 DOI: 10.6861/TANET.201810.0334

使用者長時段跨網站瀏覽資料集之蒐集與分析

陳廷睿　連丞宥　白國臻　陳弘軒
中央大學資訊工程學系
ray941216@g.ncu.edu.tw, {littlelien.peanut, ivy2350442}@gmail.com,
hhchen@ncu.edu.tw

摘要

本文將介紹一組我們蒐集之網路使用者長時段跨網站瀏覽紀錄資料集。使用者長時間的跨網站紀錄的資料不容易蒐集，且此類的開放資料集並不常見，故此蒐集並開放本資料集將能協助研究者以真實資料進行各項先進研究，如：使用者行為分析、使用者人口特徵分析、推薦引擎及線上廣告技術研發等。本文將包括以下部份： (1) 我們將說明此資料集的蒐集方式；(2) 我們將報告此資料集的基本統計資訊；(3) 我們將討論公開完整資料的顧慮，以及我們在考慮「開放」及「隱私」這兩個矛盾的議題後所採取的妥協策略；(4) 我們將介紹基於本資料集所進行的幾個實驗；(5) 我們將介紹現階段擴充此資料集的計畫。

關鍵詞：線上日誌、開放資料、使用者行為

Abstract

This paper introduces a dataset containing the logs of online users' long-term cross-website visits. Such type of open dataset is rarely-seen because collecting the logs of users' long-term cross-website visits is difficult. As a result, opening such a dataset may help the researchers conduct advanced researches, such as online user behavior analysis, user demographical information analysis, recommender systems and online advertising system development, etc. We report the following items in this paper. First, we explain the data collecting process. Second, we show the basic statistics of this dataset. Third, we discuss the concerns to release the complete dataset. Specifically, we discuss the trade-offs between "openness" and "privacy" and our current compromising sharing policy. Fourth, we introduce experiments based on this dataset. Finally, we introduce our current plan for expanding this dataset.

Keywords: online log, open data, user behavior.

1. 簡介

網路使用者長時間之跨網站瀏覽紀錄對許多研究者而言是珍貴的資料。舉例而言，人文社會科學領域的學者可能由使用者的閱讀紀錄定性或定量地研究該使用者的人格特質、心理狀態、政治傾向等；資訊科學領域的學者可能藉由此資料集發展推薦引擎技術及評估推薦成效；流行文化研究者可以藉由大眾的閱讀資料窺視流行文化的演進；傳播學的研究者也可以分析文章或議題如何由小眾的資訊演變為大眾文化。

然而，多使用者長時段跨網站之瀏覽紀錄並不容易蒐集，我們從三方面說明這類資料集的蒐集難度：(1)「跨網站」－中大型的網站經營者雖然可以從伺服器的日誌 (log) 得到使用者的站內行為，但使用者進入網站前及離開該網站後的行為卻無法得知；(2)「長時間」－某些網站會在使用者的瀏覽器留下 cookie 以紀錄使用者在其他網站的行為，但瀏覽器對於單一網站能使用的 cookie 個數或空間通常有限制[1]，故較長時間的行為不見得可以完整地獲得；(3)「多使用者」－大部份的瀏覽器允許使用者手動匯出或透過工具匯出個人的完整瀏覽紀錄，然而，要讓使用者願意匯出瀏覽紀錄並將之分享給第三方則需要足夠強的誘因。

本文將介紹一組我們蒐集並公開的網上使用者長時段跨網站瀏覽資料集。我們將介紹此資料集的蒐集方式 (第二節)，並報告目前蒐集到的資料的基本統計資訊 (第三節)，我們將說明公開此資料集之前為保持使用者的匿名性所做的資料前處理 (第四節)，並介紹我們利用此資料集所做的幾個實驗 (第五節)，我們也將說明目前計畫繼續擴充此資料集的方式以及正在進行中的工作 (第六節)，最後，我們將討論本作品的優缺點及未來可能的研究及應用方向 (第七節)。

