



National Applied Research Laboratories

National Center for High-Performance Computing

快速佈署叢集式搜尋引擎



陳威宇 郭文傑 楊順發

{ waue, rock , shunfa }@nchc.org.tw

國家高速網路與計算中心

自由軟體實驗室

August 14~15, 2010

COSCU 2010

建立專屬的搜尋引擎?

- 我的需求是...?

- 公司內部資訊網站為避免機密資料外流，需要建立內部搜尋引擎
- 個人文件可以透過自己的搜尋引擎建立索引
- 厭倦了商業搜尋引擎的搜尋結果(ex：廣告資訊)
- 某些情況下需花費許多時間過濾無意義的資訊
- 自己建一個搜尋引擎好像蠻酷的

建立專屬的搜尋引擎?

- 工具選擇

- 商業軟體 V.S. 自由軟體

	Goole企業版(<u>客戶</u>)	自由軟體
收費標準	18,000美金	免費!!! 元/公司
程式碼	不提供	稍高
系統操作技術門檻	較低	需自己修改程式碼
客製化	需洽相關技術人員	成本較低
系統維護成本	年費另計	



使用 Open Source 建搜尋引擎

- 優點

- 透明

- 擴充

- 隱私

- 客製化

免費!!!





我要選擇開源軟體建搜尋引擎!!!

NutchTutorial - Nutch Wiki - Mozilla Firefox

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

http://wiki.apache.org/nutch/NutchTutorial

最常用瀏覽 新手上路 即時新聞

waue/2009/0409 - Cloud Computing

Nutch Wiki 登入

NutchTutorial

首頁 最新更新 尋找頁面 說明 NutchTutorial

唯讀頁面 關於 附件 更多功能

Requirements

1. Java 1.4.x, either from Sun or IBM on Linux is preferred.
2. Apache's Tomcat 5.x or higher.
3. On Win32, cygwin, for shell support. (If you plan to use cygwin, you will need to have a shell that supports the 'set' command.)
4. Up to a gigabyte of free disk space, a high-speed network connection, and a good web browser.

Getting Started

First, you need to get a copy of the Nutch code. You can get it from the Nutch source code directory. Or, check out the latest source code from subversion.

Try the following command:

```
bin/nutch
```

This will display the documentation for the Nutch command.

Good! You are almost ready to crawl. You need to give the crawler some instructions. You can do this by editing the configuration file.

1. Open up \$NUTCH_HOME/conf/nutch-default.xml
2. Search for http.agent.name, and give it a value.
3. Optionally you may also set http.agent.url.

Now we're ready to crawl. There are two approaches to crawling:

1. Using the command-line to perform all the crawling.

完成

前言 環境

step 1 安裝好Hadoop叢集

step 2 下載與安裝

2.1 下載 nutch 並解壓縮

2.2 部署hadoop,nutch目錄結構

step 3 編輯設定檔

3.1 hadoop-env.sh

3.2 hadoop-site.xml

3.3 nutch-site.xml

3.4 slaves

3.5 crawl-urlfilter.txt

3.6 regex-urlfilter.txt

3.7 整個移植到另一台node

step 4 執行nutch

4.1 編輯url清單

4.2 上傳清單到HDFS

4.3 執行nutch crawl

step 5 瀏覽搜尋結果

5.1 安裝tomcat

5.1 tomcat server設定

5.3 下載crawl結果

5.4 設定nutch的搜尋引擎頁面到tomcat

5.5 設定搜尋引擎內容的來源路徑

5.6 啟動tomcat

step 6 享受結果

nutch

搜尋 標題 內文

0.9 requires Sun JDK 1.5 or higher.

you install, in the "Devel" category.)

Inpack the release and connect to its top-level

開始安裝

- 費了一番功夫後…
 - 技術文件有看沒有懂
 - 照著安裝步驟後安裝完了
 - 接下來呢…?
 - 如何驗證是否安裝成功?
 - 如何使用?
 - 怎麼解決錯誤訊息?
 -
 -
 -

糗大了 因為...

