

以時間序列加權法預測網站來訪流量

(一) 摘要

自 90 年代起，透過流量分析工具追蹤消費者線上行為已成為網站管理者的必要任務之一。隨著資訊科技不斷演進，流量分析工具可蒐集的維度與指標日益漸增，網站對於企業之重要性亦是有增無減。造成此增長趨勢的主要原因在於透過網站流量之收集可讓網站管理者更為了解訪客真實行為表徵，進而能有所依據來審視網站經營成效。回顧過去文獻發現，多數研究將焦點置放在「歷史性流量」分析，但卻少有研究針對「預測性流量」予以著墨。在多數情況下，預測性流量分析比起歷史性流量分析更能夠有效協助管理者經營其網站。再者，若任何一個網站缺乏造訪流量，則該網站便無法遂行許多商業或營運目的。有鑒於此，本計畫擬以真實網站的大量歷史性流量數據為基礎，透過時間序列加權方式賦予歷史性流量若干權重，進而預測訪客日後造訪流量之消長。此預測結果可藉由資料視覺化，針對自我網站 (within comparison) 或競爭對手網站 (between comparison) 進行流量審視並提供造訪趨勢下跌之預警。

關鍵詞: 流量分析、歷史性流量、預測性流量、時間序列、資料視覺化

(二) 研究動機與研究問題

流量分析技術約略可以追溯到 90 年代早期，最初流量分析僅是簡單地將訪客對網站伺服器請求的資料予以記錄，隨著流量分析技術不斷演進，近年來開始有了以網站日誌與網頁標籤為基礎的流量分析技術問世。流量分析主要是以 Giannandrea 等人於 1999 年所提出的專利發明為主，透過 cookies 使網站管理者得以辨識不重複訪客 (unique visitor)。時至今日，流量分析工具所提供的維度與指標種類有增無減，此演進讓許多大型網路企業 (如 Google、Facebook) 得以透過流量分析結果來優化其廣告投放策略。以人們生活中息息相關的搜尋引擎或社群媒體為例，從訪客開啟任網路連線的那一刻起，在網路上所展演出的一舉一動即被記錄下來。RealMedia 的創始人 Morgan (2018) 更指出：「大家不理解在 Facebook 中的“讚”按鈕無所不在，這表示按讚的當下也隨即被追蹤了，哪怕自己並未在 Facebook 平台上面」。這段敘述充分反應出當代網路生態，不論是企業探尋顧客接觸點方式，或是訊息接收者是否能夠獲得符合需求的客製化企業投放訊息，訊息傳遞者自此可以透過低成本流量分析工作來掌握更為真實的行為數據。王勇等學者 (2005) 亦指出，擁有大量用戶頻繁訪問的網頁其質量通常較高，因此網站經營的良窳可透過流量分析來判斷。例如一名部落客，可以透過流量分析工具觀察訪客在自己文章中的閱讀時長，藉此判斷該訪客是否對文章內容感到興趣。若是一個電商網站，則可以透過產品點擊觀察訪客對哪些商品會有較大的採購意願、哪些產品較為乏人問津等。從以上描述可知，流量分析彌補傳統問卷調查的行為資料收集方式，此特性與陳傑豪等人 (2015) 在書中指出的問題不謀而合，即問卷調查常存有霍桑效應 (Hawthorne Effect) 的偏誤。換句話說，受訪者察覺自己正在被調查，會傾向預估施測者想要的結果而改變言行。因此會導致使用問卷調查的準確度下降，也礙於資源、空間等問題，問

卷調查只能做抽樣調查，無法針對所有樣本或母體進行詢問。然而透過流量分析工具，不僅可在最低干擾情況下取得線上訪客的真實行為數據，更可以透過母體資料採集的方式獲取網站全面完整數據。

