

國 立 中 央 大 學

資訊工程學系
碩士論文

分析網路日誌中有意圖、無意圖及缺失之使用者
行為

Analyzing Intentional, Unintentional and Missing User
Behaviors in Weblogs

研 究 生：許哲芸

指 導 教 授：陳 弘 軒 博 士

中 華 民 國 一 百 零 九 年 六 月

國立中央大學圖書館學位論文授權書

填單日期：109/2/16.

2019.9 版

授權人姓名	許哲芸	學號	107522047
系所名稱	資工所	學位類別	<input checked="" type="checkbox"/> 碩士 <input type="checkbox"/> 博士
論文名稱	分析網路日誌中有意圖、無意圖及 詛咒文使用者行為	指導教授	陳弘軒

缺失之使用者行為

學位論文網路公開授權

授權本人撰寫之學位論文全文電子檔：

- 在「國立中央大學圖書館博碩士論文系統」

() 同意立即網路公開

() 同意 於西元_____年_____月_____日網路公開

() 不同意網路公開，原因是：_____

- 在國家圖書館「臺灣博碩士論文知識加值系統」

() 同意立即網路公開

() 同意 於西元_____年_____月_____日網路公開

() 不同意網路公開，原因是：_____

依著作權法規定，非專屬、無償授權國立中央大學、台灣聯合大學系統與國家圖書館，不限地域、時間與次數，以文件、錄影帶、錄音帶、光碟、微縮、數位化或其他方式將上列授權標的基於非營利目的進行重製。

學位論文紙本延後公開申請（紙本學位論文立即公開者此欄免填）

本人撰寫之學位論文紙本因以下原因將延後公開

• 延後原因

- () 已申請專利並檢附證明，專利申請案號：
 () 準備以上列論文投稿期刊
 () 涉國家機密
 () 依法不得提供，請說明：_____

• 公開日期：西元_____年_____月_____日

※繳交教務處註冊組之紙本論文(送繳國家圖書館)若不立即公開，請加填「國家圖書館學位論文延後公開申請書」

研究生簽名：許哲芸

指導教授簽名：陳弘軒

*本授權書請完整填寫並親筆簽名後，裝訂於論文封面之次頁。

國立中央大學碩士班研究生
論文指導教授推薦書

資訊工程學系碩士班 學系/研究所 許哲芸 研究生
所提之論文 分析網路日誌中有意圖、無意圖及缺失之使用者行
為
係由本人指導撰述，同意提付審查。

指導教授 許弘智 (簽章)

109 年 7 月 14 日

國立中央大學碩士班研究生
論文口試委員審定書

資訊工程學系碩士班 學系/研究所 許哲芸 研究生
所提之論文 分析網路日誌中有意圖、無意圖及缺失之使用者行
為
經由委員會審議，認定符合碩士資格標準。

學位考試委員會召集人

符江川

委

員

孫弘府
李峻峰

中 華 民 國

109. 年 7 月 7 日

1090706

分析網路日誌中有意圖、無意圖及缺失之使用者 行為

摘要

網路日誌 (Weblog) 已廣泛的用來代表使用者線上行為，然而我們發現網路日誌只記錄了使用者部分的行為，例如只記錄到使用者點擊網頁之行為卻忽略網頁分頁間切換之行為。同時可能多紀錄了非使用者自發性的行為，例如當瀏覽的網頁進行重新網址導向或者彈出廣告視窗，後面所開啟的網頁並非使用者意圖想要瀏覽的，但卻會被包含在瀏覽紀錄中。我們發現一般的網路日誌中僅記錄到使用者一半左右的瀏覽行為而且其中 5.6% 是屬於使用者可能無意識的行為。

透過建立 Google Chrome 瀏覽器中 plugin 和招募受試者下載使用，我們對有意圖的瀏覽紀錄、無意圖的瀏覽紀錄以及缺失的瀏覽紀錄進行統計，並且發現傳統的瀏覽紀錄中最常使用的網站類型之排名和加上缺失的瀏覽紀錄或者去掉無意圖的瀏覽紀錄之排名是不一樣的，也因此我們對於傳統的瀏覽紀錄是否能代表使用者瀏覽行為產生疑問，依傳統瀏覽記錄進行的分析也可能因此而產生偏誤。本文透過對傳統的瀏覽紀錄、有意圖的瀏覽紀錄及有意圖加上缺失的瀏覽紀錄三者進行分析，並使用常見的分類模型對「下次點擊的事件類型」、「下次點擊會間隔多久」及「未來的瀏覽之網站比例」進行預測，發現相較於傳統的瀏覽紀錄另外兩者皆有良好的表現。這表示網路日誌漏記的使用者行為可能含有額外的資訊且非使用者自發性但存在於網路日記中的紀錄可能雜訊大於資訊。

摘要

關鍵字：點擊流, 網路日誌分析, 使用者行為分析

Analyzing Intentional, Unintentional and Missing User Behaviors in Weblogs

Abstract

Weblogs have been widely used to represent the behavior of online users. However, we found that weblog only records part of users' behaviors. For example, traditional weblogs do not record tab switching and browser window switching. Besides, weblog may record some visits that do not come from a users' conscious actions. For instance, web pages resulted from page redirects and page pop-ups are recorded in the browsing history, but users may not have intentions to visit these pages. We discover that, on average, weblogs approximately record only half of a users' page visits and 5.6% of the visits recorded in the weblog belongs to users' unconscious actions. To collect and analyze the conscious visits, unconscious visits, and "missing" visits (i.e., the visits that are unrecorded in the traditional weblog), we created a Google Chrome plugin and recruited users to install the plugin. We reported the statistics of visits and showed that sorting the popular website categories based on the traditional weblog is different from the rankings obtained from including the missing visits or excluding the unintentional visits. Therefore, traditional weblog may be a biased representation of a user's online behaviors, and the observations or conclusions derived from weblog analysis are questionable. Additionally, we predicted users' future behaviors based on three types of training data –

Abstract

all the visits in traditional weblogs, intentional visits in weblogs, and intentional visits plus missing visits in weblogs. We applied supervised learning algorithms to make predictions. The experiment results show that using intentional visits in weblogs or intentional visits plus missing visits in weblogs usually perform better compared to using all the visits in traditional weblogs. This result indicates that missing visits in weblogs may contain additional information, and unintentional visits in weblogs may have more noise than information.

Keywords: Clickstream, Web log analysis, User behavior analysis

目錄

	頁次
摘要	ix
Abstract	xi
目錄	xiii
一、 緒論	1
1.1 研究動機	1
1.2 研究目標	2
1.3 研究貢獻	3
1.4 論文架構	4
二、 相關研究	5
2.1 網路日誌應用實例	5
2.1.1 社群網站之應用	5
2.1.2 電商網站之應用	6
2.1.3 旅遊租屋網站之應用	7
2.2 擴展點擊流：分析點擊流中缺少的使用者行為	7
三、 有意圖點擊流、無意圖點擊流及擴展點擊流	9
3.1 點擊流	9
3.2 擴展點擊流	10

四、 資料集介紹	13
4.1 原始資料集	13
4.2 資料集前處理	14
4.2.1 時間單位前處理	14
4.2.2 網站網址前處理	15
4.2.3 事件間之時間間隔前處理	16
4.3 資料統計與分析	17
五、 實驗	21
5.1 問題與想法	21
5.2 實驗設計	21
5.3 分類器選擇及介紹	22
六、 實驗結果與分析	25
6.1 實驗資料集介紹	25
6.2 模型評估標準	26
6.3 實驗結果分析討論	27
6.3.1 預測使用者下次點擊網站類型	27
6.3.2 預測使用者下次點擊間隔時間	31
6.3.3 預測使用者未來瀏覽之網站比例	32
七、 結論與未來展望	37
7.1 結論	37
7.2 未來展望	38
參考文獻	39
附錄 A 實驗之混淆矩陣	41

圖 目 錄

	頁次
1.1 使用者瀏覽行為示意圖	2
4.1 真實數據與收集資料之事件時間差異	15
4.2 原始資料集中時間間隔之事件數量統計圖	16
4.3 不同資料集中時間間隔的事件數量長條圖	17
5.1 預測網站類型之不同資料集中特徵的選擇	22
6.1 資訊集分配示意圖	25
6.2 判斷測試表	26
6.3 實驗一:XGBoost 中根據不同 max_depth 值對應之 <i>MircoF₁</i> 分數	28
6.4 XGBoost 之混淆矩陣評估 (Top 1-5)	29
6.5 XGBoost 之混淆矩陣評估 (Top 16-20)	30
6.6 實驗二:XGBoost 中根據不同 max_depth 值對應之 <i>MircoF₁</i> 分數	31
6.7 XGBoost 之混淆矩陣評估	32
6.8 實驗三:XGBoost 中根據不同 max_depth 值對應之 <i>MircoF₁</i> 分數 (Top1)	34
6.9 實驗三:XGBoost 中根據不同 max_depth 值對應之 <i>MircoF₁</i> 分數 (Top2)	34

6.10 XGBoost 之混淆矩陣評估 (Top1)	35
6.11 XGBoost 之混淆矩陣評估 (Top2)	36
A.1 實驗一：XGBoost 之混淆矩陣 (CS)	42
A.2 實驗一：XGBoost 之混淆矩陣 (ICS)	43
A.3 實驗一：XGBoost 之混淆矩陣 (ICS+ECS)	44
A.4 實驗二：XGBoost 之混淆矩陣 (CS & ICS)	45
A.5 實驗二：XGBoost 之混淆矩陣 (ICS+ECS)	46
A.6 實驗三：XGBoost 之混淆矩陣 (Top1)	47
A.7 實驗三：XGBoost 之混淆矩陣 (Top1)	48
A.8 實驗三：XGBoost 之混淆矩陣 (Top2)	48
A.9 實驗三：XGBoost 之混淆矩陣 (Top2)	49

表目錄

	頁次
3.1 有意圖點擊流、無意圖點擊流及有意圖擴展點擊流之例子	11
4.1 使用者 plugin 使用天數之統計	13
4.2 CS 及 ECS 的事件轉換類型之比例統計 (%)	14
4.3 網站類型之事件數量統計	16
4.4 原始資料集中各時間間隔中事件數量統計	17
4.5 使用者事件數量之統計	18
4.6 使用者平均每天事件數量之統計	18
4.7 不同轉換類型資料中前 20 名網站類型的事件統計	19
6.1 預測下次點擊網站類型之 $MircoF_1$	28
6.2 預測下次點擊時間間隔之 $MircoF_1$	31
6.3 預測 Top1 之網站類型比例之 $MircoF_1$	33
6.4 預測 Top2 之網站類型比例之 $MircoF_1$	33

一、緒論

1.1 研究動機

現今瀏覽網頁已經成為多數人生活中的一部份，例如查詢資料、使用社交軟體，或者是線上購物等。伴隨著瀏覽網頁而產生的點擊流也被廣泛的應用，例如分析使用者行為模式推測使用者的喜好並推播廣告 [1]–[3]。我們研究後發現一般收集到的點擊流不見得能真實的反應的使用者瀏覽網頁的狀況。我們發現一般收集到的點擊流只記錄了使用者部份的行為，因此少記了可能成為資訊的線索，點擊流甚至有可能記錄到非使用者意圖的行為，而這些無意識的行為是有可能成為資料中的雜訊，進而影響到分析的結果。

透過圖 1.1使用者的瀏覽網頁的狀況，可以看到首先使用者會開啓瀏覽器 A，接著依照自身需求進入搜尋頁面 B，然後點擊搜尋頁面 B 上的連結開啓新的分頁 C，而後分頁 C 發生網址重新導向至網頁 D。以上述的例子來看，最終傳統的網路日誌中會包含 A、B、C、D 四個瀏覽記錄，但實際上網頁 D 並非使用者有意圖而開啓的，也就是說在使用者無意識的情況下網頁 D 被記錄在瀏覽記錄內。接下來當使用者認為點開的連結不符合預期，想重新切回搜尋頁面 B，此時就會產生分頁的切換，而像這類型的分頁切換或者是視窗切換一般是不會被記錄下來的。基於這個發現，我們對於沒有被記錄到的使用者行為和使用者無意識的但被記錄到的行為很感興趣，並且認為可以探討兩者對於行為分析會產生什麼不

