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:	1.7.2018 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, December 30, 2018 . . . . . .

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 is 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, we 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 client-side user clickstream study.

Contents

List of Figures

List of Tables

LIST OF TABLES

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].

Subsequently, 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 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 chapter 7.

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 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 slightly different than clickstream since a user may use back button or switch browser tabs jump to visited web pages or parallel web pages (or say performed a visit action), a server side clickstream cannot record such detailed level of user clickstream. The term *Action path* is a generalized concept of clickstream, which replace individual URLs to user actions (with backbutton and browser tab switch effects) in the browser.

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

Let an action path is represented on directed cyclic graph, each node represents a visited page, 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:

$$\begin{aligned} E &= w_1 \frac{n_s}{N} + w_2 \frac{d_s}{D} \\ w_1 + w_2 &= 1 \end{aligned} \tag{1}$$

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 minimuzed. 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 Clickstream Overlap Ratio

3.3 URL2Vec Embedding

3.4 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 2:

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

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 is consist of a sequence of URLs $(U_1^{ij}, U_2^{ij}, ..., U_n^{ij})$ 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 the given page.

3.5 Action Path Optimization

$\arg\max_y p(o_1, o_2, ..., o_{T'}|i_1, i_2, ..., i_T)$

4 Lab Study

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 mobile laptop. All 21 subjects are recruit anonymously and randomly.

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.

4.1 Pilot Study

Initially, we developed a web crawler that downloaded the entire medien computer science website [1] .... TODO: discuss how we go here.

4.2 Study

In our lab study, we selected three mainstream websites, Amazon/Medium/Dribbble that covers categories for shopping, media consuming and design brainstorming. We manually designed nine reasonable (discussed in section 4.2.3) context-given browsing tasks (three for each website) in total to our subjects. Each task requires participant start from the homdpage 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. To simulate a real case, there is no interruption and question answering during the task.

4.2.1 Selection of Environment

The lab study uses two self provided device: 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. To eliminate the learning effect of computing devices, we use Latin square [Cochran and Cox, 1950] for the participation order of desktop and laptop to our subjects.

Considering the usage share of all browsers on the market, till the month we runs our study (November), Google Chrome wins the usage share of desktop browsers with 61.75% shares, and Apple Safari only shares 15.12% of the market. Clearly, Google Chrome dominant the desktop web browser.

Therefore, we decide to use Chrome to establish our plugin of data collection. A quesionaire after our lab study indicates the browsers usage share of all subjects, as shown in Table 4.2, which further supports our decision of browser selection.

[1] https://medien.ifi.lmu.de

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%

4.2.2 Selection of Context-given Websites

In our study, we gathered various websites covering most of the categories that people do on the Internet: Social Networks, Shopping, Email, Media Consuming, Search, Production and etc [Lori Lewis, 2018, Lori Lewis, 2017]. All selected website are listed in Table **??**.

Table 4.2: 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%

4.2.3 Selection of Designed Tasks

Before we explain the design reason of our context-given browsing task, we first present and discusses the common three types of daily browsing.

Goal-oriented Task: Typically, an Internet user opens the browser for information retrieve for bussiness work, school study, literature research and reasons that made the person has clear purpose.

Exploring Task: A user visits the browser aimlessly, explores and consumes the content lies on the Web.

Fuzzy Task:

5 Evaluation

In the chapter of evaluation, we conduct two track of analysis to our collected data, including quantitative analysis and qualitative analysis. The data are collected from 21 subjects, and 189 clickstream data in total. Each clickstream contains page-level data with a stay duration of a specific page. Each clickstream also has a subjective difficulty score voted by each of participants.

5.1 Quantitative: Subjective Task Difficulty

Figure 5.1 illusrates a normalized (raw scores are listed in Appendix C Table C.1) subjective difficulty score with respect to all tasks.

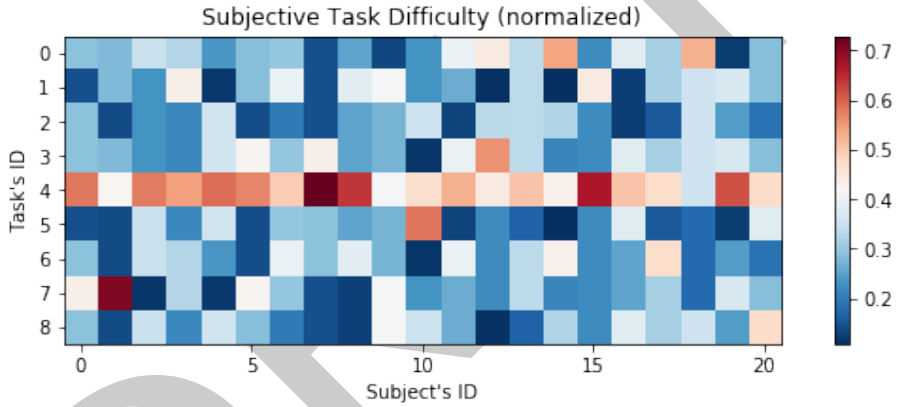

Figure 5.1: Subjective difficulty score: each column indicates an individual subject and each row indicates a browsing task. Tasks from 0 to 8 represent Amazon Goal Oriented Task, Amazon Fuzzy Task, Amazon Exploring Task; Medium Goal Oriented Task, Medium Fuzzy Task, Medium Exploring Task, Dribbble Goal 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.

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 Kolmogorov-Smirnov test [Massey Jr, 1951], 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.*

5.1.1 t-SNE

5.1.2 Prediction Accuracy

5.1.3 F1

5.2 Rationality of Designed Tasks

5.3 Explored Model Architecture Comparasion

5.4 Action Path Visualization

5.5 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

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.

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.

[2] http://creativecommons.org/licenses/by-nc-sa/4.0/
[3] https://github.com/changkun/MasterThesisHCI/blob/master/LICENSE

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

   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 any 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

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

TODO: add unselected tasks

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*.

[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.

[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.

[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.

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

[Massey Jr, 1951] Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

[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.

[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.

[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.

[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.