

資工系實務專題研究計畫成果報告

研究計畫成果摘要報告

專題編號：100-CSIE-S028

執行期限：99 年 1 學期至 100 年 1 學期

指導教授：王正豪 教授

專題計劃參與人員：黃驥、林佳志

一. 摘要

網路媒體越來越多元，而使用者的需求也越來越高，一般的新聞網站僅列出該篇新聞，至多附上同類或相關新聞，無法為使用者提供新聞中所出現人物的資訊。本專題利用資訊檢索，將文章內容之人名訊息列出，供使用者快速了解中所提到的人是誰，有哪些事蹟等等。

關鍵詞：新聞、檢索、Chrome Extension、斷章取義。

二. 緣由與目的

當一般使用者在閱讀一篇網路新聞時，往往在看到一些人名時，會想要再進一步了解“他/她是誰？”，但目前的網頁新聞僅會列出與該篇新聞同類型之其他新聞，並無法為使用者提供同步且即時的延伸訊息，甚為可惜。

本次專題即對此一問題提出改善，當使用者在閱讀新聞時，系統會將文章內容進行斷詞，分析找出人名，並做出相關之資訊檢索包含：照片、影片、維基簡介等，幫助使用者在最短的時間掌握最多的資訊，節省來回搜索的時間。

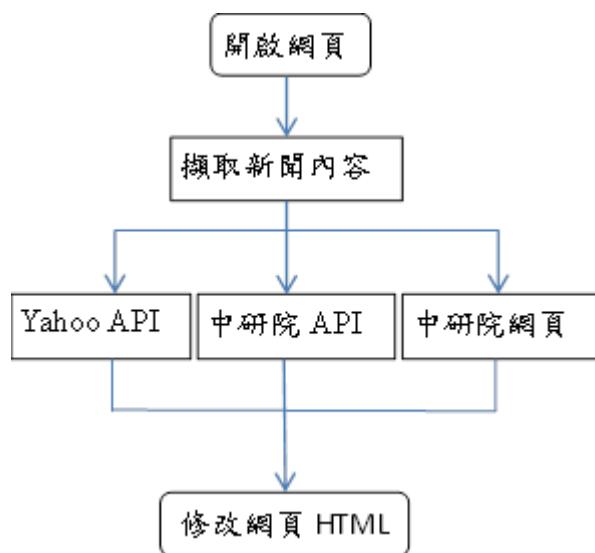
三. 系統功能簡介

本系統為開發 Chrome Extension，並結合 Yahoo 斷章取義 API、中研院斷詞系統、網頁分析並進行關鍵字之檢索，以提供相關訊息給使用者。

當使用者在瀏覽器安裝本套件後，在瀏覽新聞時，套件程式會抓取新聞內容文字，並將文章送往中研院斷詞系統以及 Yahoo

斷章取義 API，將結果交互檢查、提高準確度後，將文章人名部分標出，當使用者將滑鼠游標移到人名時，會另外跳出視窗顯示該人名的詳細資訊；若人名未被系統標出，則可利用套件的手動搜尋功能替代。

四. 程式使用流程



五. 實作重點

背景知識：

本次專題以 JavaScript、PHP 撰寫，Chrome Extension 的套件必須以 JavaScript 撰寫，且編碼須為 UTF-8，但中研院斷詞系統只接受 Big5 編碼，並為增加準確度及可靠度，同時利用三個系統，故額外撰寫 PHP 程式處理以上問題。

Yahoo 斷章取義 API：

在這個部分我們使用呼叫奇摩的 API 來對新聞進行分析，進而將這篇新聞中最高關聯度的關鍵字找出來，並回傳 xml 的檔案，讓我們可以將結果再作進一步與資

料庫中名詞比對，篩選出人名。

中研院斷詞系統：

中研院的斷詞系統又分為兩部分，其一為該系統的網頁展示版，其斷詞結果詞類較為詳細，但該系統並未提供 API，故利用程式 POST 資料過去並取回結果；另一部分為中研院提供的 API，但詞類較為精簡，故抓回來的結果要再和我們所建立的資料庫比對檢查。

替換文章：

三個系統斷完詞的結果送回來後，程式會改寫原本新聞內容的 HTML，在人名的地方加上要執行搜尋的標籤。當第一個結果收到後就會開始刷新，其它結果陸續收到後若有新的人名，會再刷新，不需等三個都收到才更改，以提高效率。

六. 軟體成果



(未使用本套件)



(使用本套件-將斷詞後的名詞加上標籤)



(游標移過去會跳出該人名的相關資訊)

七. 結論與未來展望

本專題希望使用者能利用本擴充套件，在閱讀新聞時，能夠快速地認識文章中所出現的人名。

在目前，由於斷詞中的地名詞彙只有不足 25%是屬於有效詞彙，故無法呈現地名的解析，希望在未來能夠有完整的地名解析方案，來增加新聞中名詞的完整度，進而增加套件的實用性。

八. 參考文獻

Chrome Extensions (Labs) :

<http://code.google.com/chrome/extensions/getstarted.html>

Yahoo 斷章取義 API :

<http://tw.developer.yahoo.com/cas/>

正規表示法-相關列表 :

<http://neural.cs.nthu.edu.tw/jang/books/webprog/03jscrip/reg3.asp?SessionCount=15>

PHP: JSON Functions – Manual:

<http://php.net/manual/en/ref.json.php>

YouTube APIs and Tools:

http://code.google.com/intl/zh-TW/apis/youtube/2.0/developers_guide_php.html

Facebook Developers:

<http://developers.facebook.com/docs/>