Leinbach-Reyhle (2014) 在其文章中指出：「一個企業沒有網站就相當於失去業務」，顯然網站儼然成為各家公司從事商業活動所應具備的必需品，畢竟有 63% 的消費者會透過網站查找所需資訊或與企業產生互動，甚至有高達 30% 的消費者不會在沒有網站的情況下考慮消費 (Gevelber, 2017)。除此之外，Fogg 等學者在 2002 年所進行的研究中發現，46% 的受訪者認為網站他們信任一家企業與否的最大決定因素，由此可知網站對於一間公司的重要性不言而喻。這也是為何小型企業擁有網站的比例不斷增加，而且已經有 71% 的小型企業意識到提供網站予顧客之重要性 (Clutch, 2017)。然而即便擁有網站對企業經營來說非常重要，但擁有一個網站通常只是營運成功的必要條件，網站還需要有訪客蒞臨並且網站的流量品質也會反應出訪客對於特定網站的喜惡程度，進而間接影響訪客參訪行為是否轉換成訂單。此推測可由過去許多專家學者的論述中找到驗證，例如 2017 年 Google 美洲行銷副總裁 Lisa Gevelber 於發文中表示，若行動網站的載入時間超過三秒鐘，則會有 53% 訪客放棄行動網站的參訪。該專家亦提到，網站載入時間每延遲一秒鐘，轉換率 (分子為成功達成目標的參訪數，分母為總參訪數) 就會下降 12%。上述各專家學者皆不約而同地指出了一個共同的必要性，那就是網站會影響整體公司業務之營運。當網站沒有任何訪客蒞臨時，即使企業有網站，也只不過是形同虛設的空殼而已。而要如何了解網站的好壞，可以從訪客的使用者經驗來判斷其喜惡，而使用者經驗的數據便是從流量分析中取得，此舉不但能夠判斷使用者對於網站體驗的好感度，更能夠藉由大量歷史流量預測網站未來走向。截至目前為止，多數網站流量分析研究僅針對網站歷史流量數據予以檢討。例如，陳玲霞 (2012) 在中國省級公共圖書館網站搜索引擎優化調查分析中，就依照不同的維度做歷史指標數據呈現之探討，周月娟 (2012) 在國家圖書館「數位出版品平台系統」讀者使用分析中，依照 Google Analytics 裡收集到各維度的指標從事各項歷史數據的彙整，岑榮偉等人 (2010) 在基於日誌探勘的搜尋引擎使用者行為分析中依照過去使用者搜尋之關鍵字、用戶點擊數等資料分析使用者網站行為。雖然這些研究能夠在流量分析領域中產生一定程度之貢獻，然而在網站流量預測方面略顯不足。本計畫認為，若網站管理者能夠事先預測未來流量趨勢，勢必就能夠擴充其對於過去數據呈現的分析深度。

綜合以上敘述，本研究計畫擬以真實網站分析數據為基礎，針對各網站流量從事時間序列加權運算，其後再以矩陣分類方式呈現出網站流量目前狀況及預測其未來走向，若計畫成果順利實現，將可以產生下列貢獻：(1) 當發現網站流量呈現負面或不理想狀況時能即時做出應對。(2) 在網站進行更新或維護之後，可本研究成果得知網站狀態之變化，藉以判斷當次更新或維修的適切性。(3) 以網站流量現況不變的條件下，預測網站流量未來趨勢使管理者可從事更為長遠規劃。由於目前產學兩界皆未能有一套系統可以完整地針對網站流量現況予以審視並且做出有效的預測分析，僅少部份網頁分析工具針對特定類型而非不分類型的網站進行流量審視，此舉將會導致研究成果僅適用於特定類型網站，如此變無法與其他更多網站進行必要性流量分析比較 (如競爭者網站)，因此本計畫

預計透過一套演算過程，依照時間段的不同針對各網站流量數據進行分析。將不同時間段的流量數據做加權後，將所獲得的流量數據表現轉換成分數，投射至分類矩陣上，使這套系統能有效的輔助管理者做出管理網站決策 (如圖一所示)。



圖一、研究目的示意

(三) 文獻回顧與探討

(1) 線性回歸預測

對於已標記數據的預測方法以線性回歸預測最為常見，雷紹蘭 (2006) 等學者在電力短期負荷的多變量時間序列線性回歸預測方法研究中，根據電力日負荷的各種影響因素，做一個多變量線性回歸模型，以此預測往後時間序列的負荷結果。宋雲雲等學者 (2009) 在基於多元線性回歸的發動機性能參數預測中，提到藉由對某系列民用航空發動機的 EGT 等性能參數相關性分析發現，發動機性能參數之間存在較強的線性相關性，因此採用多元線性回歸預測法來對發動機的重要性能參數進行短期預測，提高預測精度。從這些研究可以知道，線性回歸預測是一種常用且公認的預測方法，並且在預測短期數據最為常用。

