LUDWIG-MAXIMILIANS-UNIVERSITÄT AT MÜNCHEN
Department “Institut für Informatik”
Lehr- und Forschungseinheit Medieninformatik
Prof. Dr. Heinrich Hußmann

Masterarbeit

Understanding User Clickstream

Changkun Ou
hi@changkun.us

Bearbeitungszeitraum:	1.8.2018 bis 31.1.2019
Betreuer:	Malin Eiband and Dr. Daniel Buschek
Verantw. Hochschullehrer:	Prof. Dr. Heinrich Hußmann

Aufgabenstellung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbstständig angefertigt, alle Zitate als solche kenntlich gemacht sowie alle benutzten Quellen und Hilfsmittel angegeben habe.

München, January 2, 2019 . . . . . .

Acknowledgments

The author would like to thank Malin Eiband and Dr. Daniel Buschek for their great and constructive discussion, suggestion around thesis topic selection, model statements as well as their pacient of given guidance of the thesis.

Abstract

Early clickstream research emerges since the end of last century and has proliferated in the heart of our Internet world. Trades, public opinions, and almost every traces are precisely recorded on server side log files. The fundamental interaction between client and server stands immutably, despite the fact that mobile devices have governed our daily life. In this thesis, we proposed an action path translation model to characterize user action path behavior on the Web, as known as client-side clickstream. To justify our model, we first established a lab study and collected clickstream data of individuals with manually designed nine different web browsing task for three mainstream websites. Each website has three types of tasks, including goal-oriented, fuzzy and exploring browsing task. A collected clickstream of a subject consists of a timestamp based URL and the time duration of a single URL. By analyzing the subject traces from our lab study, we seek to archive these goals: 1) Understanding: to extract the common patterns between subjects and optimize the visiting clickstream pattern for a new user. 2) Prediction: with given client clickstream, present the future click path more than one step. 3) Classification: to separate and report whether a user is exploring on the web. To archive these goals, beyond justification analysis, we also developed a browser plugin as a possible application that predicts the future possible click under a visiting session and provides a score that indicates the probability of exploring. Furthermore, we generalize the design of our model and plugin communication protocol and discussed the possibility of formalizing them as standard Web APIs. To the best of our knowledge, this is the first study to client-side user clickstream modeling.

Contents

1 Introduction

1.1 The Origin of Clickstream Research

The word "clickstream" was first coined in 1995 [Friedman, Wayne and Weaver, Jane, 1995], a media comments article introduced a novel concept of tracing cyberlife of users over the nowadays "Internet". Informally, a "clickstream" contains a sequence of hyperlinks clicked by a website user over time. At the same year, the most popular server software Apache HTTP proxy on the Web was developed with a feature that records access log of entries [The Apache Software Foundation, 1995]. Afterwards, people realized the potential danger and value of tracing cyberspace, which a large discussion of clickstream influences, such as frequency based mining of clickstream [Brodwin, D., D. O'Connell, and M. Valdmanis., 1995], privacy concerns [Reidenberg, 1996], and database schema of session based time series data [Courtheoux, 2000].

Privacy discussion concludes collecting traces over net clearly offence the rights of users, the practice violates the openness and transparency of a service to a user. Serious criticism arise the tracing becomes a loss of democratic governance [Gindin, 1997].

Technologies is not guilty. After years of discussion, positive opinion proposes the rules [Reidenberg, 1996] and regulations [Skok, 1999] in cyberspace, means of protecting information privacy in cyberspace transactions [Kang, 1997], and approaches to resolve conflicting international data privacy [Reidenberg, 1999].

Meanwhile, bussiness man agilely responses to the concept and immediately initate commercial tracking of their customer to improving marketing affects [Novick, Bob, 1995], customer service and precise advertisment [Reagle and Cranor, 1999, Bucklin and Sismeiro, 2000], even measuring product success [Schonberg et al., 2000].

At the turn of this century, common reviews start accept the technology of clickstream, clickstream data has confirmed by industrial practice, which opens a new era in customer service [Walsh, John and Godfrey, Sue, 2000], most of website users start accept their click path data be aggregate analysed on the server side [Carr, 2000].

Clickstream data grows fast and becomes plentiful, researchers start convey the original concept of clickstream, tracking customer selections, into various applications, such as usability testing [Waterson et al., 2002a], understanding social network sentiment [Schneider et al., 2009], and developed visualizing technique to better interpret clickstream data [Waterson et al., 2002b].

Analysis, reports and characterizing of clickstream gains its popularity, Mobasher et al. [Mobasher et al., 2001] suggests personalize user based on association rule from their web usage data. Chatterjee et al. [Chatterjee, Patrali and Hoffman, Donna L and Novak, Thomas P, 2003] first proposed E-commerce websites should use clickstream to tracking customer navigation pattern instead of essential choice, associating and binding products for observing responses of a customer.

With the arise of characterizing and behavior understanding on clickstream data, more and more research proposes methods for the understanding of given server clickstream data. Padmanabhan et al. [Padmanabhan et al., 2001] proposed an algorithm to address personalization from incomplete clickstream data, which implies a security problem potential information leak from clickstream data. Moreover, affected by search engine indexing, Lourenco at al. [Lourenço and Belo, 2006] recommends an approach for the detection and containment of web crawler based on server side recorded visiting log file.

After a short review of clickstream history, almost all research putforwards their method based on server recorded clickstream data. Note that a daily user is always allowed accesses parallel pages simultaneously and even switching across multiple websites for a browsing purpose. An obvious missing aspect of those papers is the server log data is incomplete to a characterizing visited user, and the log data only appropriate for a specific website. As an observation, our research no longer surves server side clickstream, but focus and contributes to a client side collected

clickstream data for real visiting session of a user in a browser.

1.2 This Thesis

The main part of the thesis is structured in different chapters. Chapter 2 discusses the exitsting user behavior research based on clickstream data firstly. Then we summaried the reason of recent raise of neural approach in different scientific area and the state-of-the-art approaches for generic sequence learning, whose proposed in neural network research. Chapter 3 first defined the completion effeciency of a clickstream, then we formalizes our proposed sequence to sequence encoder/decoder model for client side clickstream as well as the training techniques for the proposed model. In subsequent chapter, chapter 4, we present our experiment for a lab study, and construe the design reason of context given web browsing tasks for our subjects. Afterwards, in chapter 5, based on SVM, t-SNE and our proposed model, we conducte a quantitative analysis with described data from our lab study, the evaluation shows a very promising result and the result suggests TODO:. Moreover, we visualizes the clickstream through directed graph, by combining our traning model outputs, we also performs a qualitative analysis to all graphs, the analysis gives evidences that further verified the correctness of our model. In chapter 6, as a consequence of our analysis, we developed a browser plugin for Google Chrome as a possible application to our model. The plugin can fairly predict the next possible visiting pages of a user. In addition, we generalize the design of our plugin architecture into a communication protocol between client and server, and then the possibilities to being a standard Web API to developers. To conclude, we finally summarize the findings of our thesis, the existing drawbacks of our study, as well as the possible future improvements and directions of the thesis in the final chapter.

2 Related works

Related works section

2.1 Client-side Clickstream

2.2 Sequence to Sequence Learning

3 Clickstream and Action-Path Models

In this chapter, we first formalize few concepts and metrics in clickstream data, and then describe a proposed clickstream model named *Action-Path model* based on recurrent nerual network that models a client side clickstream behavior. The action path is different than clickstream since a user may uses *back button* or *switch browser tabs* then jumps to visited web pages or parallel web pages, namely, a user performed a visit action. A server side clickstream does not containing such detailed level of user clickstream. The term *action path* is a generalized concept of clickstream, which replaces individual URLs to user actions (with backbutton and browser tab switch effects) in a browser. Figure 3.1 illustrates an simplified version of action path that compares vanilla clickstream.