同的影響。

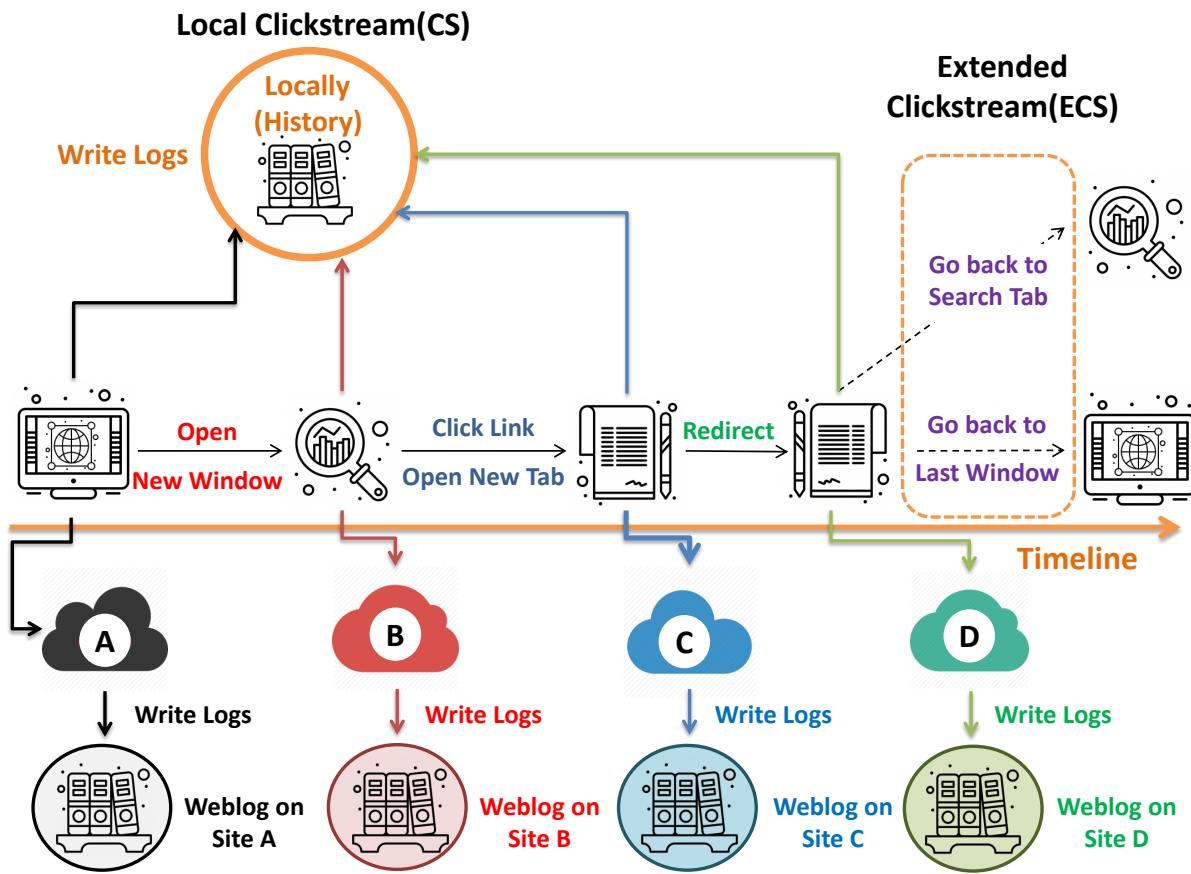


圖 1.1: 使用者瀏覽行為示意圖

1.2 研究目標

本篇論文主要的目標為探討有意圖的點擊流和傳統沒有被記錄到的額外瀏覽行為兩者資料是否會對使用者行為分析造成影響。因為相比於傳統的點擊流的資料，我們認為添加上額外瀏覽行為或是去除掉無意圖的使用者行為能更加貼近使用者的真實瀏覽行為，更加有益於進行分析。

為了能收集到完整的使用者行為資料，我們使用自製的 Google Chrome 瀏覽器中的 plugin 來蒐集並記錄使用者完整的瀏覽過程，其中包含點擊流及額外缺失的行為資料。本篇論文總共招募 300 多位使用者，

最終在資料的部分，我們選擇了 147 位擁有足夠使用天數的使用者的資料來進行實驗。資料集的部分則決定使用點擊流、有意圖點擊流及有意圖點擊流加額外缺失的行為三種不同資料集，於一般常見的分類模型中，來分別預測使用者未來點擊類型、點擊的時間間隔和未來瀏覽網站的類型比例。

1.3 研究貢獻

對於預測使用者行為之任務，很多研究廣泛的使用傳統定義上的點擊流去代表使用者線上行為，但根據本篇論文之發現，點擊流並不能完全的代表使用者真實的線上行為。點擊流中含有使用者無意識但被記錄的行為，甚至某些使用者的行為是沒被記錄到的，所以利用傳統定義上的點擊流或許不能有效的對使用者行為進行分析與預測。

本篇論文利用自製的 Google Chrome 瀏覽器中的 plugin 收集完整的使用者資料，將點擊流依觸發方式分為有意圖及無意圖的點擊流並把原先缺失的額外點擊流一併記錄下來，並對其三種點擊流資料進行數據上的分析以及統計。我們透過結合不同的點擊流之資料去預測下次點擊之事件類型、下次點擊會間隔多久以及未來的瀏覽之網站比例，去探討不同的點擊流資料集對於不同的預測項目的影響。

透過本篇論文提出來的研究成果，讓分析使用者行為之相關研究的人員能對點擊流有更多的認知，且能夠使用更加貼近使用者真實行為的資料對其做相關項目之預測，甚至可以因應不同的預測項目而選擇相對應更適合的點擊流做使用。

1.4 論文架構

本篇論文共分為六個章節，其架構如下：

第一章、說明本篇論文之研究動機、研究目標、研究貢獻。

第二章、介紹本篇論文相關的研究。

第三章、介紹有意圖點擊流、無意圖點擊流及額外點擊流。

第四章、詳細介紹原始資料集、資料前處理之過程與資料統計分析。

第五章、說明實驗之設計、挑選之特徵及分類器之選擇。

第六章、展示實驗在不同資料集及不同模型的結果並討論。

第七章、本篇論文結論與未來展望。

二、相關研究

點擊流被廣泛用於代表使用者的在線行為，而透過點擊流去分析使用者行為也隨之被廣泛應用 [4], [5]，接著根據其分析結果對使用者進行個性化的預測及推薦 [6], [7]。本章節首先將透過生活中較常接觸網站來介紹對使用者行為分析之相關應用與研究，再介紹和本篇論文相關的點擊流研究。

2.1 網路日誌應用實例

2.1.1 社群網站之應用

以前若想去分析人與人之間的互動或觀察人們的行為模式，可能需要透過大量問卷調查，或是招募自願者進行模擬實驗等方法來收集實驗相關之數據。而現今，因為人們普遍習慣使用社群網站，其中著名的包括 Twitter、Facebook、Instagram 等社群網站，用其與人分享心情及生活。漸漸的也開始出現透過社群網站的網路日誌來對網站使用者的行為分析或比較等研究 [3], [8], [9]。

廣為人知的社交平台 Facebook，擁有大量的使用群體在其網站上留下個人資訊。而 Facebook 透過使用者對於貼文表達喜愛 (Like) 的行為來收集使用者對於文章類型 (音樂、電影、運動、書籍) 的喜好，進而對使用者進行一系列之個人屬性的分析，包含年齡、性別、性向、人格特質等 [10]。接著根據對使用者分析後的結果，針對不同類型的使用者投

放特定的廣告 [11]，這是目前 Facebook 投放廣告的方法之一。

另一個生活中的常使用的社群網站就是 Youtube，許多人在該網站上傳影片及觀看影片。為了能讓使用者能快速的找到想要觀看的影片，網站中不論是首頁或是影片播放時的右側，都有推薦影片區供使用者點選。Youtube 將推薦模型分為兩部分，先用 K-nearest neighbor 挑選適合推薦給使用者的百名影片，再透過使用者觀看紀錄、搜尋資訊、個人資料、地理位置等特徵產出的嵌入對影片進行重新排名 [12]。

2.1.2 電商網站之應用

普遍認為使用者的在線行為可用點擊流代表，所以許多線上購物平台使用點擊流的資訊對消費者行為模式進行分析 [11]，並利用其結果對使用者進行商品的推薦以及廣告的投放。傳統上，基於協同過濾 (Collaborative Filtering) 應用矩陣分解 (Matrix factorization) 或其他方法，找尋共同興趣或經驗之群體的喜好做為推薦 [13]。最近，有許多研究使用深度學習的方法發現商品和使用者的嵌入，然後根據這些嵌入做預測與推薦 [14], [15]。

中國電商平台 Alibaba 其研究團隊近年來不斷的發表可應用於購物網站之推薦模型 [16]–[18]，透過對使用者的點擊流、個人資訊及商品進行嵌入再結合不同的想法和技術，包括捕捉使用者之興趣、同序列中興趣應相近、興趣之演化路徑等，進行推薦系統的模型建構。

另外在美國佔市場比例高達 47% 的電商網站 Amazon，該平台使用者數繁多也擁有許多的使用者瀏覽資訊。傳統上一般都是研究使用者未來想購買的商品，因為直觀上會認為使用者不會再購買已買過的商品，然而事實上針對生活用品 (牙膏、衛生紙、洗髮精、牙線等等) 使用者傾向於買已經買過得商品，而這類的購買被 Amazon 定義為重複購買。Amazon 基於平台擁有的豐富瀏覽資料，透過重複購買的統計分布結果加上其時間間隔因素，判斷該使用者是否會回購商品及預測推薦此商品

之最佳時機 [2]。

2.1.3 旅遊租屋網站之應用

使用網際網路解決生活中的事項已經成為常態，因應這樣的社會趨勢越來越多的線上平台出現在網路上，而其中與人們娛樂相關的便是旅遊租屋網站。人們透過網路或手機應用程式發布、搜索度假房屋資訊信息完成在線預約，常見的平台有 Agoda、Trivago 及 Airbnb。

旅遊租屋網站不同於購物網站，如果僅僅是依照客戶方的需求以及偏好推薦房間顯然是不行的，還要考慮到屋主也就是供給方的意願。基於這樣的想法 Airbnb 在 2018 年發表相關的研究 [1]，使用 Word2Vec 的模型透過使用者的點擊將商品也就是房間進行嵌入，以此來捕捉使用者短期的喜好，然後再結合使用者的訂房紀錄對使用者進行個人化的推薦。

2.2 擴展點擊流：分析點擊流中缺少的使用者行為

擴展點擊流：分析點擊流中缺少的使用者行為 [19] 中提出點擊流只能概略表示使用者部分行為，例如：分頁切換、視窗切換等介面間瀏覽行為因未與伺服器進行互動，所以不會被記錄在網路日誌中。這類型的使用者行為，被其命名為「擴展點擊流」。研究中針對擴展點擊流與傳統點擊流進行分析，並使用 GRU 元件的深度模型進行效能預測。

本研究為擴展點擊流：分析點擊流中缺少的使用者行為 [19] 之延伸，不同之處在於我們認為傳統點擊流中使用者之無意圖行為對於使用者行為分析可能雜訊大於資訊，因此我們使用有意圖點擊流之資料集和有意圖加擴展點擊流之資料集跟傳統的點擊流之資料集進行比較。

二、相關研究

三、有意圖點擊流、無意圖點擊流及擴展點擊流

本章節中，分為兩個部分。第一部分，說明我們是如何將點擊流的資料區分為使用者有意圖及無意圖行為，隨後分別對兩者資料中所包含的觸發類型進行介紹。第二部分，介紹擴展點擊流並詳細的說明其中包含的行為以及它的觸發類型定義。

3.1 點擊流

點擊流 (clickstream, CS) 是在線使用者跨網站瀏覽的日誌，一般這種類型的數據是從用戶端收集的，由於這種類型的數據集很少，所以我們用自製 Google Chrome plugin 紿受試者下載使用來收集數據。在點擊流數據中我們紀錄觸發下個網址 (URL) 的轉換類型 (transition type)，並依照轉換類型的定義¹將點擊流事件分成有意圖行為和無意圖行為。

根據定義，點擊流之轉換類型共分為 11 類。

- 鏈結 (link): 通過鏈結點擊進入頁面
- 打字 (typed): 通過輸入 URL 進入頁面
- 自動書籤 (auto_bookmark)，通過 UI 中建議圖示進入頁面
- 手動子框訊 (manual_subframe): 使用者明確請求加載子框架

¹https://developer.chrome.com/extensions/history#transition_types

三、有意圖點擊流、無意圖點擊流及擴展點擊流

- 生成 (generated): 通過網址列中輸入並選擇不像 URL 輸入進入頁面
- 關鍵字 (keyword): 該頁面是從默認搜尋提供程序以外的可替換關鍵字生成的
- 生成關鍵字 (keyword_generated): 對應於為關鍵字生成的進入
- 重新載入 (reload): 通過重新加載按鈕或網址列的 enter 重新加載頁面

由上述轉換類型所觸發的點擊流我們稱之為有意圖點擊流 (Intentional clickstream, ICS)。

另外三類則定義為非意圖的行為，因為這三類中的網頁觸發並非使用者本身而是源自於程序導致。

- 自動子訊框 (auto_subframe): 頁面自動加載到非頂層的框架中
- 表單提交 (form_submit): 頁面在提交後自動定址
- 自動頂層 (auto_toplevel): 頁面是起始頁面中的指定頁面

3.2 擴展點擊流

擴展點擊流 (Extended clickstream, ECS) 是在 2019 的研究中被提出 [19]，根據定義它是不包含在傳統點擊流中的額外點擊流資訊。其轉換類型總共五種，分別為：