(2) 矩陣分類分析

利用矩陣的形式呈現分類結果，是一種簡單且易做出應對的視覺化成果，像是黃智聰等學者 (2003) 利用策略分析中常使用之 BCG 矩陣做為分析架構，計算中國大陸自 1993 至 2001 年期間，製造業 28 個產業別各產業產值比重及其成長率。並認為運用 BCG 矩陣分析產業消長狀況尚須注意其需求面成長趨勢與進口替代之影響，如此便能有助於從另一角度來了解其產業發展策略。江向東、傅文奇 (2008) 也認為結合 AHP 和 BCG 矩陣來對進行專利評價和戰略選擇，既充分利用了層次分析法在處理多個指標綜合評價問題方面的優點。

(3) 網站用戶興趣

在過去有關網頁訪客興趣分析的研究，都不會把網頁當成同一類檢測，因為不同類型的網頁可能會有不同的特性造成無法相比，王勇等學者 (2005) 就認為用合適的函數對訪問量數據進行擬合，就可以得到用戶興趣曲線。根據用戶興趣曲線變化的不同特徵，網頁可以分為有時效性的網頁和無時效性的網頁兩類。有時效性網頁的訪客興趣會隨時間衰退，而無時效性網頁的訪客興趣則保持相對恆定。有時效性的網頁會過時，而無時效性的網頁則不會。對於訪客造訪興

趣種類也被做不同的分類探討，郭岩等學者 (2005) 就表示穩定興趣是指一個人具有持久的興趣傾向，偶然興趣是指一個人由於臨時需要或其他原因對某事物產生的偶然興趣，若可用興趣類別對網頁進行歸納，興趣存在著“穩定”和“偶然”之分，意味着為著時間的推移、規模的增加，同一個訪客所拜訪的大部分網頁將持續地落在某幾個不變的興趣類別當中，而其餘網頁則反應出不規則、不連續地分散在其它興趣類別。這代表當訪客對網站忠誠度提高時，便轉換其身分成為網站的主力訪客，自然而然地在數據中該呈現該用戶的興趣為長期且持久。

(四) 研究方法與步驟

(1) 網路追蹤數據之維度或指標

本計畫擬以 Google Analytics (GA) 做為流量收集工具，由於 GA 中資料層級可以劃分成三大類，包括：使用者、工作階段、匹配，因本研究以此三類為選取基準擇定研究所需的指標與維度如下：(名詞部分為 Google Analytics 提供之解釋)

1. 使用者層級：

- 使用者：在特定日期範圍期間，開始至少一個工作階段的使用者。
- 新使用者：在所選日期範圍內，第一次使用者的人數

2. 工作階段層級：

- 工作階段：待定期限範圍內的工作階段總數。工作階段是指使用者積極與您網站戶定的這段期間。所有使用情形資料(畫面瀏覽、事件、電子商務等)都與工作階段相關聯。
- 每個使用者的工作階段數量：每個使用者的平均工作階段數
- 跳出率：使用者未與網頁互動的單頁工作階段百分比。跳出工作階段的時間長度為 0 秒。
- 發生事件的工作階段：至少觸發了一項活動的工作階段/造訪總數

3. 匹配層級：

- 網頁瀏覽量：是網頁的總瀏覽數。重複瀏覽同一個網頁也計算在內。
- 單次工作階段頁數：是指每個工作階段的平均網頁瀏覽量，重複瀏覽同一個網頁也計算在內。
- 平均工作時間長度：工作階段的平均時間長度
- 事件總數：是事件發生的次數。
- 不重複事件：指單一使用者在單一工作階段中與內容的互動，這類互動可與網頁瀏覽或畫面瀏覽分開追蹤。
- 單次工作階段的事件數：每次工作階段的活動總數

(2) 矩陣中八個維度空間

GA 三大資料層級可以各有不同的解釋意義，當其他條件不變情況下，使用者層級的數據增加代表來訪的不重複人數增加，因此可以判斷為新客戶的增加；當其他條件不變情況下，工作階段層級的數據增加代表每人來訪次數增加，因此可以判斷為顧客忠誠度的上升；當其他條件不變情況下，匹配層級的數據增加代表每人每次來訪的互動增加，因此可以判斷為顧客活躍度的上升。我們以使用者為 x，工作階段為 y，匹配為 z，在矩陣坐標中劃分成 8 個空間，8 個空間有各自代表的意義：

