國家高速網路與計算中心

2010 開放原始碼創新應用大賽

參賽紀錄報告

參賽隊伍：NCHC 自由軟體實驗室

參賽作品：Crawlzilla-打造你專屬的搜尋引擎

參賽成員：陳威宇　郭文傑　楊順發

評選日期：2010/11/02

9　　2010 開放原始碼創新應用開發大賽為資策會所舉辦針對國內各種自由/開放原始碼軟體所設計的比賽，參賽過程及作品介紹分別描述如下：

一、活動時程

日期	項目
08/31	報名及繳交作品企劃書截止日
10/15	作品收件截止日
11/02	評選會議
12/10	公佈名次及頒獎典禮

二、作品簡介(節錄自評選會議作品介紹)

開發目的
防火牆內的網路(內網)也需要自己的搜尋引擎，透過現有的搜尋引擎(如：Google, Yahoo…等)根本無法防火牆內的資料建立索引，因此必須自己建立搜尋引擎環境，而使用商用軟體的建置成本較高，利用自由軟體又因為建置的技術門檻太高讓許多使用者無法順利建置搜尋引擎，此一專案目標為提供與商用軟體一樣好用卻不需花費軟體成本及不需太高技術門檻的自由軟體搜尋引擎工具，協助使用者為內部網路管理人員或其他使用者建立自己的搜尋引擎環境。
功能簡介
Crawlzilla 主要提供三大核心功能，安裝、管理及操作，分別描述如下： ● **安裝**：交談式介面協助使用者將 Crawlzilla 安裝於系統中，已精簡所有系統設定步驟，使用者僅需設定一組密碼及確認網路資訊即可自動完成安裝。 ● **管理**：主要以 Dialog 介面做為系統管理介面，即使不在主機前，仍可透過遠端 SSH 進行系統管理，目前管理介面提供以下功能： – 檢查 Cluster 狀態：即時檢視目前系統及其他運算節點的服務狀態 – 快速啟動所有運算服務：一個按鍵即可啟動所有叢集中運算節點及網頁伺服器的服務 – 設定 Datanode & Tasktracker：可針對全部/部份運算節點做啟

動/關閉運算服務

- 設定 Namenode & Jobtracker：針對 Namenode & Jobtracker 做啟動/關閉服務
- 設定 Tomcat：提供啟動/停止/重新啟動 Tomcat 網頁伺服器選項
- 更改 Tomcat Port：修改 Tomcat Port
- 更換語言：目前系統提供繁體中文及英文，將陸續新增其他語言以提供不同語系的使用者使用
- Client 安裝步驟提示：提供快速指令供使用者可更快速的安裝於 Client 端

- **使用**：以 Web 介面呈現，使用者不需花費太多時間適應此一操作介面，目前主要提供以下功能：
  - 網頁爬取：使用者可在此一頁面執行網頁爬取，主要提供使用者編輯爬取網址、爬取深度及索引庫名稱(即搜尋引擎名稱)相關設定。
  - 索引庫管理：爬取完成後，系統會針對不同的搜尋引擎建立不同的索引庫，此一功能主要提供使用者檢視索引庫的資訊狀態，如：建立日期、爬取深度、爬取所花費的時間、索引庫詳細的資料(爬取到的文件數、熱門的索引值...等)，並提供刪除索引庫、更新索引庫等選項供使用者操作。
  - 系統狀態：此一網頁主要提供使用者即時檢視系統運作情行及運算節點狀態是否正常。

創新性

我們分別就安裝、管理及使用介面為三大核心功能描述不同的專案創新性如下：

- **快速佈署叢集運算環境：**挑戰最快速度為使用者佈署叢集運算環境，由於 Crawlzilla 底層是交由 Hadoop 做資料運算，叢集環境佈署過程較為複雜繁瑣，根據實際安裝統計，於正常的網路環境中安裝，平均約只花費 30 秒即可為叢集環境加入一個運算節點，可大幅減少安裝時間；此外，也可透過安裝 Crawlzilla，使用 Hadoop 處理其他需要運算的資料。
- **動態新增/刪除運算節點：**承如上述，叢集環境下的系統管理較為不易，此一專案特色可減少系統管理人員負擔，隨時可視現有硬體資源動態調整運算策略，提供更彈性的硬體使用規劃。
- **即時檢視及管理搜尋引擎索引庫內容：**目前現有的搜尋引擎工具必需使

用額外的軟體才能瀏覽搜尋引擎中索引庫的資料，Crawlzilla 內建索引庫瀏覽功能，讓索引庫資訊更加透明，此外，若索引庫過於老舊，可直接進行刪除，或按下重新爬取選項，即時更新索引庫資料。

應用或商業價值

Crawlzilla 主要為提供內部網路資料索引使用，除此之外也提供文件索引功能，可支援的文件如 pdf、ppt、oepnoffice…等文件，分別列舉兩個不同類型實用案例如下：

◎ **實用案例說明(一) – 文件索引工具：**

身兼十堂專業課程的熱血助教，因為某位同學漏聽一堂課，於期末考前請助教提供有關”OepnSource”的相關講義給他，由於助教平常有整理講義到教學網站的習慣，而也不知這位同學修了哪些課程，且 OpenSource 又與多堂課程有關，除此之外，還有”開原碼”、”自由軟體”…等相關詞彙，此時，透過 Crawlzilla 建立搜尋引擎即可對每個文件檔建立索引，並提供給需要的同學進行資料搜尋，取代 Ctrl+F 手動搜尋每個文件的步驟，也可精準的搜尋到所需資料。