- 我想自己建搜尋引擎 又不想花錢 可是...
 - 技術文件看不懂
 - 不會寫程式
 - 系統維護是個問題

The logo for Crawlzilla is centered on the slide. It features a stylized spider with large, white, circular eyes and a black body, positioned on a yellow, geometric web structure. Below the spider, the word "CRAWLZILLA" is written in a large, bold, yellow font with a thick orange outline.

CRAWLZILLA

三個願望 一次滿足

Crawlzilla !?

• Crawlzilla 簡介

- 於2009推出實驗版，僅提供單機版本及基本功能
- 感謝約2,800次下載量的支持並提供寶貴意見
- 於2010加入叢集版本並命名為 Crawlzilla
- 提供簡單安裝及操作管理介面，輕鬆建立搜尋引擎的
套件工具
- 提供索引庫瀏覽功能，即時了解爬取結果



Crawlzilla

- **Crawlzilla 提供以下特色功能：**

- 簡易安裝
- 即時瀏覽資料庫資訊
- 支援叢集運算及顧全安全性
- 支援中文分詞功能
- 支援多工網頁爬取
- 解決中文亂碼及中文支援
- 網頁管理
- 多種語言
- 支援同時存在多個搜尋引擎
- 適用於泛Linux平台



Crawlzilla

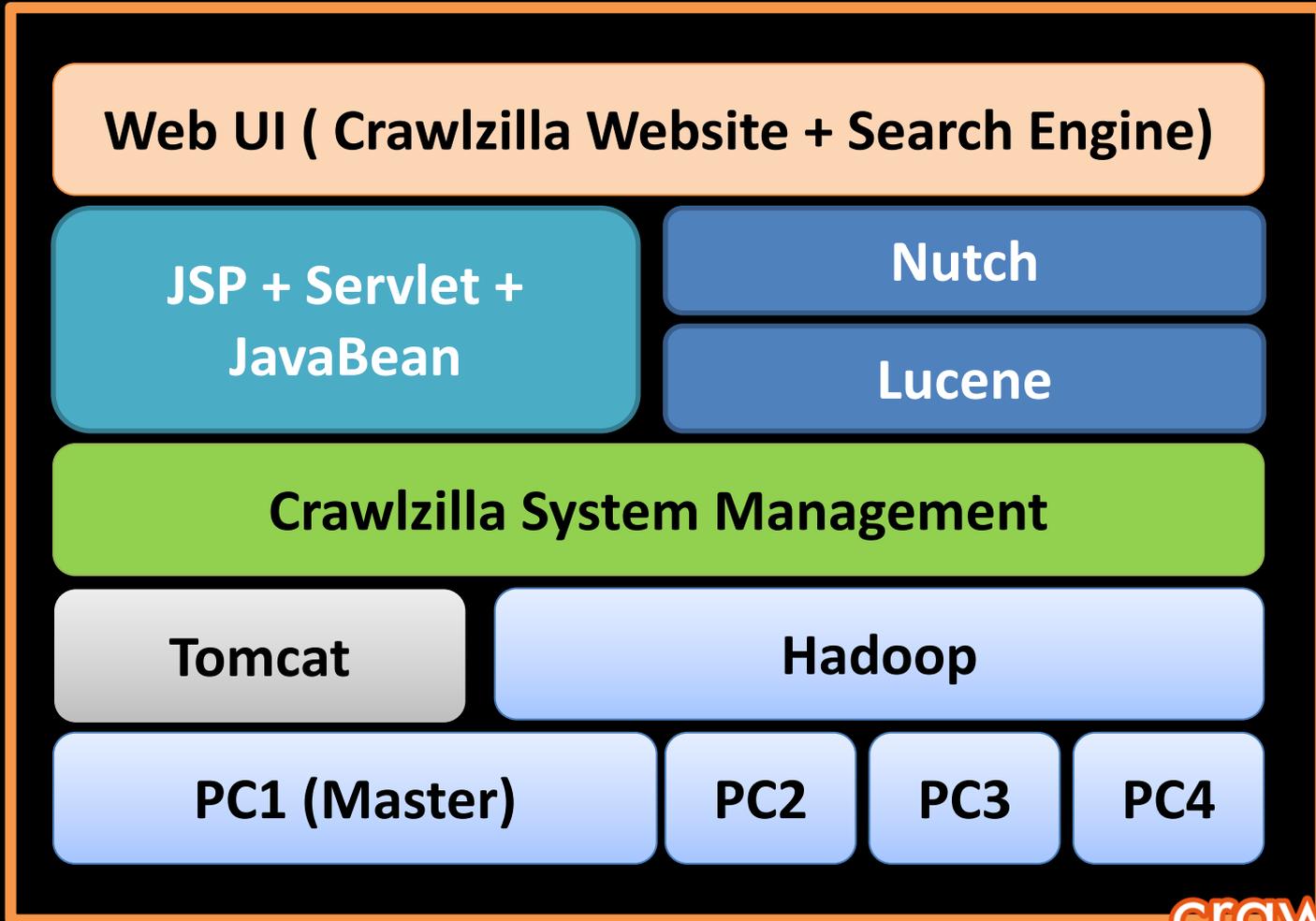
- Crawlzilla 組成

- 系統
- 管理
- 執行

- 開發者註解

- 目前Crawlzilla仍處開發階段，若系統出現不正常反應屬自然現象，不需要太恐慌並將問題回報給開發者改善後即可正常使用。

Crawlzilla 系統架構



Crawlzilla – 叢集系統需求!?

- 嫌一台電腦跑不夠快!?
- 閒置電腦太多!?
 - 讓多台電腦分工合作運算
- 架設叢集環境很麻煩?
 - 啟動Crawlzilla懶人叢集模式
 - 輕鬆安裝及移除
 - 可動態增加及刪除運算主機
 - 即時檢視叢集運算主機狀態



Crawlzilla 中文分詞

- 中文分詞是...?

- 以中英文字詞結構解釋：

- 英文以”詞”為文字的基本單位
- 中文以”字”為文字的基本單位

- 例：

- 英文 → I am a good student.
