Class Crawler課程網扒系統研發計畫

1. 計畫目標

研究與利用網扒(Web Crawler)技術，蒐集所有相關終身教育網站的課程資訊，提供報名等功能服務，規劃可行之合作機制，並依此目標訂定開發計畫，以供後續進度追踪與應用。

1. 計畫範疇

本階段擬鎖定文化大學、實踐大學、銘傳大學、世新大學、空中大學、師範大學、台北科技大學等七所學校之終身教育課程，進行搜集。未來可視需求，再新增目標學校。

1. Class Crawler系統架構

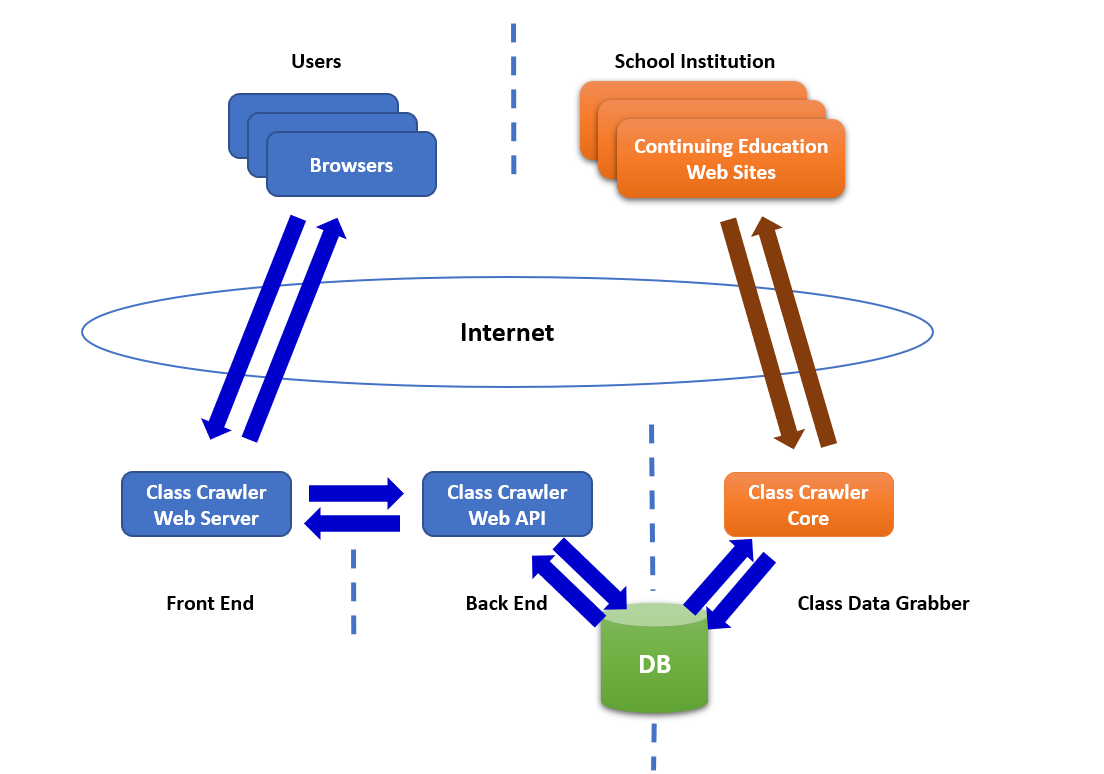


圖1: Class Crawler系統架構

如上圖所示，系統部份共可區分四大塊：Class Crawler Web Server (前端)、Class Crawler Web API (後端)、Class Crawler Core (網扒課程核心)、與資料庫端。其中，Class Crawler Core負責搜集目標終身教育網站之課程資料，經過疏理後，儲存於資料庫。最後，Class Crawler Web API再把相關課程資料，傳送到前端去呈現。

1. 合作機制

在初期為求快速導入，適合的應用模式宜仿照市面上的比價網、美食網等機制，作為各目標網站的總入口網站，然後透過流量的匯集，來爭取網路廣告收益。當此課程網扒系統之效益(亦即使用流量)愈趨明顯時，即可進行第二階段的深度整合模式(此須另訂開發計畫)。亦即，開始跟各推廣教育機構談每成交一筆課程之拆帳模式(因為有流量，才有跟其他組織談合作的空間)，同時系統也要加建購物車功能，並整合金流至本系統平台，於課程停止報名後，再將款項依所談條件，匯入至各推廣教育機構。

1. 開發環境與技術
2. 採用Web前後端分離方式開發
3. 參考網扒程式架構: DotnetCrawler  
   (https://github.com/mehmetozkaya/DotnetCrawler)
4. 採用XPath搜尋網頁字段
5. 採用HtmlAgilityPack擷取靜態網頁文字
6. 採用Selenium.PhantomJS擷取動態網頁文字(亦即，經javascript轉譯後之文字)
7. 開發語言: C# .Net Core 3.0
8. 資料庫: Open Source MariaDB
9. 開發計畫

|  |  |  |
| --- | --- | --- |
| 日期 | 階段目標 | 說明 |
| 2021/5/28 | Class Crawler系統之POC | 1. 搜集與實測適合的網扒技術 2. 確認系統技術架構 3. 以實踐大學為例，改寫1.所之網扒技術以能搜集終身教育課程資訊，並儲存於MariaDB |
| 2021/6/18 | Class Crawler後端API開發 | 1. 針對已儲存於MariaDB的資料，撰寫新增、刪除、修改、查詢之Web後端API，以提供未來Web前端之應用所需 |
| 2021/7/2 | 分析與實作文化大學之網頁課程資料擷取頁面 | 1. 針對文化大學終身教育課程網頁的特性，研究與設計能網扒該校網頁之邏輯與語法。 |
| 2021/7/16 | 分析與實作銘傳大學之網頁課程資料擷取頁面 | 1. 針對銘傳大學終身教育課程網頁的特性，研究與設計能網扒該校網頁之邏輯與語法。 |
| 2021/7/30 | 分析與實作世新大學之網頁課程資料擷取頁面 | 1. 針對世新大學終身教育課程網頁的特性，研究與設計能網扒該校網頁之邏輯與語法。 |
| 2021/8/13 | 分析與實作空中大學之網頁課程資料擷取頁面 | 1. 針對空中大學終身教育課程網頁的特性，研究與設計能網扒該校網頁之邏輯與語法。 |
| 2021/9/10 | 分析與實作師範大學之網頁課程資料擷取頁面 | 1. 針對師範大學終身教育課程網頁的特性，研究與設計能網扒該校網頁之邏輯與語法。 |
| 2021/9/24 | 分析與實作台北科技大學之網頁課程資料擷取頁面 | 1. 針對台北科技大學終身教育課程網頁的特性，研究與設計能網扒該校網頁之邏輯與語法。 |
| 2021/10/8 | 系統Web前端設計 | 1. 定義 class crawler系統前端流程 2. 設計所需之頁面 |
| 2021/11/12 | 開發Web前端 | 1. 根據10/8所定義的流程與頁面，撰寫前端程式 |
| 2021/12/31 | 測試與上線 | 1. 進行SIT、UAT測試，與上線使用 |