- 分頁 (Tabs): 同一個視窗中進行分頁的切換
- 視窗 (Windows): 同瀏覽器中進行的視窗切換
- 模糊 (Blur): 使用者切換到其它應用程序而觸發，包含關閉瀏覽器或返回桌面

- 空閒 (Idle): 使用者超過 2 分鐘沒有進行 I/O 輸入導致系統被鎖定或進入睡眠狀態
- 活動 (Active): 對應空閒類型的配對事件

表 3.1 整理出有意圖點擊流、無意圖點擊流及擴展點擊流的例子。由表中可知我們認為擴展點擊流中並不包含無意圖的行為，因為其所有觸發的轉換類型皆出於使用者自身的行為。

表 3.1: 有意圖點擊流、無意圖點擊流及有意圖擴展點擊流之例子

	intentional behavior	unintentional behavior
CS	clicks on link or bookmark URL typing on address bar	pages loaded in subframe pop-up windows page auto-redirect
ECS	tab or browser switching	-

三、有意圖點擊流、無意圖點擊流及擴展點擊流

四、資料集介紹

本章節主要討論三個部分。第一節介紹原始資料集裡的內容。第二節介紹對資料集進行前處理的過程與想法。第三節對資料進行統計與分析。

4.1 原始資料集

本篇論文所使用之資料集是取自 147 位使用者之 Google Chrome plugin 之網頁紀錄，時間從 2019 年 2 月 26 日至 2019 年 7 月 17 日。總觸發事件量為 6623178 筆，其中 CS 共 3778777 筆而 ECS 為 2844401 筆。表 4.1為使用者下載 plugin 後的使用天數統計表，而表 4.2為原始資料集中 CS 及 ECS 的事件轉換類型之比例統計表，透過表 4.2可以發現轉換類型前三名中 ECS 的轉換類型就佔兩個。

表 4.1: 使用者 plugin 使用天數之統計

	min	Q1	median	mean	Q3	max
Days	3	70.5	110	41.25	131.5	142

表 4.2: CS 及 ECS 的事件轉換類型之比例統計 (%)

CS	link	typed	auto_bookmark	manual_subframe	generated
Perc(%)	47.3726	0.8231	2.5437	0.0325	1.6619
CS	keyword	keyword_generated	reload	auto_subframe	form_submit
Perc(%)	0.0022	0.0	1.3892	0.0004	2.8147
CS	auto_toplevel				
Perc(%)	0.4135				
ECS	tabs	windows	blur	idle	active
Perc(%)	25.5639	6.8717	5.695	2.4741	2.3415

4.2 資料集前處理

4.2.1 時間單位前處理

從 CS 和 ECS 的資料中我們發現使用 plugin 收集到的事件時間之單位兩者是不一致的。CS 資料中的時間單位為秒而 ECS 資料的單位則是微秒，這個現象會在合併的資料時產生問題，使得 Idle 事件和 Active 事件無法接續出現。

圖 4.1(a) 顯示正常使用者瀏覽網頁行為，根據 ECS 的轉換類型定義，Idle 事件和 Active 事件會成對且連續出現，中間不會夾雜其它事件。但因為資料收集時對於 CS 及 ECS 的時間單位問題，導致原本應該出現在 Active 事件後的 CS 事件可能時間紀錄會早於 Active 事件，如圖 4.1(c) 中的事件時間，微秒所記錄到的時間應在 Active 後，但若用秒紀錄時間則時間會不一樣，使得事件的順序產生變化，進而出現圖 4.1(b) 的狀況。為了解決這個問題，我們對 Active 事件前一秒內的 CS 事件進行時間上的微調，將其移動到 Active 之後。

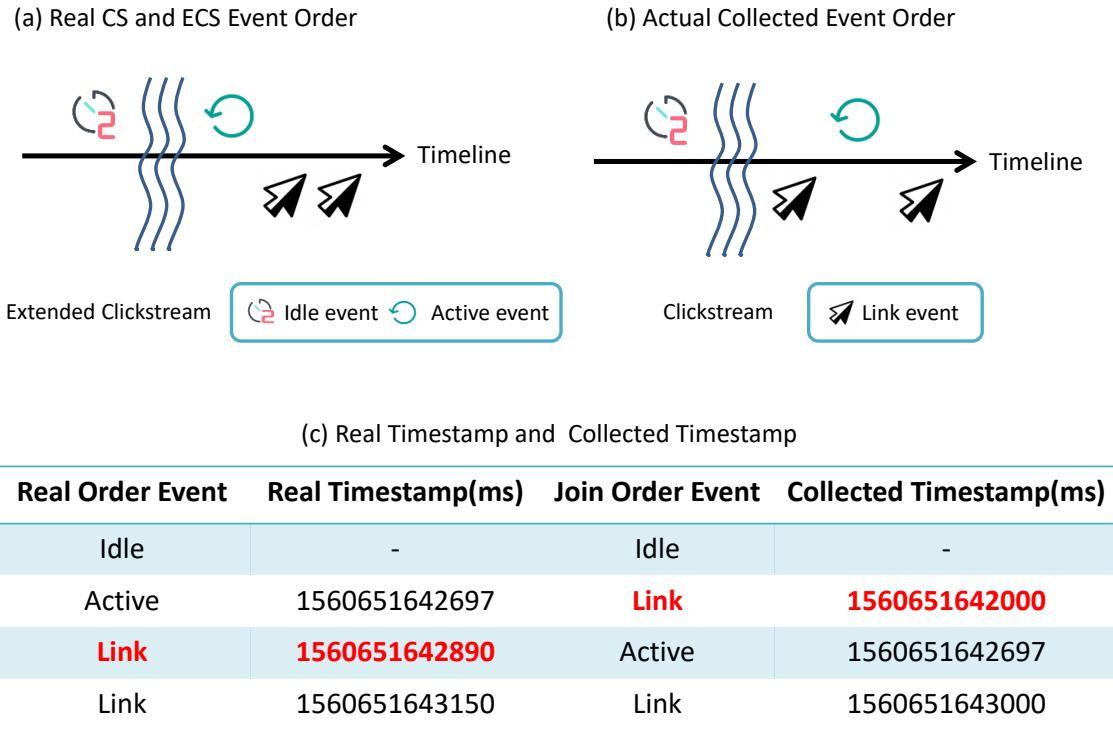


圖 4.1: 真實數據與收集資料之事件時間差異

4.2.2 網站網址前處理

網站中我們移除了表示本地主機的 URL(127.0.0.1、192.168.0.1 等)，因為我們認為這類型的網址是網頁日誌中的雜訊。我們將每一筆網址透過網頁分類器¹整理為不同網頁類型，總共將所有網頁歸納為 82 種類別。例如“google.com”歸類為“Search Engines and Portals”、“facebook.com”歸類為為“Social Networking”。表 4.3為不同資料集中網站類型之事件數量統計，Category_count 欄位表示當個資料中網站類型數量。

¹<https://fortiguard.com/webfilter>

表 4.3: 網站類型之事件數量統計

	Category_count	min	Q1	median	mean	Q3	max
CS	81	2	92.5	2251.5	46082.65	24274	608224
ICS	81	2	72	2168	43474.68	20306.75	600171
ICS + ECS	81	2	184.25	3262	78162.61	38921.25	1124145

4.2.3 事件間之時間間隔前處理

從原始資料中，我們幫每一筆事件生成與前者事件之時間間隔，並對時間間隔數據進行統計，其分佈狀況如圖 4.2。我們認為預測使用者瀏覽網頁之間隔不需要將其時間分的過細，所以最終我們將時間間隔分為五類。表 4.4顯示 5 種時間間隔並統計其含有的事件數量，圖 4.3為不同資料集中針對五種時間間隔的事件數量比較圖。

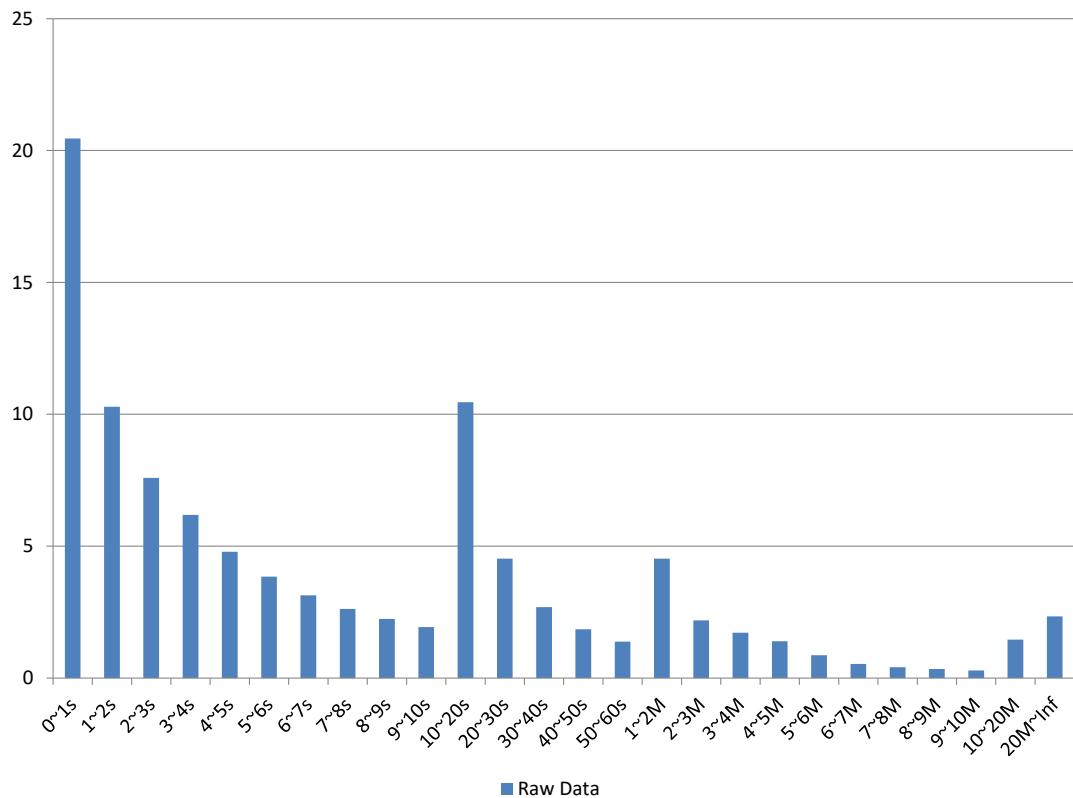


圖 4.2: 原始資料集中時間間隔之事件數量統計圖

表 4.4: 原始資料集中各時間間隔中事件數量統計

	00s_05s	05s_20s	20s_2M	2M_20M	20M_NW
Count	3126997	1471136	1155525	739855	129665
Prec(%)	47.213	22.212	17.447	11.171	1.958

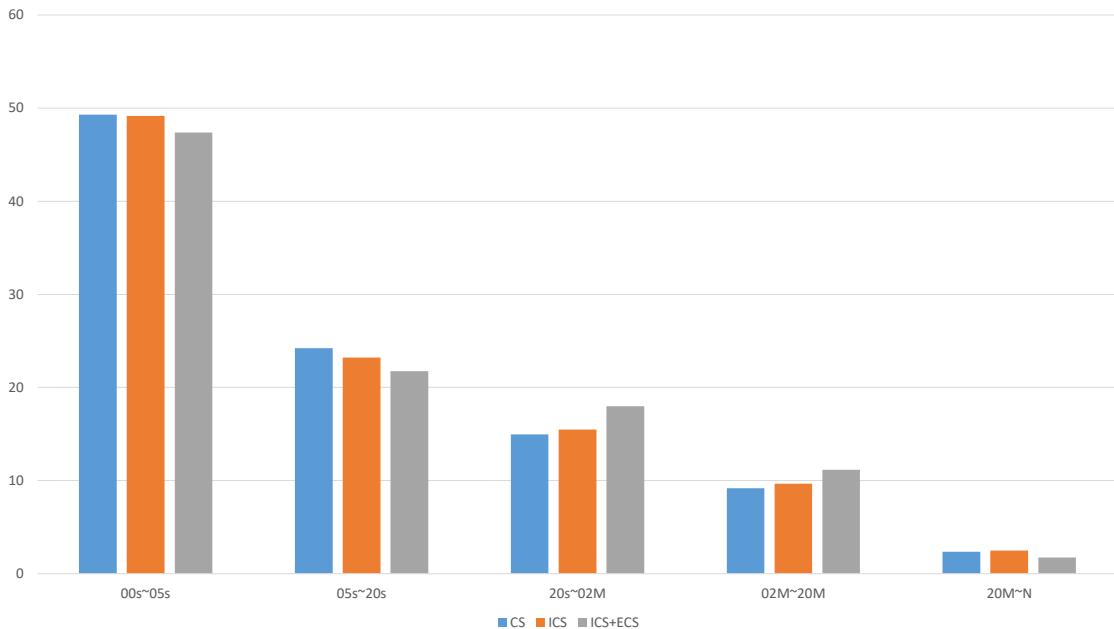


圖 4.3: 不同資料集中時間間隔的事件數量長條圖

4.3 資料統計與分析

進行完資料前處理後，我們分別對不同點擊流的數量進行統計，表 4.5和表 4.6分別說明每一位使用者觸發事件量及平均每天的觸發事件量之統計表。透過表 4.5可以發現 ICS 的事件數量約佔 CS 中 5.6%，而推算出來的 ECS 事件數量幾乎和 CS 事件數量一樣多，並且遠多於 ICS。根據此結果表示傳統的網路日誌中遺失將近了一半的使用者行為，而記錄到的行為中又有幾乎 5.6% 不是出自於使用者。因此我們認為傳統的點擊流只能代表部分的使用者行為，以此資料進行的使用者行為分析也不夠全面。