- 使用者為正、工作階段為正、匹配為正：代表新訪客人數、訪客回訪次數以及訪客活躍度皆在未來一周會呈現上升的情況。
- 使用者為負、工作階段為正、匹配為正：代表新訪客在未來一周可能不加反減，但訪客回訪次數以及訪客活躍程度皆會呈現上升的趨勢。
- 使用者為正、工作階段為負、匹配為正：代表訪客的回訪次數在未來一周可能不加反減，但新訪客以及訪客活躍程度皆會呈現上升的趨勢。
- 使用者為正、工作階段為正、匹配為負：代表新訪客跟訪客回訪的次數在未來一周呈現上升，但訪客活躍度卻下降。
- 使用者為負、工作階段為負、匹配為正：代表新訪客和訪客回訪次數呈現下降，但訪客活躍度會呈現上升。
- 使用者為負、工作階段為正、匹配為負：代表訪客回訪次數上升，但新訪客及訪客活躍度皆下降。
- 使用者為正、工作階段為負、匹配為負：代表新訪客人數增加，但訪客回訪次數以及訪客活躍度皆下降。
- 使用者為負、工作階段為負、匹配為負：新訪客人數、訪客回訪次數以及訪客活躍度皆在未來一周有下降的趨勢。

(3) 研究步驟

➤ 數據取得

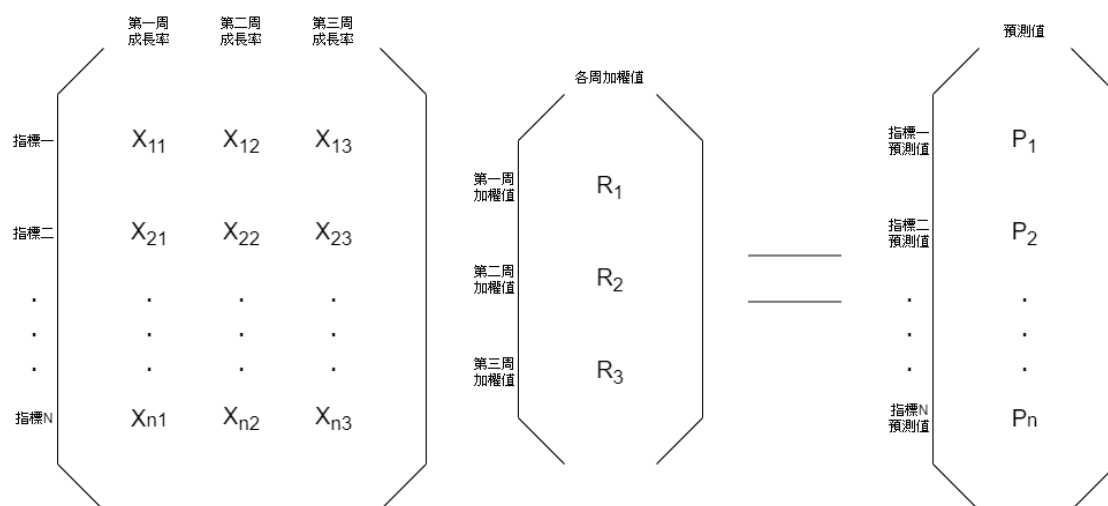
數據方面將透過網頁追蹤分析工具 Google Analytics 取得該研究網頁之數據，並將已蒐集數據透過 Google 提供之 API 或從 Google Analytics 儀表板下載 csv 檔後放到 Python 進行資料預處理及分析建模。

➤ 數據預處理

在 Python 獲取數據後，先將各項數據與前一周比較，得到當周該指標的數據成長率，將指標數據事先轉換成成長率的形式用意為 1、不因網頁規模大小影響結果的比較，有些網站的流量較大，那麼相對於流量小的網站，流量大的網站在指標數據的表現上從一開始就是力壓流量小的網站，所以在最後呈現視覺化矩陣中時，即使流量大的網站呈現負成長，在整體表現卻會仍遠遠超過流量小的網站，2、不因不同網站類型而有偏頗，有些網站使用者回訪次數少但每一次互動數卻高，有些網站卻可能相反的使用者回訪次數多但每一次互動數卻少，若不以成長率的方式比較，可能會造成不同類型的網站之間在矩陣視覺化中因相同的數據變化量有相同的效果，3、易與過去做比較，以成長率的方式表現，可以不因流量增加，而與過去的歷史預測無法直接比較。