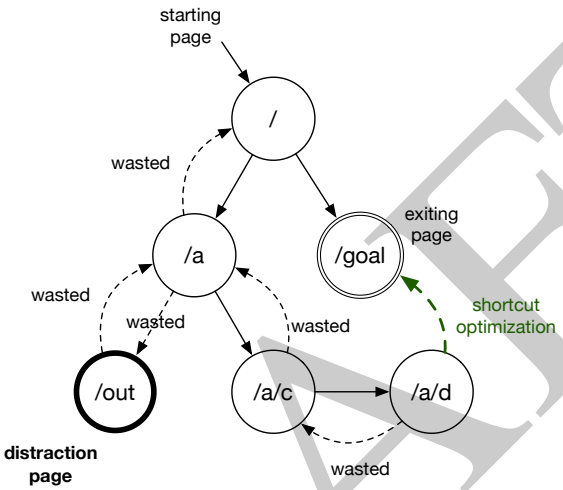

Figure 3.1: A simple action path. A user starts from the starting page, and performed a series of page click actions, ends on a exiting page. The server side records clickstream in the following order: / → /a → /out → /a/c → /a/d → /goal. However the actual user actions are: / → /a → /out → /a → /a/c → /a/d → /a/c → /a → / → /goal. The records on server side lost interaction details between users and browsers. Node that /out is a distraction page in the graph, which may located in a different website (e.g. advertisement), and dashed arrorws are wasted user actions. The /goal page may not clear in the beginning of the clickstream, one can generate a shortcut optimization navigation to the /goal page while more clickstream context be presented, i.e. an optimized user actions is / → /a → /a/c → /a/d → /goal.

For a convinience of discussion, we indiscriminate the use of term *action path* and *clickstream* in this chapter to indicate a series of user actions.

3.1 Completion Effeciency

An action path of a visiting session starts from a starting page and ends on a exiting page. Since we consider the effect of browser back button and browser tab swtich, previous page could easily be visited twice, if a user clicked the back button. Therefore, a page may directs to multiple pages. *For instance, an action path could degrade to a linked list if a user click through to different pages without using back button and switching tabs; or an action path could become a 1-to-n bipartite graph if a user use back button back to previous page after clicked a page,* as shown in Figure 3.2.

As a result, we define the *completion effeciency* based on shortest path from starting page to exiting page, and stay duration of the action path.

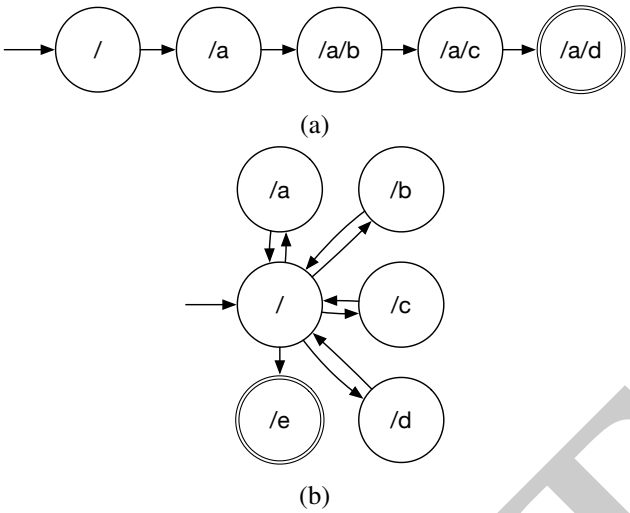

Figure 3.2: Two special case of an action path: linked list action path 3.2a, and 1-to-n bipartite graph action path 3.2b.

Let an action path is represented on directed cyclic graph, each node represents a visited page, and each edge has a weight that represents the stady duration of its tail node. Assume the total stay duration of the shortest path from starting page to existing page is d_s, and the total stay duration of the action path is D, the number of nodes in the shortest path is n_s, the total nodes in an action path is N, the *completion effeciency* E is defined as follows equation 1:

$$E = w_1 \frac{n_s}{N} + w_2 \frac{d_s}{D} \qquad w_1 + w_2 = 1 \tag{1}$$

where w_1, w_2 are hyperparameters to balancing the importance of action path and stay duration. According to the duscussion of two special case of action path, it is easy to prove the range of E is $(0,1]$. As a complement, we define *zero completion efficiency* if and only if a user cannot complete a clickstream. Therefore we have the range of E is $[0,1]$.

Remark 1. The definition of completion effeciency uses the term of shortest path, which is the problem of finding a path between the starting page and exiting page in a action path (directed cyclic graph) such taht the sum of the stay duration of its constitunent pages in minimized. The problem can be solved by Dijkstra's algorithm [Dijkstra, 1959].

Remark 2. An action path may increases with more nodes (pages) over time. The starting page of an action path is always the first page when browser was opened. However, one can always treat the current visited page is the exiting page due to we do not know when an user will exit browsing over time. Consequently, E is changing over browsing.

3.2 *url2vec* Embedding

The distributed representation of word2vec models achieve better performance in natural language processing, Mikolov et al. [Mikolov et al., 2013a] introduced continous bag-of-word (CBOW) and skip-gram model as an efficient method for learning high-quality vector representation of words, and CBOW is faster while skip-gram is slower but get better performance for infrequent words. We convey similar idea from these models and propose our *url2vec* model for client side clickstream data.

The purpose of url2vec model is to construct URL representations to better predict the surrounding URLs in a clickstream. Briefly, given a clickstream of urls $\text{URL}_1, \text{URL}_2, ..., \text{URL}_T$, the objective of url2vec is to maximize the average log softmax probability

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le i\le c, i\neq 0}\log p(\mathrm{URL}_{t+i}|\mathrm{URL}_t)$$
$$\log p(\mathrm{URL}_{t+i}|\mathrm{URL}_t)=\frac{\exp\left(v_{\mathrm{URL}_{t+i}}^{\top}v_{\mathrm{URL}_t}\right)}{\sum_{\text{all URLs}}\exp\left(v_{\mathrm{URL}_{t+i}}^{\top}v_{\mathrm{URL}_t}\right)} \tag{2}$$

where c is the size of embedding context, which is a function of starting page, v_{URL_t} is one-hot encoded representation of input URLs, and $v_{\mathrm{URL}_{t+i}}$ is the vector embedding of output representations.

Remark 1. This model described by equation 2 is actually a three layer neural network: input layer of one-hot encoded URLs, a hidden layer of feature representation and an output layer share weights to the learned embeddings of input URLs.

Remark 2. The probability in 2 is impractical due to $\nabla\log p(\mathrm{URL}_{t+i}|\mathrm{URL}_t)$ is as large as 10^5 to 10^7, two numerical optimizations based on Hofmann Tree and Negative Sampling are proposed by Mikolv et al. [Mikolov et al., 2013b].

3.3 Action-Path Model

Recurrent Neural Network (RNN) was describe by Werbos [Werbos, 1990] and Rumelhart et al. [Rumelhart et al., 1988], the original RNN generalize feedforward neural network for sequence based data.

Given a sequence of input $(i_1, i_2, ..., i_T)$, the original RNN computes a sequence of outputs $(o_1, o_2, ..., o_T)$ by interating the activation function 3:

$$o_t = W_{oh}\sigma\left(W_{hi}i_t + W_{hh}i_{t-1}\right), t = 1, 2, ..., T \tag{3}$$

where $\sigma(x) = \frac{1}{1+\exp\{-x\}}$, and W_{oh}, W_{hh}, W_{hi} are weight parameters between output, hidden and input layers.

The original RNN transfers and maps a sequence to another sequence if and only if the inputs and the outputs are aligned with equal length. Apparently, the major issue of the original RNN is the model cannot address a problem if inputs and outputs provided in different length with complicated and non-monotonic relationships.

Stutskever et al. [Sutskever et al., 2014] present a general end-to-end approach to sequence learning and estimates the conditional probability of $p(o_1, o_2, ..., o_{T'}|i_1, i_2, ..., i_T)$ where $(i_1, i_2, ..., i_T)$ is an input sequence, $(o_1, o_2, ..., o_{T'})$ is a corresponding output sequence, and T is not required to be equal with T'. Our model convey similar idea from it.

An *action path* from user i in session j consist of a sequence of *url2vec* embedded vectors $(U_1^{ij}, U_2^{ij}, ..., U_n^{ij})$ and a sequence of time duration $(d_1^{ij}, d_2^{ij}, ..., d_n^{ij})$, since each URL has a corresponding number that represents the time duration of a user spent on a given page.

3.3.1 Context Encoder

Our action path model consist a context encoder and a context decoder. Context encoder encodes URLs one by one and produces a context tensor that encodes the historical user actions, shown in Figure 3.3. In the encoder, we insert a starting mark "<SOA>" (*Start of Action*) as a sign of start feeding URLs, and a trigger mark "<COI>" (*Change of Intention*) as a sign to trigger decoder to decodes encoded context tensor.