而後對網站類型點擊數量進行排名時我們發現，CS、ICS、ICS+ECS 中的類型排名是不同的。根據表 4.7發現雖然在三種資料中前 20 名的網

表 4.5: 使用者事件數量之統計

	min	Q1	median	mean	Q3	max
CS	21	11104.5	20381	20212.71	35941	110580
ICS	21	10503.5	19821	19180.59	33590.5	103635
ICS + ECS	73	18862.3	36421	34936.55	61137.5	178473

表 4.6: 使用者平均每天事件數量之統計

	min	Q1	median	mean	Q3	max
CS	7	137.416	231.071	157.278	328.445	931.929
ICS	7	127.709	221.142	150.093	311.385	892.929
ICS + ECS	24.33	238.218	387.44	269.526	559.419	1758.357

站是一致的，但是如果只用 CS 資料進行排名會發現“Entertain”、“Web-based Application”、“Auction”等類型的網頁排名會被高估，而“Personal Website and Blogs”、“Streaming Media and Download”、“Shopping”等類型則會被低估。這樣的結果表示在前者網頁類型中，使用者不習慣將其放置在分頁並且重複關注。後者則相反，使用者傾向於開啓之後放置分頁中，並用分頁切換的方式重複瀏覽，而這類的瀏覽行為是會被傳統的網路日誌忽略錯過的。

表 4.7: 不同轉換類型資料中前 20 名網站類型的事件統計

Category	Rank(1)	ICS + ECS			CS			ICS			ECS			Rank Diff (1)-(2)
		Count	Perc(%)	CDF(%)	Rank(2)	Count	Rank	Count	Rank	Count	Rank	Count	Rank	
Streaming Media and Download	1	1124145	17.54	17.54	3	558297	2	555767	1	568378	-2			
Social Networking	2	938817	14.65	32.19	1	608224	1	600171	2	338646	1			
Search Engines and Portals	3	714769	11.15	43.34	2	559221	3	461377	5	253392	1			
Education	4	570649	8.9	52.24	5	304364	5	276313	3	294336	-1			
Information Technology	5	457948	7.15	59.39	6	200180	6	189298	4	268650	-1			
Web-based Application	6	391608	6.11	65.50	4	336886	4	333320	11	58288	2			
Games	7	386895	6.04	71.54	7	156347	7	152642	6	234253	0			
Business	8	203138	3.17	74.71	9	108060	10	99250	7	103888	-1			
Shopping	9	168830	2.63	77.34	11	94737	11	88601	8	80229	-2			
File Sharing and Storage	10	165818	2.59	79.93	10	106535	9	105062	10	60756	0			
Entertainment	11	155154	2.42	82.35	8	117181	8	115618	14	39536	3			
Reference	12	154226	2.41	84.76	12	86090	12	80408	9	73818	0			
Web-based Email	13	115595	1.8	86.56	13	68741	13	68178	12	47417	0			
News and Media	14	100603	1.57	88.13	14	67278	14	66567	17	34036	0			
Newsgroups and Message Boards	15	72541	1.13	89.26	16	35036	16	33127	15	39414	-1			
Pornography	16	69762	1.09	90.35	15	42031	15	40939	18	28823	1			
Personal Websites and Blogs	17	68312	1.08	91.56	20	25497	20	25220	13	43092	-3			
Instant Messaging	18	63289	0.99	92.42	18	29973	18	28931	16	34358	0			
Auction	19	55927	0.87	93.29	33343	33344	17	33078	20	22849	2			
Travel	20	49540	0.77	94.06	19	29955	19	25631	19	23909	1			

四、資料集介紹

五、實驗

本章節首先敘述問題與想法，接著提出三個實驗並說明預測之目標、特徵選擇。最後介紹使用之分類器。

5.1 問題與想法

與普遍認知不同，傳統的 CS 中所收集的資料並不等於使用者實際的線上行為，其中不僅多收集一部份非意圖的行為另外還缺失一部份行為。因此我們認為使用傳統的 CS 去進行使用者分析，不一定能達到最好的效果。

我們重新將使用者缺失的資料收集回來，並且進行有無意圖行為的分類，最後將資料分成 ICS 資料集及 ICS 加 ECS 資料集。對此我們認為 ICS 和 ICS 加 ECS 的資料集應該更加貼近使用者的實際行為，且認為兩者資料集應該對於使用者行為分析會有更好的效果。

我們設計三個實驗實驗去測試我們的觀點是否正確，分別是預測使用者下次點擊網站類型、預測使用者下次點擊間隔時間及預測使用者未來瀏覽之網站比例。

5.2 實驗設計

在預測網站類型及點擊間隔時間的實驗中，為了和傳統的 CS 的資料做比較，我們將預測的目標限定為由 CS 轉換類型所觸發的事件。而

在未來瀏覽之網站比例的實驗中，我們預測全部資料中 Top1、Top2 之網站類型在未來一天的點擊比例，點擊比例共分為 11 種。前者的特徵的選擇上我們使用目標 CS 事件之前 5 項行為中的網站類型、點擊之時間間隔和網站轉換類型，圖 5.1為示意圖。後者則選擇目標日期之前 5 個有紀錄的日期之網站類型前五類型比例、點擊之網站類型數量及當天之點擊數量。

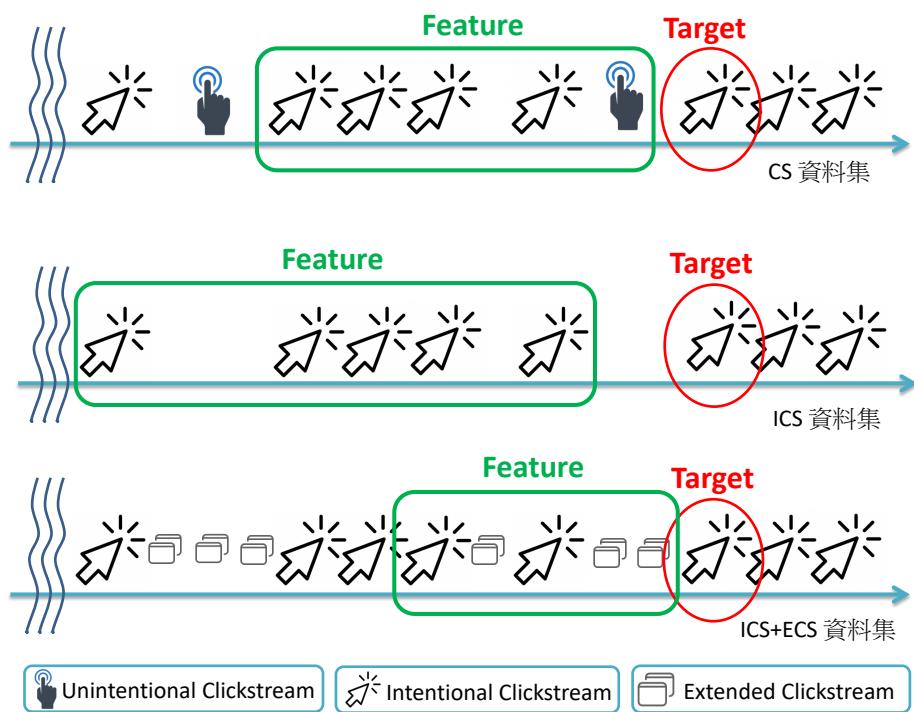


圖 5.1: 預測網站類型之不同資料集中特徵的選擇

5.3 分類器選擇及介紹

由於我們所要預測的項目皆屬於分類問題，因此選用最近鄰居法 (K-Nearest Neighbor, k-NN)[20]、隨機森林 (Random forest)[21]、極限梯度提升 (Extreme Gradient Boosting, XGBoost)[22] 以及邏輯回歸 (Logistic regression) 作為分類器進行預測。

- k-NN: 分類問題中，依據樣本特徵為座標，選定 k 個相鄰最近的樣

本類別最為預測的參考，最終以大多數相鄰的樣本類別當成最後的預測類別。

- Logistic regression: 利用多個自變數 (independent variable) 和依變數 (dependent variable) 之間其關聯性所建立的模型。與 Linear Regression 相似，差別於會將輸出透過 sigmoid 函數壓縮到 0 至 1 區間，並以 cross entropy 作為損失函數。
- Random forest: 針對相同的資料集，由隨機的方式選取部分特徵或部分資料來產生多棵不同的決策樹 (decision tree)，再將這些決策樹的結果經由簡易的方法 (平均值、多數決) 來獲得最後結果。
- XGBoost: 是一種梯度提升的決策樹。通過新加入的弱學習器，更新之前所有的弱學習器之殘差，最終再將多個學習器結果相加用來最為最終預測。預測時運算的速度快且有極佳的效果。

五、實驗

六、 實驗結果與分析

本章節首先說明如何分配資料集以及採用的模型評估標準，最後對實驗的結果提出分析與討論。

6.1 實驗資料集介紹

基於第四章的資料集，本實驗最終使用三種資料集進行預測，分別是 CS、ICS、ICS+ECS。其中我們以 ICS 的資料集中的時間為基準，用天數將使用者的瀏覽紀錄分為訓練集、驗證集以及測試集。圖 6.1為資料集分配的示意圖，可以看到測試集為使用者後 5 天的資料，而剩餘的資料則為訓練集。若使用者的訓練集大於 25 天，其訓練集中的後 5 天料為驗證集，而在我們的資料集中約有 90% 的使用者擁有超過 30 天的資料。

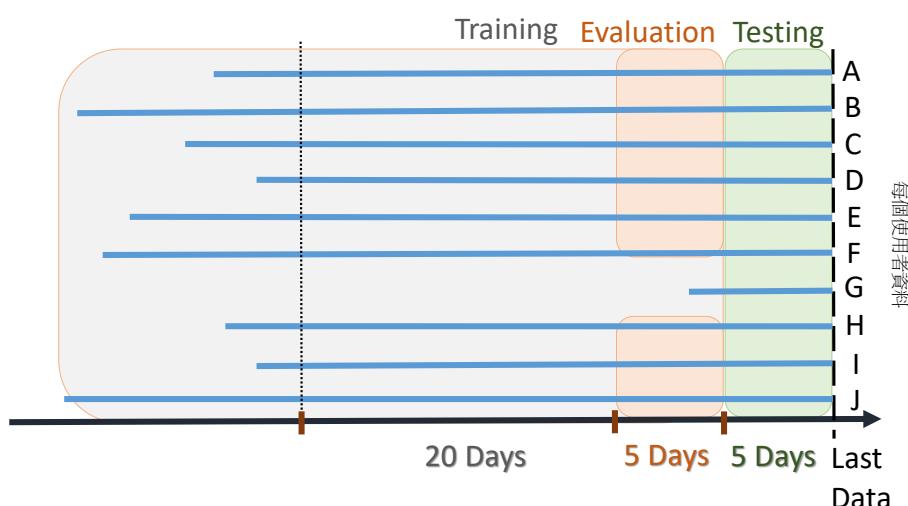


圖 6.1：資訊集分配示意圖

6.2 模型評估標準

一般的分類問題，大部分以 F_1 score 進行精準度的評估，主要在召回率 (recall) 與精確率 (precision) 之間取得平衡，透過圖 6.2 中定義的 TP、FP、FN、TN 來進行計算，其公式如下。

$$F_1 \text{ score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (6.1)$$

$$\text{Precision} = \frac{tp}{tp + fp}, \quad \text{Recall} = \frac{tp}{tp + fn} \quad (6.2)$$

		Actual	
		Positive	Negative
Prediction	Predict Positive	TP	FP
	Predict Negative	FN	TN

圖 6.2: 判斷測試表

由於本實驗中所進行的預測皆為多類別 (multi-class) 之分類問題，因此無法直接以 F_1 score 作為評估標準，必須改用 $MircoF_1$ 與 $MacroF_1$ 對多分類問題進行評估，其公式如式 6.3、6.4。 $MircoF_1$ 為將所有類別的 TP、FP、FN、TN 分別進行累加後再進行 F_1 score 的計算，而 $MacroF_1$ 則是對將每個類別的 F_1 score 算完後再平均。

