國 立 中 央 大 學

資 訊 工 程 學 系

碩 士 論 文

擴展點擊流：分析點擊流中缺少的使用者行爲

Extended Clickstream: an analysis of the missing user behaviors in the Clickstream

研 究 生：陳廷睿

指導教授：陳弘軒 博士

中 華 民 國 一百零八 年 七 月

國立中央大學圖書館
碩博士論文電子檔授權書

(104 年 5 月最新修正版)

本授權書授權本人撰寫之碩/博士學位論文全文電子檔(不包含紙本、詳備註 1 說明)，在「國立中央大學圖書館博碩士論文系統」。(以下請擇一勾選)

(　　)同意 (立即開放)

(　　)同意 (請於西元 _____年_____月_____日開放)

(　　)不同意，原因是：_____

在國家圖書館「臺灣博碩士論文知識加值系統」

(　　)同意 (立即開放)

(　　)同意 (請於西元 _____年_____月_____日開放)

(　　)不同意，原因是：_____

以非專屬、無償授權國立中央大學、台灣聯合大學系統圖書館與國家圖書館，基於推動「資源共享、互惠合作」之理念，於回饋社會與學術研究之目的，得不限地域、時間與次數，以紙本、微縮、光碟及其它各種方法將上列論文收錄、重製、與利用，並得將數位化之上列論文與論文電子檔以上載網路方式，提供讀者基於個人非營利性質之線上檢索、閱覽、下載或列印。

研究生簽名: _____　　學號：_____

論文名稱:擴展點擊流：分析點擊流中缺少的使用者行爲

指導教授姓名：陳弘軒

系所　：資工所　☐博士班　☐碩士班

填單日期：_____

備註：

1. 本授權書之授權範圍僅限**電子檔**，紙本論文部分依著作權法第 15 條第 3 款之規定，採推定原則即預設同意圖書館得公開上架閱覽，如您有申請專利或投稿等考量，不同意紙本上架陳列，須另行加填申請書，詳細說明與紙本申請書下載請至本館數位博碩論文網頁。
2. 本授權書請填寫並**親筆**簽名後，裝訂於各紙本論文封面後之次頁（全文電子檔內之授權書簽名，可用電腦打字代替）。
3. 讀者基於個人非營利性質之線上檢索、閱覽、下載或列印上列論文，應遵守著作權法規定。

國立中央大學碩士班研究生

論文指導教授推薦書

資工 學系/研究所 陳廷睿 研究生所提之論文 擴展點擊流：分析點擊流中缺少的使用者行爲係由本人指導撰述，同意提付審查。

指導教授_____(簽章)

____年____月____日

國立中央大學碩士班研究生

論文口試委員審定書

資工 學系/研究所 陳廷睿 研究生

所提之論文

擴展點擊流：分析點擊流中缺少的使用者行爲

經本委員會審議，認定符合碩士資格標準。

學位考試委員會召集人 [signature]

委 員 [signature]

[signature]

[signature]

中 華 民 國 108 年 7 月 11 日

101.06.15

擴展點擊流：分析點擊流中缺少的使用者行爲

摘要

一般認為使用者的點擊流 (clickstream) 可以代表使用者的線上瀏覽行為，然而，我們發現點擊流只能概略表示使用者的部份行為，例如：分頁切換、視窗切換等介面間的瀏覽行為因為沒有產生與伺服器的互動，所以不會出現在點擊流或日誌 (log) 中，但使用者仍然在瀏覽網頁。本文將這些行為收集並命名為「擴展點擊流」(extended clickstream)。透過建設完整的系統服務並招募受試者來同步蒐集點擊流和擴展點擊流，並對兩者進行比較分析及建構深度學習模型。我們使用含有 GRU 元件的深度學習模型，對點擊流和擴展點擊流這類型的時序資料進行「使用者下次會去什麼類型的網站」、「下次點擊會間隔多久」的多目標預測。實驗結果顯示：融合點擊流和擴展點擊流可以增進預測效能。除此之外，本文發現點擊流會因為部分網站的運作機制而多計入了使用者沒有意圖執行的行為；另外，我們也可以透過融合點擊流及擴展點擊流來區分出來自不同裝置的單一使用者。

關鍵字： 點擊流，日誌分析，使用者行為分析，時序資料回歸預測，Clickstream, log analysis, User Behavior Model, Time-Series Recurrent Prediction

Extended Clickstream: an analysis of the missing user behaviors in the Clickstream

Abstract

Nowadays, people often use clickstream to represent the behavior of online users. However, we found that clickstream only represents part of users' browsing behaviors. For instance, clickstream does not include tab switching and browser window switching. We collect these kinds of behaviors and named as "extended clickstream". This thesis builds a service to capture both of clickstream and extended clickstream, also provides an analysis of the differences between above. We use a Multi-Task learning model with GRU components to perform multi-objective predictions of "what kind of website the user will go next time" and "how long the interval of clicks will be" for the time series of clickstreams and extended clickstreams. Our experimental results show that combining clickstream and extended clickstream can improve the prediction performance. In addition, this article finds that the clickstream will record unintended clicks due to the operation mechanism of certain websites. Moreover, we can differentiate the single user from several devices by combining the clickstream and extended clickstream.

Keywords: Clickstream, log analysis, Web mining, web usage mining, User Behavior Model, Time-Series Recurrent Prediction

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis, we propose a new type of clickstream that is collected to extend the lack of intention part of clickstream. We call it Extended Clickstream(ECS).

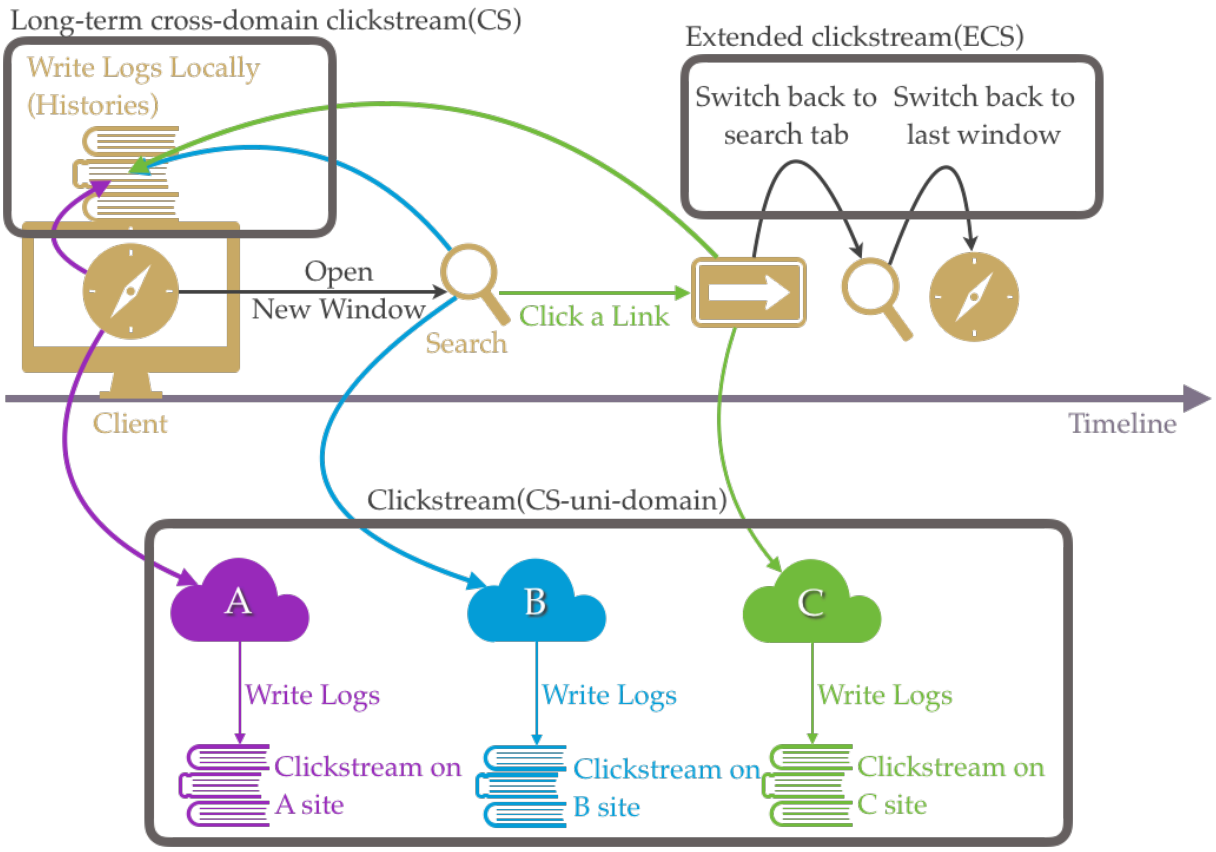


Figure 1.1: An overview of two types of CS and ECS.

There are two types of clickstream, one collected from the server log (CS-uni-domain) and the other collected from the client (CS). Here we focus on the latter type of clickstream which is cross-domain. We built a service to collect clickstream and monitored users' behaviors on the browser. We found certain behaviors, such as switching tabs, browser windows, idle/return on tabs/windows, or switching to other applications but still browsing the site, were not recorded in the clickstream (The black arrow in the Figure 1.1). Therefore, we completed our system to collect clickstream (CS) and extended clickstream (ECS). We started the service on February 26, 2019, and analyzed the data until June 17, 2019. We found the number of ECS events generated per day were more than CS, and distribution of URLs was different. By adding ECS to the CS, we could detect unintentional events on websites and time periods the same user was using on different devices. Next, we used a Multi-Task learning model with GRU components to perform multi-objective predictions of "what kind of website the user will go next time" and "how long the interval of clicks will be" for the time series of clickstream and extended clickstream.

In Chapter 2, we review previous work on the difference between cross-domain and uni-domain clickstreams. We show the experiments we did in the past research and the techniques used in this research. In Chapter 3, we introduce the structure, collecting scenario, and advantages of Extended Clickstream(ECS). In Chapter 4, we show the phase of this research from collecting data to build a Multi-Task learning model to model behaviors of users. In Chapter 5, we present the statistical reports of the CS and ECS data we collected, a case study of what we can learn from the combination of ECS and CS data, and the results of the deep learning model. In Chapter 6, we discuss the difficulties of collecting ECS and CS, and what we can do in the future if we can overcome them.

Chapter 2

Related Work

In this chapter, we will begin with introducing the difference between common clickstream dataset and long-term cross-domain clickstream dataset. Next, we will show the dataset we have already collected and made it as an open dataset. Then we will show you several experiments we have done. Furthermore, we decided to extend our dataset. Since the dataset lacks the source code, we build a whole new system to collect data and recruit new 150 individual users. Lastly, we guide several skills or tricks from the past research.

2.1 Clickstream & Long-Term Cross-Domain Clickstream

Clickstream (CS-uni-domain) always points to the data consist of summaries of HTTP header information for the networking traffic exchanged logs between servers and users. It is a completely server-side data. This type of data can model user behavior in online service[1]–[4], detect Sybil Attack[5] and detect web intrusions[6].

Long-term cross-domain clickstream (CS) is the logs of long-term cross-website visits for online users. This type of data is collected from **user-side**(include a part of server-side information and hole client-side information if we want). Due to this characteristic, such type of dataset is rare. It is difficult to collect the long-term cross-website logs from users; therefore, many researchers take this type of dataset as a treasure. For instance, social scientists may use the dataset to apply qualitative or quantitative research on the users demographically information analysis,psychology information analysis, and political spectrum, etc.; computer scientists may use the dataset on recommender systems and online advertising system development; popular culture researcher can inspect the dataset to get the information about popular culture evolution path; communication scientists may use the dataset by analyzing topics of articles to know how the subculture become a popular culture. Thus, collecting the type of dataset is difficult but valuable. To collect Long-term cross-domain clickstream(CS) from user side, we may need to overcome at least three problems as follows:

1. **Cross-domain**: Although a website or an application service manager can fetch logs from their server, they can only know who interact with their services. They cannot know what their users do before entering or even after leaving their service.

2. **Long-term**: Although certain websites store several cookies in browsers to record users behaviors, browsers always limit the storage of cookies. Therefore, the long-term behavior may be missing.

3. **Multiple users**: Although browsers allows users to export their histories in complete, it needs lots of benefits for users to exchange their data with third-party.

2.2 Post-collected Dataset

We recruited 508 individuals as target users. Specifically, we collected the full browsing history of these users stored in Google Chrome. Most of these browsing history records were recorded from August 2016 to December 2016. All these individuals report that they are familiar with the Internet and experienced in online shopping. Additionally, these users are with detailed demographic information(Gender: Male=45%, Female=54%, Other=1% ; Age: 0-20: 1%, 21-30: 59%, 31-40: 31%, 40+: 9%; Relationship status: Single=47%, In a relationship=33%, Married=18%, Other=2%). After preprocessing a series of data, we verified that almost 672 users have histories in our dataset. These users contribute a total of 12,837,216 histories. Table 2.1 shows a statistical summary of the number of visited URLs per user.

Table 2.1: A statistical summary of the number of visited URLs per user

MIN	1^{st} Quartile	Median	Mean	3^{rd} Quartile	MAX
44	4,239	13,335	19,103	26,698	130,992

Experiment 1 - Predicting Users' Demographic Information and Personality Through Browsing History[7]: We choose common supervise learning methods as models include k-Nearest Neighbors (kNN), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine(SVM). In addition, we use CS data to predict gender, age, and relationship status. It is possible to predict users' demographic information and personality by using CS data.

Experiment 2 - Predicting User's Browsing Tendency During Holidays[8]: According to the past research, online shopping behavior of users changed over time. However, we would like to know if it is possible to predict the change point of behaviors before the behavior changes. We try to model this problem with our CS dataset by choosing the festival and common supervise learning methods on training. After evaluation, we found we can predict the change point of behaviors with modeling CS data.
There was a phenomenon we found after experiment 2. When we use more data in the training phase, we get higher AUC in the testing phase. This result shows that our dataset should be extended the amount to improving the predicting power.
As we know the type of our dataset is rare on the internet, we publish as an open dataset by anonymizing any sensitive information. Hope the dataset could help researchers to work on real data from users.

