Understanding and Predicting Client-side User Clickstream
Introduction

“Hello world!”
Background & Motivation

- Clickstream research: e.g. Privacy issue [SKOK, 1999]; Compromise Detection [WELLER, 2018], etc.

“Clickstream” [FRIEDMAN, 1995]

branching [HUANG, 2010] and backtrace [HUANG, 2012]

Distraction
Research Questions

1. Understanding:
   • Why collecting clickstream on client-side differs from server-side collecting?
   • What are the most significant, identifiable user behaviors or activity patterns can be observed or detected in the context of web browsing that indicates information needs,
   • in which form of quantitative data can characterize a definitive boundary to distinguish browsing behaviors of a user?

2. Classification:
   • How accurate or how affirmative we can model or identify the proposed browsing behaviors progressively that makes an intelligent system serves proactively?

3. Prediction:
   • How much future movements of a user can be accurately inferred from the context of web browsing, and how much context is required for the prediction?
Agenda

Approaches ➔ Evaluations ➔ Applications
Approaches

“Stop talking. Just coding.”
Data

- An action path (URL+Duration):

\[(URL_1, d_1, URL_2, d_2, ..., URL_n, d_n)\]

where \(n\) is the total number of action in a browsing session.

- Special cases:
Model: *Action-Path* Model Overview

- Recurrent unit is **not** a standard GRU or LSTM;
- Special tokens `<SOA>`/`<COI>`/`<SOP>`/`<EOA>`/`<MIS>`/`<PAD>` are used in practice.
Model: url2vec Embedding

- url2vec construct URL presentation to better predict surrounding URLs

\[ p(\text{URL}_{t+i}|\text{URL}_t) = \frac{\exp(v_{\text{URL}_{t+i}}^T v_{\text{URL}_t})}{\sum_{i \in \text{all URLs}} \exp(v_i^T v_{\text{URL}_t})} \]
Model: *Recurrent Unit*

- Improved from **GRU or LSTM**;
- Feed not only URL embeddings but also stay duration

\[
I_t = \sigma(P^{(I)} U_t^{ij} + Q^{(I)} h_{t-1} + \frac{d_t^{ij}}{d_t^{ij} + 1})
\]

\[
F_t = \sigma(P^{(F)} U_t^{ij} + Q^{(F)} h_{t-1} + b^{(F)})
\]

**LSTM-like:**

\[
O_t = \sigma(P^{(O)} U_t^{ij} + Q^{(O)} h_{t-1})
\]

\[
C_t = F^{(t)} \circ C_{t-1} + I_t \circ \tanh(P^{(C)} U_t^{ij} + Q^{(C)} h_{t-1})
\]

\[
h_t = O_t \circ \tanh(C_t)
\]

\[
Z_t = \sigma(P^{(Z)} U_t^{ij} + Q^{(Z)} h_{t-1} + \frac{d_t^{ij}}{d_t^{ij} + 1})
\]

**GRU-like:**

\[
R_t = \sigma(P^{(R)} U_t^{ij} + Q^{(R)} h_{t-1})
\]

\[
h_t = (1 - Z_t) \circ \tanh(P^{(H)} U_t^{ij} + Q^{(H)} h_{t-1}) + Z_t \circ h_{t-1}
\]
User Study Design

“Data is priceless.”
Information Behavior Theory on the Web

- Comparison (All based on Wilson’s theory [WILSON, 1997] and Ellis’s Model [ELLIS, 1989]):

<table>
<thead>
<tr>
<th>Author</th>
<th>Terminologies</th>
<th>Terminologies</th>
<th>Terminologies</th>
<th>Main Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>[CHOO et al, 1999]</td>
<td>Formal search</td>
<td>Conditioned viewing; Informal search</td>
<td>Undirected viewing</td>
<td>Moves of information seeking</td>
</tr>
<tr>
<td>[JOHNSON, 2017]</td>
<td>Directed browsing</td>
<td>Semi-directed browsing</td>
<td>Undirected browsing</td>
<td>Actions; Objectives</td>
</tr>
<tr>
<td></td>
<td>Known-item search</td>
<td>Explorative seeking</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>“You don’t know what you need”</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Re-finding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>This thesis</td>
<td>Goal-oriented</td>
<td>Fuzzy</td>
<td>Exploring</td>
<td>Information needs (purpose); Chaining; Differentiating, etc.</td>
</tr>
</tbody>
</table>

- We conservatively grouped three distinguishable (evidence shows later) behaviors:
  - **Goal-oriented**: browsing on the web deterministically with systematically specified purpose
  - **Exploring**: browsing aimlessly without information needs and use
  - **Fuzzy**: “chaining” without “differentiating" from “starting”
Task Design & Collected Data

- 5~10 min/task, no visit restriction
- 9 tasks from 35 tasks, simulate three browsing behavior.