兩者之間最明顯的差異在於，有無考慮每種類別的樣本數量平衡問題。 $MacroF_1$ 的計算方式下，樣本數較少的類別對於整體的分數的影響

會等同於其他類別，因此使用的時候要考慮到樣本類別間數量是否平均。而本實驗所用的資料集，每種類別之樣本數較不平衡，因此最後選擇以 $MircoF_1$ 為評估標準。

$$MicroF_1 = \frac{2 \cdot \sum_{i=1}^C TP_i}{2 \cdot \sum_{i=1}^C TP_i + \sum_{i=1}^C FP_i + \sum_{i=1}^C FN_i} \quad (6.3)$$

$$MacroF_i = \frac{1}{C} \sum_{i=1}^C \frac{2 \cdot TP_i}{2 \cdot TP_i + FP_i + FN_i} \quad (6.4)$$

6.3 實驗結果分析討論

本章節將介紹的 4 種分類器：k-NN、Logistic regression、Random forest 以及 XGBoost，透過混淆矩陣 (confusion matrix) 以及 $MircoF_1$ 之分數比較 CS、ICS 與 ICS+ECS 三者間預測結果。

所有實驗中，我們一律使用驗證集對不同的 k-NN 之 k 值、Random forest 和 XGBoost 之 max_depth 及 Logistic regression 之 C 值計算 $MircoF_1$ ，找出不同資料集中最佳的參數，再使用其參數進行測試集之預測。

6.3.1 預測使用者下次點擊網站類型

本實驗經由驗證集找出對於 CS、ICS、ICS+ECS 最佳參數，再給予測試集預測的結果如表 6.1。其中可發現樹狀結構之分類器中 ICS+ECS 之效果會最好，而在 k-NN 和 Logistic regression 中則分別是 ICS 和 CS 之資料集效果更好。由結果發現新生成的 ICS+ECS 和 ICS 資料集在大部分分類器中效果皆比傳統 CS 好，代表著 ICS 和 ICS+ECS 的資料集確實比傳統 CS 資料集更加貼近使用者真實行為，也可以對分析帶來更好的效果。

表 6.1: 預測下次點擊網站類型之 $MircoF_1$

分類器	CS	ICS	ICS+ECS
Logistic regression	0.407	0.402	0.398
k-NN	0.729	0.738	0.727
Random forest	0.765	0.767	0.767
XGBoost	0.772	0.775	0.777

圖 6.3為 XGBoost 使用驗證集對不同最大深度 (max_depth) 之 $MircoF_1$ 預測效果比較，由驗證集的結果顯示對 CS、ICS 和 ICS+ECS 之資料集，最大深度分別在 7、9、8 時達到最佳，其 $MircoF_1$ 分別為 0.7816、0.7793、0.7853。

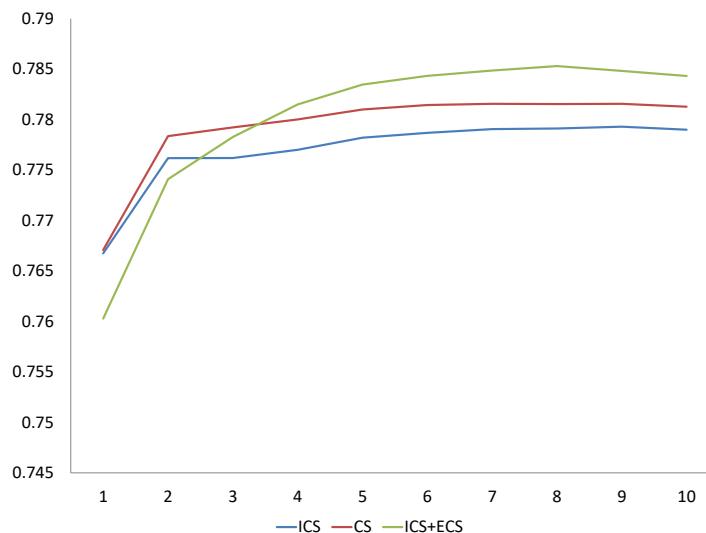


圖 6.3: 實驗一: XGBoost 中根據不同 max_depth 值對應之 $MircoF_1$ 分數

圖 6.4、圖 6.5為 XGBoost 對三個資料集所生成的混淆矩陣，欄位的數字對應全部資料集中網站類型的排名。上半部為資料集前五名排名，下半部為 16 到 20 名，紅色是準確度高於傳統 CS 的地方，完整之混淆矩陣將於附錄中補充。每格中上面的數字代表類別中預測結果為該類別的數量，下面則是該數量占類別的比例，因此對角線值為每個類別正確預測的數量以及類別的準確度。由此可以發現 ICS 和 ICS+ECS 中樣本數較少的類別其準確度會更較佳。

true/predict	1	2	3	4	5
1	19141 (79.82)	1232 (5.14)	804 (3.35)	466 (1.94)	327 (1.36)
2	1184 (4.78)	20799 (83.98)	533 (2.15)	395 (1.59)	242 (0.98)
3	1238 (6.30)	749 (3.81)	12958 (65.97)	834 (4.25)	931 (4.74)
4	374 (3.95)	341 (3.60)	687 (7.25)	7050 (74.41)	229 (2.42)
5	233 (3.65)	228 (3.57)	792 (12.39)	207 (3.24)	4210 (65.87)

ICS

true/predict	1	2	3	4	5
1	19033 (80.97)	1206 (5.13)	821 (3.49)	410 (1.74)	281 (1.20)
2	1328 (5.42)	20575 (83.94)	545 (2.22)	371 (1.51)	217 (0.89)
3	1472 (7.56)	753 (3.87)	12884 (66.17)	808 (4.15)	859 (4.41)
4	415 (4.59)	336 (3.72)	718 (7.95)	6678 (73.92)	216 (2.39)
5	236 (3.94)	210 (3.50)	791 (13.20)	190 (3.17)	3939 (65.72)

ICS+ECS

圖 6.4: XGBoost 之混淆矩陣評估 (Top 1-5)

六、實驗結果與分析

true/predict	16	17	18	19	20
16	1961 (88.69)	1 (0.05)	0 (0.00)	3 (0.14)	0 (0.00)
17	0 (0.00)	223 (22.32)	0 (0.00)	5 (0.50)	17 (1.70)
18	0 (0.00)	1 (0.15)	420 (64.81)	0 (0.00)	1 (0.15)
19	1 (0.06)	2 (0.12)	0 (0.00)	1359 (83.58)	1 (0.06)
20	1 (0.25)	4 (1.00)	0 (0.00)	1 (0.25)	231 (57.75)

ICS

true/predict	16	17	18	19	20
16	1916 (87.81)	1 (0.05)	0 (0.00)	3 (0.14)	0 (0.00)
17	0 (0.00)	222 (23.05)	1 (0.10)	4 (0.42)	15 (1.56)
18	0 (0.00)	1 (0.16)	416 (64.70)	0 (0.00)	1 (0.16)
19	0 (0.00)	1 (0.06)	0 (0.00)	1337 (83.41)	1 (0.06)
20	1 (0.26)	3 (0.79)	1 (0.26)	1 (0.26)	228 (60.00)

ICS+ECS

圖 6.5: XGBoost 之混淆矩陣評估 (Top 16-20)

6.3.2 預測使用者下次點擊間隔時間

本實驗主要預測下一次使用者點擊的間隔，透過不同的資料集其預測的結果如表 6.2。透過表 6.2 發現三者資料集之效果排名為 ICS+ECS > CS \approx ICS，其中 XGBoost 的分類器最優。

表 6.2: 預測下次點擊時間間隔之 $MircoF_1$

分類器	CS	ICS	ICS+ECS
k-NN	0.511	0.512	0.519
Logistic regression	0.487	0.488	0.499
Random forest	0.534	0.533	0.544
XGBoost	0.537	0.535	0.547

圖 6.6為 XGBoost 使用驗證集對不同最大深度 (max_depth) 之 $MircoF_1$ 預測效果比較，由驗證集的結果顯示對 CS、ICS 和 ICS+ECS 之資料集，最大深度分別在 12、11、12 時達到最佳，其 $MircoF_1$ 分別為 0.543、0.540、0.552。

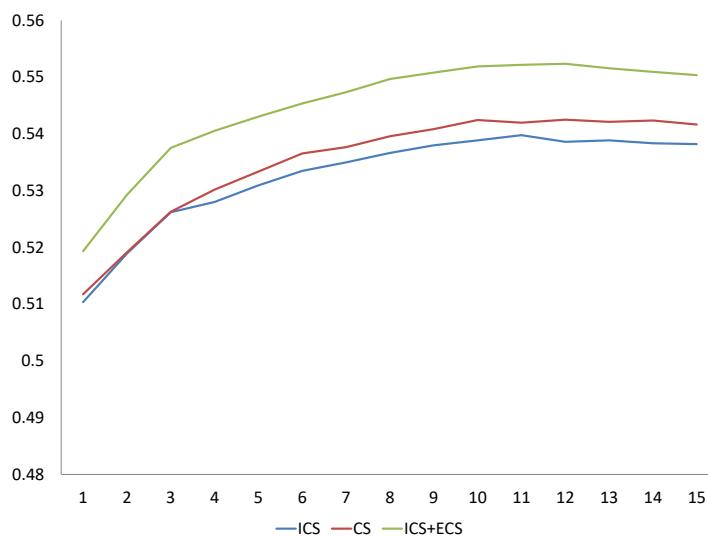


圖 6.6: 實驗二: XGBoost 中根據不同 max_depth 值對應之 $MircoF_1$ 分數

根據圖 6.7發現 ICS+ECS 資料集，在預測樣本數較少的類別時其準確度相對於資料少的資料集 (ICS、CS) 更好，進而我們推測在預測時間

六、實驗結果與分析

間隔的任務中，使用者行為資料越完整越多的狀況下，較少樣本數之類別的預測狀況會越好。

True/Predict	1	2	3	4	5
1	61098 (83.06)	9036 (12.28)	2179 (2.96)	1204 (1.64)	38 (0.05)
2	19036 (51.26)	13293 (35.08)	3219 (8.67)	1544 (4.16)	41 (0.11)
3	10751 (44.10)	7513 (30.82)	4228 (17.34)	1854 (7.60)	33 (0.14)
4	5809 (39.38)	2914 (19.75)	2153 (14.60)	3828 (25.95)	47 (0.32)
5	1848 (47.09)	860 (21.92)	490 (12.49)	579 (14.76)	147 (3.75)

CS

True/Predict	1	2	3	4	5
1	58269 (84.12)	7435 (10.73)	2304 (3.33)	1224 (1.77)	34 (0.05)
2	17585 (52.51)	10874 (32.47)	3446 (10.29)	1550 (4.63)	36 (0.11)
3	10563 (44.00)	6885 (28.68)	4678 (19.48)	1839 (7.66)	44 (0.18)
4	5724 (39.13)	2687 (18.37)	2331 (15.93)	3837 (26.23)	50 (0.34)
5	1811 (46.32)	843 (21.56)	553 (14.14)	561 (14.35)	142 (3.63)

ICS

True/Predict	1	2	3	4	5
1	57971 (84.27)	7458 (10.84)	2411 (3.50)	929 (1.35)	25 (0.04)
2	17775 (53.33)	10785 (32.36)	3526 (10.58)	1218 (3.65)	25 (0.08)
3	10210 (42.94)	6626 (27.87)	5119 (21.53)	1705 (7.17)	116 (0.49)
4	4788 (36.74)	2285 (17.54)	2527 (19.39)	3282 (25.19)	149 (1.14)
5	1014 (37.07)	510 (18.65)	500 (18.28)	443 (16.20)	268 (9.80)

ICS+ECS

圖 6.7: XGBoost 之混淆矩陣評估

6.3.3 預測使用者未來瀏覽之網站比例

本實驗中我們將網站比例共分為五類，其中依序為 0-1%、1-5%、5-20%、20-60%、60-100%。表 6.3、6.4為四個分類器分別對三個資料集其預測的結果，參數方面皆使用上述用驗證集找到之最佳參數，由結果可以發現 Top1 為 Random forest 最佳而 Top2 為 XGBoost。

圖 6.8為 XGBoost 使用驗證集對不同最大深度 (max_depth) 之 $MircoF_1$ 預測效果比較，由驗證集的結果顯示對 CS、ICS 和 ICS+ECS 之資料集，最大深度分別在 2、3、3 時達到最佳，其 $MircoF_1$ 分別為 0.520、0.517、0.512。圖 6.9為 XGBoost 使用驗證集對不同最大深度

表 6.3: 預測 Top1 之網站類型比例之 $MircoF_1$

	Top1		
	CS	ICS	ICS+ECS
k-NN	0.337	0.341	0.318
Logistic regression	0.416	0.404	0.398
Random forest	0.497	0.502	0.519
XGBoost	0.484	0.497	0.515

表 6.4: 預測 Top2 之網站類型比例之 $MircoF_1$

	Top2		
	CS	ICS	ICS+ECS
k-NN	0.369	0.371	0.338
Logistic regression	0.402	0.401	0.372
Random forest	0.503	0.509	0.497
XGBoost	0.507	0.505	0.512

(max_depth) 之 $MircoF_1$ 預測效果比較，由驗證集的結果顯示對 CS、ICS 和 ICS+ECS 之資料集，最大深度分別在 1、1、2 時達到最佳，其 $MircoF_1$ 分別為 0.518、0.510、0.513。