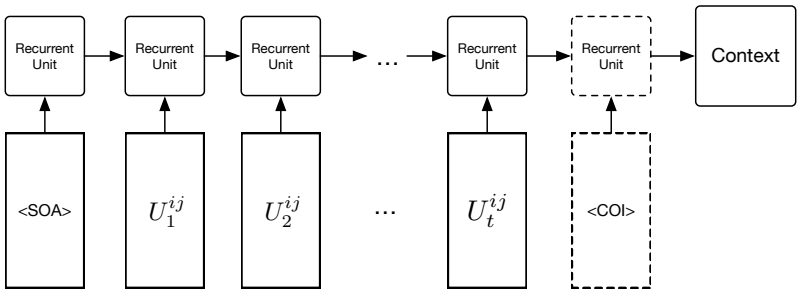

Figure 3.3: Context encoder of Action Path Model. In the encoder, a starting mark "<SOA>" is used as a sign of start feeding URLs, and a trigger mark "<COI>" as a sign to trigger decoder to decodes encoded context tensor. The trigger mark is automatically inserted after the k-th URL in the end of encoder model over time, k is increasing over time. In addition, the recurrent unit is not detailly describeed in the figure but afterwards.

3.3.2 Context Decoder

Context decoder decodes the context tensor produced by encoder. We feed a prediction mark "<SOP>" (*Start of Prediction*)as a sign to initiate the decoding of encoded context. In the end of decoder, decoder produces an ending mark "<EOA>" (*End of Action*) that terminates the decoding process.

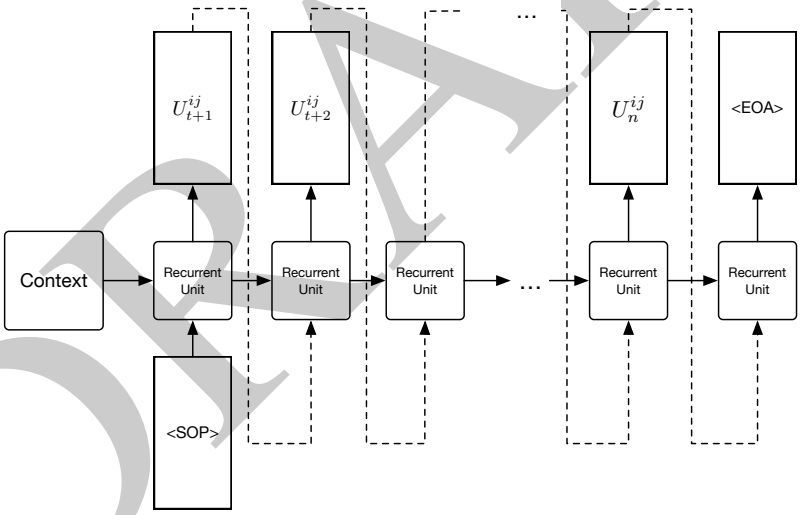

Figure 3.4: Context decoder of Action Path Model. In the decoder, a prediction mark "<SOP>" is used to initiate decoding process, and an ending mark "<EOA>" as a sign to terminate decode process. The output of decoder uses a softmax intermediate operation to magnify and normalize the probability of predicted URL embedding. In addition, the recurrent unit is not detailly describeed in the figure but afterwards.

Note that the decoder model in training phase and prediction phase is different. In the training phase, teacher forcing algorithm [Williams and Zipser, 1989] is used, it supplying observed user actions as inputs. In the testing phase, decoder uses the output from recurrent unit as input, shown through dashed lines in Figure 3.4.

3.3.3 Recurrent Unit

The recurrent unit in the Action Path model is not as standard as original Long Short-term Memory unit (LSTM) [Hochreiter and Schmidhuber, 1997] or Gated Recurrent unit (GRU) [Cho et al.,

2014].

When using LSTM as recurrent unit, we feed time duration $(d_1^{ij}, d_2^{ij}, ..., d_n^{ij})$ into input gate I_t, and others (forget gate F_t, output gate O_t, memory cell C_t and hidden state h_t) remains the same:

$$\begin{aligned} I_t &= \sigma(P^{(I)} U_t^{ij} + Q^{(I)} h_{t-1} + \frac{d_t^{ij}}{d_t^{ij}+1})) \\ F_t &= \sigma(P^{(F)} U_t^{ij} + Q^{(F)} h_{t-1} + b^{(F)}) \\ O_t &= \sigma(P^{(O)} U_t^{ij} + Q^{(O)} h_{t-1}) \\ C_t &= F^{(t)} \circ C_{t-1} + I_t \circ \tanh(P^{(C)} U_t^{ij} + Q^{(C)} h_{t-1}) \\ h_t &= O_t \circ \tanh(C_t) \end{aligned} \tag{4}$$

where $t = 1, 2, ..., n; P^{(I)}, Q^{(I)}, P^{(F)}, Q^{(F)}, P^{(O)}, Q^{(O)}$ are shared weight parameters, $b^{(F)}$ is a bias in forget gate F_t, $\circ$ represents element-wise product of two matrices.

When using GRU as recurrent unit, we feed time duration $(d_1^{ij}, d_2^{ij}, ..., d_n^{ij})$ in to update gate Z_t, and others (reset gate R_t, hidden state h_t) stay the same:

$$\begin{aligned} Z_t &= \sigma(P^{(Z)} U_t^{ij} + Q^{(Z)} h_{t-1} + \frac{d_t^{ij}}{d_t^{ij}+1}) \\ R_t &= \sigma(P^{(R)} U_t^{ij} + Q^{(R)} h_{t-1}) \\ h_t &= (1 - Z_t) \circ \tanh(P^{(H)} U_t^{ij} + Q^{(H)} h_{t-1}) + Z_t \circ h_{t-1} \end{aligned} \tag{5}$$

where $t = 1, 2, ..., n; P^{(Z)}, Q^{(Z)}, P^{(R)}, Q^{(R)}, P^{(H)}, Q^{(H)}$ are shared weight parameters, $\circ$ represents element-wise product of two matrices.

Remark 1. The units we described in this section is neither LSTM nor GRU since the input gate I_t or update gate Z_t introduces time duration d_t^{ij} as input, which completely different than introduce learnable bias in these gates. It is worth mentioning that adding bias to the gates are helpful to improve learning performance in LSTM [Jozefowicz et al., 2015], we also use the trick in our model.

Remark 2. The term $\frac{d_t^{ij}}{d_t^{ij}+1}$ is a squashing mechanism, it normalizes d_t^{ij} from $(0, \infty)$ to $(0, 1)$.

3.3.4 Ending Mark Interpretation

In context decoder, we mentioned an ending mark "<EOA>" that indicates the termination decoding process. However, the ending mark is different than other marks, since in practice, "<EOA>" is represented in different symbols of categorical clickstream, which as a label to involve classification of user actions.

Assume action paths are labeled by ont-hot encoded ending marks $\text{EOA}_1, \text{EOA}_2, ..., \text{EOA}_m$ and the last output of decoder hidden state is h_n, we have:

$$\begin{aligned} \hat{y} &= \text{argmax}(\text{softmax}(W^{(M)} h_n)) \\ \hat{y} &\in \{\text{EOA}_1, \text{EOA}_2, ..., \text{EOA}_m\} \end{aligned} \tag{6}$$

where $W^{(M)}$ is a weight parameter, and m is the number of ending mark categories.

3.4 Action Path Optimization

In traditional classification models, the arguments of the maxima (argmax) is used to select labels with highest probability, scilicet, argmax selects predicted pages with highest probability of user action from decoder outputs. However, this method is under the condition of all outputs are independent in probability, which is not suitable to our senario.