<table>
<thead>
<tr>
<th>Starting Point</th>
<th>Goal-oriented task</th>
<th>Fuzzy task</th>
<th>Exploring task</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.amazon.com">www.amazon.com</a></td>
<td>Assume your smartphone was broken and you have <strong>1200 euros as your budget</strong>. You want to buy an iPhone, a protection case, and a wireless charging dock. Look for these items and add them to your cart.</td>
<td>You want to buy <strong>birthday present(s)</strong> for your best friend Add three items to your cart.</td>
<td>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.</td>
</tr>
</tbody>
</table>

- Equipment of experiment
  - Software: Chrome; Hardware: Desktop & Laptop
  - Latin square
- 21 subjects, 189 action paths (clickstreams)
  - Age & Gender: min=18, max=29, median=23, mean=23.04, SD=3.22, male=10, female=11
Evaluations

“History tells our future.”
Subjective Tasks Difficulty Score

- Subjective difficulty score from participants

- One-tailed Mann-Whitney U test

- H0: the difficulty of fuzzy task is not greater than exploring task, \( p < 0.05 \), reject H0.

- For **difficulty:**
  - Fuzzy task > Goal-oriented task > Exploring task
“Efficiency”, Length, Duration: Classification?

- Feature Importance (RDT approach): \textbf{Length} > \textbf{Duration} > \textbf{Efficiency}
- One-tailed Mann-Whitney U test:
  - completion efficiency: \{Goal,Fuzzy\} > Exploring
  - browsing action length: \{Fuzzy,Exploring\} > Goal
  - total stay duration: Exploring > \{Goal, Fuzzy\}
Model Performance: on Prediction and Classification

- One-layer model
  - 90323 params, 1500 epochs, 10 latent dim, 32 batch size, Adam optimizer
  - categorical cross-entropy loss, L1&L2 regularizer in decoder with early stopping
  - Total samples: 189, train/val/test: 132/38/19 (0.7/0.2/0.1 ratio)

- Split ratio >0.9 achieves excellent performance, >60% accurate (still can be optimized)
- Best prediction steps = average length of sequence * 0.95, in our tiny dataset ≈ 3~5 steps
Model Performance: More Observations

- **Classification** is a special case of prediction in the model (predict last step)
  - Ending marks as classes: <EOA_GOAL>, <EOA_FUZZY>, <EOA_EXPLORE>
  - Classification accuracy: ~100.0%
  - Prediction accuracy: >60%
- Validation loss:

![Graph of categorical cross-entropy validation loss](image)

**Classification**

![Graph of categorical cross-entropy validation loss](image)

**3 future steps prediction**
Patterns: Goal-oriented Tasks

Example: Amazon’s goal oriented task
Patterns: Fuzzy & Exploring Tasks

Example: Two participants, Amazon and Dribbble’s fuzzy and exploring task
Patterns: Overlaps

- Example: Medium goal-oriented task
- 4 participants (non-black nodes)
- **Overlap ratio: 9.43%**
- Highest 11.84%, lowest 0.00%
Patterns: Overlaps

- Example: Dribbble exploring task
  - 4 participants same as previous
  - **Overlap ratio: 1.15%**
  - Highest 11.84%, lowest 0.00%

![Overlapped page diagram]
Applications

“Let’s make a story”
Possible Applications: Browser Plugin & Standardization

Understanding and Predicting Client-side User ClickstreamApplicationsShow Cases
Conclusion

“To the higher truth.”
Summary

- A study regarding **action path**, a.k.a. **client-side** clickstreams; contributes to:
  - **Understanding of browsing behavior on client-side:**
    - Three classifiable behavior: **goal-oriented**, **fuzzy**, and **exploring**
    - **Historical URLs** & **stay duration** are the most important features
    - Browsing intents tend to distributed and clustered individually
    - Browsing behavior is **not user-specific** but intersection
      - “cluster”/“hesitation”/“ring”/“star”/“overlaps” patterns
    - Goal-oriented behavior tent to overlap because no overlaps
  - One model to learn them all (Classification & Accuracy):
    - Browsing behavior **classification**: ~100.0% accurate
    - **Page-level** 5 future steps universal **prediction**: >60% accurate
  - Possible Applications:
    - As browser plugin proactively improves user actions
    - As standard API helps developers improve their product
  - Limitations & Future works:
    - Privacy & Trustiness & Security Issues
    - More data for more precise behavior categories
    - Performance optimization & & more applications
References

- .... And More!
“You found a secret place.”
A Brief History of “Clickstream”

- A clickstream [FRIEDMAN, 1995] (informally) contains a sequence of hyperlinks clicked by users on the web over time. Early clickstream research emerges for privacy discussion [SKOK, 1999], customer services [WALSH et al., 2000], business decisions [SCHONBERG, 2000], personalization web services [MOBASHER et al., 2001], remote usability testing [WATERSOn et al., 2002], etc.