圖 6.10、圖 6.11為 XGBoost 對三個資料集所生成的混淆矩陣，完整之混淆矩陣將於附錄補充。透過混淆矩陣可以發現，在這個任務中三個資料集普遍對於最少樣本類別之預測效果較不佳，但 ICS 和 ICS+ECS 之資料集對於樣本數量較少的類別效果大多還是比 CS 稍好。

六、實驗結果與分析

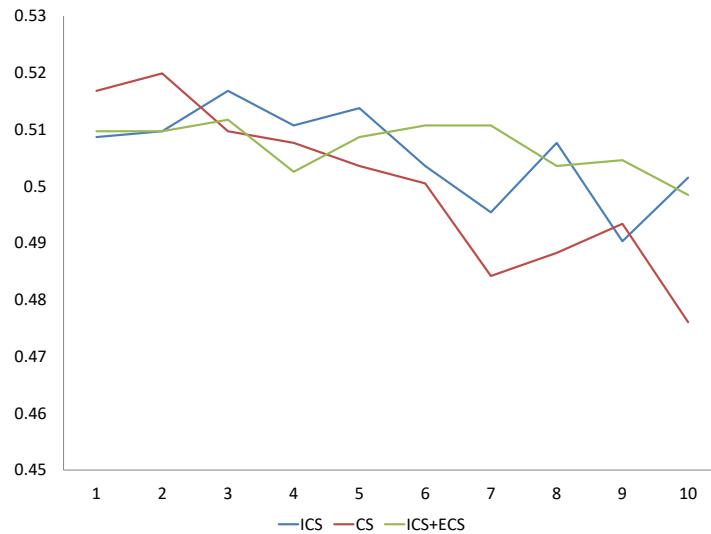


圖 6.8: 實驗三: XGBoost 中根據不同 max_depth 值對應之 Mirco F_1 分數 (Top1)

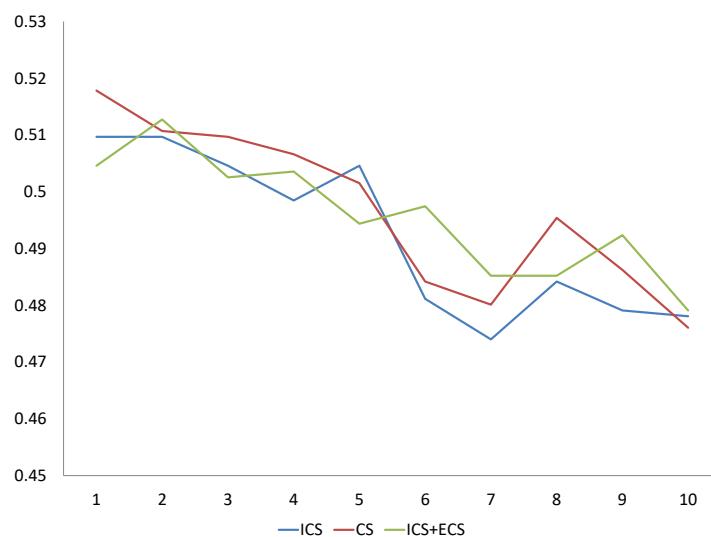


圖 6.9: 實驗三: XGBoost 中根據不同 max_depth 值對應之 Mirco F_1 分數 (Top2)

True/Predict	1	2	3	4	5
1	134 (55.60)	0 (0.00)	78 (32.37)	19 (7.88)	10 (4.15)
2	27 (25.23)	0 (0.00)	67 (62.62)	11 (10.28)	2 (1.87)
3	39 (10.71)	0 (0.00)	234 (64.29)	85 (23.35)	6 (1.64)
4	21 (6.76)	0 (0.00)	122 (39.23)	144 (46.30)	24 (7.72)
5	14 (9.52)	0 (0.00)	55 (37.41)	54 (36.73)	54 (36.73)

CS

True/Predict	1	2	3	4	5
1	147 (60.49)	1 (0.41)	59 (24.28)	25 (10.29)	11 (4.53)
2	32 (32.99)	0 (0.00)	47 (48.45)	17 (17.53)	1 (1.03)
3	37 (10.85)	0 (0.00)	199 (58.36)	98 (28.74)	7 (2.05)
4	31 (9.42)	0 (0.00)	94 (28.57)	169 (51.37)	35 (10.64)
5	16 (10.00)	0 (0.00)	18 (11.25)	60 (37.50)	66 (41.25)

ICS

True/Predict	1	2	3	4	5
1	155 (60.78)	1 (0.39)	67 (26.27)	25 (9.80)	7 (2.74)
2	32 (33.33)	0 (0.00)	48 (50.00)	14 (14.58)	2 (2.08)
3	40 (11.80)	0 (0.00)	202 (59.59)	93 (27.43)	4 (1.18)
4	26 (8.02)	0 (0.00)	87 (26.85)	179 (55.25)	32 (9.88)
5	20 (12.82)	0 (0.00)	15 (9.62)	55 (35.26)	66 (42.31)

ICS+ECS

圖 6.10: XGBoost 之混淆矩陣評估 (Top1)

六、實驗結果與分析

True/Predict	1	2	3	4	5
1	234 (71.56)	0 (0.00)	71 (21.71)	19 (5.81)	3 (0.92)
2	54 (37.24)	1 (0.69)	80 (55.17)	10 (6.90)	0 (0.00)
3	46 (13.94)	0 (0.00)	205 (62.12)	77 (23.33)	2 (0.61)
4	23 (8.78)	0 (0.00)	106 (40.46)	122 (46.56)	11 (4.19)
5	14 (13.21)	0 (0.00)	19 (17.92)	42 (39.62)	31 (29.25)

CS

True/Predict	1	2	3	4	5
1	231 (71.52)	0 (0.00)	68 (21.05)	21 (6.50)	3 (0.93)
2	53 (37.86)	0 (0.00)	77 (55.00)	10 (7.14)	0 (0.00)
3	49 (15.17)	0 (0.00)	188 (58.20)	85 (26.32)	1 (0.31)
4	23 (8.42)	0 (0.00)	97 (35.53)	142 (52.01)	10 (3.66)
5	13 (11.71)	0 (0.00)	20 (18.02)	48 (43.24)	30 (27.03)

ICS

True/Predict	1	2	3	4	5
1	243 (72.75)	2 (0.60)	67 (20.06)	18 (5.39)	4 (1.20)
2	55 (39.86)	0 (0.00)	77 (55.80)	6 (4.35)	0 (0.00)
3	51 (15.94)	2 (0.63)	176 (55.00)	89 (27.81)	2 (0.63)
4	26 (9.63)	1 (0.37)	85 (31.48)	147 (54.44)	11 (4.07)
5	15 (13.89)	0 (0.00)	21 (19.44)	39 (36.11)	33 (30.56)

ICS+ECS

圖 6.11: XGBoost 之混淆矩陣評估 (Top2)

七、結論與未來展望

7.1 結論

一般使用點擊流代表使用者的瀏覽行為，但本篇論文所收集到的資料集中發現傳統點擊流會缺少紀錄使用者行為，也會記錄到少數不屬於使用者自身意圖的行為。因此我們透過去除無意圖的行為和加入缺少沒被紀錄到的行為，產生兩個更貼近使用者真實瀏覽行為之新資料集 (ICS、ICS+ECS)。對新資料集與傳統 CS 資料集做比較、分析與預測，想進而證明新資料集在本篇實驗中相比於傳統 CS 資料對於使用者行為的分析預測更加的準確。

本篇論文之實驗分為預測下個點擊網頁類型、預測下次點擊的間隔以及預測網站類型比例三個部分。在實驗一中透過混淆矩陣可以發現 ICS+ECS 的資料集對於樣本數量極端之類別 (最多和最少)，預測之準確度相對傳統 CS 高。我們認為因為 ICS+ECS 之資料集為更加完整的使用者真實線上行為，所以才會提高樣本數量較少之類別的準確度。由整體的 $MircoF_1$ 數值來看，ICS+ECS 的資料集的結果是比較好的。

實驗二中 ICS+ECS 表現最佳，但不同於實驗一，ICS 之資料集效果在樹狀結構的分類器中比傳統 CS 差。透過混淆矩陣也可以發現 ICS+ECS 在較少類別時效果皆最佳，也發現此種結果在樹狀結構的分類器中更加明顯，此結果與實驗一吻合。

實驗三中我們分別對網站類型進行預測，Top1 中的結果為 ICS+ECS 較好，而 Top2 則是 ICS 表現較好。由混淆矩陣可以發現 ICS+ECS 和

ICS 對於樣本數之較少類別的預測較好，與實驗一的預測結果類似。在點擊類型預測中，也是在樣本數量較少之類別的效果會更加突出，因此在網頁類型比例之預測方面，我們也得出和實驗一相似的結論。

綜合三個實驗的結果說明在整體的預測目標上，ICS+ECS 之資料集因為本身更貼近使用者的行為，所以其整體預測表現上會比傳統 CS 更好，特別是在樣本數之較少類別，而傳統的 CS 之資料集更適合於某些特別的任務。

7.2 未來展望

由於本實驗所使用的資料集中在 CS 及 ECS 之收集的時間單位上有差異，因此對於資料合併順序上可能存在問題。也因為資料集中使用者人數較少（142 人），在進行測試集預測時，容易受到資料量較多的使用者影響導致預測結果偏向較活躍的使用者。未來若能夠重新收集更多且更詳細的資訊，對於分析預測上將會有更大的幫助，也能夠提升預測的效果。

另一方面，因為使用者之網頁瀏覽資料紀錄是具有時序性特徵。若能結合深度學習，例如 LSTM、GRU 等結構之類神經網路進行預測，對於較貼近的使用者行為資料的 ICS+ECS 之資料集應該會得到更加突出的效果。

參考文獻

- [1] M. Grbovic and H. Cheng, “Real-time personalization using embeddings for search ranking at airbnb,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, 2018.
- [2] R. Bhagat, S. Muralidharan, A. Lobzhanidze, and S. Vishwanath, “Buy it again: Modeling repeat purchase recommendations,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, 2018.
- [3] A. Kumar, V. Ahirwar, and R. K. Singh, “A study on prediction of user behavior based on web server log files in web usage mining,” *International Journal of Engineering and Computer Science*, 2017.
- [4] G. Wang, X. Zhang, S. Tang, H. Zheng, and B. Y. Zhao, “Unsupervised clickstream clustering for user behavior analysis,” *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.
- [5] N. B. Pawar, M. Gaikwad, S. Kalyani, and M. Savla, “Analysis and prediction of e-customers behaviour by mining clickstream data using naive bayes,” *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 4, pp. 2427–2430, 2018.
- [6] J. Liu, P. Dolan, and E. R. Pedersen, “Personalized news recommendation based on click behavior,” in *IUI '10*, 2010.
- [7] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, “Analyzing user modeling on twitter for personalized news recommendations,” in *UMAP'11*, 2011.
- [8] K. R. Suneetha and R. Krishnamoorthi, “Identifying user behavior by analyzing web server access log file,” 2009.
- [9] F. Buccafurri, G. Lax, S. Nicolazzo, and A. Nocera, “Comparing twitter and facebook user behavior: Privacy and other aspects,” *Comput. Hum. Behav.*, vol. 52, pp. 87–95, 2015.
- [10] M. Kosinski, D. Stillwell, and T. Graepel, “Private traits and attributes are predictable from digital records of human behavior.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110 15, pp. 5802–5, 2013.