◎ **實用案例說明(二) – 在對的索引庫找你要的資料：**

以 3C 資訊用品店家資訊網站為例，許多詳細產品資訊並無於店家網站提供，必須透過外部廠商網站才能找到詳細的產品資訊，透過 Crawlzilla 將所有廠商的網站整合於爬取清單，所建立的索引庫可以很精準搜尋到使用者所需的資料。

參、作品相關連結

- Crawlzilla @ Google Code Project Hosting (中文說明頁)
  http://code.google.com/p/crawlzilla/
- Crawlzilla @ Source Forge (Toturial in English)
  http://sourceforge.net/p/crawlzilla/home/
- Crawlzilla User Group @ Google
  http://groups.google.com/group/crawlzilla-user
- NCHC Cloud Computing Research Group
  http://trac.nchc.org.tw/cloud

四、評選會議

評選會議日期：2010/11/02

● 圖爲評選會議前於休息區做系統最後測試

評選會議過程由郭文傑做作品簡報，依序簡報內容為 Crawlzilla 簡介、相關應用、創新性、商業價值分析、系統架構及相關技術，最後展示作品相關功能，分別就安裝、管理及使用三大面向做為展示主軸，由於時間控制得宜，並補充 Crawlzilla 相關的額外資訊供評審參考，最後，評審亦提出相關問題結束評選會議發表。

● 圖爲參賽者於會場中合影

五、頒獎典禮

日期：2010/12/10

地點：世貿一館

得獎名單如下：

組別	名次	團隊名稱	成員名單	共計
學生組	冠軍	AndroidVG	指導教授：衛信文 老師、石維寬 老師 組員:陳碩鴻、林筱玫、陳易成、王恩儷、陳聖裕、洪徹易、賴政皜、陳增益、黃珮琪、李鈺文	12人
	特優	DM430	指導教授：陳彥仰老師 組員：郭冠宏、劉于凡、蕭方儀、張孟修、蔡典哲、陳冠名	7人
	優等	SDRC	杜秉穎、劉峻利、蔡明諺、曹昌盛、邵俊棋、尹培鑫、周楷傑、鄧閔仁、張淵智、江宜芳	10人
	佳作	我要打十個	曹祐嘉	1人
	佳作	NTUT-VCT	指導教授：楊士萱 老師 范瓊文、陳有成、朱嘉玲	4人
	百資科技贊助獎	源泉混混	施詠翔	1人
	百資科技贊助獎	酷爾斯迷思	丁君廷、趙仲尉、曹瑋桓、蔡志龍	4人
職業組	冠軍	NCHC自由軟體實驗室	郭文傑、陳威宇、楊順發	3人
	特優	Android-x86	黃志偉	1人
	優等	PCMan	洪任諭	1人
	佳作	BTPastry	吳偉碩	1人
	佳作	TTR	郭宗賢	1人

● 圖爲 **Crawlzilla** 團隊於頒獎典禮上合影

六、參賽心得

一年一度的開放原始碼創新應用大賽是自由軟體者的兵家必爭之地,在今年共三十多個隊伍的角逐之下,能在短短幾個月內開發出專案、不斷改進功能、修改錯誤、到端出成品參加比賽,直到得到首獎的那一刻都不感相信是真的,除了是團隊成員大家一起腦力激盪、互助合作下得到這個成果,也要歸功於副組長王耀聰與組長蕭志棋給我們技術上的指導與鼓勵,也感謝黃副主任與中心各單位的幫忙才能有此產出。

當然還要分享一下如何才能在比賽中脫穎而出：

1. 系統分析：要開發的產品最好是現階段最火熱的主題，才能讓評審第一印象就眼睛一亮；要已經有相關的知識，開發起來才能事半功倍。
2. 功能規劃：要將開發的專案依功能分析拆解成模組化，如此才能協同開發並容易除錯。
3. 簡易安裝：適用平台廣泛、安裝的過程要簡單，否則使用者裝到一半卡住就不玩了。
4. 簡單操作： UI 是使用者對此軟體溝通最重要的橋樑，因此操作 UI 一定要簡單明瞭、但使用者想知道的訊息要越詳細越好。
5. 官方網站： 官方網站除了要有中英文版之外，說明、安裝、使用的資訊都要完整，最好要能有系統畫面，使用者才能在第一時間就明白此專案的用途，並清楚是否使用及如何使用。
6. 盡善呈現： 入選決賽的口頭報告很重要，短短的 15 分鐘內要將整個專案都介紹過並實況 demo;報告者要口才最好最幽默、投影片做的最美麗，而實況 demo 要想好感動人心的方式，才能得到高的印象分數。
7. 不斷改進： 即使決賽的口頭報告結束，也要不斷地更新 svn 專案，把它當作自己的孩子一樣照顧。

以上分享完畢，希望能對之後參賽的人有幫助

威宇

七、參考連結

- 2010 開放原始碼創新應用開發大賽大會網站
  http://www.oss.org.tw/contest_2010/index.html