應用資料探勘分析訂房網站評論資料之研究

林建成[1]，蔡燿旭[2]，*林建成
[1,2]旅遊與觀光學系,中華大學,新竹市,中華民國
*旅遊與觀光學系,中華大學, 新竹市,中華民國

1. 研究背景與目的

網路爬蟲在資料搜索和資料探勘過程中扮演著重要的角色，對網路爬蟲的研究開始於上個世紀就開始，目前爬蟲技術已趨於成熟。網路爬蟲通過自動抓取網頁的方式完成下載網頁的工作，實現大規模資料的下載，省去諸多人工繁瑣的工作。基於上述研究背景，本研究希望透過python 語言編寫網路爬蟲，來分析網路上的訂房網站內遊客評論資料，進而分析遊客評論對於訂房網站評論所給予的評分和權重。綜合上述，本研究希望能達成以下目的：
一、整理過去於這個領域所提出的研究和文獻，讓之後的研究者可以參考。
二、應用python 網路爬蟲抓取網站上旅遊評論，進而分析使用者評論的評分和模擬使用爬蟲程式及人工抓取評論資料數量及所消耗的時間。

2. 研究方法

本研究選擇了一個非常常見的資料探勘目標。透過各個方向對該此目標進行分析、比較、說明。特別說明，本研究中所有資料均來自於網際網路上公開資料，無需登入就可以查訊閱覽。本研究的資料來源是：booking中文官方網站，網址為：https://www.booking.com/。
Scrapy是一個為了爬取網站資料、擷取結構性資料而撰寫的應用框架。在這個框架之下只需要寫入較少的原始碼就可以進行快速的抓取網頁資料。Scrapy使用了Twisted非同步網路框架，Twisted是一個用python語言寫的網路框架，支援很多種網路協議，包括UDP,TCP,TLS和其他應用協議可以加速網路資料的爬取速度。