➤ 矩陣列運算

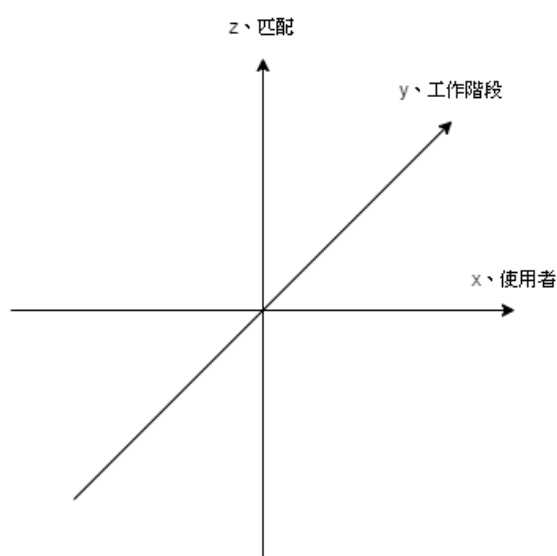
在本研究中，以前三周的數據去做預測下一周結果的根據，因為我們認為，時間距離越遠的數據，影響預測結果的能力越小，而事先以三周為例。按照使用者、工作階段、匹配分別計算出下周的預測值，如圖二。計算完成後相加，以使用者為 x 、工作階段為 y 、匹配為 z ，得到一座標點 (x,y,z) 。



圖二、利用各周加權預測網頁流量算式示意圖

➤ 資料結果視覺化

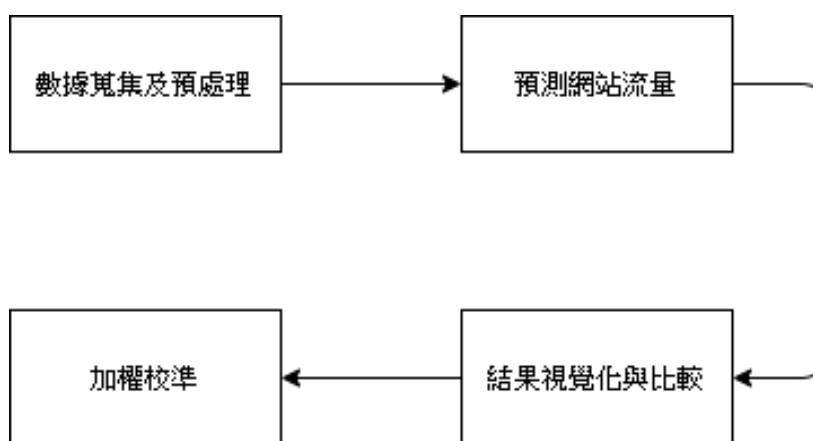
得到預測結果後，網站擁有未必能簡單清楚明瞭的知道自己網站現在的處境，也無法有效做比較，因此在此資料視覺化的步驟中，此部分將以 x 軸 y 軸 z 軸切出 8 個象限，並承接上一步矩陣列運算後的座標結果，對映到此座標系統中，完成後就可以簡單的了解網站的狀況，與此同時也可以將過去或其他網站的結果引進，清楚的做比較，如圖三所示。



圖三、視覺化矩陣示意圖

➤ 線性預測加權調整

每過完一周都會有新的真實數據產生，而我們可以藉由新數據來使我們的預測更精確。具體方法如下，每一周過去後會分別以使用者、工作階段、匹配的數據產生三個新點，按照原先矩陣列運算之方式，將前三周加權值換成 W_1 、 W_2 、 W_3 ，預期結果的部分改為當周真實的成長值，如此新增加的三個點與過去產生的點，找出一條回歸直線，而回歸直線的係數就是 W_1 、 W_2 、 W_3 的預測加權值。