2.2.1 Published as an Open Dataset

We gathered this dataset[9] and we anonymized the privacy information. Besides, we published this dataset which contains 4 columns: user-id, website category(query from a online website classifier), transition type, and visit time.

2.3 Discretize the intervals between events in Time-Series data

We use the skill, proposed by Wang et al. [5], to transform the intervals into time events. Based on this task, we can make all time-series data into events sequence data. The task map all the intervals between events into

categorical bucket, such as the interval from A event to B event is 10 seconds and assigned to the bucket of range 10 seconds to 11 seconds.

2.4 Multi-Task Learning(MTL)

Multi-Task Learning(MTL)[10] is a skill that has led to successes in many tasks. We use the hard parameter sharing of MTL structure. It can reduce the risk of overfitting original task and predict better than the models fitting individual tasks. Zhou et al. [11] used MTL skill as an auxiliary loss to make recurrent unit learn better.

Chapter 3

Extended Clickstream(ECS)

In this chapter, we will first introduce what ECS is and how it differs from CS. Second, we will list certain of the advantages of ECS.

3.1 What is Extended Clickstream(ECS)

We design the system to collect CS data from user side. During system implementation, we found that a range of behavioral data could be recorded by the same system but never recorded in Google Chrome. In addition, the behavior is user-side data, which records users switching the **tabs/windows** on browser, making the browser **blur**(switch to other applications), **idle**, or **active**. All of these behaviors or trigger events are defined by the browser. We design our system to capture the events and collect a series of them as Extended Clickstream(ECS).

As the name of 'extended' clickstream, we are not replacing but expanding CS data to make up the deficiency. We can track switching behavior of users from ECS records. Besides, it would be better if we indicate user online behavior by merging ECS with CS.

ECS data structure include the **title** of events occurred page, **URL**, **domain** name, event **type**, and **timestamp** of events with the precision to microsecond. These event types are all triggered by users with **actual intent**. The triggering scenario of events:

1. **Tabs** events triggered by a user switch the focusing tab to another tab in the same window.

2. **Windows** events triggered by a user switch the focusing window to another window in the same browser.

3. **Blur** events triggered by a user switch to other applications include closing browser or returning to desktop.

4. **Idle** events are defined as a user does not have any I/O input over 2 minutes, makes the operating system be locked, or turns to sleep mode.

5. **Active** events are ideally defined paired with 'Idle' events. This type of events are opposite to idle events

ECS data characteristic is triggered by actual intention of users as listed below. We find that CS data transition types[12] include both intentional and unintentional actions as compared below.
According to the definition, these CS transition types are ideally **unintentional** actions:

1. auto_subframe: This is any content automatically loaded in a non-top-level frame.

2. form_submit: The user filled out values in a form and submitted it.

3. auto_toplevel: The page was specified in the command line or the start page.

These CS transition types are ideally **intentional** actions:

1. link: The user get to this page by clicking a link on another page.

2. typed: The user get this page by typing the URL in the address bar, also used for other explicit navigation actions.

3. auto_bookmark: The user get to this page through a suggestion in the UI such as bookmark.

4. manual_subframe: For subframe navigations explicitly re-quested by the user, they can generate new navigation entries in the back/ forward list.

5. keyword: The URL is generated from a replaceable keyword other than the default search provider.

6. generated: The user get to this page by typing in the address bar and selecting an entry unlike a URL.

7. keyword_generated: Corresponds to a visit generated for a keyword.

Finally, the CS transition type 'reload' is mixed with **intentional** and **unintentional** actions.
As the definition, "The user reloaded the page, either by clicking the reload button or by pressing Enter in the address bar.", it seems to be an **intentional** action but nowadays Google Chrome with a memory management trick which unloads the tabs in background and automatically reloads it when users back to the tab. The trick makes this transition type mix with

intentional and **unintentional** actions.

Figure 3.1 we visualized the CS and ECS data we collected in a short period.

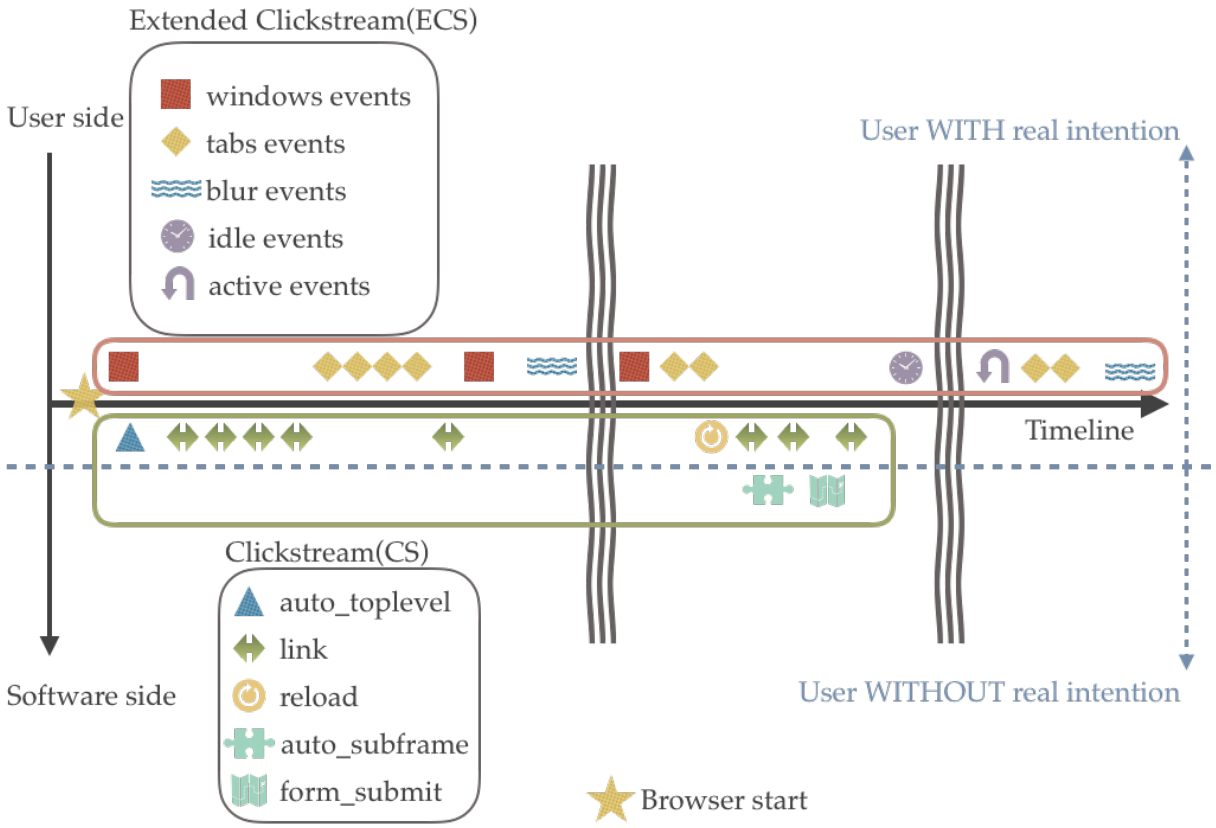


Figure 3.1: ECS merge into CS on the timeline in a short period.
As browser start, the first event always comes a **windows** event. This timeline is from a user who prefers starting browser with homepage, so the next event is **auto_toplevel**. Then the user opens several links in the tab, switches to another **tabs** for 4 times, brings out several actions and **blur** the browser to other applications. Over a time interval, the user gets back to the browser and does several actions. Without having any I/O input for 2 minutes the user **idle**s in a tab. After an idle interval, user returns and does several actions then **blur**s again.

3.2 Merits of ECS

In this section, we summarize these merits of collecting ECS data. We believe there are more merits will be found in the future. Here is the start of ECS.

3.2.1 Easy to understand

Since all ECS events are real and simple actions, we can easily know the intention of users and display ECS as raw data. By visualizing the ECS data combined with CS(if Figure 3.1 mixed with metadata), we can better understand the causal relationship of user behavior and even know what users do between the period.

3.2.2 Make CS more useful

Through adding ECS to CS data, we can get complete behavior data of how users interact with browser and full footprint from beginning to end while browsing. Besides, we can accurately calculate how long users stays on a tag, window, or even a URL. Based on this merit, we may design a system for digital health to protect users not to excessively use the internet.

3.2.3 Enhance the predictive power of modeling user behavior

We verified CS and CS+ECS data on deep learning frameworks. Adding ECS data can perform better on both tasks of predicting categories users will want to browse next and intervals events will trigger next.

Chapter 4

Methods

In this chapter, we start from how we collect data from users including the recruiting process. Then we show the process of data preprocessing for the next step. Lastly, we show the model we used to verify adding ECS to CS is better than using CS alone.

4.1 Phase I. - Data Collecting

Before we built the extension, we analyzed the system requirement and did the market analysis. After the process, we figured out the solution to build our system. During the system implementation period, we found a missing part of the CS data.

4.1.1 System Requirement

First, we want to expand the post-collected dataset, so the database should be designed to align the past we did. Second, we need to solve 3 problems for collecting CS data. Third, since people focus on privacy and security nowadays, we need to use modern encryption skill to make data transport in a safe way.

4.1.2 Market Analysis

Most of our users are students, the rest are office workers. In general, students spend a whole day browsing to get the information from the Internet. As for office workers, the time they spend on browsing is less than students.

4.1.3 Solution

Here we show the solutions for the task.

1. **Cross-domain**: To solve this problem we set up an application on the top of the browsing interface. We build the system by providing a Google Chrome extension as a service to serve users and collect user-side CS data and ECS data. We can provide users on-query personalized histories report by the extension. Consequently, we have a good reason for collecting necessary data from users. We show the system structures and workflow in the appendix. Benefit from this solution, we could easily collect the cross-domain data.

2. **Long-term**: Since users installed our extension, the API of Google Chrome extension would provide us histories of users in the past three months at most. Besides, we can keep collecting histories as the users installed our extension. In order to make the features for both students and office workers, we designed a personalized histories visualization tool as a service to our users. We recruit users which are from all walks of life and all over Taiwan on the Facebook and BBS site.

Modern web is usually designed with the AJAX(Asynchronous JavaScript and XML) method on content loading, which can make the page load the

new content without reloading. It will directly influence the user behavior on browsing the Internet. Users are more likely to open many pages and just browse them by scrolling, switching tabs or windows. As a result, we propose the complete behavior should include these actions with ECS. After finishing the system to collect CS, we find that we can make up the deficiency part of CS data with ECS in the same system.

4.2 Phase II. - Data Preprocessing

After we start the service, we try to make features for deep learning model. We want to predict next action of users on web browsing. Besides, we set our goal to predict what kind of pages users will go in the next events.

4.2.1 Filter unintentional event

Since the unintentional events, which we discussed in Section 3.1, 'reload', 'auto_subframe', 'auto_toplevel', and 'form_submit' may confuse the model, we filtered them for modeling the behavior.

4.2.2 Session split

After time mapping process, we have tried to split sessions for modeling the session-level prediction. To split the sessions, we followed these rules as below,

1. We took the blur event and idle event in ECS as a session end.

2. For any period without ECS data, we check every interval after CS event that is less than 20 minutes. We assume a session ends on the

CS event that is over 20 minutes before the interval.

We split the sessions, and we found several problem in the sessions. We launched a case study in Section 5.2.2.

4.2.3 Time Mapping

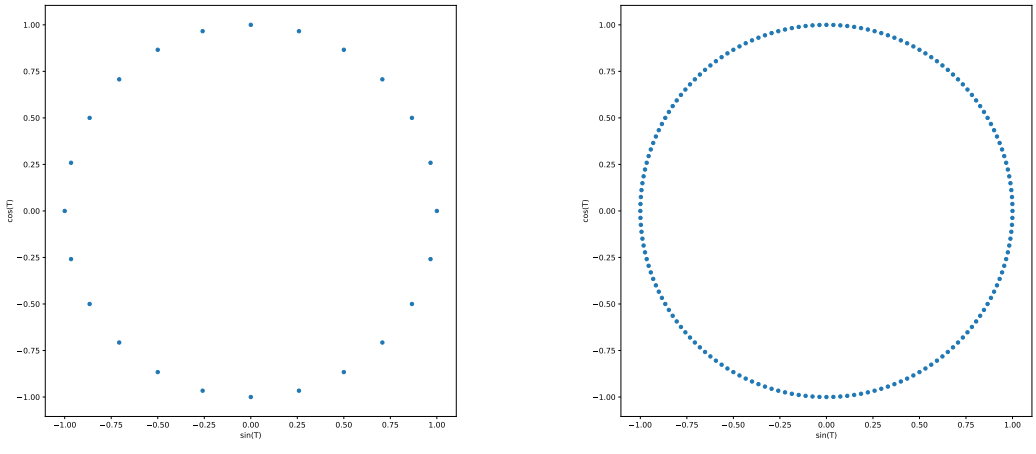
First of all, we think date and time should be transformed to the contiguous values which represent the position of the cycle. We apply a simple idea on the datetime column and transform it into two cycling relationship.

1. **Time in a day**: Initially, we calculate every every datetime column of events into N_{day} seconds of a day, which is the absolute position of a day. Next, we apply Equation 4.1 to successfully put the point onto the 2-D coordinate system. There are total 86400 seconds in a day. Figure 4.1a plots the 24 hours after this transform.

$$(x,\ y) = (\ \sin(2\pi \times \frac{N_{day}}{86400}),\ \cos(2\pi \times \frac{N_{day}}{86400})\ ) \tag{4.1}$$

2. **Time in a week**: We apply similar procedure as Time in a day on the datetime column. We calculate every datetime column of events into N_{week} seconds of a week, which means the absolute position in a week. Next, we apply almost the same Equation 4.2 to finish this task. Because of encoding relationship of weekday and daytime, we extend the Equation4.1 by multiplying 7 on 86400 seconds to represent total seconds of a week. Figure 4.1b plots the 24 hours per day of a week after this transform.

$$(x,\ y) = (\ \sin(2\pi \times \frac{N_{week}}{86400 \times 7}),\ \cos(2\pi \times \frac{N_{week}}{86400 \times 7})\ ) \tag{4.2}$$



(a) The 24 hours transformed by Equation 4.1 onto the 2-D coordinate system and plot.