In previous sections, our model feeds an input clickstream $(U_1^{ij}, U_2^{ij}, ..., U_t^{ij})$, and produce an output $(o_1, o_2, ..., o_m)$ that expect close to actual clickstream $(U_{t+1}^{ij}, U_{t+2}^{ij}, ..., U_n^{ij})$. Then the probability of expected clickstream is a conditional probability under the input clickstream, i.e. we need to solve an optimization problem

$$\begin{aligned} &\operatorname*{argmax}_o p(o_1, o_2, ..., o_m | U_1^{ij}, U_2^{ij}, ..., U_t^{ij}) \\ =&\operatorname*{argmax}_o \prod_{k=1}^{m} p(o_k | U_1^{ij}, ..., U_t^{ij}, o_1, ..., o_{k-1}) \\ =&\operatorname*{argmax}_o \sum_{k=1}^{m} \log p(o_k | U_1^{ij}, ..., U_t^{ij}, o_1, ..., o_{k-1}) \end{aligned} \tag{7}$$

Heuristic approach can solve the optimization problem efficiently, namely beam search [Graves, 2012]. In each step of decoder output, we reserve the top-k best combinations of pages and eliminate the rest of pages from evaluation, and finally selects k best clickstreams. The pseudocode is given in Algorithm 1.

Algorithm 1: Output Clickstream Search

```
input : Decoder outputs (o_1, o_2, ..., o_m),
        Number of candidates k
output: k clickstream candidates with highest probability
begin
    Initialize empty clickstreams list
    for o ∈ (o_1, o_2, ..., o_m) do
        Initialize empty candidates list
        for clickstream ∈ clickstreams do
            for page ∈ o) do
                candidates.append([clickstream.append(page),
                  log(p(clickstream)) + log(p(page))])
            end
        end
        ordered = descending order sort candidates by score
        clickstreams = ordered[:k]
    end
end
```

Remark. The algorithm produces an heuristic output with given clickstream context. Combining with *url2vec* model, the prediction can heuristically optimize the click path of a specific user since the embeddings are trained over all possible action path. For instance, a distraction advertisement page will not appear after optimization because the embedding of advertisement page is far from a desired page if embeddings are learned correctly.

4 Experiment

The lab study took place during the last two weeks of November, from 14/11/2018 to 29/11/2018 in Frauenlobstrasse 7a, a faculty building of Ludwig-Maximillians-Universitaet Muenchen. Data was collected from the mainstream browser Google Chrome on a self provided desktop computer and a laptop.

The following of this chapter, we present the process of our lab study then construe the purpose of context given web browsing tasks for our subjects.

In lab study, we select three mainstream websites, Amazon/Medium/Dribbble that covers categories for shopping, media consuming and design brainstorming. We manually designed 9 reasonable (discussed in section 4.2) context-given browsing tasks (three for each website) in total to our subjects. Each task requires participant start from a starting page of a given website, and all tasks do not restrict participants use the given website, but also allow they access websites outside the landing page to help they complete the task. Participants start browsing after they completely understand the requirements of each task, and no interruption or question answering during the task except exceeding time limit of a task, however subjects can either acquire more time to accomplish the task or give up directly.

The study is designed as a within-subject study, thus every participant performs all tasks. To eliminate the learning effect due to the long time of using same websites, we use Latin square [Cochran and Cox, 1950] for the device and tasks participation order to our participants.

21 participants with a mean age of 23.04 (standard deviation of 3.216, min=18 and max=19) took part in the study, 10 male and 11 female, whom are recruit anonymously and randomly via a mailing list.

4.1 Environment

The lab study uses two self provided devices: a desktop computer and a mobile laptop. The reason of choose two morphology of computing device is our study requires recording a complete clickstream of during the study.

A major issue of mobile devices is the operating system installed in mobile phone does not open the permission to allow us to collect data precisely over pages. Though Android device can overpass system permission to privilge, the user behavior between iOS and Android device is still completely different with different models. Subjects shows abnormal behavior when they use a newly provided device. Therefore we stick our study environment to desktop devices, which empower us easily collects the clickstream data from plugin-supported browsers.

Although all modern browsers support plugin development, however considering the usage share of all browsers on the market, Google Chrome obtains 61.7% market shares of desktop browsers, and Apple Safari only shares 15.0% of the market [StatCounter, 2018]. Clearly, Google Chrome dominant the desktop web browser.

Therefore, we decide to use Chrome to establish our plugin of data collection. The quesionaire after our lab study indicates the browsers usage share of all subjects, as shown in Table 4.1, which further supports our decision of browser selection.

Table 4.1: Browser Usage Shares of Lab Study Subjects

	Google Chrome	**Apple Safari**	**Mozilla Firefox**	**Microsoft Edge**
Number	11	5	3	2
Percentage	52.38%	23.81%	14.29%	9.52%

4.2 Tasks Design

Before we explain the design reason of our context-given browsing task, we first present and discusses the common three types of daily browsing behavior: goal-oriented, exploring and fuzzy.

Goal-oriented Task: A user opens a browser caused by a determined purpose for bussiness work, communication, school study, literature research and etc. For instance, a college student has a goal of downloading latest lecture slide, the student then access college website and navigates to the lecture homepage. Finally, the student exit browsing after download the slides.

Exploring Task: A user opens a browser aimlessly, the person explores and consumes the content on the Web. For instance, a person who accesses an utility web application, he/she explores what functions are provided and what he/she can do while using the application.

Fuzzy Task: A user opens a browser with a fuzzy expectation, the clearness of the access purpose is less than goal-oriented task but stronger than exploring task. For instance, a researcher heared a new technique proposed in other field that may influence he/she's study, then the person opens a search engine to seek existing follow up researches.

Note that these three type of tasks sometimes implies the familiarity of a user to the website.

We designed 35 browsing tasks, after conduct a pilot study, 9 tasks are selected for three websites: Amazon.com, Medium.com and Dribbble.com because of the following reasons:

1. These three websites all have coresponding tasks to the three type of browsing behavior;
2. Each of the task can be finished around 5 to 10 minutes;
3. All these websites are mainstream websites, they do not require massive professional domain knowledge for using.

Moreover, the unselected tasks are listed in Appendix B.3.

4.2.1 Goal-oriented Task

We designed and selected an appropriate goal-oriented task for selected websites respectively, and each task is designed with three determined goal.

- **Amazon.com**: *Assume your smartphone was broken and you have 1200 euros as your budget. You want to buy an iPhone, a protection case, and a wireless charging dock. Look for these items and add them to your cart.*

  This task contains three determined purpose since a subject is required to add three items to the cart. There are few hidden consideration behind the task, which makes the task more realistic: a) There is a budget of this task, which requires subjects must consider the price of items instead of simply add the first recommended item to cart; b) the starting page is amazon.com instead of amazon.de. This decision requires subjects must also consider the currency rate between US dollars and Euros for budget. c) There are some items cannot be shipped to Germany (the study took place in Germany). As a result, subjects cannot add these items to cart and they should find other alternatives.

- **Medium.com**: *Assume you were making plans for your summer vacation. You want to visit Tokyo, Kyoto, and Osaka. You want to find out what kind of experience other people made when traveling to these three places in Japan. Your task is to find three posts for traveling tips regarding these cities. Elevate a post if it is one of your choices.*

  This task contains three determined purpose since there are three fixed traveling destination. The task also implies few considerations that increase the required interaction of the task to subjects: a) The website only offers English version, some Japanese character may appear

in an article, thus, a translation util may be used while the study; b) An article may apears numerous noun, such as toponym. Search engine may used while the study; c) the articles, those require a membership to unlock reading, cannot be elevated.

- **Dribbble.com**: *You are hired to a Cloud Computing startup company. You get an assignment to designing the logo of the company. Search for existing logos for inspiration and download three candidate logos you like the most.*

  The task also has three determined prupose since subjects are quired to download three candicate trademarks. While the participation, subjects still need take few implicit facts in to account: a) Subjects who unfamiliar with the term "Cloud Computing" need visit other explainations to figure out the vision and mission of this type of company, and subjects whom already familiar with the term still need to compares the designed made by other competitors. b) Subjects should aware some of the designs shared on the website are not suitable for trademark or icon design.

4.2.2 Exploring Task

Exploring tasks simply do not provides any deterministic objective, and all websites has a designed exploring task for subjects.

- **Amazon.com**: *Look for a product category that you are interested in and start browsing. Add three items to your cart that you would like to buy.*

  Although the task do not require any specific items to the subjects, the task remains three different purpose because participants need add three items to the cart. This designed task is aimlessly because: all tasks is not formerly informed to participants, they either do not have needs of buying items or formerly exist needs of buying a specific category but do not have a product candicate yet. Besides, the description of the task ask participants start from a product category, which avoids goal-oriented buying a specific product.

