國立臺北大學電機工程學系
103 學年度學生專題製作計畫書

運輸節能之資料探勘

Data Mining of Transportation and energy-saving

組員：

學號：410185014　　姓名：曾柏諺

指導老師：林嘉淦　老師

中　華　民　國　１０３　年　１０　月　２６　日

壹、計畫摘要

本計畫已資料探勘之方式，將龐大的資料庫取其有用之資源。根據納智捷公司所題共的資料庫，使用 PHP、HTML 網頁程式技術，使一千多筆的資料轉為 execl 檔案，已至其他工程師能夠直接取用。

貳、背景及目的

由於現今服務業眾多，公司都會收集客戶的資料，以及一些消費行為等，去做評估以及規劃。例如：7-11 大型連鎖企業，店員會紀錄每個客人消費的時間、物品、地點等，對於每家分店做出不同的物流規劃。而在這龐大的資料中，我們必須取捨其中的資料，做出正確的數據來判斷。

資料挖掘是因為海量有用資料快速增長的產物。使用計算機進行歷史資料分析，1960 年代數字方式採集資料已經實現。1980 年代，關聯式資料庫隨著能夠適應動態按需分析資料的結構化查詢語言發展起來。資料倉儲開始用來存儲大量的資料。

因為面臨處理資料庫中大量資料的挑戰，於是資料挖掘應運而生，對於這些問題，它的主要方法是資料統計分析和人工智慧搜尋技術。

文本挖掘(Text mining)，也稱為文本數據挖掘，大致相當於文本分析。高質量的信息通常是通過手段的模式和趨勢的制定衍生而來，例如統計模式學習。文本挖掘通常涉及構成該輸入文本，所述結構化數據中推導模式的過程中，最後的評估和解釋的輸出。“高品質”的文本挖掘通常是指相關性，新穎性和趣味性的某種組合。典型的文本挖掘任務包括文本分類，文本聚類，概念/實體提取，生產粒狀分類法，情感分析，文獻綜述，以及實體關係模型。

此專題目的是在將網路上的資料抓取下來，做完統整在交給公司做客戶後端的處理。由於這是一個新的資料庫，所有資料進來才能在進一步做其他處理，因此這過程式必須的。

參、研究方法及進行步驟

(劃圖)

1. 其系統可分為前端與後端處理。前端包括網頁設計，廠商、學校、法人如何註冊納智捷的會員，與其一起合作開方案件。而後端處理的是他們申請的資料，如何做好會整的工作。例如：每一次開會，便可以從上次的會議記錄去尋找，如此一來便不必耗費心神，去擔憂哪些事項已經處理過，而不用在討論了。
2. 為此到達目的，我們使用 php 以及 html 語法，透過 xampp 軟體(此包含 Apache、Mysql、Filezilla 等)寫入程式語言在根目錄執行，便能將資料抓下來做匯整。
3. 因此，必須先熟悉 php 的語法，並且先對其他網頁資料做測試。目前我們先把所有要的字串傳回 Mysql 資料庫，再將之取出轉成 execl 檔。

4. 由於這資料庫是屬於公司內部，因此必須到公司內部的網路 IP 才能取得相關資料，並且在抓取的時候，還被內部擋了下來。
   所以我們從原本直接取，間接轉成先存取到 Mysql 的方式，看此方法是否可行。
5. 網路上有介紹修改 php.ini，其中檔案裡有兩個選項:allow_url_fopen =on(表示可以通過 url 打開遠程文件)，user_agent="PHP"（表示通過哪種腳本訪問網絡，默認前面有個" ; " 去掉即可。）重啟遊覽器即可。
   但是有些還是會有這個警告，那就得設置 php.ini 裡面的 user_agent，php 默認的 user_agent 是 PHP，我們把它改成 Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0) 來模擬瀏覽器就可以了
   user_agent=”Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)”

肆、儀器設備需求表

使用者自備筆記型電腦或桌上型電腦即可

伍、預期完成之工作項目及具體成果

1. 在九、十月份取出公司其中一筆資料，並對其做深入研究，並且觀察其他網站的結構，嘗試以 PHP 抓取資料下來。
2. 現在對於龐大的資料庫中，我們學到能將之如何取資料下來，並且能夠做彙整資料的處理
3. 除了學校有 C 與 C++的課程訓練之外，透過這次專題讓我學到使用資料庫的方式，以及如何使用 PHP 和 HTML 的應用。

陸、預定進度甘梯圖

月 次 工作項目	第 1 月	第 2 月	第 3 月	第 4 月	第 5 月	第 6 月	第 7 月	第 8 月	第 9 月	第 10 月	第 11 月	第 12 月
R 和 MYSQL 學習	∨	∨										
PHP 和 HTML 學習		∨	∨									
PHP 計算機		∨	∨									
網路爬蟲		∨	∨	∨	∨							
進度累計百分比(%)												

柒、參考文獻

[1] http://en.wikipedia.org/wiki/Text_mining