(b) The 7x24 hours transformed by Equation 4.2 onto the 2-D coordinate system and plot.

Figure 4.1: Transform the datetime feature to plot on the X-Y coordinate system.

4.2.4 Time Precision Alignment

After split sessions, we found a big problem on the column which recorded the CS events occur time. While merging the ECS into CS, the precision between ECS and CS is not aligned. It may lead to wrong sequences. The CS data records the precision to second, but the ECS data records the precision to micro-second. Intuitively, it may not cause any problem before combining the two type of data. However, we join the data and we observe the error shifting on active events with minus 1-second range. We figure out the problem is made by our data collecting policy. We collect the CS from the API built-in function which round the time precision to second, but we collect ECS from our functions.

To solve this problem, especially on the active events with minus 1 second range, we go through all active events minus 1-second range and change the CS in the range into active events time plus 1 micro-second.

4.2.5 Summary of Data Preprocessing

The summary of our data preprocessing workflow shown below:

1. Query ECS and CS data from the database.

2. Filter unintentional events from both of data. The details are described in Section 4.2.1.

3. Merge two types of data.

4. Align Time Precision. The details are described in Section 4.2.4.

5. Discretize the time interval into buckets(categories). The details are described in Section 2.3.

6. Transform Time(Mapping to 2-D coordinate system), and output a temp file.

7. Target on the temp file, because there are concurrent events, we use One-Hot to encode the categorical columns (event types, URL category, and time buckets), group by timestamp, and output this ECS+CS preprocessed data to file.

8. Target on the temp file, Filtered ECS events, One-Hot encoded the categorical columns(event types, URL category, and time buckets) and grouped by timestamp. Output this as CS preprocessed data to file.

9. By using CS preprocessed data, we can slide a window to make the sequences of features and target label to cach all these sequences to file named CS_seq.#

10. By using ECS+CS preprocessed data aligned with all the CS_seq periods, we add ECS into every sequence of features. However, the target is not changed, which means that features are ECS+CS but the label is CS events. We cache all these sequences to file named ECS_to_CS_seq.$_{\#}$

11. By using ECS+CS preprocessed data, we slide a window to make the sequence of features and target label cached all these sequences to file named ECS+CS_seq.$_{\#}$

Note: # is the files we using on training and testing.

4.3 Phase III. - Model the User Behavior

Deep Learning Model is designed on a tiny scale in order to fit the CS data(almost half the quantity of ECS+CS dataset) for performing the fair comparison. Figure 4.2 shows the structure of our model. We expand the time-series part of Input Layer and GRU[13] Layer. Finally, we have two outputs of URL categories probability and intervals probability of intervals in next events. We use GRU to capture the user behavior rather than LSTM[14], because GRU can overcome the problem of gradients vanishing and it is faster than LSTM.

Train Test data split task is the process before we start training our model. Figure 4.3 showed the example of the process we did. We first split our data into the training part and testing part by setting the period of testing part to last 5 days for every user. If the user has data less than 5 days, all of the user data will be assigned to the testing part. Next, we split evaluation part from training part, the training part of users who

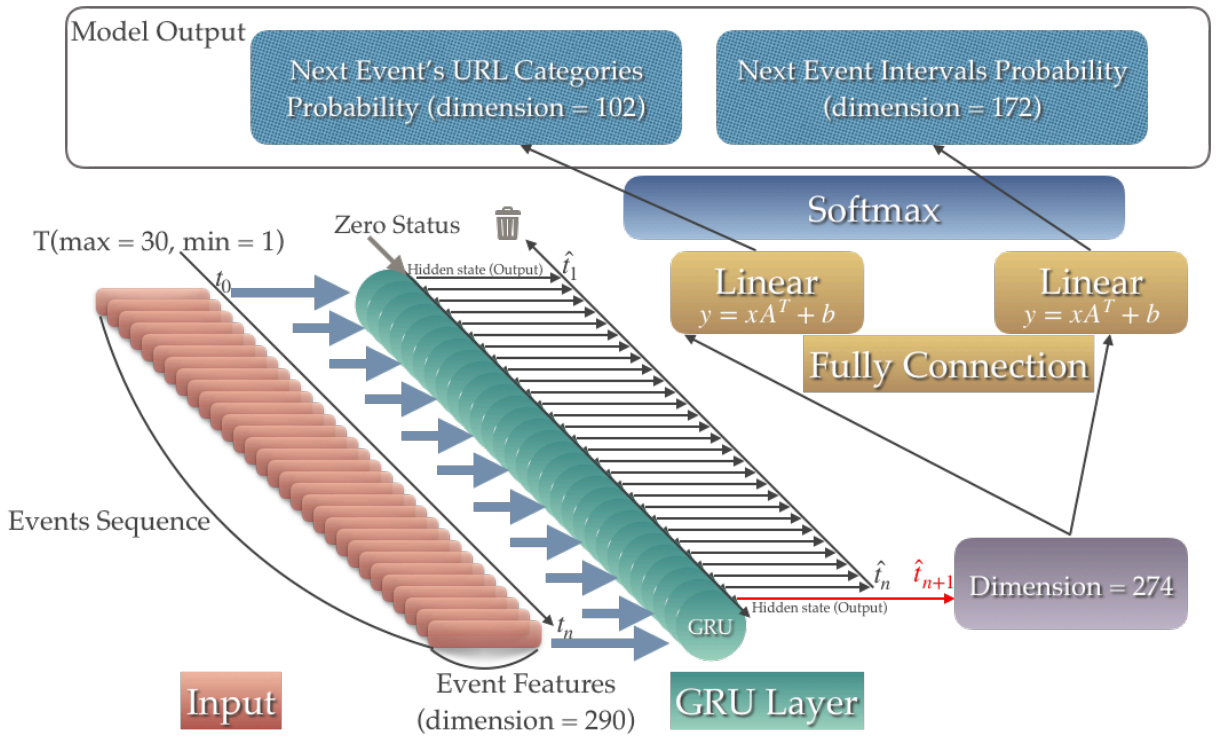


Figure 4.2: Our Model's data flow is Input -> GRU -> FC -> Output. We are using Multi-Task Learning skill to predict categories and intervals information at the same time because we think the category of a URL is highly related to the interval. The figure is shown the data of CS_seq and ECS+CS_seq that every input sequences are all in the window size of range 1 to 30. Note that ECS+CS_seq data are made from these CS_seq by adding ECS events into the CS_seq so ECS_to_CS_seq window size is in the range of 1 to unknown depends on the period that user has how many ECS events.

has greater or equal 25 days data will be split into 5 days evaluation part and N days training part(N >= 20). Note that our users almost use our extension for more than 28 days.

Training Models are used the batch training method with Adam[15] optimizer(set the initial learning rate = 0.001). Due to decaying the training time, we apply the trick[16] of increasing batch size between training period.

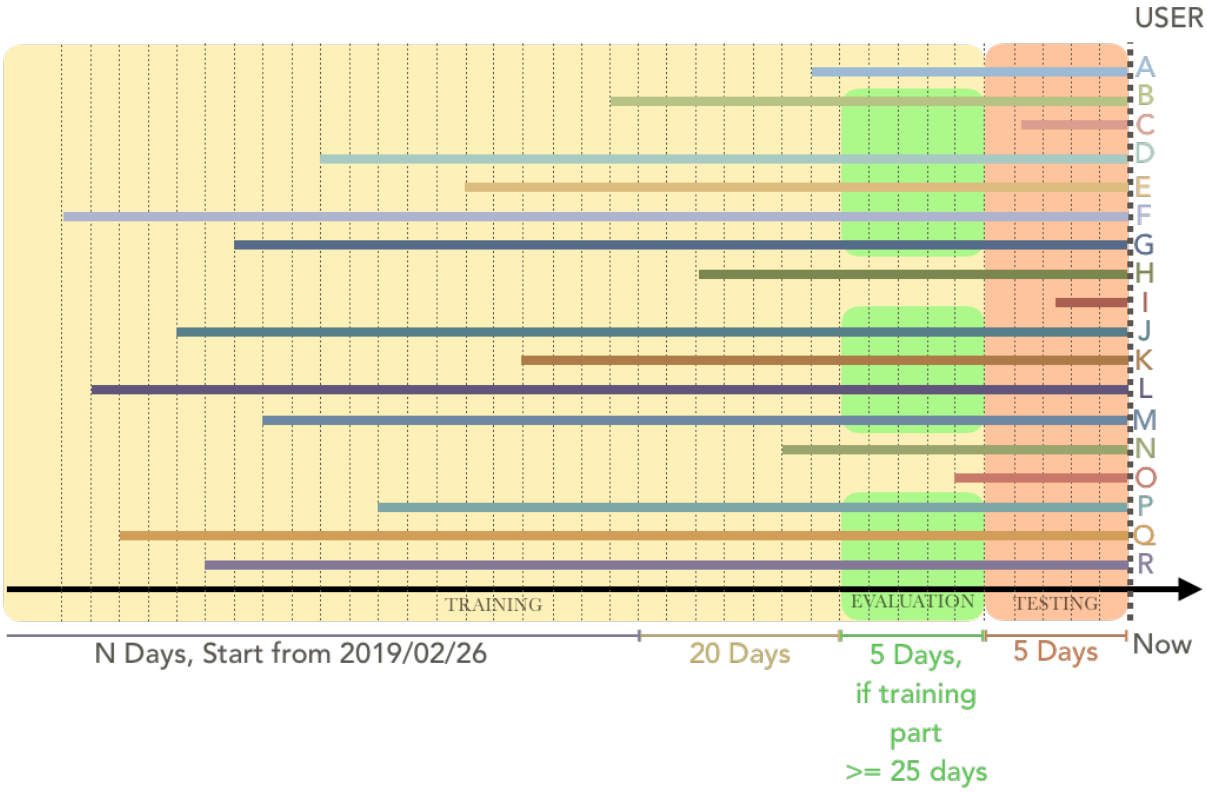


Figure 4.3: This is an example about out train test split process. **User- A, H, N, O** are the users who have the **testing part** and the **training part**. Because these users data after splitting, the remaining data less than 25 days; **User- C, I** are the users who only have less than 5 days data. We assigned these users only in the **testing part**. Because these users are the whole new user in the online environment of service, we want to serve them, too; Other users are regular users. In our dataset, over 80% are this type of user. They all have the **training part**, the **evaluation part**, and the **testing part**.

Chapter 5

Results

5.1 Collected Data

Our extension service started since February 26 in 2019. We verified **150 users** contributes over 4,657,434 CS data and over 2,815,167 ECS data until June 17, 2019. Noted that ECS data could be collected ever since users installed the extension, but CS data could only be collected within three months.

Table 5.1 is the raw data of ECS we collected from the client. There are 7 columns for ECS table in our database. The UUID column is the user unique identity which is hashed token by using SHA3-512 algorithm ideally without collision. The type column is the types of ECS events. The dt column is the event happened that timestamps precision to micro-second. The url and domain columns are the events happened of pages.

Table 5.2 is the raw data of CS we collected from the client. There are 8 columns for CS table in our database. The UUID, title, url, and domain columns are defined the same as the ECS table. The visit_id column is the ID generated by Chrome. The visit_time is the datetime format that records the date and time of user visiting the page. The type column is

the transition type of events.

Table 5.1 and Table 5.2 are sampled from the same period and user. You can clearly see the difference. First is the quantity of events, second is the diversity of events, and last is the intuitive of information. ECS can be easily converted to know what user actually did at the period.

Table 5.1: A sample of ECS raw data we collected(drop a column names 'incognito' which records the tab in an incognito window or not)

	UUID	title	url	domain	type	dt
0	6d0...867	グランブル...	http://game...	game....	tabs	155...824
1	6d0...867	グランブル...	http://game...	game....	windows	155...824
2	6d0...867	グランブル...	http://game...	game....	idle	155...012
3	6d0...867	グランブル...	http://game...	game....	active	155...044
4	6d0...867	グランブル...	http://game...	game....	blur	155...762
5	6d0...867	グランブル...	http://game...	game....	windows	155...541
6	6d0...867	グランブル...	http://game...	game....	blur	155...604
7	6d0...867	グランブル...	http://game...	game....	windows	155...953
8	6d0...867	グランブル...	http://game...	game....	blur	155...530
9	6d0...867	グランブル...	http://game...	game....	windows	155...692
10	6d0...867	グランブル...	http://game...	game....	idle	155...228

Table 5.2: A sample of CS raw data we collected(drop a column names 'log_time' which records when the log be logged)

	UUID	visit_id	visit_time	title	url	type	domain
0	6d0...867	384399	2019-05-25 20:40:16	グランブル...	http://game...	link	game....
1	6d0...867	384400	2019-05-25 20:40:19	グランブル...	http://game...	link	game....
2	6d0...867	384401	2019-05-25 20:40:24	グランブル...	http://game...	link	game....
3	6d0...867	384402	2019-05-25 20:40:24	グランブル...	http://game...	link	game....
4	6d0...867	384403	2019-05-25 20:45:56	グランブル...	http://game...	link	game....
5	6d0...867	384404	2019-05-25 20:46:32	グランブル...	http://game...	link	game....
6	6d0...867	384405	2019-05-25 20:46:38	グランブル...	http://game...	link	game....
7	6d0...867	384406	2019-05-25 20:46:46	グランブル...	http://game...	link	game....
8	6d0...867	384407	2019-05-25 20:46:46	グランブル...	http://game...	link	game....

5.2 Data Analysis

We perform a series of data preprocessing and find a special phenomenon on merging ECS and CS together. First, we do a survey of the two types of data distribution. Then we try to make the time information contiguous. Next we focus on a case study for the special phenomenon.

5.2.1 Statics Analysis

1. We perform a statistic on both of ECS and CS data for records counts per user after installing extension in the Table 5.3.

Table 5.3: A statistical summary of the number of events per user.

