

資工系實務專題研究計畫成果報告

專題編號：100CSIE- S027

執行期限：99 年 1 學期至 100 年 1 學期

指導教授：王正豪

專題參與人員：97590451 楊俊仁

一、 中文摘要

不管您是要找休閒人文或是藝文展覽，還是適合全家人一起參與的親子活動，或是好奇最近有什麼特賣會活動可以撿便宜，都能提供您所在地區的最新活動資訊。並提供該活動地點之週邊相關生活景點，讓您參加活動之餘能順便體會該地方之風俗民情。

本組利用爬蟲來蒐集資訊、並以群聚分群方式[1]來匯整網路上各種活動資訊來源：包含活動地點、內容、性質，提供在 Android 手機上，並利用相關 API 使其內容跟豐富化與更強大的地理關聯滿足使用者對資訊充足的訴求。支援行事曆整合功能，讓您可以立即將有興趣的活動通通加入行事曆中！或者分享至 Facebook 中。

關鍵詞：

Crawler、Android, Google APIs,
Yahoo APIs, Facebook APIs,
CKIP API,
TCP/IP programming,
Hierarchical clustering

二、 緣由與目的

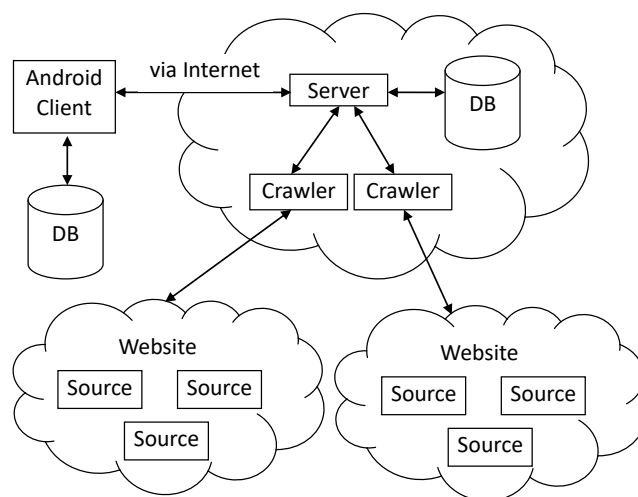
不知道您有沒發生過下面的情況：週休二日時，與家人外出吃餐聚，回程時看到不遠的週遭似乎有大型戶外活動，坐在車內的您，好奇、打量、端詳著，這是什麼樣的活動？免費嗎？現在去會不會要結束了？好像好有趣呢？在種種可能考量情況下，往往綠燈一來，油門一踩，心想算了，離開了。或許因此錯過一次與家人活動之回憶。

為了避免上述之憾事，若能在出門前，手機一輸入要餐聚或與家人活動地點，馬上得知週遭是否有正在進行的活動，能讓您戶外活動能夠延續。讓您的生活不再單調，而是活在資訊充足的當下。

三、系統簡介

本系統希望以更人性化[2]的介面並提供更多元的資訊訊息給 android 手機用戶。

本系架構如下：



在上圖中，您可以看到 Server 端定期透過 Crawler 將各 website 的有用資訊經由一定篩選、分類、匯整後 insert/update 至 DB 中。而 Android clients 您可以看到本身也有獨立一個 DB 而主要作用是用來同步 Server 端之 DB，此做法是使 user 能夠在無網路的環境下，也能夠離線查看到各種活動資訊。

接下來，將介紹期末已完成之重點功能

四、實作重點

Server 功能實作：這部分在此作業流程中運算量最大的部分，分別為：

(1)**活動爬蟲：**可依管理員的設定而彈性化設定取得各項活動細節

(2)地理周圍資訊取得:將特地資料之位置分別利用 Google Geocoding API 及 Yahoo LifeStyle API 將周圍活動資料整合及存入資料庫中

(3)活動分群演算法實作:使用聚合式階層分群法(HAC)方式,先將活動資訊之名詞取出,將個名詞樹狀結構的底部開始層層聚合,依據其同類型名詞出現頻率最為主要合併因素,因其同一群特性與文件內重反覆出現之名詞可知其重要性。運算完後將其活動分群結果存入資料庫中。

(4)提供 client 之資訊請求:由於目前的 android 並無法直接與 MYSQL Server 做直接溝通傳遞資訊,因此本組製作了 MYSQL 代理人,作為 Android 與 Server 之通訊橋梁,負責處理資料庫之傳遞功能。

Client 端功能實作:主要為將取得的資料人性的呈現給使用者,細說如下:

(1)與遠端伺服器活動資料同步:透過 MYSQL 代理人而取得 SQL Server 端資料,並將取得的活動各項資訊存入 SQLite 資料庫中,如此一來即可在沒有網路的情境下進行活動檢索。

(2)Android 端各功能實作:這部分包含活動資訊檢索之呈現、結合 Facebook 發佈、整合行事曆、活動地點地圖呈現及周遭景點呈現、分群活動結果之顯現、進階搜尋活動等。〈此部分詳細說明於簡報上〉

五、主要使用技術與工具

JavaHtml Parser:用來解析 Html 檔案,並加以轉換、抽取正文。

Yahoo LifeStyle API:將會提供相關的商家搜尋功能。利用關鍵字以及位置搜尋附近商家,或僅使用關鍵字查詢相關的商店資料。

Google APIs:(1)Map:用來獲取經緯度座標(並將其直給 Yahoo LifeStyle API,以及呈現於手機上面)(2)Calendar:同步遠端的行事曆(3)Geocoding:將地標轉成 GeoPoint 以供其他相依 API 需求。

同步兩端資料庫:透過 Java.net 相關套件實作 TCP 程式部分

CKIP 斷詞系統:用於將文章內的名詞分別抽取出,再將其統計其頻率之結果作為 HAC 運算的分群依據。

六、未來預計完成目標

目前想要進一步蒐集使用者的操作資訊,將這些點選活動、查詢、加入日歷作整合分析,類似作資料探勘應用,以便作日後相關參考使用,例如將更適合使用者的廣告置入,以增加廣告點取之機率。

七、註解

- [1] 先採用文件取出的特徵計算出來的相似度值進行初步群集分析,本研究將採用階層式群集分析方法 – hierarchical agglomerative clustering 演算法進行初步群集分析。此方法基本的概念是:初始將每個文件視為各自一群,每回合將最相似的兩群文件合併成同一群集。反覆重複直到最後將所有文件分群完。
- [2] 為了補足網站資訊的不足性、較弱的地方周遭資訊性、手機用戶操作方便性。
- [3] 一個台灣知名開放性活動網站,提供各式活動資訊來源:藝文、休閒人文、展覽、消費好康等。

八、參考文獻

- [1] 自己動手寫網路爬蟲
作者:羅剛、王振東,清華大學出版社
- [2] Android SDK 開發大全
作者:余志龍、陳昱勛等五人
悅知文化出版社
- [3] Android 案例開發完全講義
作者:李寧 基峯出版社
- [4] 統計式中文新聞摘要
作者:楊政遠 台北科技大學碩士論文
- [5] 網路程式設計初學指引-使用 Java
作者:湯秉翰 博碩文化出版社
- [6] Yoshida, M., Ikeda, M., Ono, S., Sato, I., and Nakagawa, H. Person name disambiguation by bootstrapping. In Proceedings of SIGIR. 2010, 10-17.