2. 資料蒐集方式

我們利用 Google Chrome 瀏覽器的插件 (plug-in) 將瀏覽器中所有的觀看歷史及書籤中所有的網頁存下並上傳至伺服器供事後分析。本資料所蒐集的使用者透過線上網站及論壇招募而來，使用者在初次安裝該插件時，該插件會要求使用者輸入一些基本資訊，如：性別、年齡、感情狀態、郵件地址等。我們給予使用者小額的獎金或獎品做為使用本插件的誘因。另外，若使用者使用本插件超過一定的週數，我們再給予額外的小獎勵。

所有參與的使用者都被告知插件會紀錄使用者過去的瀏覽歷史以及插件安裝期間的瀏覽網頁，也被告知這些紀錄將會被上傳至伺服器進行額外的分析。

[1] Cookie limit test: http://www.ruslog.com/tools/cookies.html

DOI: 10.6861/TANET.201810.0334

3. 資料集基本統計

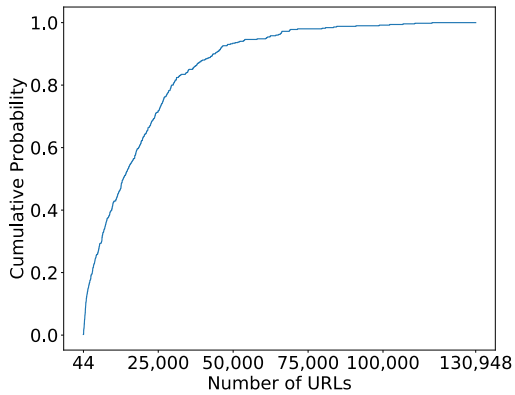

圖 1 使用者瀏覽網頁數量之經驗分布函數

本資料集經過初步刪除有問題之使用者後共留下 672 位使用者之瀏覽紀錄，大部分的網頁瀏覽歷史紀錄之紀錄期間為 2016 年 8 月至 2016 年 12 月。所有使用者之瀏覽量為 12,837,216 筆，圖 1 與表 1 分別展示每一位使用者瀏覽網頁數量之經驗分布函數 (empirical distribution function) 與綜合統計表格。

表 1 個別使用者瀏覽網頁總量之綜合統計

最小值	Q1	Q2	平均值	Q3	最大值
44	4,239	13,335	19,103	26,698	130,992

另外，全部具有網頁瀏覽紀錄的使用者之中共有 508 位使用者具有較詳細的個人資訊 (demographic information)，個人資訊包含：性別 (男：45%、女：54%、其他：1%)、年齡 (0-20: 1%、21-30: 59%、31-40: 31%、40+: 9%)、及感情狀態 (單身: 47%、交往中: 33%、已婚: 18%、其他: 2%)。

4. 公開完整資料集之顧慮及妥協

由於原始資料集中除了郵件外並沒有太多敏感的個人資訊，我們原先認為公開除了郵件外的完整的資料集並不會有隱私權的顧慮。然而，在仔細查看使用者的瀏覽紀錄後，我們發現某些網頁瀏覽資訊可能可以拼湊出瀏覽者本身 (例如：某些人喜歡在網路上搜尋自己的名字)；另外，美國的著名網路服務商 AOL 曾經公佈二千萬筆去識別化後的搜尋紀錄，但卻被記者根據搜尋紀錄找到編號 4417749 的搜尋者 – 一位住在喬治亞州的 62 歲寡婦[2]。為了為避免類似的事件發生，我們在公開資料前進一步做了以下的前處理。