Data type	MIN	1^{st} Quartile	$Median$	$Mean$	3^{rd} Quartile	MAX	$Std.$
ECS	11	5323.75	11096.5	14653.11	19997	61769	13072.41
CS	5	4120.50	9441.5	12920.13	19113	81062	12095.50

2. We perform a statistic on the data for how many days using our extension per user in the Table 5.4.

Table 5.4: A statistical summary of the number of days per user

	MIN	1^{st} Quartile	$Median$	$Mean$	3^{rd} Quartile	MAX	$Std.$
Days	0.000961	42.340304	50.766186	50.94735	67.086444	96.666597	24.016554

3. We perform a statistic on the data for every user(not included using less than 1 day) generating events per day in the Table 5.5.

Table 5.5: A statistical summary of the number of events per user generates per day. There are 138 users using over 1 day.

Data type	MIN	1^{st} Quartile	$Median$	$Mean$	3^{rd} Quartile	MAX	$Std.$
ECS	0.9999	146.4713	248.9599	298.7828	406.5702	1471.5511	229.6235
CS	0.8181	118.0636	219.1643	258.0534	357.0189	943.3539	190.1575

4. We built the extension that can auto request the online website classifier for the categories of every domain name. Table 5.6 is the category distribution in both types of data. Table 5.7 shows the top 25% counts of ECS categories compared with CS categories. We can find the critical difference, two types of data are shown very different behavior on web browsing.

Table 5.6: A statistical summary of the URL category events counting distribution in both types of data .

Data type	Category Count #	*MIN* #	1^{st} Quartile #	*Median* #	*Mean* #	3^{rd} Quartile #	*MAX* #	*Std.* #
ECS	82	2	114	2023	26816.0732	16436.0	384973	64013.8399
CS	83	1	83	1198	23301.6627	13710.5	303999	59319.6952

Table 5.7: The top 25% counts of ECS URL categories compared with CS URL categories.

Category	ECS Rank	ECS events #	ECS events %	ECS CDF%	CS Rank	CS events #	CS events %	CS CDF%
Streaming Media and Download	1	465433	17.30	17.30	3	344329	14.44	14.44
Social Networking	2	279536	10.39	27.69	1	381929	16.02	30.46
Web Browser Application	3	250883	9.33	37.02	20	17109	0.72	31.18
Education	4	243564	9.05	46.07	5	178708	7.49	38.67
Information Technology	5	224919	8.36	54.43	6	124528	5.22	43.89
Search Engines and Portals	6	210840	7.84	62.27	2	347674	14.58	58.47
Games	7	178580	6.64	68.91	7	107947	4.53	63.00
Local Host	8	102049	3.79	72.70	12	54532	2.29	65.29
Business	9	87012	3.23	75.93	10	62566	2.62	67.91
Shopping	10	67879	2.52	78.45	11	58941	2.47	70.38
Reference	11	61166	2.27	80.72	13	52865	2.22	72.60
File Sharing and Storage	12	49639	1.85	82.57	9	65109	2.73	75.33
Web-based Applications	13	41392	1.54	84.11	4	198447	8.32	83.65
Web-based Email	14	40163	1.49	85.60	14	40514	1.70	85.35
Personal Websites and Blogs	15	36807	1.37	86.97	23	15997	0.67	86.02
Entertainment	16	32285	1.20	88.17	8	72985	3.06	89.08
Newsgroups and Message Boards	17	31624	1.18	89.35	17	20384	0.85	89.93
Instant Messaging	18	28980	1.08	90.43	18	20017	0.84	90.77
News and Media	19	27646	1.03	91.46	15	36385	1.53	92.30
File	20	25622	0.95	92.41	30	5032	0.21	92.51
Pornography	21	20964	0.78	93.19	16	24589	1.03	93.54
Others out of Top 25%	-	183164	6.81	100.00	-	154018	6.46	100.00

5. We perform a statistic on both of the data for event types distribution. Table 5.8 shows the CS event types distribution and Table 5.9 shows the ECS event types distribution.

Table 5.8: A statistical summary of the percentage of CS events types distribution.

Event type	link	form_submit	auto_bookmark	reload	generated
percentage%	83.025	4.795	4.443	2.813	2.769
Event type	typed	auto_toplevel	manual_subframe	keyword	auto_subframe
percentage%	1.333	0.763	0.053	0.006	0.001

Table 5.9: A statistical summary of the percentage of ECS events types distribution.

Event type	tabs	windows	blur	idle	active
percentage%	59.145	15.694	14.622	5.42	5.119

6. After we discretized the interval of events, which method described in 2.3, we perform a statistic. Table 5.10 and Figure 5.1 show the cumulative distribution function of CS intervals and ECS intervals. There is a weird concurrent phenomenon in CS data. According to Table 5.10, we observed that over 20% CS events are concurrent, while there are only 0.38% concurrent events in ECS data. We do more research on this and find the possible reason is the timestamp precision of CS data is on second-scale. There are lots of users like to open many links at once or set browser reopen the last browsing state as starting browsing. Furthermore, we also found there are over 23% ECS events are less than 1 second and we guess it is related to the CS event.

Table 5.10: The cumulative distribution function of ECS intervals and CS intervals. (data until June 17, 2019)

	ECS events			CS events		
Interval	#	%	CDF%	#	%	CDF%
0S(Concurrent)	14945	0.380929	0.380929	446550	20.428927	20.428927
0S~0.5S	557791	14.217375	14.598304	0	0.000000	20.428927
0.5S~1S	341172	8.696036	23.294340	0	0.000000	20.428927
1S~2S	443516	11.304653	34.598992	235268	10.763124	31.192051
2S~3S	257358	6.559725	41.158717	165649	7.578169	38.770220
3S~4S	182871	4.661147	45.819864	126636	5.793389	44.563609
4S~5S	142119	3.622431	49.442294	99547	4.554111	49.117720
5S~6S	116058	2.958169	52.400463	79191	3.622858	52.740578
6S~7S	96962	2.471437	54.871900	64258	2.939698	55.680276
7S~8S	83079	2.117577	56.989477	53911	2.466340	58.146615
8S~9S	73066	1.862358	58.851835	47028	2.151454	60.298069
9S~10S	64151	1.635127	60.486962	40543	1.854776	62.152844
10S~20S	398182	10.149147	70.636109	225758	10.328057	72.480901
20S~30S	204946	5.223810	75.859919	100739	4.608643	77.089545
30S~40S	132624	3.380415	79.240334	61065	2.793623	79.883168
40S~50S	94842	2.417401	81.657735	41907	1.917176	81.800344
50S~1M	73683	1.878085	83.535820	31485	1.440387	83.240731
1M~2M	247728	6.314268	89.850088	102663	4.696663	87.937394
2M~3M	134175	3.419948	93.270036	50617	2.315644	90.253039
3M~4M	68877	1.755586	95.025623	38320	1.753077	92.006116
4M~5M	36339	0.926234	95.951857	31551	1.443406	93.449522
5M~6M	23402	0.596487	96.548344	19759	0.903942	94.353464
6M~7M	16608	0.423317	96.971660	12307	0.563025	94.916489
7M~8M	12607	0.321336	97.292997	9524	0.435707	95.352196
8M~9M	9998	0.254836	97.547833	7827	0.358072	95.710268
9M~10M	7781	0.198328	97.746160	6676	0.305416	96.015684
10M~20M	35348	0.900975	98.647136	34046	1.557548	97.573233
20M~30M	13240	0.337471	98.984606	13725	0.627896	98.201129
30M~40M	6575	0.167588	99.152194	7878	0.360406	98.561535
40M~50M	4199	0.107027	99.259221	4554	0.208338	98.769873
50M~1H	2778	0.070808	99.330029	2921	0.133631	98.903503
1H~2H	7485	0.190783	99.520812	8016	0.366719	99.270222
2H~4H	5420	0.138149	99.658961	4732	0.216481	99.486703
4H~8H	4428	0.112864	99.771825	3672	0.167988	99.654691
8H~12H	4342	0.110672	99.882497	3601	0.164740	99.819431
12H~1D	3646	0.092932	99.975429	3213	0.146989	99.966421
1D~2D	611	0.015574	99.991002	500	0.022874	99.989295
2D~3D	157	0.004002	99.995004	116	0.005307	99.994602
3D~4D	80	0.002039	99.997043	52	0.002379	99.996981
4D~5D	29	0.000739	99.997782	22	0.001006	99.997987
5D~6D	18	0.000459	99.998241	13	0.000595	99.998582
6D~1W	24	0.000612	99.998853	11	0.000503	99.999085
>= 1W	45	0.001147	100.000000	20	0.000915	100.000000

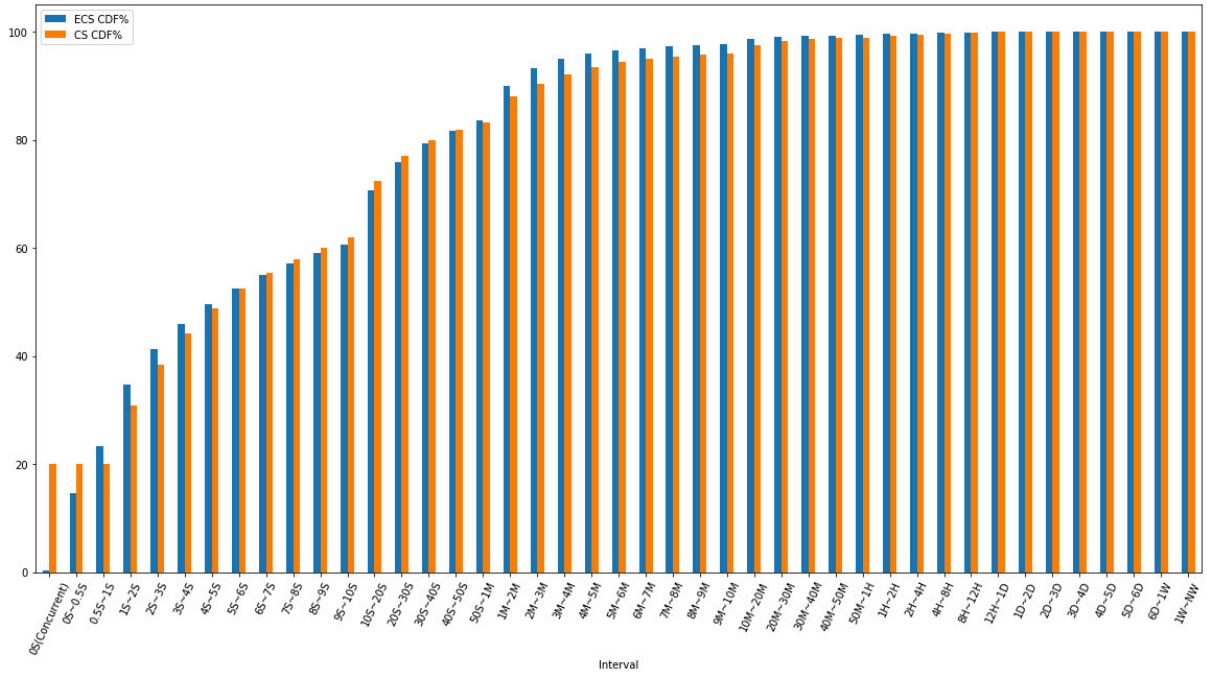


Figure 5.1: The cumulative distribution function of CS intervals and ECS intervals.

5.2.2 Case Study - Multi-device detection and Unintentional events in CS

Following the rule described in Section 4.2.2, after we finish the task, we find the weird phenomenon against our definition of the idle event. Idle events are defined as a user does not have any I/O input over 2 minutes; while active event are defined oppositely to idle events. By definition, the active event should always be followed by idle event like Figure 3.1. We first check the collector(extension code) but there is no bug in our code. Next, we inspect into the raw data (CS data merged with ECS data ordered by the datetime). We find there are other events between the idle and active events.

We started quick research on this phenomenon and observed many events between the idle-active interval were the mobile-site URL. Regarding definition, these events may happen on other device but sync to the google account. So the active event may occur after lots of events. How-

ever, we observed that the events in the idle-active interval would appear YouTube, Google Document and News Website a lot. We monitored these sites and found these sites reloaded by linking a URL even user has been idled. The links events were unintentional but recorded in the data.

5.3 Model Evaluate

Table 5.11 shows the representative of our models and the detail of their data usage on training and testing phase. The representative of features is preprocessing sequences, the same as Section 4.2.5. We use KL divergence loss[17] as our criterion to optimize our models. Because there are concurrent events in our data, both of the label and the outputs of our models may lead to being a multi-class probability vector. To compute the distance of two probability distributions, the KL divergence loss is the best choice. Table 5.12 shows the detail of testing loss of our models. Note

Table 5.11: The representative of our models. Note the ECS_to_CS_seq is at least contain 2 - 30 CS events, by adding ECS events into CS_seq, so Sequence Length is shown 1 - 30^{+}.

Model		Training	Phase			Testing	Phase	
Representative	Features	Label	Sequence Length	Sequence Count	Task	Features	Label	Sequence Length
CS_CS	CS_seq	CS	1 - 30	~1360000	Both	CS_seq	CS	1 - 30
ECSCS_CS	ECS_to_CS_seq	CS	1 - 30^{+}	~1360000	Both	ECS_to_CS_seq	CS	1 - 30^{+}
CS_ECSCS	CS_seq	CS	1 - 30	~1360000	URL Categories	ECS+CS_seq	ECS+CS	1 - 30
					Interval class	ECS+CS_seq	CS	1 - 30
ECSCS_ECSCS	ECS+CS_seq	ECS+CS	1 - 30	~2800000	Both	ECS+CS_seq	ECS+CS	1 - 30

that the CS_CS and ECSCS_CS models in the part of the interval class of next event have recalculated the loss by focusing on the one-hot encoded label of CS part.

We compute the accuracy based on the macro-averaged score (we named "macro accuracy") and the micro-averaged score (we named "micro accuracy"). Specifically, the macro accuracy is computed by averaging the accuracy score of each class, as shown in Equation 5.1. Consequently, the

accuracy score of every class is equally important. We compute the micro accuracy by dividing the number of the correct predictions for all classes to the total number of predictions, as shown in Equation 5.2. Therefore, the result is highly influenced by the accuracy scores of the classes that have many instances.

$$Macro\ Accuracy = \frac{\sum_i^C (\frac{the\ number\ of\ correct\ prediction\ of\ the\ class_i}{the\ number\ of\ the\ class_i})}{C}, \quad C = the\ number\ of\ classes \tag{5.1}$$

$$Micro\ Accuracy = \frac{\sum_i^C the\ number\ of\ correct\ prediction\ of\ the\ class_i}{\sum_i^C the\ number\ of\ the\ class_i}, \quad C = the\ number\ of\ classes \tag{5.2}$$

Compared CS_CS with ECSCS_CS models, we observe little improvement on both tasks of the model adding ECS into CS sequences. Table 5.13, 5.14, 5.15, 5.16 show the CS_CS and ECSCS_CS models testing hit

Table 5.12: Testing Loss of the four models.