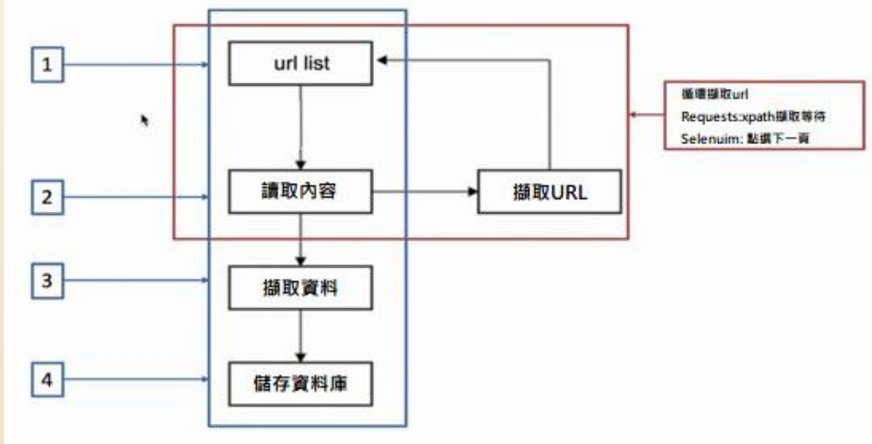

圖1：爬蟲流程圖

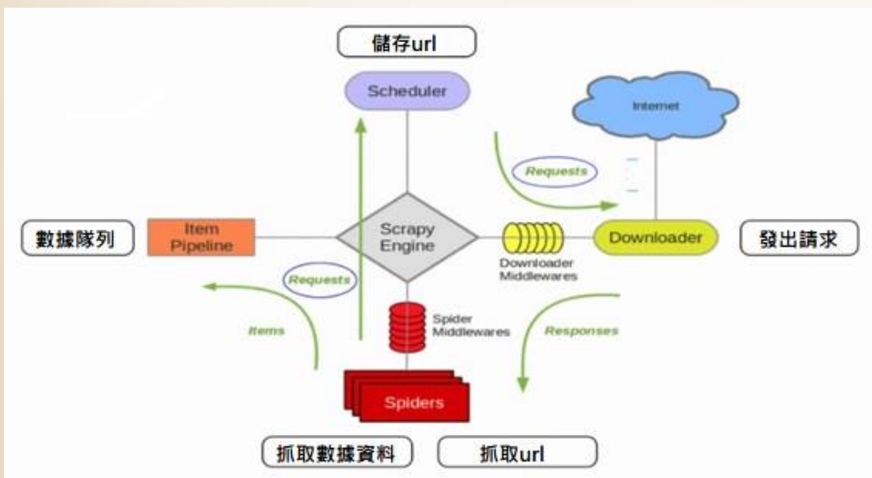

圖2：Scrapy流程圖

Scrapy engine (核心)	控制中心：負責數據資料和訊息在不同模組間的傳遞。
Scheduler (調節器)	排列順序：存放從控制中心發送過來的request請求。
Downloader (下載器)	下載數據資料：下載從控制中心發送過來的request 請求，並送回控制中心。
Spider (爬蟲)	處理數據資料：處理從控制中心發送過來的request內容，並讀取數據及url，送回控制中心。
Item Pipeline (通道)	處理數據資料：處理從控制中心發送過來數據資料，逕行儲存。
Downloader Middleware (下載中介)	設定下載中介代理：可以自行定義下載的方式，例如可以設定代理IP下載。
Spider MiddlewareSpider (爬蟲中介)	設定爬蟲中介代理：可以自行定義requests 請求和進行responser 過濾。

圖3：Scrapy框架詳解圖

3. 結果與討論

完成原始碼之後選擇了所有臺灣（台北、高雄、台中、台南、埔里、小琉球島、羅東鎮、台東、基隆、礁溪、桃園、宜蘭）共十二個城市進行項目驗證。使用類比訪問的方式獲取僅擷取所有單位（總計26561行）主要為台北市的相關評論，耗時約5小時；使用Scrapy爬取的方式擷取包括所有評論資料（總計82278行）在內的資料耗時在57分鐘以內。詳見表所示。對於爬取效率而言，在爬取資料比較完整而且沒有反爬蟲機制的情況下，使用Scrapy方法比類比訪問方式高一個數量級以上。
從程式執行結果可以發現，雖然因為運算資源和時間的限制無法將結果最佳化，但可以發現不管是中文評論或是英文評論在使用python 網路爬蟲模擬真人抓取訂房網站評論內容及評分相關性表現上均優於使用人工抓取評論資料數量及所消耗的時間，更重要的是使用python網路爬蟲可以同時針對網路上不同的訂房網站進行抓取評論來進一步分析研究，這樣可以讓研究的應用更加廣泛。

表1：性能比較表

方法名稱	執行緒列數	擷取資料量	耗時
類比訪問	1	26561行	5小時
模擬訪問	4	82278行	57分鐘

表2：爬蟲抓取結果

4. 結論

4.1 研究結果

本研究試圖解決以往想分析訂房網站上評論內容及評分但所遇到資料數量龐大的問題，運用python網路爬蟲，希望透過模擬真人抓取訂房網站評論內容及評分，提供使用者在獲得網站評論及評分內容中所潛藏的面向等更深入的決策資訊。程式執行結果中我們可以看到使用python 網路爬蟲模擬真人抓取訂房網站評論內容及評分相關性表現上都表現的比使用人工抓取評論資料數量及所消耗的時間來的優異。
此外，我們也整理過去研究者在資料探勘（Date mining）方面的研究。讓對於此方面研究有興趣的研究者可以更進一步的深入探討相關的議題。

4.2 未來研究方向

在應用python 網路爬蟲模擬真人抓取評論後，未來研究可以朝三個方向繼續深入探討相關議題。第一方面是嘗試最佳化python 網路爬蟲程式碼的各項參數，雖然最佳化參數並非本研究之目標，但若能在運算資源和時間的許可下增加爬取評論的數量，讓整個研究的結果可以表現的更好；第二個方向是更深入的將研究中的模型應用於個人化商品推薦、客製化資訊檢索等領域；第三個方向是針對其它相關旅遊訂房網站等進行更進一步的分析研究。