對於每一筆瀏覽紀錄，我們將其網址 (URL) 透過一個網頁分類線上服務進行分類[3]，該分類器能將輸入的網址轉為對應的網頁類型，如：輸入網址為 https://www.google.com/，則輸出之網頁類型為 “Search Engines and Portals”，以及輸入網址為 https://www.facebook.com/，則輸出之網頁類型為 “Social Networking services”。將所有使用者網頁瀏覽紀錄之網址進行分類後，共包含 88 種網頁類型，統計後發現前五大熱門的網頁類型為 “Social Networking services” (29%)、 “Search Engines and Portals” (15%)、“Web-based Email” (8%)、“Media” (7%)、及 “Shopping” (7%)。最後我們並沒有公佈使用者瀏覽的網頁 URL，我們公佈的是使用者瀏覽的網頁之類型。

5. 實驗分析

以下說明我們利用此資料集做的兩個實驗。

5.1 使用者基本屬性預測

表 2 基本屬性預測之 MicroF1 分數比較表

模型	年齡	性別	感情狀態
baseline	0.388	0.545	0.474
k-NN	0.427	0.594	0.478
RF	0.453	0.697	0.488
LR	0.427	0.697	0.476
SVM	0.388	0.591	0.474

給定使用者之瀏覽紀錄，能否預測其個人基本屬性，如：性別、年齡等？倘若可預測，這同時代表巨大的商機及可能的隱私權危機。

我們選擇幾個常見的監督式分類器做為訓練模型，包括：k-Nearest Neighbors (kNN)、Random Forest (RF)、 Logistic Regression (LR)、及 Support Vector Machine (SVM)。以上方法的細節及參數設定如 [1] [2] 所述。另外，我們加上一個多數決演算法當作 baseline，此演算法總是預測使用者數量最多的類別。我們預測的使用者基本屬性包括：年齡、性別、及感情狀態。

我們採用兩大類的特徵來訓練模型：(1) 使用者在各個類型的網頁的瀏覽比例；(2) 使用者一天中各個時段瀏覽網頁的比例。

我們比較各方法之 MicroF1 分數，採用 MicroF1 的原因是：(1) 當分類問題的目標類別大於二時，常見的 F1 分數無法直接使用；(2) 相較於 MacroF1，MicroF1 的分數不會被出現次數較少的類別的結果所支配。

表格 2 展示各個模型在三個使用者基本屬性的預測效果。結果顯示：各個方法在測試集上的表現均優於基礎模型。我們在 [1] 中另外建構了一個新的演算法且其表現大多優於上述的各監督式模型。然而，本論文的目的並非介紹新的演算法，故我們不在此對該演算法做介紹。

[2] https://www.nytimes.com/2006/08/09/technology/09aol.html

[3] http://www.fortiguard.com/webfilter

DOI: 10.6861/TANET.201810.0334

5.2 使用者在特殊事件前的行為變化分析

研究顯示：使用者在線上的購物行為可能會隨著時間而變化 [3] 。然而，使用者在特殊事件前，瀏覽習慣是否會發生變化呢？倘若我們能在特殊事件發生前預測使用者的行為變化，其影響將第 5.1 節的實驗般，可能同時代表巨大的商機及隱私權的危機。

我們選定「節日」做為特殊事件，預測使用者在節日來臨前夕至節日間的將會增加或減少拜訪購物網站的頻率。倘若節日間比平時更常造訪，則為正樣本，反之為負樣本。我們實驗的假期為 2016 年的中秋節 (9/15)、單身節 (11/11)、及聖誕節 (12/25)。

我們選定的監督式分類器依然包括 kNN、RF、LR、及 SVM。另外，我們選定的特徵包括兩大類：(1) 使用者基本屬性 (性別、年齡、感情狀態)；(2) 平時各類型網站的觀看比例。各方法的細節及參數設定如 [4] 所述。

表 3 假日前購物網站拜訪行為變化之 AUC 分數比較表

模型	中秋節	單身節	聖誕節
k-NN	0.55	0.63	0.72
RF	0.60	0.60	0.68
LR	0.65	0.61	0.73
SVM	0.64	0.64	0.77