Task	The URL categories of next event		
Models	CS_CS	ECSCS_CS	ECSCS_ECSCS
Experiment count	4	4	11
Mean of Loss	0.010580	0.010533	0.011004
Standard Deviation	0.000022	0.000090	0.000441
Macro Accuracy(%)	69.982029	70.354430	64.569433
Micro Accuracy(%)	72.881802	72.830589	67.194181
Task	The interval class of next event		
Models	CS_CS	ECSCS_CS	ECSCS_ECSCS
Experiment count	4	4	11
Mean of Loss	0.063585	0.059752	0.016629
Standard Deviation	0.004193	0.004169	0.000395
Macro Accuracy(%)	15.012181	15.184085	32.664327
Micro Accuracy(%)	24.227752	24.461621	41.059520

map in both of tasks. These tables selected the top events count classes and aggregated the others as a class named ‘Others’.

Table 5.13: Model CS_CS Hit map on task of the Top 25% event counts of CS URL categories prediction

True Label (Total)	Social Networking (82080)	Search Engines and Portals (73040)	Streaming Media and Download (75316)	Web-based Applications (10044)	Education (37816)	Information Technology (23400)	Games (27684)	Entertainment (13028)	File Sharing and Storage (12028)	Business (12604)	Shopping (11584)	Local Host (6208)	Reference (12160)	Web-based Email (10468)	News and Media (8784)	Pornography (7644)	Newsgroups and Message Boards (4884)	Instant Messaging (4936)	Auction (6440)	Web Browser Application (3572)	Not Rated (1288)	Others (26100)
Social Networking	66886 (81.49%)	3462 (4.74%)	4800 (6.37%)	552 (5.50%)	1397 (3.69%)	925 (3.95%)	914 (3.30%)	456 (3.50%)	479 (3.98%)	896 (7.11%)	322 (2.78%)	187 (3.01%)	307 (2.52%)	842 (8.04%)	977 (11.12%)	119 (1.56%)	378 (7.74%)	285 (5.77%)	141 (2.19%)	285 (7.98%)	84 (6.52%)	1332 (5.10%)
Search Engines and Portals	1854 (2.26%)	46735 (63.99%)	2606 (3.46%)	189 (1.88%)	2777 (7.34%)	2552 (10.91%)	917 (3.31%)	545 (4.18%)	360 (2.99%)	1891 (15.00%)	1177 (10.16%)	138 (2.22%)	1400 (11.51%)	515 (4.92%)	950 (10.82%)	150 (1.96%)	812 (16.63%)	97 (1.97%)	322 (5.00%)	151 (4.23%)	91 (7.07%)	3445 (13.20%)
Streaming Media and Download	4582 (5.58%)	5022 (6.88%)	58406 (77.55%)	274 (2.73%)	1365 (3.61%)	782 (3.34%)	1396 (5.04%)	558 (4.28%)	506 (4.21%)	524 (4.16%)	239 (2.06%)	93 (1.50%)	672 (5.53%)	510 (4.87%)	265 (3.02%)	124 (1.62%)	366 (7.49%)	427 (8.65%)	133 (2.07%)	459 (12.85%)	117 (9.08%)	1291 (4.95%)
Web-based Applications	459 (0.56%)	435 (0.60%)	293 (0.39%)	7917 (78.82%)	155 (0.41%)	71 (0.30%)	28 (0.10%)	17 (0.13%)	169 (1.41%)	53 (0.42%)	16 (0.14%)	40 (0.64%)	20 (0.16%)	141 (1.35%)	12 (0.14%)	4 (0.05%)	8 (0.16%)	74 (1.50%)	0 (0.00%)	32 (0.90%)	12 (0.93%)	162 (0.62%)
Education	1533 (1.87%)	3307 (4.53%)	1383 (1.84%)	231 (2.30%)	28443 (75.21%)	670 (2.86%)	318 (1.15%)	84 (0.64%)	224 (1.86%)	441 (3.50%)	118 (1.02%)	38 (0.61%)	493 (4.05%)	326 (3.11%)	145 (1.65%)	66 (0.86%)	183 (3.75%)	115 (2.33%)	69 (1.07%)	128 (3.58%)	75 (5.82%)	921 (3.53%)
Information Technology	775 (0.94%)	2972 (4.07%)	977 (1.30%)	133 (1.32%)	810 (2.14%)	15832 (67.66%)	449 (1.62%)	81 (0.62%)	105 (0.87%)	599 (4.75%)	263 (2.27%)	183 (2.95%)	259 (2.13%)	170 (1.62%)	179 (2.04%)	105 (1.37%)	203 (4.16%)	67 (1.36%)	26 (0.40%)	102 (2.86%)	83 (6.44%)	861 (3.30%)
Games	946 (1.15%)	1494 (2.05%)	1627 (2.16%)	22 (0.22%)	438 (1.16%)	477 (2.04%)	22447 (81.08%)	403 (3.09%)	95 (0.79%)	214 (1.70%)	68 (0.59%)	19 (0.31%)	184 (1.51%)	42 (0.40%)	65 (0.74%)	74 (0.97%)	68 (1.39%)	49 (0.99%)	0 (0.00%)	77 (2.16%)	14 (1.09%)	500 (1.92%)
Entertainment	363 (0.44%)	439 (0.60%)	452 (0.60%)	19 (0.19%)	70 (0.19%)	83 (0.35%)	201 (0.73%)	10429 (80.05%)	18 (0.15%)	128 (1.02%)	5 (0.04%)	51 (0.82%)	44 (0.36%)	26 (0.25%)	28 (0.32%)	63 (0.82%)	30 (0.61%)	171 (3.46%)	8 (0.12%)	17 (0.48%)	9 (0.70%)	355 (1.36%)
File Sharing and Storage	434 (0.53%)	436 (0.60%)	434 (0.58%)	284 (2.83%)	211 (0.56%)	55 (0.24%)	87 (0.31%)	12 (0.09%)	9518 (79.13%)	43 (0.34%)	11 (0.09%)	75 (1.21%)	39 (0.32%)	124 (1.18%)	28 (0.32%)	34 (0.44%)	43 (0.88%)	41 (0.83%)	7 (0.11%)	150 (4.20%)	12 (0.93%)	103 (0.39%)
Business	663 (0.81%)	1179 (1.61%)	466 (0.62%)	59 (0.59%)	267 (0.71%)	362 (1.55%)	91 (0.33%)	76 (0.58%)	60 (0.50%)	6216 (49.32%)	272 (2.35%)	20 (0.32%)	138 (1.13%)	198 (1.89%)	122 (1.39%)	17 (0.22%)	80 (1.64%)	49 (0.99%)	39 (0.61%)	46 (1.29%)	32 (2.48%)	488 (1.87%)
Shopping	359 (0.44%)	1432 (1.96%)	256 (0.34%)	15 (0.15%)	227 (0.60%)	258 (1.10%)	141 (0.51%)	15 (0.12%)	20 (0.17%)	353 (2.80%)	8452 (72.96%)	0 (0.00%)	40 (0.33%)	38 (0.36%)	49 (0.56%)	20 (0.26%)	52 (1.06%)	34 (0.69%)	293 (4.55%)	22 (0.62%)	0 (0.00%)	226 (0.87%)
Local Host	223 (0.27%)	287 (0.39%)	122 (0.16%)	32 (0.32%)	24 (0.06%)	125 (0.53%)	12 (0.04%)	57 (0.44%)	76 (0.63%)	26 (0.21%)	0 (0.00%)	5259 (84.71%)	5 (0.04%)	3 (0.03%)	1 (0.01%)	0 (0.00%)	2 (0.04%)	11 (0.22%)	0 (0.00%)	12 (0.34%)	35 (2.72%)	97 (0.37%)
Reference	419 (0.51%)	1351 (1.85%)	763 (1.01%)	12 (0.12%)	441 (1.17%)	206 (0.88%)	187 (0.68%)	43 (0.33%)	29 (0.24%)	189 (1.50%)	35 (0.30%)	10 (0.16%)	8266 (67.98%)	77 (0.74%)	77 (0.88%)	10 (0.13%)	66 (1.35%)	19 (0.38%)	5 (0.08%)	22 (0.62%)	3 (0.23%)	268 (1.03%)
Web-based Email	635 (0.77%)	554 (0.76%)	520 (0.69%)	106 (1.06%)	323 (0.85%)	244 (1.04%)	33 (0.12%)	32 (0.25%)	130 (1.08%)	251 (1.99%)	62 (0.54%)	8 (0.13%)	61 (0.50%)	7084 (67.67%)	52 (0.59%)	23 (0.30%)	20 (0.41%)	53 (1.07%)	16 (0.25%)	89 (2.49%)	12 (0.93%)	233 (0.89%)
News and Media	481 (0.59%)	557 (0.76%)	290 (0.39%)	10 (0.10%)	79 (0.21%)	105 (0.45%)	40 (0.14%)	62 (0.48%)	28 (0.23%)	110 (0.87%)	10 (0.09%)	1 (0.02%)	31 (0.25%)	48 (0.46%)	5582 (63.55%)	10 (0.13%)	37 (0.76%)	21 (0.43%)	12 (0.19%)	41 (1.15%)	0 (0.00%)	138 (0.53%)
Pornography	116 (0.14%)	234 (0.32%)	133 (0.18%)	2 (0.02%)	36 (0.10%)	89 (0.38%)	77 (0.28%)	68 (0.52%)	16 (0.13%)	12 (0.10%)	14 (0.12%)	1 (0.02%)	4 (0.03%)	23 (0.22%)	17 (0.19%)	6675 (87.32%)	22 (0.45%)	14 (0.28%)	8 (0.12%)	17 (0.48%)	24 (1.86%)	166 (0.64%)
Newsgroups and Message Boards	140 (0.17%)	346 (0.47%)	247 (0.33%)	20 (0.20%)	81 (0.21%)	72 (0.31%)	70 (0.25%)	43 (0.33%)	61 (0.51%)	58 (0.46%)	33 (0.28%)	2 (0.03%)	32 (0.26%)	44 (0.42%)	11 (0.13%)	21 (0.27%)	2339 (47.89%)	2 (0.04%)	15 (0.23%)	9 (0.25%)	0 (0.00%)	103 (0.39%)
Instant Messaging	276 (0.34%)	142 (0.19%)	338 (0.45%)	45 (0.45%)	101 (0.27%)	38 (0.16%)	55 (0.20%)	146 (1.12%)	50 (0.42%)	80 (0.63%)	7 (0.06%)	0 (0.00%)	29 (0.24%)	52 (0.50%)	19 (0.22%)	10 (0.13%)	5 (0.10%)	3357 (68.01%)	0 (0.00%)	23 (0.64%)	1 (0.08%)	49 (0.19%)
Auction	137 (0.17%)	454 (0.62%)	132 (0.18%)	0 (0.00%)	75 (0.20%)	14 (0.06%)	4 (0.01%)	10 (0.08%)	0 (0.00%)	34 (0.27%)	265 (2.29%)	4 (0.06%)	4 (0.03%)	12 (0.11%)	13 (0.15%)	0 (0.00%)	46 (0.94%)	0 (0.00%)	5349 (83.06%)	0 (0.00%)	0 (0.00%)	62 (0.24%)
Web Browser Application	176 (0.21%)	191 (0.26%)	303 (0.40%)	30 (0.30%)	93 (0.25%)	62 (0.26%)	33 (0.12%)	16 (0.12%)	38 (0.32%)	53 (0.42%)	10 (0.09%)	16 (0.26%)	5 (0.04%)	72 (0.69%)	12 (0.14%)	7 (0.09%)	16 (0.33%)	12 (0.24%)	0 (0.00%)	1879 (52.60%)	3 (0.23%)	41 (0.16%)
Not Rated	32 (0.04%)	97 (0.13%)	42 (0.06%)	6 (0.06%)	30 (0.08%)	43 (0.18%)	6 (0.02%)	2 (0.02%)	0 (0.00%)	12 (0.10%)	2 (0.02%)	6 (0.10%)	12 (0.10%)	12 (0.11%)	4 (0.05%)	0 (0.00%)	2 (0.04%)	0 (0.00%)	4 (0.06%)	2 (0.06%)	639 (49.61%)	21 (0.08%)
Others	895 (1.09%)	2234 (3.06%)	914 (1.21%)	138 (1.37%)	521 (1.38%)	487 (2.08%)	238 (0.86%)	181 (1.39%)	102 (0.85%)	629 (4.99%)	295 (2.55%)	61 (0.98%)	159 (1.31%)	225 (2.15%)	216 (2.46%)	124 (1.62%)	166 (3.40%)	78 (1.58%)	37 (0.57%)	41 (1.15%)	54 (4.19%)	15642 (59.93%)

Table 5.14: Model ECSCS_CS Hit map on task of the Top 25% event counts of CS URL categories prediction