參考文獻

- [11] S. C. Matz, M. Kosinski, G. Nave, and D. Stillwell, “Psychological targeting as an effective approach to digital mass persuasion,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, pp. 12 714–12 719, 2017.
- [12] P. Covington, J. L. Adams, and E. Sargin, “Deep neural networks for youtube recommendations,” *Proceedings of the 10th ACM Conference on Recommender Systems*, 2016.
- [13] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, “Neural collaborative filtering,” *Proceedings of the 26th International Conference on World Wide Web*, 2017.
- [14] H. Bang and J.-H. Lee, “Collective matrix factorization using tag embedding for effective recommender system,” *2016 Joint 8th International Conference on Soft Computing and Intelligent Systems (SCIS) and 17th International Symposium on Advanced Intelligent Systems (ISIS)*, pp. 846–850, 2016.
- [15] J. Tang and K. Wang, “Personalized top-n sequential recommendation via convolutional sequence embedding,” *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018.
- [16] G. Zhou, C. Song, X. Zhu, X. Ma, Y. Yan, X. Dai, H. Zhu, J. Jin, H. Li, and K. Gai, “Deep interest network for click-through rate prediction,” *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, 2018.
- [17] G. Zhou, N. Mou, Y. Fan, Q. Pi, W. Bian, C. Zhou, X. Zhu, and K. Gai, “Deep interest evolution network for click-through rate prediction,” *ArXiv*, vol. abs/1809.03672, 2019.
- [18] Y. Feng, F. Lv, W. Shen, M. Wang, F. Sun, Y. Zhu, and K. Yang, “Deep session interest network for click-through rate prediction,” in *IJCAI*, 2019.
- [19] T.-R. Chen, “Extended clickstream: An analysis of the missing user behaviors in the clickstream,” Master’s thesis, NCU, 2019.
- [20] N. Altman, “An introduction to kernel and nearest-neighbor nonparametric regression,” *American Statistician*, 1992.
- [21] T. K. Ho, “Random decision forests,” in *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, 1995.
- [22] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

附錄 A 實驗之混淆矩陣

以下將依照實驗的順序，呈現 XGBoost 對三個資料集 (CS、ICS、ICS+ECS) 所產生之混淆矩陣。

實驗一：預測下次點擊網站類型中，紅色字體代表的是和 CS 資料集相比預測準確度更高的部分。橫軸為預測之網站類型的排名，縱軸為真實之網站類型的排名，網站類型名稱可對照表 4.7 中 Rank(1)。表格中上方數字代表樣本數量，下方則是佔真實類別之比例。

實驗二：預測下次點擊時間間隔，橫軸為真實之時間間隔，縱軸為預測之時間間隔。表格中上方數字代表樣本數量，下方則是佔真實類別之比例。

實驗三：預測網站類型比例，在此將點擊比例劃分為 5 個區間種類，呈現對 Top1 和 Top2 網站之預測產生之混淆矩陣。橫軸為預測類別，縱軸為真實類別。表格中上方數字代表樣本數量，下方則是佔真實類別之比例。

其他模型結果圖：https://github.com/eleceel/DART_analyze_user_behavior.git

A、實驗之混淆矩陣

true/predict	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	19125 (79.31)	1245 (5.16)	1122 (4.65)	444 (1.84)	286 (1.19)	142 (0.59)	411 (1.70)	110 (0.46)	90 (0.37)	95 (0.39)	150 (0.62)	176 (0.73)	109 (0.25)	61 (0.21)	50 (0.18)	43 (0.05)	50 (0.21)	43 (0.18)	43 (0.11)		
2	1166 (4.66)	21005 (83.92)	692 (2.76)	385 (1.54)	23 (0.89)	130 (0.52)	307 (1.23)	160 (0.64)	67 (0.22)	104 (0.42)	80 (0.32)	93 (0.37)	144 (0.58)	77 (0.31)	41 (0.16)	27 (0.11)	20 (0.08)	55 (0.22)	44 (0.18)	19 (0.08)	
3	1353 (5.65)	879 (3.67)	16803 (70.21)	815 (3.41)	1040 (4.35)	155 (0.65)	400 (1.16)	278 (1.93)	462 (0.61)	145 (0.49)	118 (1.22)	293 (0.64)	152 (0.52)	125 (0.43)	103 (0.36)	85 (0.24)	58 (0.24)	34 (0.14)	101 (0.42)	63 (0.26)	
4	365 (3.39)	343 (3.18)	193 (11.08)	8000 (74.28)	176 (1.63)	43 (0.40)	134 (1.24)	53 (0.49)	23 (0.21)	44 (0.41)	9 (0.08)	58 (0.54)	76 (0.71)	13 (0.12)	11 (0.10)	25 (0.23)	9 (0.08)	13 (0.12)	13 (0.12)	7 (0.06)	
5	232 (3.45)	232 (3.83)	1269 (1.82)	150 (2.23)	4209 (62.62)	38 (0.57)	89 (1.32)	55 (1.23)	83 (0.82)	40 (0.60)	21 (0.31)	56 (0.36)	34 (0.83)	16 (0.51)	19 (0.24)	16 (0.28)	14 (0.21)	6 (0.09)	9 (0.13)	8 (0.12)	
6	118 (0.77)	156 (1.01)	92 (0.60)	37 (0.24)	14698 (95.47)	37 (0.08)	12 (0.10)	15 (0.04)	6 (0.68)	104 (0.01)	1 (0.10)	16 (0.25)	38 (0.03)	4 (0.01)	2 (0.03)	4 (0.01)	1 (0.01)	9 (0.06)	5 (0.03)	6 (0.04)	
7	345 (4.59)	288 (3.83)	437 (1.72)	97 (0.21)	5901 (78.53)	28 (0.37)	28 (0.37)	25 (0.33)	83 (0.80)	25 (0.36)	60 (0.24)	18 (0.08)	6 (0.32)	24 (0.32)	27 (0.36)	8 (0.11)	2 (0.03)	2 (0.03)	2 (0.03)	0 (0.00)	
8	112 (3.52)	166 (5.22)	625 (9.67)	72 (2.27)	128 (4.63)	16 (0.50)	12 (1.10)	15 (0.54)	6 (0.83)	104 (0.47)	15 (0.79)	58 (0.44)	14 (0.38)	12 (0.38)	12 (0.94)	14 (0.44)	8 (0.25)	1 (0.19)	5 (0.16)	5 (0.38)	
9	48 (1.39)	48 (1.80)	62 (14.46)	499 (0.58)	20 (0.17)	65 (0.38)	13 (1.51)	52 (0.47)	2 (0.27)	0 (0.03)	1 (0.01)	18 (0.25)	2 (0.03)	1 (0.01)	2 (0.03)	2 (0.01)	1 (0.01)	5 (0.06)	6 (0.03)	5 (0.04)	
10	98 (2.46)	100 (2.51)	118 (2.96)	56 (1.40)	47 (1.18)	63 (0.58)	24 (0.60)	14 (0.35)	2 (0.05)	3337 (83.72)	5 (0.13)	15 (0.38)	27 (0.68)	5 (0.13)	27 (0.63)	25 (0.15)	6 (0.08)	3 (0.11)	6 (0.15)	1 (0.03)	4 (0.10)
11	135 (2.34)	105 (1.82)	177 (3.06)	18 (0.31)	21 (0.36)	1 (0.26)	87 (0.15)	15 (0.26)	2 (0.03)	6 (0.10)	5107 (88.34)	1 (0.10)	10 (0.07)	4 (0.07)	4 (0.17)	10 (0.12)	7 (0.02)	1 (0.19)	18 (0.31)	4 (0.07)	0 (0.00)
12	156 (4.66)	83 (2.48)	395 (17.77)	81 (2.42)	44 (1.31)	24 (0.72)	30 (0.90)	28 (0.84)	2 (0.06)	39 (0.39)	15 (0.09)	2208 (65.93)	13 (0.48)	12 (0.12)	4 (0.21)	7 (0.03)	1 (0.09)	3 (0.15)	5 (0.09)	3 (0.09)	3 (0.09)
13	114 (4.47)	215 (8.44)	162 (6.36)	77 (3.02)	46 (1.81)	42 (1.65)	42 (1.51)	13 (1.96)	50 (0.59)	15 (1.14)	29 (0.20)	12 (0.47)	1686 (66.17)	10 (0.39)	5 (0.20)	8 (0.31)	8 (0.08)	2 (0.08)	14 (0.20)	5 (0.20)	
14	148 (2.60)	149 (8.08)	220 (11.94)	13 (0.71)	42 (0.33)	8 (0.43)	12 (0.65)	12 (0.11)	2 (0.22)	4 (0.27)	5 (0.27)	7 (0.38)	1281 (69.51)	3 (0.16)	2 (0.11)	3 (0.16)	3 (0.16)	1 (0.16)	1 (0.00)	0 (0.00)	
15	67 (3.91)	73 (4.26)	297 (17.34)	30 (1.75)	50 (0.41)	7 (1.58)	27 (1.34)	23 (0.41)	7 (0.41)	10 (0.58)	10 (0.47)	8 (0.29)	5 (0.65)	1039 (60.65)	8 (0.47)	5 (0.29)	8 (0.66)	1 (0.06)	1 (0.06)	1 (0.06)	0 (0.00)
16	29 (1.26)	22 (0.96)	68 (2.97)	24 (1.05)	4 (1.05)	130 (0.16)	28 (0.17)	8 (0.35)	1 (0.04)	8 (0.35)	5 (0.22)	3 (0.13)	4 (0.17)	2 (0.09)	2 (0.31)	7 (0.17)	1 (0.04)	0 (0.04)	3 (0.13)	0 (0.00)	
17	30 (2.94)	53 (5.19)	503 (49.27)	38 (3.72)	88 (0.39)	4 (1.76)	18 (1.67)	17 (1.37)	5 (0.49)	14 (0.69)	2 (0.20)	293 (0.29)	5 (0.49)	3 (0.29)	0 (0.00)	3 (18.32)	0 (0.00)	0 (0.39)	4 (0.137)	0 (0.37)	
18	52 (7.28)	59 (8.26)	27 (3.78)	20 (2.80)	13 (1.82)	9 (1.68)	12 (0.98)	1 (0.42)	4 (0.84)	10 (0.70)	1281 (0.84)	7 (0.52)	42 (0.42)	3 (0.14)	2 (0.00)	1 (0.14)	1 (0.00)	1 (0.14)	1 (0.14)	0 (0.14)	
19	39 (2.38)	39 (6.71)	110 (0.73)	9 (0.55)	2 (0.12)	1 (0.12)	1 (0.06)	4 (0.24)	36 (0.20)	0 (0.00)	2 (0.12)	4 (0.12)	1 (0.12)	2 (0.12)	1 (0.06)	1 (0.06)	1 (0.06)	1 (0.06)	1 (0.06)	0 (0.00)	
20	17 (3.95)	17 (22.33)	96 (1.40)	6 (2.09)	9 (0.93)	4 (0.23)	8 (1.86)	1 (0.23)	8 (0.23)	1 (0.47)	2 (0.47)	7 (0.70)	163 (0.63)	1 (0.23)	1 (0.23)	1 (0.23)	0 (0.00)	0 (0.00)	1 (0.23)	0 (0.23)	

圖 A.1: 實驗一: XGBoost 之混淆矩陣 (CS)

true/predict	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	19141 (5.14)	1232	804	466	327	141	409	137	96	90	165	215	111	79	57	45	28	40	43	23	
2	1184 (4.78)	20799 (83.98)	533	395	(1.94)	(1.36)	(0.59)	(1.71)	129	310	171	71	99	83	109	155	85	44	22	29	57
3	1238	749	12958	834	931	147	397	290	416	127	113	283	137	124	99	76	85	27	95	60	
4	374	341	687	7050 (65.97)	229	46	138	73	46	45	9	87	74	16	15	25	21	11	15	6	
5	233	228	792	207	4210 (3.24)	36	(1.42)	(0.49)	(1.46)	(0.47)	(0.47)	(0.49)	(0.78)	(0.92)	(0.78)	(0.17)	(0.16)	(0.22)	(0.12)	(0.16)	
6	120	151	72	39	44	14639	14	17	7	104	2	18	41	4	4	2	2	9	4	6	
7	355	281	241	130	117	16	5809 (1.78)	42	28	24	62	42	19	9	25	25	12	0	3	0	
8	113	149	407	98	146	16	46	1625 (1.54)	70	13	26	32	63	13	37	13	16	6	5	13	
9	59	64	346	41	91	4	19	54	2453 (1.25)	2	1	3	19	3	4	1	8	5	35	1	
10	95	99	103	54	45	60	23	17	2	3272 (1.15)	5	16	27	5	23	6	5	6	1	6	
11	140	103	118	29	25	2	88	16	1	5	5087 (0.51)	10	4	7	8	9	1	18	4	0	
12	170	86	383	121	53	25	43	34	3	11	7	2229 (0.64)	15	5	6	2	9	4	3	4	
13	120	215	146	74	49	42	13	47	15	29	5	14	1667	13	18	13	0.06	(0.12)	(0.09)	(0.02)	
14	47	129	148	24	49	6	7	13	3	4	5	10	9	1310 (0.81)	3	2	7	3	3	1	
15	67	68	183	50	67	6	39	31	8	10	10	19	5	7	953	6	4	1	2	2	
16	31	18	43	23	23	3	27	5	2	7	6	3	4(0.18)	2	8	1961 (0.40)	1	0	3	0	
17	35	59	323	65	126	5	20	23	24	5	7	16	4	10	5	0	223 (3.50)	0	5	17	
18	41	56	17	20	8	12	6	3	6	5	13	6	17	4	1	0	420 (6.33)	0	0.50	(1.70)	
19	49	43	78	15	10	2	1	7	49	0	2	4	2	1	1	2	1359 (2.46)	0	0.06	(0.06)	
20	23	16	57	11	10	4	2	14	1	4	2	3	6	2	1	1	4	0	1	231 (5.75)	

圖 A.2: 實驗一: XGBoost 之混淆矩陣 (ICS)