- **Medium.com**: *Visit a category you are interested in and elevate three post you like.*

  Similar reason as discussed in Amazon.com's exploring task. It is well to be remined that Medium is a media website, visiting a specific article formerly read before participation is relatively difficult since all contents showed to users are daily updated. Thus the task can be directly consider as an exploring task.

- **Dribbble.com**: *Explore dribbble and download three images you like the most while you browse.*

  Dribbble illustrates designs by using image gallery. The major difference between Dribbble and Google Image Search is dribbble is a user-centered content aggregation website, but Google Image Search is a simple content aggregation engine. As a result, there will be two different interaction in Dribbble: exploring designs based on keywords and categories, or exploring designs based on users. The latter can helps its user finding similar designing style. The task is aimlessly since the task simply describes nothing and completely let participants explore their preferences.

4.2.3 Fuzzy Task

Each of our selected websites also has an fuzzy task respectively, and there are three major goals per task.

- **Amazon.com**: *You want to buy a gift for your best friend as a birthday present. Add three items to your cart as candidate.*

The clearness of the task is stronger than exploring task but weaker than goal-oriented task, because The task restricts participants adding items for a specific purpose (birthday present) but not points any specific product.

- **Medium.com**: *Assume you got an occasion to visit China for business. You are free to travel to China for a week. You want to make a travel plan for touring China within a week. Your task is to find out what kind of experience other how people made when going to secondary cities or towns in China, then decide on three cities you want to visit (excluding Beijing, Shanghai, Guangzhou, and Shenzhen). Elevate if a post helped you make a decision.*

  The clearness of the task is stronger than exploring task, because it asks a participant to exploring a non-deterministic direction of looking for secondary cities. But the clearness of the task is weaker than goal-oriented task due to secondary cities described in Medium's user posts is unclear, participants suppose to make decision themselves. Furthermore, this ask is asking regarding traveling China around a week. Cities cannot be randomly selected because to make traveling plan requires consider geographic location of the city.

- **Dribbble.com**: *You are preparing a presentation and need one picture for each of these animals: cat, dog, and ant. Download the three pictures you like the most.*

  The task has three purpose of downloading images of animals, which restrict participant to a specific direction, thus, the clearness of the task is stronger than exploring task. However, the task describes a scenario of using these images in a presentation, and hence participants must consider continuity of design style, which makes the clearness of the task is weaker than goal-oriented task.

5 Evaluation and Discussion

In this chapter, we conduct evaluations to our collected data. The data is collected from 21 subjects, and 189 clickstream data are collected in total. Each clickstream contains action-level data with a stay duration of a specific page, for instance, we still collect an URL as a step of clickstream if a participant uses back button rollback to a previous visited page without requesting server. A clickstream also has a subjective difficulty score from questionaire (shown in Appendix B) after the completion of each task.

5.1 Subjective Task Difficulty

This section discusses the subjective task difficulty qualitatively and quantitatively. Figure 5.1 illusrates a normalized (raw scores are listed in Appendix C Table C.1) subjective difficulty score with respect to all tasks.

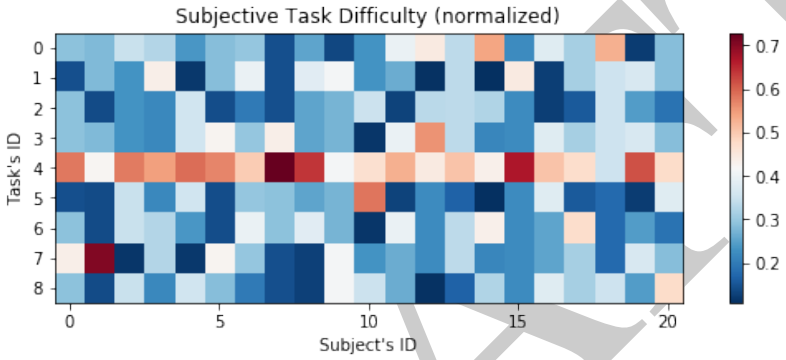

Figure 5.1: Subjective difficulty score: each column indicates an individual subject and each row indicates a browsing task. Tasks from 0 to 8 represent Amazon Goal Oriented Task, Amazon Fuzzy Task, Amazon Exploring Task; Medium Goal Oriented Task, Medium Fuzzy Task, Medium Exploring Task, Dribbble Goal Oriented Task, Dribbble Fuzzy Task and Dribbble Exploring Task respectively. From this heat map, we clearly observes Medium Fuzzy Task is the most difficulty task according to the subjects voted subjective difficulty, a significant test confirmed this observation. Further, Mann-Whitney U significant test justifies our result.

To generalize the task difficulty, the null hypothesis (H_0): the difficulty of fuzzy task is not greater than exploring task and alternative hypothesis (H_1): the difficulty of fuzzy task is greater than exploring task. We conduct non-parametric one-tailed Mann-Whitney U test [Mann and Whitney, 1947], under null hypothesis, $p = 2.54 \times 10^{-5} < 0.05$, reject H_0. Similarly, we compare difficulty score on goal oriented task and exploring task (with corresponding hypothesis, $p = 0.00534 < 0.05$), difficulty score on fuzzy task and goal oriented task (with corresponding hypothesis, $p = 0.0145 < 0.05$), all rejects H_0. Therefore we concludes the task difficulty is ordered as follows: *difficulty of fuzzy task* > *difficulty of goal oriented task* > *difficulty of exploring task*, which means exploring tasks have lower effort in clickstream, and effort of doing fuzzy task gains highest effort.

5.2 Browsing Behavior Classification

As discussed in section 4.2, we described three type of browsing behavior. In this section, we provides two type of evaluations to interpret the browsing behavior classification.

First, we evaluate the indication of general features browsing behavior, features including difficulty of task, number of actions in a clickstream as well as the total stay duration in a clickstream.

Then we evaluate our action path model by using the action-level clickstream data and stay duration of each page, which was described in section 3.3.3 and 3.3.4.

5.2.1 Interpretation of General Features

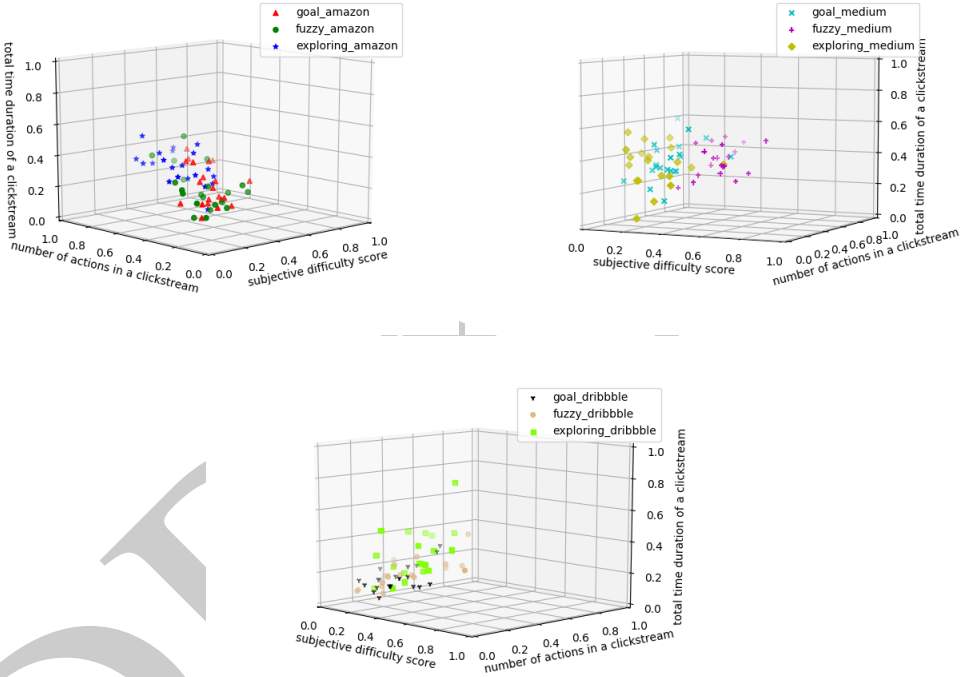

Figure 5.2: XXXXX

5.3 Intepretation of Action Path

To use full capacity of our data, this section uses the entire clickstream and its corresponding page-level stay time duration as input, three ending mark (<EOA_GOAL>, <EOA_FUZZY>, and <EOA_EXPLORE>) as classification outputs, and then implements a single GRU layer action path model to classify the three type of tasks.

