

資工系實務專題研究計畫成果報告

中文專有名詞辨識系統

Chinese Named Entity Recognition System

專題編號：096CSIE-S024

執行期限：95 年 1 學期至 96 年 1 學期

指導教授：王正豪 教授

專題計劃參與人員：林承翰 陳玉雯

一、摘要

本專題設計了一介面提供使用者辨識中文的文章裡出現的人名、地名、國家名、以及各種詞類以便於使用者更方便了解文章或網頁內容，呈現方式則針對詞類以不同的方式表達。中文斷詞部分使用中研院研發的「中央研究院資訊科學所詞庫小組中文斷詞線上服務」來對中文文章做處理，然後再針對回傳的資訊做分類。例如地名、國家名部分使用 Google Map 標示位置；人名以及各種詞類用 Wikipedia 連結提供解釋。

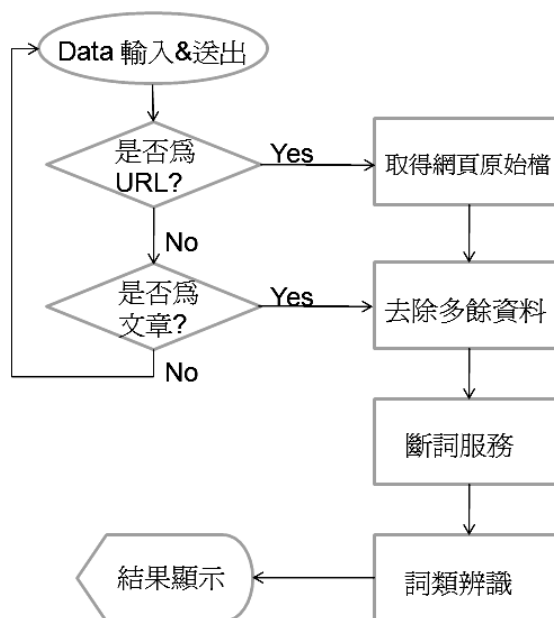
關鍵詞：

Named Entity Recognition、Google Map、Wikipedia、CKIP、PHP。

二、緣由與目的

網路發展自今，資訊成長速度從爆發式演變成倍增式的增加，各種資訊充斥於網路中，如部落格的日記到介紹文章、討論區的詢問到精闢探討、專業網站評論、學術文獻等。搜尋並掌握適切的資訊，已成為現今科學研究的基石；而便於使用的資訊搜尋系統，更是奠定基石的背後功臣。但反觀現今的搜尋系統，仍以英文搜尋為主，針對特定地區與文化進行資訊辨識的系統仍依需求而設計；本專題則以中文專有名詞為辨識題目，針對文章或網頁中的中文人名、地名、國家名稱、各類詞性等進行分類與辨識以提供使用者容易了解文章/網頁內容。

三、系統架構



(流程圖)

系統架構分為四個部份：

(一) 資料輸入

資料來源：可以是文章或者是網址

資料格式：繁體中文+英文+符號組成

因中研院中文斷詞線上服務不支援簡體中文，所以無法辨識。欲斷詞的資料在送出前會去除符號、英文並且限制在 600 個由 UTF-8 組成的字以內，以增加斷詞服務的處理效率。

(二) 斷詞服務

將處理過的文字送到中文斷詞線上服務交大 mirror 站做斷詞服務，因交大 mirror 站需要處理的資料量龐大，所以資料回傳的時間在三分鐘內不等的時間，此系統花費最多的時間即為此處。

(三) 詞類辨識

將斷詞系統回傳的結果依照詞類 Tag 判斷，並分別與資料庫中的資料做比對。地名部分先對曾查詢過的地名資料庫做比對，才對整個地名資料庫比對；人名部分先判斷此名詞是否為二到三字組成，然後拿第一個字到百家姓資料庫中比對；其餘部分則依據所屬的 Tag 做分類。

(四) 結果顯示

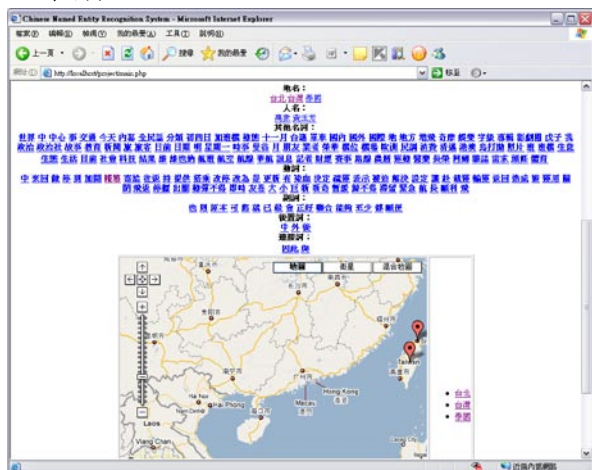
分類後的結果是以 wikipedia 超連結方式分別顯示，地名部分額外在底下用 Google map 標示位置，除了地名、人名、其他名詞外的詞類並沒有直接顯示出來，必須再按下顯示其他分類結果後才會顯示。

四、成果

(一) 輸入網址作為辨識來源

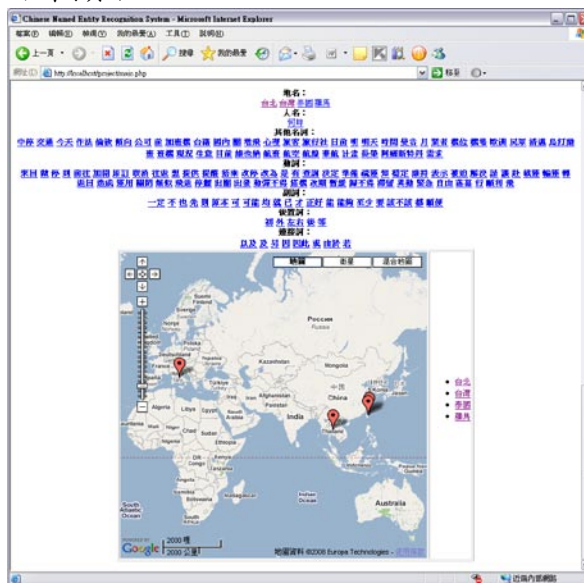
<http://tw.news.yahoo.com/article/url/d/a/081130/2/1aced.html>

結果顯示：



(二) 輸入文章作為辨識來源

文章以(一)網址內的新聞做為輸入。結果顯示：



五、結論

本系統提供使用者能更快速的了解一篇中文的文章與網站，系統結合 Google Map 讓使用者知道發生的地點，連結到 Wikipedia 與 Google Define 更深入了解使用者不明白的專有名詞。未來可增加時間軸來對多篇文章做發生時間分類，以及支援多國語言同時處理。

六、相關技術與工具

程式語言：Html, PHP5, Java script

Web Server：Apache + MySQL

斷詞分析：中研院線上中文斷詞服務

地圖顯示：Google Map API

七、參考資料

網站：

[1] PHP Example and Source Code.

<http://phpbb-tw.net/phpbb/>

[2] CKIP Client Example

<http://ckipclient.sourceforge.net/>

[3] Google Map Client Example

<http://www.phpinsider.com/php/code/GoogleMapAPI/>

[4] PHP manual

<http://en.php-resource.de/handbuch/index.htm>

