到的網址:

排名	內容	引用次數	排名	內容	引用次數
----	----	------	----	----	------

完成

搜尋引擎快速連結

CrawlZilla 搜尋引擎範例

[nchc-en_3](#)

[nchc-tw_3](#)

系統功能

[修改管理者密碼](#)

相關資源

[CrawlZilla@GoogleCode](#)