Our training parameters are: The GRU latent dimension is 10, the training evaluate the classification accuracy with 80%-20% cross validation split and propagates 500 epochs with 32 batch size. In the end of training, We archieved 100% accuracy on our validation set.

The validation loss during the training is as shown in Figure 5.3.

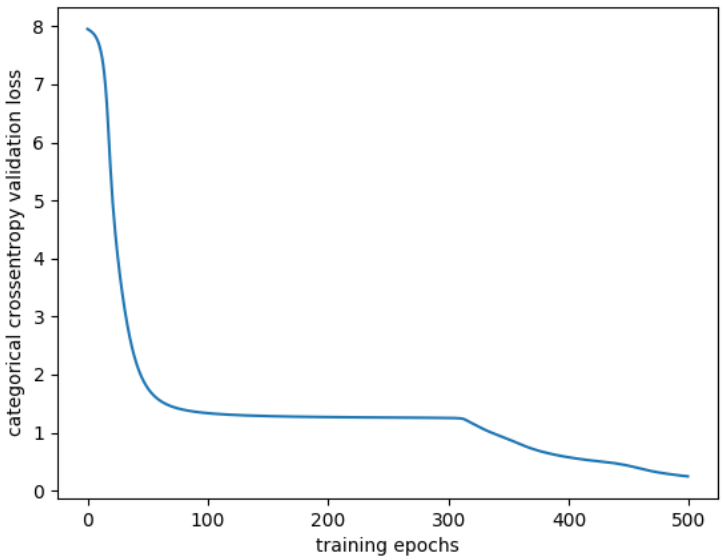

Figure 5.3: Categorical Cross-Entropy Validation loss curve while 500 epoches. The curves indicates the training process is not an overfitting since the loss is not increasing.

One can observed that the training process is not an overfit, and the validation loss is still not increase after 500 epoches, thus, single GRU layer action path model remains a large expressive performance (100% of three class classification in this dataset) when we have more data.

In addition, the action path model feeds the entire clickstream and time duration as inputs, therfore the entire clickstream contains informations regarding the number of visit actions as well as completion effeciency and etc.

5.4 Task Completion Effeciency

5.4.1 t-SNE

5.4.2 Prediction Accuracy

5.4.3 F1

5.5 Explored Model Architecture Comparasion

5.6 Action Path Visualization

5.7 Discussion

6 Applications

6.1 Client Side Browser Plugin

6.1.1 Client-side architecture

6.1.2 Server-side architecture

6.1.3 Model Evolution Automation Architecture

6.2 Standard Browser Web APIs

6.2.1 Communication Protocol

7 Conclusions

7.1 Summary

This thesis proposed an action path model which exposit and construct the

7.2 Future Works

Appendix

All resources relates to the thesis are open source, they can be found publicly in:

- Thesis homepage: https://changkun.us/thesis/;
- GitHub repostory: https://github.com/changkun/MasterThesisHCI/.

A Content of enclosed USB

1. */documents/* - TODO

B Tasks and Questionnaire in Lab Study

B.1 Phase 1: Browsing Task

This section approximately takes 80 minutes.

In this study, you are asked to accomplish a series of tasks provided in the table below. Please read the following tips carefully before you do the task [3].

1. **Please start from the given starting page.** You can then visit any other page. For instance, if you find a task too difficult, you can visit any other websites that help you accomplish the task (e.g. Google as a search engine), but you should only use the browser.
2. The tasks are designed to take **5 10 minutes**. Do not feel stressed if you spend more time because you have 80 minutes in total to **do the 9 tasks**. You will be notified if you spend more than 10 minutes on a task. You can decide to go to the next task or spend some to accomplish the unfinished task.
3. **Close the browser before you start working on the next task.**
4. **Unfortunately, questions cannot be answered while doing the tasks. Please ask them before starting a task if something is not clear.**

B.1.1 Task Group 1: Amazon.com

Task Category: Shopping

1. Assume your smartphone was broken and you have 1200 euros as your budget. You want to buy an iPhone, a protection case, and a wireless charging dock. Look for these items and add them to your cart.

   Requirement to Finish: Click "Proceed to checkout" when you finished, exit the browser when you see the "sign in" page.

[1] http://creativecommons.org/licenses/by-nc-sa/4.0/
[2] https://github.com/changkun/MasterThesisHCI/blob/master/LICENSE
[3] The order of the tasks are rearranged through Latin square, this section only illustrate one possible order of tasks

2. You want to buy a gift for your best friend as a birthday present. Add three items to your cart as candidate.

   Requirement to Finish: Click "Proceed to checkout" when you finished, exit the browser when you see the "sign in" page.

3. Look for a product category that you are interested in and start browsing. Add three items to your cart that you would like to buy.

   Requirement to Finish: Clicked "Proceed to checkout" when time is up, exit the browser when you see the "sign in" page.

How difficult was the task? (1 5, 1 means very easy, 5 means very difficult)

_____, _____, _____

B.1.2 Task Group 2: Medium.com

Task Category: Media

1. Assume you were making plans for your summer vacation. You want to visit Tokyo, Kyoto, and Osaka. You want to find out what kind of experience other people made when traveling to these three places in Japan. Your task is to find three posts for traveling tips regarding these cities. Elevate a post if it is one of your choices.

   Requirement to Finish: Write down three tips. Close the browser when you are finished.

2. Assume you got an occasion to visit China for business. You are free to travel to China for a week. You want to make a travel plan for touring China within a week. Your task is to find out what kind of experience other how people made when going to secondary cities or towns in China, then decide on three cities you want to visit (excluding Beijing, Shanghai, Guangzhou, and Shenzhen). Elevate if a post helped you make a decision.

   Requirement to Finish: Write down the names of the cities you decided. Close the browser when you are finished.

3. Visit a category you are interested in and elevate the post you like.

   Requirement to Finish: Close the browser when time is up.

How difficult was the task? (1 5, 1 means very easy, 5 means very difficult)

_____, _____, _____

B.1.3 Task Group 3: Dribbble.com

Task Category: Design

1. You are hired to a Cloud Computing startup company. You get an assignment to designing the logo of the company. Search for existing logos for inspiration and download three candidate logos you like the most.

   Requirement to Finish: Close the browser when you finished the download.

2. You are preparing a presentation and need one picture for each of these animals: cat, dog, and ant. Download the three pictures you like the most.

   Requirement to Finish: Close the browser when you finished the download.

3. Explore dribbble and download images you like the most while you browse.

   Requirement to Finish: Close the browser when you finished the download.

How difficult was the task? (1 5, 1 means very easy, 5 means very difficult)

_____, _____, _____

B.2 Phase 2: Questionnaire

This section approximately takes 10 minutes.

1. Age: _____
2. Gender: Female / Male
3. What is your study program or occupation?
4. What are the websites that you access mostly? List your top-5 (max 10, including private use).
5. What do you usually do when you access these websites? Shortly answer your case for all the websites you listed in above and name two common reasons, ordered by frequency. (For example, for YouTube, the most common reason could be “Just for fun”, the second most common reason “Looking for tutorial”. Then write as “Mostly for fun, sometimes for learning” below.)
6. Do you use bookmarks to save webpages that you have found through a search engine? If so, why?
7. Which browser do you use mainly on your PC or Mac? Chrome / Safari / IE / Microsoft Edge / Firefox / Others, the name is: _____
8. Would you like to participate in a follow-up study? The study will ask you to install a browser plugin for a week which anonymously records your browsing history. Yes / No
9. Do you have any feedback on this questionnaire?

B.3 Unselected Tasks

This section lists all designed tasks but unselected to lab study.