表 3 展示各方法在三個節日的預測結果，我們使用 ROC 曲線 (Receiver Operating Characteristic curve) 的曲線下面積 (Area Under Curve，簡稱 AUC) 分數做為各方法比較的指標。由於這裡的問題是一個二元分類問題且類別並不平均，我們認為採用 AUC 是個較合理的指標。實驗結果顯示：透過使用者的網頁瀏覽歷史紀錄及個人基本屬性，可以在一定程度上預測使用者在特殊節日前購物習慣的改變。因此，此應用將能幫助電子商務平台針對不同使用者給予特定的市場行銷手段，有機會節省廣告成本，同時使廣告效果最大化。

6. 資料集擴充計畫

以下說明我們需要擴充資料集的原因，以及擴充計畫的概要。

6.1 動機

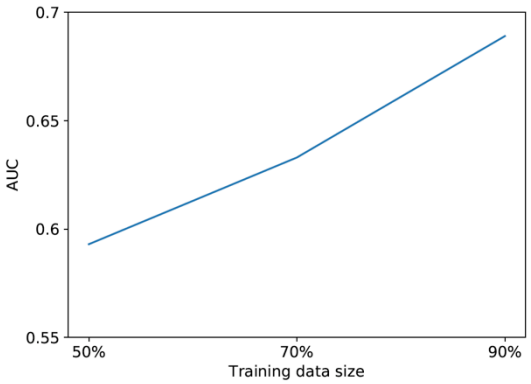

圖 2 使用50%、70%、90%可用資料當作訓練資料集時，測試集 AUC 的表現

我們在預測使用者在假日前拜訪購物網站的行為變化時，發現預測的效果會隨著訓練筆數的增加而提升 [2]：如圖 2 所示，當使用 50%、70%、及 90% 的可用資料做訓練以預測剩餘的資料中使用者的行為時，其 ROC 曲線的 AUC 分數明顯變高。