A、實驗之混淆矩陣

true/predict	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	19033 (80.97)	1206 (5.13)	821 (3.49)	410 (1.74)	281 (1.20)	123 (0.52)	376 (1.60)	116 (0.49)	88 (0.37)	77 (0.33)	144 (0.61)	190 (0.81)	107 (0.46)	68 (0.29)	48 (0.20)	41 (0.17)	17 (0.07)	38 (0.16)	36 (0.15)	22 (0.09)	
2	1328 (5.42)	20575 (83.94)	545 (2.22)	371 (1.51)	217 (0.89)	120 (1.22)	299 (0.66)	162 (0.28)	68 (0.37)	91 (0.31)	77 (0.36)	88 (0.58)	141 (0.29)	70 (0.17)	41 (0.08)	20 (0.11)	26 (0.11)	55 (0.16)	40 (0.16)	22 (0.09)	
3	1472 (7.56)	753 (66.17)	12884 (4.15)	808 (4.41)	859 (0.70)	136 (1.83)	357 (0.70)	268 (0.38)	397 (0.04)	106 (0.56)	109 (0.54)	266 (0.37)	126 (0.60)	116 (0.45)	87 (0.45)	73 (0.37)	26 (0.39)	92 (0.13)	56 (0.47)	56 (0.29)	
4	415 (4.59)	336 (3.72)	718 (7.95)	6678 (73.92)	216 (0.40)	36 (1.36)	123 (0.76)	69 (0.49)	44 (0.42)	38 (0.12)	11 (0.82)	74 (0.77)	70 (0.12)	11 (0.14)	13 (0.19)	17 (0.11)	21 (0.23)	10 (0.12)	11 (0.12)	6 (0.07)	
5	236 (3.94)	210 (3.50)	791 (13.20)	190 (3.17)	3939 (65.72)	31 (0.52)	96 (1.60)	89 (1.48)	65 (1.08)	26 (0.43)	19 (0.50)	30 (1.00)	60 (0.53)	32 (0.55)	21 (0.55)	11 (0.55)	34 (0.18)	5 (0.57)	7 (0.08)	10 (0.12)	(0.17)
6	139 (0.91)	156 (0.92)	83 (0.54)	29 (0.19)	35 (0.23)	14587 (0.545)	14 (0.69)	14 (0.09)	5 (0.03)	106 (0.69)	2 (0.11)	17 (0.23)	35 (0.03)	4 (0.03)	4 (0.03)	1 (0.01)	2 (0.01)	8 (0.05)	4 (0.03)	6 (0.04)	
7	435 (6.27)	270 (3.89)	234 (3.37)	93 (1.34)	91 (1.31)	12 (0.17)	5112 (79.42)	33 (0.48)	23 (0.33)	23 (0.33)	57 (0.82)	19 (0.52)	36 (0.20)	14 (0.13)	9 (0.23)	16 (0.36)	25 (0.13)	9 (0.13)	0 (0.00)	0 (0.04)	0 (0.00)
8	115 (3.99)	140 (4.86)	381 (13.22)	94 (3.26)	142 (4.93)	13 (0.45)	40 (1.39)	1601 (55.55)	72 (2.50)	12 (0.42)	24 (0.83)	30 (1.04)	59 (2.05)	11 (0.38)	11 (1.21)	13 (0.45)	13 (0.45)	5 (0.17)	4 (0.14)	6 (0.38)	
9	72 (2.26)	60 (1.88)	335 (10.49)	36 (1.13)	86 (2.69)	4 (0.13)	13 (0.41)	52 (1.63)	52 (0.66)	13 (0.06)	2 (0.13)	33 (0.13)	14 (0.50)	14 (0.03)	14 (0.13)	16 (0.03)	9 (0.19)	0 (0.16)	0 (0.13)	0 (0.00)	
10	108 (2.83)	95 (2.49)	50 (2.32)	50 (1.31)	31 (0.81)	52 (1.36)	21 (0.55)	21 (0.42)	16 (0.05)	21 (0.05)	3227 (84.54)	4 (0.10)	17 (0.45)	25 (0.65)	4 (0.10)	19 (0.50)	4 (0.10)	19 (0.10)	0 (0.16)	0 (0.00)	5 (0.13)
11	150 (2.65)	99 (1.75)	25 (0.44)	19 (0.34)	25 (0.04)	2 (0.33)	75 (0.25)	14 (0.02)	14 (0.07)	8920 (89.20)	4 (0.11)	6 (0.05)	3 (0.12)	7 (0.18)	10 (0.18)	10 (0.18)	10 (0.02)	14 (0.25)	3 (0.05)	0 (0.00)	
12	201 (6.24)	83 (2.58)	395 (12.27)	115 (3.57)	55 (1.71)	34 (0.75)	34 (1.06)	34 (0.09)	3 (0.28)	9 (0.16)	2178 (67.64)	15 (0.47)	4 (0.12)	8 (0.25)	8 (0.06)	15 (0.10)	8 (0.25)	8 (0.06)	3 (0.09)	4 (0.09)	4 (0.12)
13	137 (5.49)	209 (8.38)	149 (5.97)	64 (2.57)	41 (1.64)	11 (1.72)	43 (0.44)	11 (0.48)	47 (0.72)	18 (1.04)	5044 (66.41)	6 (0.24)	12 (0.48)	1657 (66.41)	9 (0.36)	7 (0.20)	7 (0.28)	7 (0.20)	7 (0.28)	3 (0.12)	3 (0.00)
14	56 (3.13)	130 (7.26)	153 (8.54)	21 (1.17)	38 (0.212)	6 (0.34)	7 (0.39)	15 (0.84)	4 (0.22)	2 (0.11)	3228 (50.50)	9 (0.50)	9 (0.39)	1293 (72.19)	3 (0.17)	2 (0.11)	7 (0.39)	3 (0.17)	3 (0.17)	1 (0.06)	1 (0.06)
15	76 (5.05)	57 (3.79)	184 (12.23)	46 (3.06)	56 (6.72)	7 (0.47)	32 (2.13)	31 (0.66)	5 (0.33)	8 (0.53)	10 (0.66)	16 (1.06)	4 (0.27)	6 (0.40)	918 (61.00)	6 (0.40)	3 (0.20)	3 (0.13)	2 (0.20)	3 (0.13)	2 (0.13)
16	49 (2.25)	18 (0.82)	47 (2.15)	25 (1.15)	18 (0.82)	3 (0.14)	30 (1.37)	8 (0.09)	2 (0.27)	6 (0.23)	5 (0.05)	1 (0.18)	4 (0.14)	3 (0.22)	1916 (87.81)	1 (0.05)	0 (0.00)	3 (0.14)	0 (0.00)	0 (0.00)	
17	47 (4.88)	54 (5.61)	301 (31.26)	63 (6.54)	126 (13.08)	3 (0.31)	15 (1.36)	21 (2.18)	24 (2.49)	5 (0.52)	106 (0.62)	6 (1.56)	15 (0.21)	9 (0.93)	7 (0.73)	222 (23.05)	1 (0.10)	1 (0.10)	4 (0.42)	1 (1.56)	0 (0.00)
18	59 (9.18)	54 (4.5)	19 (2.37)	17 (0.81)	13 (0.69)	7 (0.06)	1 (0.66)	1 (0.56)	9 (47)	0 (0.00)	1 (0.66)	1 (0.12)	1 (0.25)	2 (0.12)	1 (0.06)	0 (0.00)	1 (0.16)	416 (64.70)	0 (0.00)	0 (0.16)	
19	24 (6.32)	13 (3.42)	53 (13.95)	7 (1.84)	7 (0.79)	1 (0.26)	3 (3.42)	1 (0.26)	3 (0.79)	1 (0.53)	1 (0.79)	1 (1.58)	1 (0.53)	1 (0.26)	1 (0.79)	1 (0.26)	1 (0.79)	1 (0.26)	228 (60.00)	0 (0.00)	

圖 A.3: 實驗一: XGBoost 之混淆矩陣 (ICS+ECS)

True/predict	1	2	3	4	5
1	61098 (83.06)	9036 (12.28)	2179 (2.96)	1204 (1.64)	38 (0.05)
2	19036 (51.26)	13293 (35.80)	3219 (8.67)	1544 (4.16)	41 (0.11)
3	10751 (44.10)	7513 (30.82)	4228 (17.34)	1854 (7.60)	33 (0.14)
4	5809 (39.38)	2914 (19.75)	2153 (14.60)	3828 (25.95)	47 (0.32)
5	1848 (47.09)	860 (21.92)	490 (12.49)	579 (14.76)	147 (3.75)

CS

True/predict	1	2	3	4	5
1	58269 (84.12)	7435 (10.73)	2304 (3.33)	1224 (1.77)	34 (0.05)
2	17585 (52.51)	10874 (32.41)	3446 (10.29)	1550 (4.63)	36 (0.11)
3	10563 (44.00)	6885 (28.68)	4678 (19.48)	1839 (7.66)	44 (0.18)
4	5724 (39.13)	2687 (18.37)	2331 (15.93)	3837 (26.26)	50 (0.34)
5	1811 (46.32)	843 (21.56)	553 (14.14)	561 (14.35)	142 (3.63)

ICS

圖 A.4: 實驗二: XGBoost 之混淆矩陣 (CS & ICS)

A、實驗之混淆矩陣

True/predict	1	2	3	4	5
1	57971 (84.27)	7458 (10.84)	2411 (3.50)	929 (1.35)	25 (0.04)
2	17775 (53.33)	10785 (32.36)	3526 (10.58)	1218 (3.65)	25 (0.08)
3	10210 (42.94)	6626 (27.87)	5119 (21.53)	1705 (7.17)	116 (0.49)
4	4788 (36.74)	2285 (17.54)	2527 (19.39)	3282 (25.19)	149 (1.14)
5	1014 (37.07)	510 (18.65)	500 (18.28)	443 (16.20)	268 (9.80)

ICS+ECS

圖 A.5: 實驗二: XGBoost 之混淆矩陣 (ICS+ECS)

True/Predict	1	2	3	4	5
1	134 (55.60)	0 (0.00)	78 (32.37)	19 (7.88)	10 (4.15)
2	27 (25.23)	0 (0.00)	67 (62.62)	11 (10.28)	2 (1.87)
3	39 (10.71)	0 (0.00)	234 (64.29)	85 (23.35)	6 (1.64)
4	21 (6.76)	0 (0.00)	122 (39.23)	144 (46.30)	24 (7.72)
5	14 (9.52)	0 (0.00)	55 (37.41)	54 (36.73)	54 (36.73)

CS

True/Predict	1	2	3	4	5
1	147 (60.49)	1 (0.41)	59 (24.28)	25 (10.29)	11 (4.53)
2	32 (32.99)	0 (0.00)	47 (48.45)	17 (17.53)	1 (1.03)
3	37 (10.85)	0 (0.00)	199 (58.36)	98 (28.74)	7 (2.05)
4	31 (9.42)	0 (0.00)	94 (28.57)	169 (51.37)	35 (10.64)
5	16 (10.00)	0 (0.00)	18 (11.25)	60 (37.50)	66 (41.25)

ICS

圖 A.6: 實驗三: XGBoost 之混淆矩陣 (Top1)

A、實驗之混淆矩陣

True/Predict	1	2	3	4	5
1	155 (60.78)	1 (0.39)	67 (26.27)	25 (9.80)	7 (2.74)
2	32 (33.33)	0 (0.00)	48 (50.00)	14 (14.58)	2 (2.08)
3	40 (11.80)	0 (0.00)	202 (59.59)	93 (27.43)	4 (1.18)
4	26 (8.02)	0 (0.00)	87 (26.85)	179 (55.25)	32 (9.88)
5	20 (12.82)	0 (0.00)	15 (9.62)	55 (35.26)	66 (42.31)

ICS+ECS

圖 A.7: 實驗三: XGBoost 之混淆矩陣 (Top1)

True/Predict	1	2	3	4	5
1	234 (71.56)	0 (0.00)	71 (21.71)	19 (5.81)	3 (0.92)
2	54 (37.24)	1 (0.69)	80 (55.17)	10 (6.90)	0 (0.00)
3	46 (13.94)	0 (0.00)	205 (62.12)	77 (23.33)	2 (0.61)
4	23 (8.78)	0 (0.00)	106 (40.46)	122 (46.56)	11 (4.19)
5	14 (13.21)	0 (0.00)	19 (17.92)	42 (39.62)	31 (29.25)

CS

圖 A.8: 實驗三: XGBoost 之混淆矩陣 (Top2)

True/Predict	1	2	3	4	5
1	231 (71.52)	0 (0.00)	68 (21.05)	21 (6.50)	3 (0.93)
2	53 (37.86)	0 (0.00)	77 (55.00)	10 (7.14)	0 (0.00)
3	49 (15.17)	0 (0.00)	188 (58.20)	85 (26.32)	1 (0.31)
4	23 (8.42)	0 (0.00)	97 (35.53)	142 (52.01)	10 (3.66)
5	13 (11.71)	0 (0.00)	20 (18.02)	48 (43.24)	30 (27.03)

ICS

True/Predict	1	2	3	4	5
1	243 (72.75)	2 (0.60)	67 (20.06)	18 (5.39)	4 (1.20)
2	55 (39.86)	0 (0.00)	77 (55.80)	6 (4.35)	0 (0.00)
3	51 (15.94)	2 (0.63)	176 (55.00)	89 (27.81)	2 (0.63)
4	26 (9.63)	1 (0.37)	85 (31.48)	147 (54.44)	11 (4.07)
5	15 (13.89)	0 (0.00)	21 (19.44)	39 (36.11)	33 (30.56)

ICS+ECS

圖 A.9: 實驗三: XGBoost 之混淆矩陣 (Top2)