B.3.1 Goal-oriented Task

1. **www.github.com**: You are comparing three most popular frontend desktop frameworks: Electron / NW.js / ReactNative Desktop. Your goal is to find out the latest release download link.
2. **www.medien.ifi.lmu.de**: You are a fresh medieninformatik student major in HCI program. You wants to find out recommended first semester study plan provided by the program, then select "Human-Computer Interaction II" opened in WS18/19 and check previous "Human-Computer Interaction I" opened in SS18 and SS17.
3. **www.en.uni-muenchen.de**: You are a international student who want to apply economics program for your master study at LMU. Find the page for application requirement.
4. **www.ielts.org**: You live in Munich, you want to participate to IELTS test next year on Feburary. Looking for the entrace to register the examination. You must keep seeking and stop when you selected the first track of Feburary test.

5. **www.bloomberg.com**: You somehow heared about Bloomberg reported a news about China use tiny chips infiltrate U.S companies. You wants to find the article.

6. **www.reddit.com**: You are a fan of Marvel comics, you want to view some spoilers regarding a comming moive "The Avengers 4". Find latest three post that spoilers The Avengers 4.

7. **www.facebook.com**: You are a facebook user, and you have a wide social. However you don't wants to see parenting information in your timeline, you wish to turn them off for a year from your timeline; then recently you start interested in ping pong, you want to join a related local group.

8. **www.twitter.com**: You lost your phone and phone number, and you bought a new one. However the old phone number was registered in your twitter account, you want to change it for your account safety. Please find the entrace to change your phone number and password. Then you becomes curious on twitter's settings. You want to know how twitter use your data and prevent twitter collect your data.

9. **www.youtube.com**: You want to be a Youtuber. You wants to know how to earn money from making videos, and what should you concern when you publishing a video.

10. **www.google.com**: You can't access your gmail. You want to findout whether gmail are current malfuntioning or not. Contact instance messaging support.

B.3.2 Fuzzy Task

1. **www.github.com**: You were a senior developer. Your boss wants you write a report regarding the tends of current development techniques. You want to find the most three popular (top-3 stars) web backend Go frameworks and access their repository, write their name down on a paper when you decided.

2. **www.medien.ifi.lmu.de**: You are a fresh medieninformatik student. You wants to select three lectures, one seminar and one practicum for your study in WS18/19.

3. **www.arxiv.org**: Find the most recent published a overview paper for these three topics respectively: affective computing, convolutional neural networks, distributed consistency algorithm.

4. **www.google.com**: You want to know how google profiling you based on your history. Find your personality profile that created by Google.

5. **www.bloomberg.com**: You want to find the relevant news regarding the progress of China use tiny chips infiltrate U.S companies.

B.3.3 Exploring Task

1. **www.github.com**: Browsing github and select three github repository your most interested in.

2. **www.medien.ifi.lmu.de**: Browsing the website until time is up.

3. **www.en.uni-muenchen.de**: Browsing the website until time is up.

4. **www.ielts.org**: Browsing the website to see what you can do except register to examination.

5. **www.bloomberg.com**: Browsing the website until time is up.

6. **www.reddit.com**: Browsing the website until time is up.
7. **www.facebook.com**: Browsing the website until time is up.
8. **www.twitter.com**: Browsing the website until time is up.
9. **www.youtube.com**: Browsing the website until time is up.
10. **www.arxiv.org**: Browsing the website for categories you interested in until time is up.
11. **www.google.com**: Browsing google until time is up.

C Raw Data Illustration

C.1 Subjective Difficulty Score from Lab Study

Table C.1: Subjective task difficulty from lab study

Subject ID	Amazon.com	Medium.com	Dribbble.com
1	2, 1, 2	2, 4, 1	2, 3, 2
2	2, 2, 1	2, 3, 1	1, 5, 1
3	3, 2, 2	2, 5, 3	3, 1, 3
4	3, 4, 2	2, 5, 2	3, 3, 2
5	2, 1, 3	3, 5, 3	2, 1, 3
6	2, 2, 1	3, 4, 1	1, 3, 2
7	3, 4, 2	3, 5, 3	4, 3, 2
8	1, 1, 1	3, 5, 2	2, 1, 1
9	2, 3, 2	2, 5, 2	3, 1, 1
10	1, 3, 2	2, 3, 2	2, 3, 3
11	2, 2, 3	1, 4, 5	1, 2, 3
12	3, 2, 1	3, 4, 1	3, 2, 2
13	4, 1, 3	5, 4, 2	2, 2, 1
14	2, 2, 2	2, 3, 1	2, 2, 1
15	5, 1, 3	2, 4, 1	4, 2, 3
16	1, 2, 1	1, 3, 1	1, 1, 1
17	3, 1, 1	3, 4, 3	2, 2, 3
18	2, 2, 1	2, 3, 1	3, 2, 2
19	3, 2, 2	2, 2, 1	1, 1, 2
20	1, 3, 2	3, 5, 1	2, 3, 2
21	3, 3, 2	3, 5, 4	2, 3, 5

Bibliography

References

[Amo Filvá et al., 2018] Amo Filvá, D., Alier Forment, M., García Peñalvo, F. J., Fonseca Escudero, D., and Casany Guerrero, M. J. (2018). Learning analytics to assess students behavior with scratch through clickstream. In *Proceedings of the Learning Analytics Summer Institute Spain 2018: León, Spain, June 18-19, 2018*, pages 74–82. CEUR-WS. org.

[Baumann et al., 2018] Baumann, A., Haupt, J., Gebert, F., and Lessmann, S. (2018). The price of privacy: An evaluation of the economic value of collecting clickstream data. *Business and Information Systems Engineering*.

[Benevenuto et al., 2009] Benevenuto, F., Rodrigues, T., Cha, M., and Almeida, V. (2009). Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, IMC '09, pages 49–62, New York, NY, USA. ACM.

[Brodwin, D., D. O'Connell, and M. Valdmanis., 1995] Brodwin, D., D. O'Connell, and M. Valdmanis. (1995). Mining the Clickstream. pages 101–106.

[Bucklin and Sismeiro, 2000] Bucklin, R. E. and Sismeiro, C. (2000). How sticky is your web site? modeling site navigation choices using clickstream data. Technical report, Working paper, Anderson School UCLA.

[Carr, 2000] Carr, N. G. (2000). Hypermediation: commerce as clickstream. *Harvard Business Review*, 78(1):46–47.

[Cavoukian, 2000] Cavoukian, A. (2000). Should the oecd guidelines apply to personal data online. In *A report to the 22nd international conference of data protection commissioners*.

[Chatterjee, Patrali and Hoffman, Donna L and Novak, Thomas P, 2003] Chatterjee, Patrali and Hoffman, Donna L and Novak, Thomas P (2003). Modeling the clickstream: Implications for web-based advertising efforts. *Marketing Science*, 22(4):520–541.

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

[Cho et al., 2014] Cho, K., van Merrienboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

[Cochran and Cox, 1950] Cochran, W. G. and Cox, G. M. (1950). Experimental designs.

[Courtheoux, 2000] Courtheoux, R. J. (2000). Database marketing connects to the internet. *Interactive Marketing*, 2(2):129–137.

[Dijkstra, 1959] Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271.

[Friedman, Wayne and Weaver, Jane, 1995] Friedman, Wayne and Weaver, Jane (1995). Calculating cyberspace: tracking "clickstreams.".

[Gindin, 1997] Gindin, S. E. (1997). Lost and found in cyberspace: Informational privacy in the age of the internet. *San Diego L. Rev.*, 34:1153.

[Goldfarb, 2002] Goldfarb, A. (2002). Analyzing website choice using clickstream data. In *The Economics of the Internet and E-commerce*, pages 209–230. Emerald Group Publishing Limited.

[Graves, 2012] Graves, A. (2012). Sequence transduction with recurrent neural networks. *CoRR*, abs/1211.3711.

[Gundala and Spezzano, 2018] Gundala, L. A. and Spezzano, F. (2018). Readers' demanded hyperlink prediction in wikipedia. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 1805–1807, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9:1735–80.

[Jozefowicz et al., 2015] Jozefowicz, R., Zaremba, W., and Sutskever, I. (2015). An empirical exploration of recurrent network architectures. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2342–2350. JMLR.org.

[Kammenhuber et al., 2006] Kammenhuber, N., Luxenburger, J., Feldmann, A., and Weikum, G. (2006). Web search clickstreams. In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, IMC '06, pages 245–250, New York, NY, USA. ACM.

[Kang, 1997] Kang, J. (1997). Information privacy in cyberspace transactions. *Stan. L. Rev.*, 50:1193.