- 中文(無分詞時) → 我 | 是 | 一 | 位 | 好 | 學 | 生
- 中文(加入分詞) → 我 | 是 | 一位 | 好學 | 學生

Crawlzilla 加入中文分詞效果



簡介 常見問題

電影

Search help

Hits **1-11** (out of about **43** total matching pages):

[從《喜羊羊》到《唐山》中國電影 打造自己的阿凡達](#)

... 羊》到《唐山》中國電影 打造自己的阿凡 ... 童的心。中國電 ...

<http://www.cw.com.tw/article/index.jsp?id=41547> (cached) (explain) (anchors) (more from www.cw.com.tw)

[開心的城市就該有人跳舞 電影](#)

... 相關關鍵字: 電影 看了幾部街 ... 的街舞 (與街舞電 ...

<http://www.cw.com.tw/article/index.jsp?id=37344> (cached) (explain) (anchors) (more from www.cw.com.tw)

<http://blog.xuite.net/cwbook/blog>

... 台灣、娜娜、319鄉、電影、旅遊、美食、設計 ...

<http://blog.xuite.net/cwbook/blog> (cached) (explain) (anchors)

[天下雜誌: 行動綠生活, 台灣不碳氣!](#)

(25615 bytes) - [View as Plain Text](#)

... 他愛看藝術電影, 高職時迷上執 ... movie (專播藝術電影 ...

http://green.cw.com.tw/Words/wordcw15_1.aspx (cached) (explain) (anchors) (more from green.cw.com.tw)

[天下網路書店 - 紅皇后精神](#)

... 瑪雅預言都以電影呈現虛擬實 ...

<http://www.cwbook.com.tw/common/book.jsp?productID=4326> (cached) (explain) (anchors) (more from www.cwbook.com.tw)

<http://www.cwbook.com.tw/epaper/cwbook.jsp> (cached) (explain) (anchors) (more from www.cwbook.com.tw)



Why ik-analyzer?