True Label (Total)	Social Networking (81880)	Search Engines and Portals (73012)	Streaming Media and Download (75056)	Web-based Applications (10020)	Education (37704)	Information Technology (23400)	Games (27684)	Entertainment (13028)	File Sharing and Storage (12012)	Business (12600)	Shopping (11584)	Local Host (6208)	Reference (12160)	Web-based Email (10468)	News and Media (8732)	Pornography (7640)	Newsgroups and Message Boards (4884)	Instant Messaging (4936)	Auction (6440)	Web Browser Application (3572)	Not Rated (1284)	Others (26008)
Social Networking	66388 (81.08%)	3282 (4.50%)	4575 (6.10%)	511 (5.10%)	1316 (3.49%)	887 (3.79%)	845 (3.05%)	410 (3.15%)	462 (3.85%)	868 (6.89%)	297 (2.56%)	189 (3.04%)	305 (2.51%)	797 (7.61%)	961 (11.01%)	83 (1.09%)	384 (7.86%)	264 (5.35%)	123 (1.91%)	282 (7.89%)	72 (5.61%)	1257 (4.83%)
Search Engines and Portals	2380 (2.91%)	46621 (63.85%)	3460 (4.61%)	361 (3.60%)	3275 (8.69%)	2397 (10.24%)	938 (3.39%)	591 (4.54%)	526 (4.38%)	1751 (13.90%)	1131 (9.76%)	97 (1.56%)	1359 (11.18%)	506 (4.83%)	989 (11.33%)	122 (1.60%)	715 (14.64%)	72 (1.46%)	313 (4.86%)	142 (3.98%)	115 (8.96%)	3379 (12.99%)
Streaming Media and Download	4673 (5.71%)	5067 (6.94%)	57719 (76.90%)	259 (2.58%)	1286 (3.41%)	805 (3.44%)	1393 (5.03%)	643 (4.94%)	485 (4.04%)	559 (4.44%)	247 (2.13%)	91 (1.47%)	711 (5.85%)	475 (4.54%)	269 (3.08%)	124 (1.62%)	347 (7.10%)	424 (8.59%)	113 (1.75%)	497 (13.91%)	101 (7.87%)	1291 (4.96%)
Web-based Applications	429 (0.52%)	425 (0.58%)	287 (0.38%)	7818 (78.02%)	142 (0.38%)	65 (0.28%)	29 (0.10%)	16 (0.12%)	133 (1.11%)	54 (0.43%)	11 (0.09%)	39 (0.63%)	17 (0.14%)	136 (1.30%)	8 (0.09%)	4 (0.05%)	8 (0.16%)	69 (1.40%)	0 (0.00%)	41 (1.15%)	10 (0.78%)	159 (0.61%)
Education	1446 (1.77%)	3155 (4.32%)	1313 (1.75%)	210 (2.10%)	27964 (74.17%)	665 (2.84%)	284 (1.03%)	81 (0.62%)	222 (1.85%)	404 (3.21%)	111 (0.96%)	36 (0.58%)	450 (3.70%)	319 (3.05%)	141 (1.61%)	57 (0.75%)	180 (3.69%)	108 (2.19%)	64 (0.99%)	137 (3.84%)	78 (6.07%)	882 (3.39%)
Information Technology	794 (0.97%)	3124 (4.28%)	1000 (1.33%)	129 (1.29%)	810 (2.15%)	16039 (68.54%)	453 (1.64%)	76 (0.58%)	91 (0.76%)	600 (4.76%)	227 (1.96%)	181 (2.92%)	266 (2.19%)	164 (1.57%)	203 (2.32%)	80 (1.05%)	222 (4.55%)	72 (1.46%)	24 (0.37%)	107 (3.00%)	71 (5.53%)	907 (3.49%)
Games	904 (1.10%)	1468 (2.01%)	1611 (2.15%)	23 (0.23%)	432 (1.15%)	450 (1.92%)	22563 (81.50%)	371 (2.85%)	92 (0.77%)	209 (1.66%)	68 (0.59%)	16 (0.26%)	182 (1.50%)	51 (0.49%)	53 (0.61%)	62 (0.81%)	70 (1.43%)	42 (0.85%)	1 (0.02%)	84 (2.35%)	14 (1.09%)	451 (1.73%)
Entertainment	354 (0.43%)	410 (0.56%)	411 (0.55%)	8 (0.08%)	61 (0.16%)	73 (0.31%)	195 (0.70%)	10403 (79.85%)	19 (0.16%)	133 (1.06%)	6 (0.05%)	53 (0.85%)	43 (0.35%)	28 (0.27%)	22 (0.25%)	59 (0.77%)	30 (0.61%)	156 (3.16%)	8 (0.12%)	16 (0.45%)	11 (0.86%)	334 (1.28%)
File Sharing and Storage	418 (0.51%)	417 (0.57%)	412 (0.55%)	259 (2.58%)	190 (0.50%)	44 (0.19%)	92 (0.33%)	14 (0.11%)	9431 (78.51%)	36 (0.29%)	9 (0.08%)	78 (1.26%)	35 (0.29%)	113 (1.08%)	26 (0.30%)	20 (0.26%)	42 (0.86%)	52 (1.05%)	8 (0.12%)	142 (3.98%)	11 (0.86%)	86 (0.33%)
Business	656 (0.80%)	1200 (1.64%)	439 (0.58%)	70 (0.70%)	285 (0.76%)	381 (1.63%)	85 (0.31%)	69 (0.53%)	59 (0.49%)	6368 (50.54%)	278 (2.40%)	17 (0.27%)	139 (1.14%)	204 (1.95%)	112 (1.28%)	22 (0.29%)	76 (1.56%)	48 (0.97%)	41 (0.64%)	56 (1.57%)	29 (2.26%)	463 (1.78%)
Shopping	330 (0.40%)	1511 (2.07%)	278 (0.37%)	14 (0.14%)	243 (0.64%)	316 (1.35%)	157 (0.57%)	17 (0.13%)	20 (0.17%)	367 (2.91%)	8593 (74.18%)	0 (0.00%)	41 (0.34%)	41 (0.39%)	32 (0.37%)	19 (0.25%)	50 (1.02%)	37 (0.75%)	296 (4.60%)	16 (0.45%)	1 (0.08%)	225 (0.87%)
Local Host	214 (0.26%)	318 (0.44%)	129 (0.17%)	41 (0.41%)	29 (0.08%)	137 (0.59%)	12 (0.04%)	45 (0.35%)	82 (0.68%)	19 (0.15%)	0 (0.00%)	5300 (85.37%)	8 (0.07%)	4 (0.04%)	0 (0.00%)	3 (0.04%)	12 (0.25%)	7 (0.14%)	0 (0.00%)	12 (0.34%)	20 (1.56%)	104 (0.40%)
Reference	403 (0.49%)	1390 (1.90%)	712 (0.95%)	11 (0.11%)	407 (1.08%)	200 (0.85%)	167 (0.60%)	41 (0.31%)	27 (0.22%)	185 (1.47%)	28 (0.24%)	9 (0.14%)	8267 (67.99%)	83 (0.79%)	76 (0.87%)	6 (0.08%)	60 (1.23%)	20 (0.41%)	4 (0.06%)	22 (0.62%)	7 (0.55%)	253 (0.97%)
Web-based Email	617 (0.75%)	570 (0.78%)	505 (0.67%)	107 (1.07%)	333 (0.88%)	234 (1.00%)	30 (0.11%)	30 (0.23%)	130 (1.08%)	254 (2.02%)	62 (0.54%)	8 (0.13%)	65 (0.53%)	7192 (68.70%)	40 (0.46%)	22 (0.29%)	16 (0.33%)	47 (0.95%)	16 (0.25%)	82 (2.30%)	10 (0.78%)	234 (0.90%)
News and Media	420 (0.51%)	573 (0.78%)	281 (0.37%)	12 (0.12%)	92 (0.24%)	102 (0.44%)	44 (0.16%)	54 (0.41%)	15 (0.12%)	130 (1.03%)	3 (0.03%)	1 (0.02%)	34 (0.28%)	42 (0.40%)	5538 (63.42%)	6 (0.08%)	25 (0.51%)	16 (0.32%)	16 (0.25%)	37 (1.04%)	0 (0.00%)	153 (0.59%)
Pornography	133 (0.16%)	236 (0.32%)	135 (0.18%)	1 (0.01%)	43 (0.11%)	76 (0.32%)	74 (0.27%)	69 (0.53%)	18 (0.15%)	10 (0.08%)	19 (0.16%)	0 (0.00%)	4 (0.03%)	25 (0.24%)	13 (0.15%)	6816 (89.21%)	31 (0.63%)	12 (0.24%)	7 (0.11%)	24 (0.67%)	27 (2.10%)	180 (0.69%)
Newsgroups and Message Boards	154 (0.19%)	378 (0.52%)	277 (0.37%)	17 (0.17%)	84 (0.22%)	80 (0.34%)	51 (0.18%)	38 (0.29%)	69 (0.57%)	63 (0.50%)	39 (0.34%)	0 (0.00%)	28 (0.23%)	48 (0.46%)	12 (0.14%)	23 (0.30%)	2447 (50.10%)	4 (0.08%)	14 (0.22%)	8 (0.22%)	0 (0.00%)	115 (0.44%)
Instant Messaging	259 (0.32%)	148 (0.20%)	350 (0.47%)	54 (0.54%)	106 (0.28%)	38 (0.16%)	51 (0.18%)	132 (1.01%)	50 (0.42%)	73 (0.58%)	8 (0.07%)	0 (0.00%)	31 (0.25%)	50 (0.48%)	15 (0.17%)	8 (0.10%)	8 (0.16%)	3441 (69.71%)	0 (0.00%)	20 (0.56%)	0 (0.00%)	58 (0.22%)
Auction	138 (0.17%)	448 (0.61%)	130 (0.17%)	0 (0.00%)	73 (0.19%)	8 (0.03%)	5 (0.02%)	5 (0.04%)	0 (0.00%)	33 (0.26%)	257 (2.22%)	4 (0.06%)	6 (0.05%)	11 (0.11%)	15 (0.17%)	0 (0.00%)	38 (0.78%)	0 (0.00%)	5386 (83.63%)	0 (0.00%)	0 (0.00%)	59 (0.23%)
Web Browser Application	158 (0.19%)	168 (0.23%)	254 (0.34%)	23 (0.23%)	80 (0.21%)	48 (0.21%)	18 (0.07%)	10 (0.08%)	35 (0.29%)	44 (0.35%)	10 (0.09%)	14 (0.23%)	1 (0.01%)	61 (0.58%)	9 (0.10%)	5 (0.07%)	16 (0.33%)	10 (0.20%)	0 (0.00%)	1820 (50.95%)	2 (0.16%)	41 (0.16%)
Not Rated	27 (0.03%)	75 (0.10%)	47 (0.06%)	1 (0.01%)	38 (0.10%)	38 (0.16%)	5 (0.02%)	3 (0.02%)	0 (0.00%)	12 (0.10%)	0 (0.00%)	8 (0.13%)	11 (0.09%)	12 (0.11%)	4 (0.05%)	0 (0.00%)	3 (0.06%)	0 (0.00%)	4 (0.06%)	9 (0.25%)	654 (50.93%)	18 (0.07%)
Others	889 (1.09%)	2348 (3.22%)	919 (1.22%)	143 (1.43%)	563 (1.49%)	469 (2.00%)	253 (0.91%)	218 (1.67%)	102 (0.85%)	636 (5.05%)	272 (2.35%)	71 (1.14%)	201 (1.65%)	222 (2.12%)	234 (2.68%)	111 (1.45%)	164 (3.36%)	75 (1.52%)	46 (0.71%)	50 (1.40%)	63 (4.91%)	15763 (60.61%)

Table 5.15: Model CS_CS Hit map on task of the 96% of event counts of CS interval class prediction