[Lin et al., 2012] Lin, M., Lin, M., and Kauffman, R. J. (2012). From clickstreams to searchstreams: Search network graph evidence from a b2b e-market. In *Proceedings of the 14th Annual International Conference on Electronic Commerce*, ICEC '12, pages 274–275, New York, NY, USA. ACM.

[Lori Lewis, 2017] Lori Lewis (2017). What Your Audience Is Doing When They're Not Listening To You. https://www.allaccess.com/merge/archive/26034/what-your-audience-is-doing-when-they-re-not. Accessed: 2018-12-28.

[Lori Lewis, 2018] Lori Lewis (2018). What Happens In An Internet Minute: 2018 Update. https://www.allaccess.com/merge/archive/28030/2018-update-what-happens-in-an-internet-minute. Accessed: 2018-12-28.

[Lourenço and Belo, 2006] Lourenço, A. G. and Belo, O. O. (2006). Catching web crawlers in the act. In *Proceedings of the 6th International Conference on Web Engineering*, ICWE '06, pages 265–272, New York, NY, USA. ACM.

[Lyons and Henderson, 2005] Lyons, B. and Henderson, K. (2005). Opinion leadership in a computer-mediated environment. *Journal of Consumer Behaviour: An International Research Review*, 4(5):319–329.

[Mandese, 1995] Mandese, J. (1995). Clickstreams' in cyberspace. *Advertising Age*, 66(12):18–18.

[Mann and Whitney, 1947] Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

[Meier and Elsweiler, 2016] Meier, F. and Elsweiler, D. (2016). Going back in time: An investigation of social media re-finding. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 355–364, New York, NY, USA. ACM.

[Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

[Mikolov et al., 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

[Mobasher et al., 2001] Mobasher, B., Dai, H., Luo, T., and Nakagawa, M. (2001). Effective personalization based on association rule discovery from web usage data. In *Proceedings of the 3rd International Workshop on Web Information and Data Management*, WIDM '01, pages 9–15, New York, NY, USA. ACM.

[N and Ravindran, 2018] N, C. T. and Ravindran, B. (2018). A neural attention based approach for clickstream mining. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, CoDS-COMAD '18, pages 118–127, New York, NY, USA. ACM.

[Novick, Bob, 1995] Novick, Bob (1995). Internet Marketing: The Clickstream. http://www.im.com/archives/9503/0375.htmlhttp:/www.im.com/archives/9503/0375.html. Accessed: 2018-12-10.

[Padmanabhan et al., 2001] Padmanabhan, B., Zheng, Z., and Kimbrough, S. O. (2001). Personalization from incomplete data: What you don't know can hurt. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pages 154–163, New York, NY, USA. ACM.

[Park et al., 2017] Park, J., Denaro, K., Rodriguez, F., Smyth, P., and Warschauer, M. (2017). Detecting changes in student behavior from clickstream data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, LAK '17, pages 21–30, New York, NY, USA. ACM.

[Reagle and Cranor, 1999] Reagle, J. and Cranor, L. F. (1999). The platform for privacy preferences. *Communications of the ACM*, 42(2):48–55.

[Reidenberg, 1996] Reidenberg, J. R. (1996). Governing networks and rule-making in cyberspace. *Emory LJ*, 45:911.

[Reidenberg, 1999] Reidenberg, J. R. (1999). Resolving conflicting international data privacy rules in cyberspace. *Stan. L. Rev.*, 52:1315.

[Rumelhart et al., 1988] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1988). Neurocomputing: Foundations of research. chapter Learning Representations by Back-propagating Errors, pages 696–699. MIT Press, Cambridge, MA, USA.

[Sadagopan and Li, 2008] Sadagopan, N. and Li, J. (2008). Characterizing typical and atypical user sessions in clickstreams. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 885–894, New York, NY, USA. ACM.

[Schneider et al., 2009] Schneider, F., Feldmann, A., Krishnamurthy, B., and Willinger, W. (2009). Understanding online social network usage from a network perspective. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, IMC '09, pages 35–48, New York, NY, USA. ACM.

[Schonberg et al., 2000] Schonberg, E., Cofino, T., Hoch, R., Podlaseck, M., and Spraragen, S. L. (2000). Measuring success. *Communications of the ACM*, 43(8):53–57.

[Shimada et al., 2018] Shimada, A., Taniguchi, Y., Okubo, F., Konomi, S., and Ogata, H. (2018). Online change detection for monitoring individual student behavior via clickstream data on e-book system. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, LAK '18, pages 446–450, New York, NY, USA. ACM.

[Skok, 1999] Skok, G. (1999). Establishing a legitimate expectation of privacy in clickstream data. *Mich. Telecomm. & Tech. L. Rev.*, 6:61.

[StatCounter, 2018] StatCounter (2018). Usage share of web browsers. http://gs.statcounter.com/browser-market-share#monthly-201811-201811-bar. Accessed: 2018-12-29.

[Sun and Xin, 2017] Sun, Y. and Xin, C. (2017). Using coursera clickstream data to improve online education for software engineering. In *Proceedings of the ACM Turing 50th Celebration Conference - China*, ACM TUR-C '17, pages 16:1–16:6, New York, NY, USA. ACM.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

[The Apache Software Foundation, 1995] The Apache Software Foundation (1995). About Apache: How Apache Came to Be. http://httpd.apache.org/ABOUT_APACHE.html. Accessed: 2018-12-10.

[Ting et al., 2005] Ting, I.-H., Kimble, C., and Kudenko, D. (2005). Ubb mining: Finding unexpected browsing behaviour in clickstream data to improve a web site's design. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '05, pages 179–185, Washington, DC, USA. IEEE Computer Society.

[Vassio et al., 2018] Vassio, L., Drago, I., Mellia, M., Houidi, Z. B., and Lamali, M. L. (2018). You, the web, and your device: Longitudinal characterization of browsing habits. *ACM Trans. Web*, 12(4):24:1–24:30.

[Walsh, John and Godfrey, Sue, 2000] Walsh, John and Godfrey, Sue (2000). The Internet: a new era in customer service. *European Management Journal*, 18(1):85–92.

[Wang et al., 2017] Wang, G., Zhang, X., Tang, S., Wilson, C., Zheng, H., and Zhao, B. Y. (2017). Clickstream User Behavior Models. *ACM Trans. Web*, 11(4):21:1–21:37.

[Wang et al., 2016] Wang, G., Zhang, X., Tang, S., Zheng, H., and Zhao, B. Y. (2016). Unsupervised clickstream clustering for user behavior analysis. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 225–236, New York, NY, USA. ACM.

[Waterson et al., 2002a] Waterson, S., Landay, J. A., and Matthews, T. (2002a). In the lab and out in the wild: Remote web usability testing for mobile devices. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '02, pages 796–797, New York, NY, USA. ACM.

[Waterson et al., 2002b] Waterson, S. J., Hong, J. I., Sohn, T., Landay, J. A., Heer, J., and Matthews, T. (2002b). What did they do? understanding clickstreams with the webquilt visualization system. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '02, pages 94–102, New York, NY, USA. ACM.

[Weller, 2018] Weller, T. (2018). Compromised account detection based on clickstream data. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 819–823, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

[Werbos, 1990] Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

[Williams and Zipser, 1989] Williams, R. J. and Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

[Yamakami, 2009] Yamakami, T. (2009). Inter-service revisit analysis of three user groups using intra-day behavior in the mobile clickstream. In *Proceedings of the 2009 International Conference on Hybrid Information Technology*, ICHIT '09, pages 340–344, New York, NY, USA. ACM.

[Yang et al., 2014] Yang, Z., Wilson, C., Wang, X., Gao, T., Zhao, B. Y., and Dai, Y. (2014). Uncovering social network sybils in the wild. *ACM Trans. Knowl. Discov. Data*, 8(1):2:1–2:29.

[Zaloudek, 2018] Zaloudek, J. (2018). User Behavior Clustering and Behavior Modeling Based on Clickstream Data. Master's thesis, Czech Technical University in Prague, Faculty of Electrical Engineering Department of Computer Science.

[Zhang et al., 2016] Zhang, X., Brown, H.-F., and Shankar, A. (2016). Data-driven personas: Constructing archetypal users with clickstreams and user telemetry. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 5350–5359, New York, NY, USA. ACM.