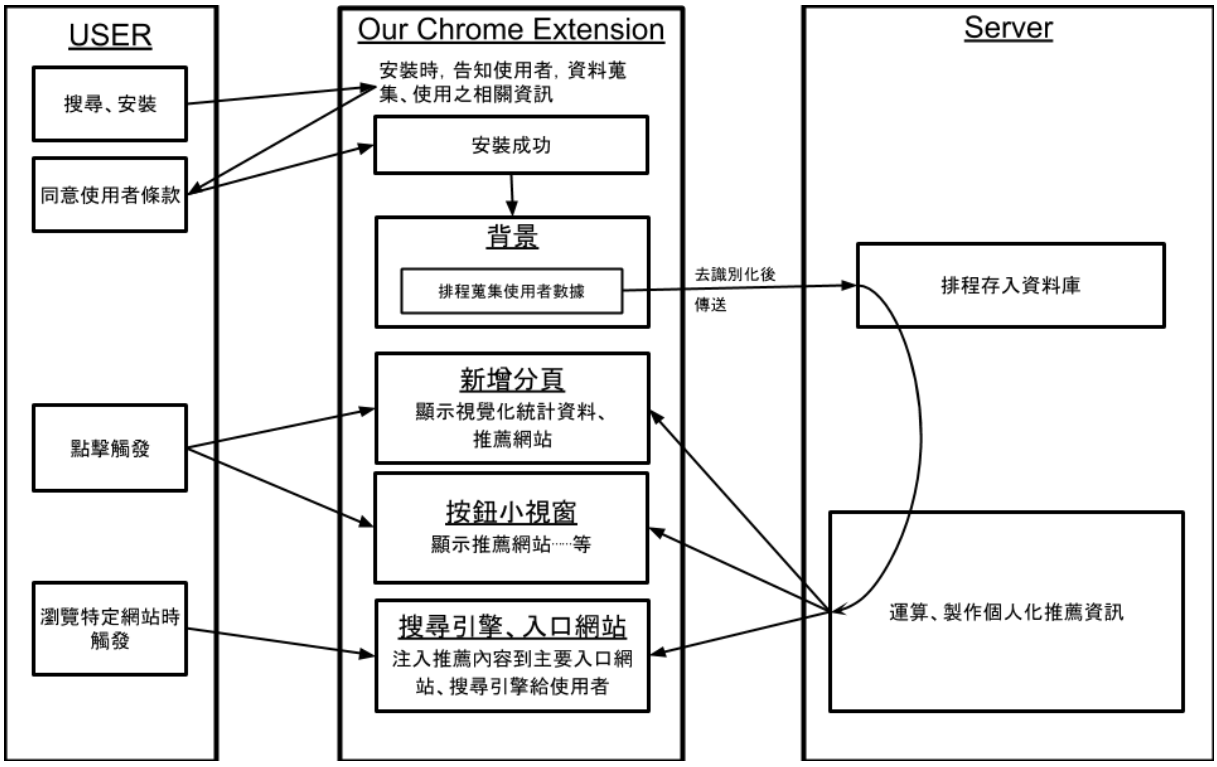

圖 3 基礎架構圖

DOI: 10.6861/TANET.201810.0334

這表示目前的訓練模型可能由於訓練資料的不足而有過適 (overfitting) 的現象，因此我們決定蒐集更多的使用者瀏覽紀錄以改善模型，我們也希望未來公開更大的資料集能讓其他研究者從中受惠。

6.2 計畫概敘

由於原始的 Google Chrome 插件原始碼已遺失 ，且原始方案需要以金錢吸引使用者參與，我們決定重新設計插件，此插件將根據使用者的瀏覽歷史紀錄完成某些功能，讓使用者願意為這些功能而分享其瀏覽歷史紀錄。

我們目前計畫的功能有以下二者：(1) 透過機器學習推薦個人化的建議閱讀清單給使用者，使其能更方便瀏覽到需要的網站；(2) 提供使用者網路使用的數據視覺化報告，讓使用者能知道自己的上網習性 (如：上網時段統計、每週花多少時間等)，以及在各種類型的網頁上花了多少時間。於提供服務的同時，蒐集使用者網路使用之行為數據以擴充本資料集相關的部分。

6.3 進行中之事項

我們計畫的基本服務架構如圖 3 所示。Google Chrome 插件將排程蒐集使用者的瀏覽紀錄，在前端去識別化後傳送至伺服器並排程存入資料庫，伺服器端根據該使用者的瀏覽紀錄提供個人化網頁閱讀推薦清單給使用者。同時，插件在前端可直接利用個人瀏覽歷史展示網路使用數據的視覺化報表，幫助使用者瞭解自己的網路使用習慣。使用者介面規劃如圖 4 所示。

我們目前已完成後台使用者資訊蒐集的部份功能，並透過 Chrome 提供的去識別化標籤將使用者資訊傳送至我們的伺服器，目前資料庫部分僅部份介接完成，待能完整介接後，將可開始推廣並蒐集新的資料。

7. 討論

本論文介紹一組新的公開資料集，此資料集包含超過六百位使用者的中長期瀏覽紀錄。為了保證資料的匿名性，我們除了將使用者去識別化，同時也將網站的 URL 轉為網站的類型。

我們對此資料集做了基本統計分析。此外，我們也介紹了兩個基於此資料集進行的實驗－使用者基本屬性預測及使用者的線上行為變化。這些實驗顯示：即使只使用網站的類型而不使用原始 URL 當作特徵，也能產生有效的預測。在隱私權的考量下，我們決定只公佈使用者的瀏覽網站類別而非網站的 URL。

我們計畫繼續擴充此資料集，目前計畫透過提供資料視覺化服務以及推薦功能來換取資料的方式，讓使用者能夠在獲得個人化服務的同時，我們也能小成本地蒐集資料，並不定時發佈去識別化後的資料供研究者使用，創造三贏模式。

同時，我們也希望藉由此系統發展新的推薦引擎技術，能對使用者進行客製化的推薦，並藉由實際運作的平台發現實務上的難題，以科學化的方式解決這些問題。

致謝

我們感謝工研院巨資中心的支援讓我們有足夠的資源招募使用者使用插件，也感謝崔文博士及鍾筑安女士在歷次的討論中給予的建議，我們感謝台大心理系的黃從仁教授及其研究團隊開發初版的 Google Chrome 插件及資料蒐集及分析過程給予的建議及指導。我們感謝科技部對本計畫的補助 (MOST 107-2221-E-008-077-MY3)。

參考文獻

[1] 連丞宥, “透過網頁瀏覽紀錄預測使用者之個人資訊與性格特質,” 碩士論文, 2018.