True Label (Total)	0S(Concurrent) (55456)	0S~0.5S (0)	0.5S~1S (0)	1S~2S (61132)	2S~3S (43012)	3S~4S (32936)	4S~5S (26752)	5S~6S (21188)	6S~7S (17104)	7S~8S (14536)	8S~9S (12400)	9S~10S (10844)	10S~20S (60664)	20S~30S (28048)	30S~40S (17056)	40S~50S (11736)	50S~1M (9124)	1M~2M (29628)	2M~3M (13968)	3M~4M (10264)	4M~5M (8532)	5M~6M (5668)	6M~7M (3472)	7M~8M (2848)	8M~9M (2244)	9M~10M (1904)	Others (24704)
0S(Concurrent)	26379 (47.57%)	0 (0.00%)	0 (0.00%)	9709 (15.88%)	4954 (11.52%)	3305 (10.03%)	2362 (8.83%)	1946 (9.18%)	1572 (9.19%)	1371 (9.43%)	1141 (9.20%)	947 (8.73%)	5348 (8.82%)	2274 (8.11%)	1088 (6.38%)	776 (6.61%)	598 (6.55%)	1758 (5.93%)	793 (5.68%)	585 (5.70%)	489 (5.73%)	272 (4.80%)	221 (6.37%)	205 (7.20%)	221 (9.85%)	149 (7.83%)	1845 (7.47%)
0S~0.5S	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
0.5S~1S	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
1S~2S	27961 (50.42%)	0 (0.00%)	0 (0.00%)	38916 (63.66%)	16659 (38.73%)	8474 (25.73%)	5378 (20.10%)	3961 (18.69%)	2992 (17.49%)	2306 (15.86%)	1954 (15.76%)	1560 (14.39%)	8136 (13.41%)	3574 (12.74%)	1965 (11.52%)	1377 (11.73%)	1191 (13.05%)	3407 (11.50%)	1390 (9.95%)	771 (7.51%)	552 (6.47%)	331 (5.84%)	340 (9.79%)	265 (9.30%)	169 (7.53%)	172 (9.03%)	2114 (8.56%)
2S~3S	14752 (26.60%)	0 (0.00%)	0 (0.00%)	8962 (14.66%)	7054 (16.40%)	4569 (13.87%)	3276 (12.25%)	2288 (10.80%)	1645 (9.62%)	1229 (8.45%)	996 (8.03%)	874 (8.06%)	4177 (6.89%)	1608 (5.73%)	792 (4.64%)	582 (4.96%)	441 (4.83%)	1442 (4.87%)	602 (4.31%)	462 (4.50%)	298 (3.49%)	207 (3.65%)	137 (3.95%)	124 (4.35%)	109 (4.86%)	77 (4.04%)	1189 (4.81%)
3S~4S	3016 (5.44%)	0 (0.00%)	0 (0.00%)	1151 (1.88%)	1889 (4.39%)	1989 (6.04%)	1562 (5.84%)	1103 (5.21%)	825 (4.82%)	550 (3.78%)	479 (3.86%)	401 (3.70%)	1585 (2.61%)	492 (1.75%)	299 (1.75%)	162 (1.38%)	118 (1.29%)	401 (1.35%)	174 (1.25%)	118 (1.15%)	73 (0.86%)	56 (0.99%)	35 (1.01%)	34 (1.19%)	45 (2.01%)	24 (1.26%)	408 (1.65%)
4S~5S	819 (1.48%)	0 (0.00%)	0 (0.00%)	258 (0.42%)	353 (0.82%)	491 (1.49%)	579 (2.16%)	392 (1.85%)	265 (1.55%)	220 (1.51%)	145 (1.17%)	156 (1.44%)	403 (0.66%)	73 (0.26%)	52 (0.30%)	29 (0.25%)	22 (0.24%)	52 (0.18%)	27 (0.19%)	28 (0.27%)	9 (0.11%)	11 (0.19%)	6 (0.17%)	7 (0.25%)	5 (0.22%)	0 (0.00%)	80 (0.32%)
5S~6S	114 (0.21%)	0 (0.00%)	0 (0.00%)	29 (0.05%)	23 (0.05%)	34 (0.10%)	46 (0.17%)	20 (0.09%)	12 (0.07%)	15 (0.10%)	11 (0.09%)	8 (0.07%)	37 (0.06%)	8 (0.03%)	2 (0.01%)	2 (0.02%)	2 (0.02%)	4 (0.01%)	1 (0.01%)	0 (0.00%)	0 (0.00%)	2 (0.04%)	0 (0.00%)	1 (0.04%)	0 (0.00%)	1 (0.05%)	20 (0.08%)
6S~7S	16 (0.03%)	0 (0.00%)	0 (0.00%)	3 (0.00%)	1 (0.00%)	1 (0.00%)	2 (0.01%)	2 (0.01%)	0 (0.00%)	2 (0.01%)	0 (0.00%)	0 (0.00%)	7 (0.01%)	1 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
7S~8S	11 (0.02%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.01%)	1 (0.00%)	1 (0.01%)	0 (0.00%)	2 (0.02%)	0 (0.00%)	1 (0.00%)	0 (0.00%)	1 (0.01%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.01%)
8S~9S	1 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.01%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
9S~10S	2 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.01%)	0 (0.00%)	2 (0.01%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
10S~20S	24539 (44.25%)	0 (0.00%)	0 (0.00%)	13330 (21.81%)	15656 (36.40%)	15010 (45.57%)	13526 (50.56%)	11339 (53.52%)	9439 (55.19%)	8508 (58.53%)	7334 (59.15%)	6522 (60.14%)	37345 (61.56%)	17508 (62.42%)	10551 (61.86%)	7027 (59.88%)	5074 (55.61%)	16348 (55.18%)	7272 (52.06%)	4692 (45.71%)	3293 (38.60%)	2360 (41.64%)	1538 (44.30%)	1247 (43.79%)	1029 (45.86%)	882 (46.32%)	10465 (42.36%)
20S~30S	1742 (3.14%)	0 (0.00%)	0 (0.00%)	148 (0.24%)	126 (0.29%)	101 (0.31%)	85 (0.32%)	54 (0.25%)	76 (0.44%)	74 (0.51%)	30 (0.24%)	47 (0.43%)	268 (0.44%)	173 (0.62%)	112 (0.66%)	48 (0.41%)	53 (0.58%)	153 (0.52%)	60 (0.43%)	34 (0.33%)	19 (0.22%)	26 (0.46%)	20 (0.58%)	1 (0.04%)	12 (0.53%)	7 (0.37%)	87 (0.35%)
30S~40S	10 (0.02%)	0 (0.00%)	0 (0.00%)	1 (0.00%)	2 (0.00%)	1 (0.00%)	1 (0.00%)	2 (0.01%)	1 (0.01%)	0 (0.00%)	0 (0.00%)	1 (0.01%)	2 (0.00%)	3 (0.01%)	4 (0.02%)	2 (0.02%)	1 (0.01%)	6 (0.02%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.01%)
40S~50S	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
50S~1M	3 (0.01%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.01%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.01%)	0 (0.00%)	1 (0.01%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
1M~2M	4757 (8.58%)	0 (0.00%)	0 (0.00%)	1440 (2.36%)	1347 (3.13%)	1053 (3.20%)	968 (3.62%)	753 (3.55%)	683 (3.99%)	580 (3.99%)	545 (4.40%)	477 (4.40%)	3401 (5.61%)	2133 (7.60%)	1689 (9.90%)	1310 (11.16%)	1177 (12.90%)	4404 (14.86%)	2068 (14.81%)	1129 (11.00%)	883 (10.35%)	523 (9.23%)	351 (10.11%)	276 (9.69%)	179 (7.98%)	186 (9.77%)	1881 (7.61%)
2M~3M	110 (0.20%)	0 (0.00%)	0 (0.00%)	16 (0.03%)	17 (0.04%)	25 (0.08%)	9 (0.03%)	12 (0.06%)	10 (0.06%)	8 (0.06%)	5 (0.04%)	4 (0.04%)	34 (0.06%)	23 (0.08%)	13 (0.08%)	14 (0.12%)	10 (0.11%)	53 (0.18%)	135 (0.97%)	102 (0.99%)	44 (0.52%)	21 (0.37%)	3 (0.09%)	3 (0.11%)	6 (0.27%)	2 (0.11%)	24 (0.10%)
3M~4M	273 (0.49%)	0 (0.00%)	0 (0.00%)	32 (0.05%)	49 (0.11%)	73 (0.22%)	56 (0.21%)	46 (0.22%)	34 (0.20%)	37 (0.25%)	27 (0.22%)	32 (0.30%)	202 (0.33%)	92 (0.33%)	69 (0.40%)	41 (0.35%)	64 (0.70%)	220 (0.74%)	324 (2.32%)	609 (5.93%)	506 (5.93%)	194 (3.42%)	58 (1.67%)	32 (1.12%)	19 (0.85%)	17 (0.89%)	155 (0.63%)
4M~5M	310 (0.56%)	0 (0.00%)	0 (0.00%)	52 (0.09%)	53 (0.12%)	65 (0.20%)	80 (0.30%)	43 (0.20%)	44 (0.26%)	33 (0.23%)	34 (0.27%)	33 (0.30%)	220 (0.36%)	120 (0.43%)	90 (0.53%)	100 (0.85%)	67 (0.73%)	329 (1.11%)	284 (2.03%)	815 (7.94%)	1414 (16.57%)	812 (14.33%)	211 (6.08%)	102 (3.58%)	69 (3.07%)	49 (2.57%)	312 (1.26%)
5M~6M	48 (0.09%)	0 (0.00%)	0 (0.00%)	4 (0.01%)	0 (0.00%)	0 (0.00%)	3 (0.01%)	2 (0.01%)	6 (0.04%)	0 (0.00%)	2 (0.02%)	0 (0.00%)	9 (0.01%)	0 (0.00%)	1 (0.01%)	1 (0.01%)	1 (0.01%)	13 (0.04%)	15 (0.11%)	22 (0.21%)	30 (0.35%)	30 (0.53%)	14 (0.40%)	19 (0.67%)	6 (0.27%)	1 (0.05%)	21 (0.09%)
6M~7M	2 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
7M~8M	1 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.04%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
8M~9M	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
9M~10M	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Others	6046 (10.90%)	0 (0.00%)	0 (0.00%)	1959 (3.20%)	1893 (4.40%)	1829 (5.55%)	1695 (6.34%)	1432 (6.76%)	1321 (7.72%)	1087 (7.48%)	943 (7.60%)	857 (7.90%)	5251 (8.66%)	2539 (9.05%)	1544 (9.05%)	1073 (9.14%)	872 (9.56%)	3134 (10.58%)	1698 (12.16%)	1629 (15.87%)	1510 (17.70%)	1239 (21.86%)	818 (23.56%)	779 (27.35%)	591 (26.34%)	565 (29.67%)	8195 (33.17%)

Table 5.16: Model ECSCS_CS Hit map on task of the 96% of event counts of CS interval class prediction

True Label (Total)	0S(Concurrent) (55384)	0S~0.5S (0)	0.5S~1S (0)	1S~2S (60988)	2S~3S (42916)	3S~4S (32900)	4S~5S (26724)	5S~6S (21176)	6S~7S (17080)	7S~8S (14520)	8S~9S (12392)	9S~10S (10832)	10S~20S (60508)	20S~30S (27964)	30S~40S (17036)	40S~50S (11728)	50S~1M (9120)	1M~2M (29580)	2M~3M (13960)	3M~4M (10232)	4M~5M (8524)	5M~6M (5640)	6M~7M (3464)	7M~8M (2848)	8M~9M (2244)	9M~10M (1904)	Others (24688)
0S(Concurrent)	26610 (48.05%)	0 (0.00%)	0 (0.00%)	9722 (15.94%)	4905 (11.43%)	3307 (10.05%)	2482 (9.29%)	1988 (9.39%)	1581 (9.26%)	1368 (9.42%)	1139 (9.19%)	959 (8.85%)	5465 (9.03%)	2339 (8.36%)	1083 (6.36%)	788 (6.72%)	615 (6.74%)	1754 (5.93%)	803 (5.75%)	588 (5.75%)	435 (5.10%)	231 (4.10%)	232 (6.70%)	174 (6.11%)	222 (9.89%)	144 (7.56%)	1827 (7.40%)
0S~0.5S	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
0.5S~1S	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
1S~2S	26761 (48.32%)	0 (0.00%)	0 (0.00%)	37923 (62.18%)	15615 (36.39%)	7934 (24.12%)	4923 (18.42%)	3538 (16.71%)	2731 (15.99%)	2067 (14.24%)	1754 (14.15%)	1358 (12.54%)	7112 (11.75%)	3174 (11.35%)	1624 (9.53%)	1161 (9.90%)	1015 (11.13%)	2816 (9.52%)	1101 (7.89%)	677 (6.62%)	457 (5.36%)	292 (5.18%)	248 (7.16%)	234 (8.22%)	128 (5.70%)	147 (7.72%)	1755 (7.11%)
2S~3S	15756 (28.45%)	0 (0.00%)	0 (0.00%)	9657 (15.83%)	7867 (18.33%)	5161 (15.69%)	3601 (13.47%)	2625 (12.40%)	1828 (10.70%)	1354 (9.33%)	1209 (9.76%)	960 (8.86%)	4685 (7.74%)	1743 (6.23%)	914 (5.37%)	637 (5.43%)	504 (5.53%)	1626 (5.50%)	737 (5.28%)	492 (4.81%)	348 (4.08%)	230 (4.08%)	173 (4.99%)	144 (5.06%)	113 (5.04%)	97 (5.09%)	1328 (5.38%)
3S~4S	1784 (3.22%)	0 (0.00%)	0 (0.00%)	710 (1.16%)	1347 (3.14%)	1444 (4.39%)	1192 (4.46%)	675 (3.19%)	508 (2.97%)	348 (2.40%)	280 (2.26%)	239 (2.21%)	773 (1.28%)	221 (0.79%)	154 (0.90%)	66 (0.56%)	54 (0.59%)	189 (0.64%)	81 (0.58%)	75 (0.73%)	29 (0.34%)	23 (0.41%)	12 (0.35%)	12 (0.42%)	10 (0.45%)	20 (1.05%)	195 (0.79%)
4S~5S	477 (0.86%)	0 (0.00%)	0 (0.00%)	174 (0.29%)	233 (0.54%)	407 (1.24%)	473 (1.77%)	315 (1.49%)	208 (1.22%)	188 (1.29%)	126 (1.02%)	100 (0.92%)	276 (0.46%)	45 (0.16%)	36 (0.21%)	18 (0.15%)	9 (0.10%)	28 (0.09%)	8 (0.06%)	8 (0.08%)	12 (0.14%)	2 (0.04%)	1 (0.03%)	4 (0.14%)	3 (0.13%)	3 (0.16%)	53 (0.21%)
5S~6S	81 (0.15%)	0 (0.00%)	0 (0.00%)	30 (0.05%)	25 (0.06%)	49 (0.15%)	53 (0.20%)	50 (0.24%)	24 (0.14%)	21 (0.14%)	20 (0.16%)	20 (0.18%)	39 (0.06%)	1 (0.00%)	2 (0.01%)	1 (0.01%)	3 (0.03%)	3 (0.01%)	1 (0.01%)	0 (0.00%)	0 (0.00%)	1 (0.02%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	6 (0.02%)
6S~7S	33 (0.06%)	0 (0.00%)	0 (0.00%)	8 (0.01%)	9 (0.02%)	8 (0.02%)	7 (0.03%)	11 (0.05%)	6 (0.04%)	4 (0.03%)	6 (0.05%)	3 (0.03%)	7 (0.01%)	3 (0.01%)	1 (0.01%)	0 (0.00%)	0 (0.00%)	3 (0.01%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.02%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
7S~8S	9 (0.02%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.00%)	1 (0.00%)	4 (0.01%)	6 (0.03%)	0 (0.00%)	2 (0.01%)	3 (0.02%)	0 (0.00%)	6 (0.01%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.01%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.00%)
8S~9S	11 (0.02%)	0 (0.00%)	0 (0.00%)	4 (0.01%)	0 (0.00%)	0 (0.00%)	4 (0.01%)	3 (0.01%)	1 (0.01%)	2 (0.01%)	0 (0.00%)	0 (0.00%)	3 (0.00%)	5 (0.02%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.01%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.01%)
9S~10S	5 (0.01%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.00%)	1 (0.00%)	1 (0.00%)	2 (0.01%)	1 (0.01%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (0.00%)	1 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
10S~20S	25635 (46.29%)	0 (0.00%)	0 (0.00%)	13925 (22.83%)	16509 (38.47%)	15593 (47.40%)	14012 (52.43%)	11889 (56.14%)	9913 (58.04%)	8837 (60.86%)	7533 (60.79%)	6785 (62.64%)	38603 (63.80%)	17842 (63.80%)	10954 (64.30%)	7186 (61.27%)	5284 (57.94%)	16643 (56.26%)	7342 (52.59%)	4703 (45.96%)	3339 (39.17%)	2339 (41.47%)	1580 (45.61%)	1286 (45.15%)	1100 (49.02%)	871 (45.75%)	10636 (43.08%)
20S~30S	2275 (4.11%)	0 (0.00%)	0 (0.00%)	221 (0.36%)	200 (0.47%)	163 (0.50%)	124 (0.46%)	100 (0.47%)	114 (0.67%)	92 (0.63%)	75 (0.61%)	62 (0.57%)	450 (0.74%)	274 (0.98%)	157 (0.92%)	75 (0.64%)	77 (0.84%)	261 (0.88%)	94 (0.67%)	53 (0.52%)	47 (0.55%)	47 (0.83%)	28 (0.81%)	11 (0.39%)	16 (0.71%)	5 (0.26%)	162 (0.66%)
30S~40S	44 (0.08%)	0 (0.00%)	0 (0.00%)	11 (0.02%)	5 (0.01%)	8 (0.02%)	4 (0.01%)	10 (0.05%)	1 (0.01%)	5 (0.03%)	3 (0.02%)	4 (0.04%)	31 (0.05%)	11 (0.04%)	5 (0.03%)	5 (0.04%)	10 (0.11%)	17 (0.06%)	7 (0.05%)	1 (0.01%)	2 (0.02%)	0 (0.00%)	0 (0.00%)	1 (0.04%)	1 (0.04%)	1 (0.05%)	12 (0.05%)
40S~50S	5 (0.01%)	0 (0.00%)	0 (0.00%)	1 (0.00%)	1 (0.00%)	0 (0.00%)	1 (0.00%)	1 (0.00%)	1 (0.01%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.00%)	3 (0.01%)	0 (0.00%)	1 (0.01%)	2 (0.02%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.01%)
50S~1M	1 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.01%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
1M~2M	4811 (8.69%)	0 (0.00%)	0 (0.00%)	1514 (2.48%)	1379 (3.21%)	1106 (3.36%)	958 (3.58%)	741 (3.50%)	709 (4.15%)	638 (4.39%)	551 (4.45%)	550 (5.08%)	3491 (5.77%)	2332 (8.34%)	1688 (9.91%)	1422 (12.12%)	1208 (13.25%)	4599 (15.55%)	2174 (15.57%)	1238 (12.10%)	977 (11.46%)	572 (10.14%)	395 (11.40%)	339 (11.90%)	178 (7.93%)	188 (9.87%)	2086 (8.45%)
2M~3M	118 (0.21%)	0 (0.00%)	0 (0.00%)	12 (0.02%)	14 (0.03%)	14 (0.04%)	10 (0.04%)	12 (0.06%)	5 (0.03%)	6 (0.04%)	3 (0.02%)	5 (0.05%)	28 (0.05%)	19 (0.07%)	22 (0.13%)	18 (0.15%)	12 (0.13%)	89 (0.30%)	146 (1.05%)	109 (1.07%)	33 (0.39%)	19 (0.34%)	4 (0.12%)	3 (0.11%)	6 (0.27%)	6 (0.32%)	37 (0.15%)
3M~4M	262 (0.47%)	0 (0.00%)	0 (0.00%)	28 (0.05%)	25 (0.06%)	46 (0.14%)	19 (0.07%)	26 (0.12%)	33 (0.19%)	31 (0.21%)	26 (0.21%)	19 (0.18%)	147 (0.24%)	71 (0.25%)	71 (0.42%)	39 (0.33%)	54 (0.59%)	240 (0.81%)	316 (2.26%)	663 (6.48%)	528 (6.19%)	195 (3.46%)	58 (1.67%)	22 (0.77%)	14 (0.62%)	14 (0.74%)	155 (0.63%)
4M~5M	301 (0.54%)	0 (0.00%)	0 (0.00%)	29 (0.05%)	50 (0.12%)	45 (0.14%)	56 (0.21%)	37 (0.17%)	34 (0.20%)	30 (0.21%)	31 (0.25%)	32 (0.30%)	196 (0.32%)	106 (0.38%)	114 (0.67%)	84 (0.72%)	69 (0.76%)	343 (1.16%)	290 (2.08%)	764 (7.47%)	1360 (15.95%)	806 (14.29%)	203 (5.86%)	99 (3.48%)	73 (3.25%)	44 (2.31%)	330 (1.34%)
5M~6M	75 (0.14%)	0 (0.00%)	0 (0.00%)	2 (0.00%)	6 (0.01%)	2 (0.01%)	1 (0.00%)	0 (0.00%)	1 (0.01%)	1 (0.01%)	1 (0.01%)	2 (0.02%)	20 (0.03%)	5 (0.02%)	4 (0.02%)	5 (0.04%)	5 (0.05%)	17 (0.06%)	21 (0.15%)	59 (0.58%)	100 (1.17%)	80 (1.42%)	41 (1.18%)	34 (1.19%)	13 (0.58%)	10 (0.53%)	37 (0.15%)
6M~7M	1 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
7M~8M	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
8M~9M	1 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.01%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
9M~10M	1 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.00%)
Others	5711 (10.31%)	0 (0.00%)	0 (0.00%)	1857 (3.04%)	1762 (4.11%)	1694 (5.15%)	1679 (6.28%)	1355 (6.40%)	1205 (7.06%)	1005 (6.92%)	880 (7.10%)	810 (7.48%)	4936 (8.16%)	2345 (8.39%)	1423 (8.35%)	1029 (8.77%)	767 (8.41%)	3045 (10.29%)	1711 (12.26%)	1534 (14.99%)	1444 (16.94%)	1218 (21.60%)	769 (22.20%)	733 (25.74%)	583 (25.98%)	582 (30.57%)	8159 (33.05%)