	Paoding	indict	mmseg4j	ik-analyzer
使用者自定詞庫	支援	不支援	支援	支援
每秒處理文字 (基於官方介紹)	100萬字	約26萬字	15萬字~24萬字	50萬字
相關支援文件	很少,幾乎沒有	很少, 幾乎沒有	MMSeg演算法 說明文件	提供一個詳細的使用範 例及設定說明文件
其他	提供詞庫自動 更新檢測	基於ictclas 模組開發, 分詞效率佳	實作最多分詞, 但技術仍不成 熟,待改善	針對Lucene全文檢索提 供最佳化的查詢分析器- IKQueryParser,能提 高Lucene檢索的命中率

Crawlzilla網頁管理介面



The screenshot shows the Crawlzilla web management interface. At the top, there is a blue header with the title "CrawlZilla 網頁管理介面" and a navigation menu with buttons for "HOME", "Crawl", "資料庫管理", "系統狀態", "使用者設定", and "登出系統". Below the header, the main content area is divided into two columns. The left column, titled "管理介面功能介紹", lists several features: "抓取網頁設定", "資料庫管理", "系統狀態", "管理者設定", and "修改密碼". The right column contains a sidebar with sections for "搜尋引擎快速連結" (listing 3@books, 3@businessweekly, and 3@yahoo), "系統功能" (with a link for "修改管理者密碼"), and "相關資源" (with a link for "CrawlZilla@GoogleCode").

CrawlZilla 網頁管理介面

管理介面功能介紹

- * **抓取網頁設定** : 透過網頁介面建立搜尋引擎索引
- * **資料庫管理** : 查詢、刪除已爬取的索引資料庫
- * **系統狀態** : 瀏覽目前系統狀態
- * **管理者設定** : 包含個人化與語系設定
- * **修改密碼**

搜尋引擎快速連結

- CrawlZilla 搜尋引擎範例
- 3@books
- 3@businessweekly
- 3@yahoo

系統功能

- 修改管理者密碼

相關資源

- CrawlZilla@GoogleCode



Crawlzilla網頁管理介面 – 資料庫管理

資料總覽

起始URL	http://iso.nchc.org.tw/document/		
本機索引路徑	/home/crawler/crawlzilla/archieve/4@nchc-document/index		
總共文字數	13593	文件檔數量	303
資料庫更新日期	Wed Aug 11 23:38:41 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:www2.nchc.org.tw	130	1	site:iso.nchc.org.tw	96
2	site:www.nchc.org.tw	39	3	site:service.nchc.org.tw	20
4	site:itf.nchc.org.tw	4	5	site:edu.nchc.org.tw	2
6	site:intra.nchc.org.tw	2	7	site:www.medicalgrid.org	1
8	site:wlanrc.nchc.org.tw	1	9	site:ecogrid.nchc.org.tw	1
10	site:volunteer.nchc.org.tw	1	11	site:www.floodgrid.nchc.org.tw	1
12	site:elib.nchc.org.tw	1	13	site:www.narl.org.tw	1
14	site:pccluster.nchc.org.tw	1	15	site:colife.nchc.org.tw	1
16	site:bioinfo.nchc.org.tw	1			

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:application	202	1	type:application/pdf	200
2	type:pdf	200	3	type:html	101
4	type:text	101	5	type:text/html	101
6	type:msword	1	7	type:vnd.ms-powerpoint	1
8	type:application/vnd.ms-powerpoint	1	9	type:application/msword	1



Crawlzilla 叢集與伺服器管理介面

檔案(F) 編輯(E) 檢視(V) 終端機(T) 求助(H)

[管理功能選項]

請選擇：

cluster_status	檢查 Cluster 狀態
cluster_setup	設定 datanode & tasktracker
server_setup	設定 namenode & jobtracker
tomcat_switch	啟動/停止/重新啟動 Tomcat
tomcat_port	更改 Tomcat port
lang_switch	更換語言
client_install	Client 安裝步驟
exit	結束

< 確定 >

< 取消 >



Crawlzilla 系統 Demo

- 系統安裝 (Demo Video also on [Youtube](#))





Crawlzilla系統Demo

- 網頁操作
 - 執行網頁爬取任務
 - 資料庫管理
 - 搜索引擎展示
- 系統管理





Q & A





mail list

- 陳威宇 waue@nchc.org.tw
- 郭文傑 rock@nchc.org.tw
- 楊順發 shunfa@nchc.org.tw





相關資源

- **Crawlzilla @ Google Code Project Hosting**
 - <http://code.google.com/p/crawlzilla/>
- **NCHC Cloud Computing Research Group**
 - <http://trac.nchc.org.tw/cloud>