- Recent works take clickstream on web user behavior analysis for social characterizing [SCHEIDER et al., 2009], media re-finding [MEIER et al., 2016], user clustering [WANG et al., 2016], account security [WELLER, 2018] and more...

- Closest research to ours is branching [HUANG, 2010] and backtrace [HUANG, 2012], but still a server side analysis and individually analyzed.

Questions:
- Making difference if collect from user side?
- Branching & backtrace effect?
- Matters to user?
- Benefits to user?

For business and on server side:
+ branching [HUANG, 2010]
+ backtrace [HUANG, 2012]
Information Behavior Theory

- Four information behavior [CHOO et al. 1999] on the Web are discussed for information need, information seeking and information use and uses Ellis’ Model [ELLIS, 1989].

<table>
<thead>
<tr>
<th>Author</th>
<th>Passive Attention</th>
<th>Passive Search</th>
<th>Active Search</th>
<th>Ongoing Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>[WILSON, 1997]</td>
<td>No info-seeking intended, but acquisition take place</td>
<td>Occasionally relevant to individuals</td>
<td>Actively seeking</td>
<td>Update basic framework of ideas</td>
</tr>
<tr>
<td>[CHOO et al. 1999]</td>
<td>Undirected viewing</td>
<td>Conditioned viewing</td>
<td>Formal search</td>
<td>Informal search</td>
</tr>
</tbody>
</table>

- More browsing patterns are discussed [JOHNSON, 2017], and we selected three categories:
  - Directed browsing
  - Semi-directed browsing
  - Undirected browsing
  - Known-item search
  - Exploratory seeking
  - “You-don’t-know-what-you-need”
  - Re-finding
### Analysis based on Wilson and Ellis’ Model

<table>
<thead>
<tr>
<th>Behaviors / Patterns</th>
<th>Information Need</th>
<th>Information Seeking</th>
<th>Information Use</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Starting</td>
<td>Chaining</td>
</tr>
<tr>
<td>Goal-oriented</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Fuzzy</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Exploring</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>“cluster”</td>
<td>Observed</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>“star”</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>“ring”</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>“hesitation”</td>
<td>Observed</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>“overlap”</td>
<td>Observed</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Model: **Context Encoder**

Ref. #ch 3.4.1
Model: **Context Decoder**

Ref. #ch 3.4.1

\[ U_{ij}^{t+1} \quad U_{ij}^{t+2} \quad U_{ij}^{t+n} \]

Context

Recurrence Unit

\[ <SOP> \]

Recurrence Unit

Recurrence Unit

Recurrence Unit

Recurrence Unit

Recurrence Unit

Recurrence Unit

Recurrence Unit

Recurrence Unit

Recurrence Unit

Recurrence Unit
# Task Design: Collect Data from User

- Participants are allowed to visit any pages during the task, even outside the starting point domain.
- 5~10 minutes for each of the task, 80min in total.

<table>
<thead>
<tr>
<th>Starting Point</th>
<th>Goal-oriented task</th>
<th>Fuzzy task</th>
<th>Exploring task</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.amazon.com">www.amazon.com</a></td>
<td>Assume your smartphone was broken and you have <strong>1200 euros as your budget</strong>. You want to buy <strong>an iPhone</strong>, a <strong>protection case</strong>, and a <strong>wireless charging dock</strong>. Look for these items and add them to your cart.</td>
<td>You want to buy a <strong>gift</strong> for your <strong>best friend</strong> as a birthday present. Add three items to your cart.</td>
<td>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.</td>
</tr>
<tr>
<td><a href="http://www.medium.com">www.medium.com</a></td>
<td>Assume you were making plans for your <strong>summer vacation</strong>. You want to visit <strong>Tokyo, Kyoto, and Osaka</strong>. You want to find out what kind of experience other people made when traveling to these three places in Japan. Your task is to <strong>find three posts for traveling tips</strong> regarding these cities. <strong>Elevate (👏) a post if it is one of your choices.</strong></td>
<td>Assume you got an occasion to visit <strong>China for business</strong>. You are <strong>free to travel</strong> to China for a <strong>week</strong>. 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. <strong>Elevate (👏) if a post helped you make a decision.</strong></td>
<td>Visit a category you are interested in and elevate (👏) the post you like.</td>
</tr>
<tr>
<td><a href="http://www.dribbble.com">www.dribbble.com</a></td>
<td>You are hired to a <strong>Cloud Computing</strong> startup company. You get an assignment to designing the logo of the company. Search for existing logos for inspiration and <strong>download three candidate logos you like the most.</strong></td>
<td>You are preparing a presentation and need one picture for each of these animals: <strong>cat</strong>, <strong>dog</strong>, and <strong>ant</strong>. Download the three pictures you like the most.</td>
<td>Explore dribbble and download images you like the most while you browse.</td>
</tr>
</tbody>
</table>