Chapter 6

Conclusion & Discussion

We found that, by combining both ECS and CS as the input features of a deep learning model, we can predict the intention and online behavior of users more accurately. However, it is costly to collect both ECS and CS. We need to overcome at least three difficulties, privacy in particular. Even the services can make users healthier or feel better, it is still hard to make people trust plugins or services with ECS and CS tracking user preferences and their private website footprints. According to ECS, researchers can understand user behavior more easily by merging ECS into CS, companies can provide personalized AD generated by tracking intention of users, browser maker can provide digital health from tracking the usage time of users, or personalized recommendation website playlist service similar to the YouTube playlist. While you start browsing, you can get the recommended website on a personalized list from the user interface. We think this is the earliest beginning on researching Extended clickstream. In the future, we can look at how users end the browsing period, especially on video casting sites. If we figure this out, we can know exactly how long the user stays in the tag/window.

Bibliography

[1] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC '09, Chicago, Illinois, USA: ACM, 2009, pp. 49–62, ISBN: 978-1-60558-771-4. DOI: 10.1145/1644893.1644900. [Online]. Available: http://doi.acm.org/10.1145/1644893.1644900.

[2] Y. Chi, T. Jiang, D. He, and R. Meng, "Towards an integrated clickstream data analysis framework for understanding web users' information behavior," *iConference 2017 Proceedings*, 2017.

[3] Z. S. Zubi and M. Raiani, "Using web logs dataset via web mining for user behavior understanding," *Int J Comput Comm*, vol. 8, pp. 103–111, 2014.

[4] Y. Wang, N. Law, E. Hemberg, and U.-M. O'Reilly, "Using detailed access trajectories for learning behavior analysis," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, ser. LAK19, Tempe, AZ, USA: ACM, 2019, pp. 290–299, ISBN: 978-1-4503-6256-6. DOI: 10.1145/3303772.3303781. [Online]. Available: http://doi.acm.org/10.1145/3303772.3303781.

[5] G. Wang, X. Zhang, S. Tang, C. Wilson, H. Zheng, and B. Y. Zhao, "Clickstream user behavior models," *ACM Trans. Web*, vol. 11, no. 4, 21:1–21:37, Jul. 2017, ISSN: 1559-1131. DOI: 10.1145/3068332. [Online]. Available: http://doi.acm.org/10.1145/3068332.

[6] K. Ma, R. Jiang, M. Dong, Y. Jia, and A. Li, "Neural network based web log analysis for web intrusion detection," in *Security, Privacy, and Anonymity in Computation, Communication, and Storage*, G. Wang, M. Atiquzzaman, Z. Yan, and K.-K. R. Choo, Eds., Cham: Springer International Publishing, 2017, pp. 194–204, ISBN: 978-3-319-72395-2.

[7] C.-Y. Lien, *Predicting Users' Demographic Information and Personality Through Browsing History*. 2018. [Online]. Available: https://github.com/ncu-dart/Lab-Publications/raw/master/Thesis2018_Cheng_You_Lien.pdf.

[8] G.-J. Bai, *Predicting Users' Browsing Tendency During Holidays by Matrix Factorization based Multi-objective Method*. 2018. [Online]. Available: https://github.com/ncu-dart/Lab-Publications/raw/master/Thesis2018_Guo_Jhen_Bai.pdf.

[9] T.-R. Chen, *Clickstream open dataset*. [Online]. Available: https://ncu-dart.github.io/#CS_open_dataset.

[10] S. Ruder, "An overview of multi-task learning in deep neural networks," *CoRR*, vol. abs/1706.05098, 2017. arXiv: 1706.05098. [Online]. Available: http://arxiv.org/abs/1706.05098.

[11] 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," *CoRR*, vol. abs/1809.03672, 2018. arXiv: 1809.03672. [Online]. Available: http://arxiv.org/abs/1809.03672.

[12] Google, *Chrome.history*. [Online]. Available: https://developer.chrome.com/extensions/history#transition_types.

[13] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *CoRR*, vol. abs/1412.3555, 2014. arXiv: 1412.3555. [Online]. Available: http://arxiv.org/abs/1412.3555.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980.

[16] S. L. Smith, P. Kindermans, and Q. V. Le, "Don't decay the learning rate, increase the batch size," *CoRR*, vol. abs/1711.00489, 2017. arXiv: 1711.00489. [Online]. Available: http://arxiv.org/abs/1711.00489.

[17] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, Mar. 1951. DOI: 10.1214/aoms/1177729694. [Online]. Available: https://doi.org/10.1214/aoms/1177729694.

Appendix A

Data Collect System

Figure A.1 shows the system workflow we built for collecting ECS and CS. Based on the workflow, we collect the integrated data with a personalized report service. All event listeners are provided in the Google Chrome Extension API. The Square shape represents the event and the rectangle shape represents the procedure we designed. The same color on every part represents the same session or the same process. This workflow shows how the data sends and processes. We use Heroku service for handling load-balance and https problem. We store data locally for security.

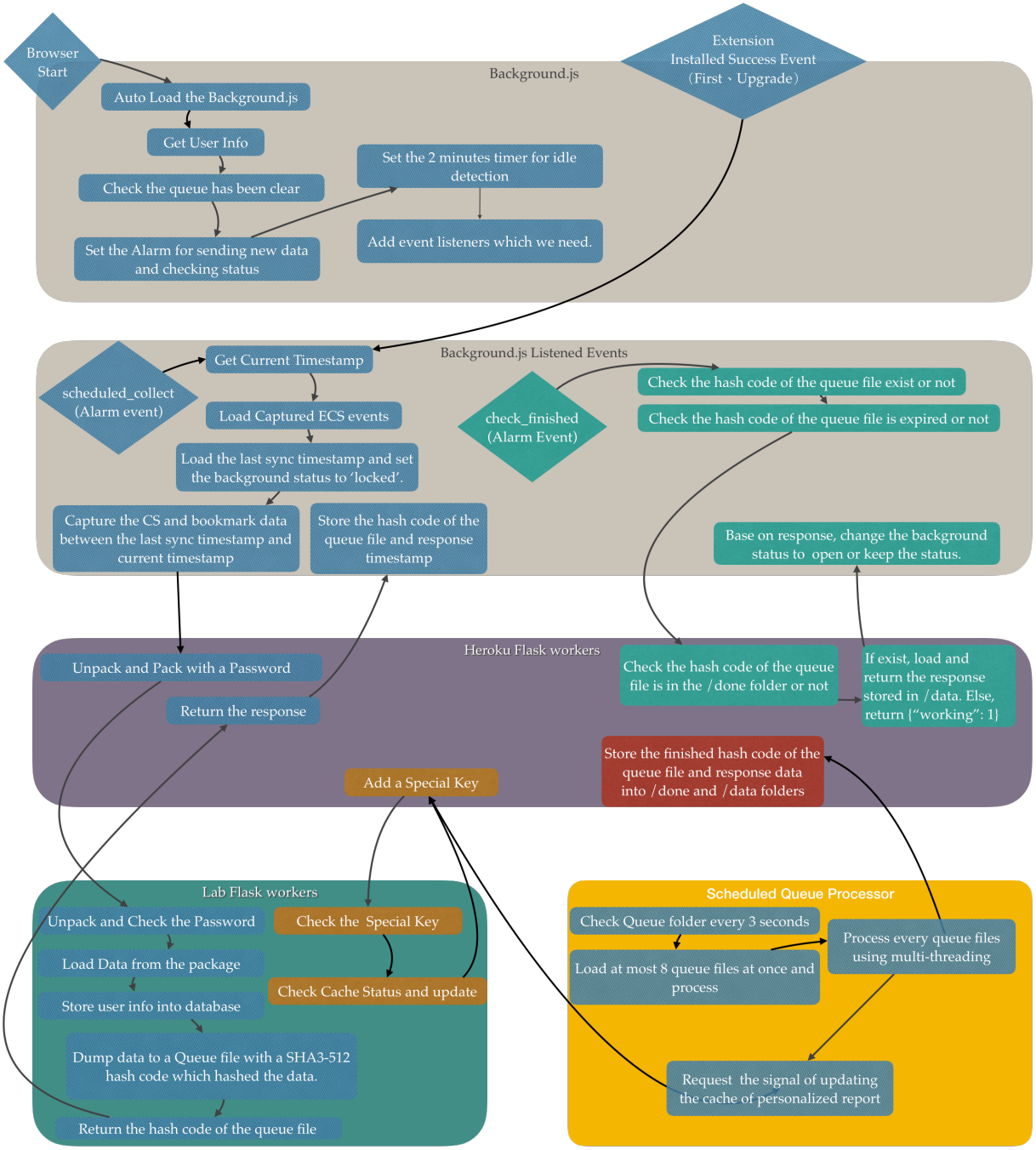


Figure A.1: The main system workflow of the ECS and CS collect system. The Background.js is the client part in the Chrome extension. Others are the servers that we maintain to collect data and serve users. This workflow is the scenario of data collection.