圖四、研究步驟

(五) 預期結果

藉由上述步驟，本計畫預計能夠預測未來一周的網站流量趨勢，在一象限矩陣圖中呈現網站對於訪客三個方面在未來一周的成長情況與幅度，此方式能夠讓網站管理者能夠在一張圖表上與過往多筆資料或其他網站的流量趨勢做比較，並且每周透過新產生的數據來對預測加權校準。具體而言，本計畫預期結果包含：

- 透過過往的數據預測未來網站流量的趨勢
- 將預測結果視覺化呈現，且給予各象限不同代表的意義
- 隨著每周數據的增加預測更精準化

(六) 參考文獻

※ 中文部分

1. 王勇, 劉奕群, 張敏, 馬少平, & 茹立雲. (2008). 基於用戶興趣分析的網頁生命週期建模. 中文資訊學報, 22(2), 76-80.

2. 江向東, & 傅文奇. (2008). 十年來我國數字圖書館智慧財產權研究論文的統計分析. 情報科學, 4, 580-585.
3. 宋雲雪, 張科星, & 史永勝. (2009). 基於多元線性回歸的發動機性能參數預測 (Doctoral dissertation).
4. 岑榮偉, 劉奕群, 張敏, 茹立雲, & 馬少平. (2010). 基於日誌挖掘的搜尋引擎使用者行為分析. 中文資訊學報, 24(3), 49-55.
5. 周月娟. (2012). 國家圖書館 [數位出版品平臺系統] 讀者使用分析. 國家圖書館館訊, (134), 41-45.
6. 郭岩, 白碩, 楊志峰, & 張凱. (2005). 網路日誌規模分析和使用者興趣挖掘 (Doctoral dissertation).
7. 陳玲霞. (2012). 我國省級公共圖書館網站搜尋引擎優化調查分析. 圖書情報工作, 56(21), 88-91.
8. 陳傑豪. (2015). 大數據玩行銷, 臺北, 30 雜誌
9. 黃智聰, 高安邦, & 陳子芸. (2003). 中國大陸製造業的產業發展策略——BCG 矩陣的應用. 遠景基金會季刊, 4(4), 117-153.
10. 雷紹蘭, 孫才新, 周淥, & 張曉星. (2006). 電力短期負荷的多變數時間序列線性回歸預測方法研究. 中國電機工程學報, (2).

※ 英文部分

1. Clutch (2017). Small Business Websites in 2017: Survey. Available from <https://clutch.co/web-designers/resources/small-business-2017-website-survey>. Last accessed: 2020/03/06
2. Fogg, B. J., Soohoo, C., Danielsen, D., Marable, L., & Stanford, J. (2002). How do people evaluate a web sites credibility.
3. Gevelber, L. (2017). Micro-Moments Now: Why Expectations for 'Right Now' Are On the Rise. Available from <https://www.thinkwithgoogle.com/consumer-insights/consumer-immediate-need-mobile-experiences>. Last accessed: 2020/03/06.
4. Giannandrea, J., Bina, E. J., & Montulli, L. J. (1999). *U.S. Patent No. 5,978,817*. Washington, DC: U.S. Patent and Trademark Office.
5. Leinbach-Reyhle, N. Forbes, (Septemer 29, 2014,). 3 Reasons Websites Are Vital for Small Businesses. Available from <https://www.forbes.com/sites/nicoleleinbachreyhle/2014/09/29/websites-for-small-businesses/#6210e4f72026>. Last accessed: 2020/03/06.
6. Morgan, D. (2018) The Cookie That Ate The World. Available from <https://techonomy.com/2018/12/cookie-ate-world>. Last accessed: 2020/03/06.

(七) 需要指導教授指導內容

由於這是學生第一次親手撰寫計畫，對於學術上仍有許多不足夠的地方需要指導教授 不吝協助，特別是在實驗設計安排與資料分析方面。雖然學生目前才大三，也在系上修習了巨量資料概論、程式設計、網路行銷與搜尋引擎最佳化等課程，但是對於如何嚴謹地將 「實驗妥善安排與設計」 以及如何運用 「機器學習分析」 技巧來增加網路流量預測的準確度，仍有所不

解需仰賴指導教授指導。此外，由於學生對「成果報告撰寫」方面較不具有經驗，雖然深知學術報告具有一定的要求和規範，但對於報告撰寫的各種細節上仍需仰賴指導教授多年來的學術研究經驗，期盼能夠從中得到許多寶貴經驗。