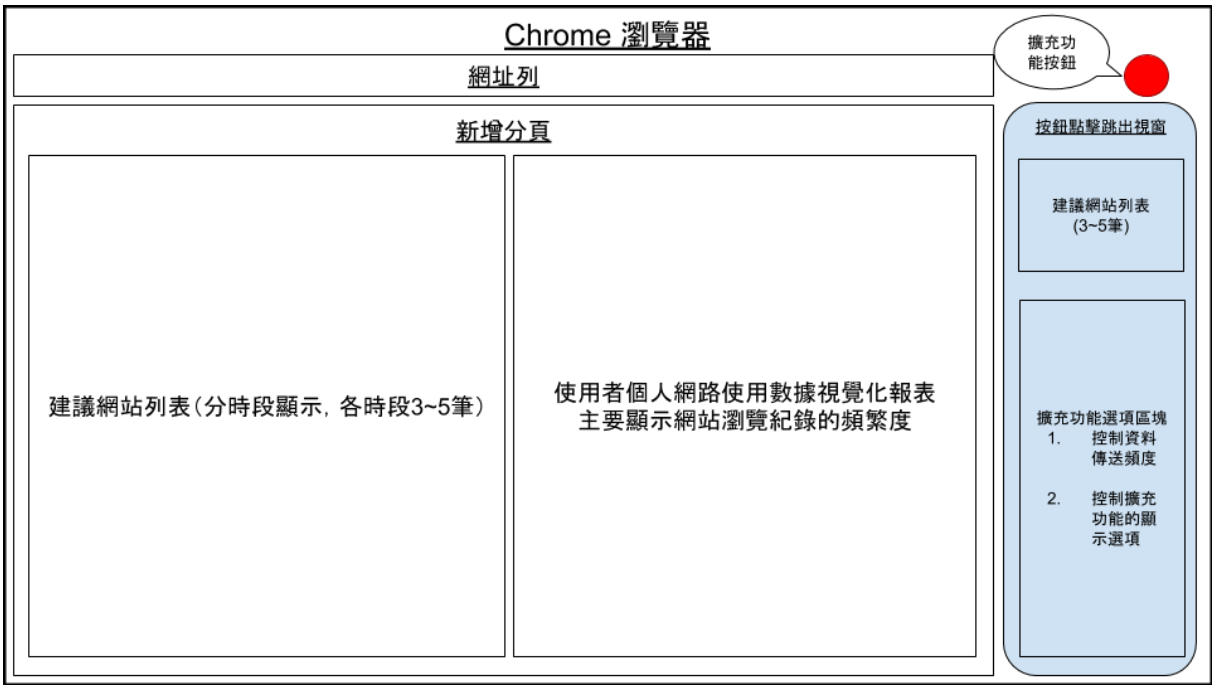

圖 4 使用者介面設計

DOI: 10.6861/TANET.201810.0334

[2] C.-Y. Lien, G.-J. Bai, T.-R. Chen and H.-H. Chen, "Predicting User's Online Shopping Tendency During Shopping Holidays," in The 2017 Conference on Technologies and Applications of Artificial Intelligence, Hsinchu, 2017.

[3] C. Lo, D. Frankowski and J. Leskovec, "Understanding behaviors that lead to purchasing: A case study of pinterest," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016.

[4] 白國臻, "透過矩陣分解之多目標預測方法預測使用者於特殊節日前之瀏覽行為變化," 碩士論文, 2018.