From Skimming to Reading: A Two-stage Examination Model for Web Search

Yiqun Liu, Chao Wang, Ke Zhou, Jianyun Nie, Min Zhang, Shaoping Ma
Tsinghua University; Yahoo Labs; Université de Montréal
November, 2014
Background: Search Engine Result Ranking

• SEO report: 100+ signals
• Yahoo LTR task: 700+ signals
  - Hyperlink, Content relevance, **User behavior**, Page structure, Freshness, Service stability,
  ……
• Basic assumption
  - Results that are clicked more tend to be more relevant
Background: User Implicit Feedback

• A simple solution: user click = relevance voting
  • Tsinghua University => www.tsinghua.edu.cn

• Problem: position bias
  • Users tend to click more on higher-ranked results

![User click v.s. Search Ranking](http://hubdesignsmagazine.com/2011/03/27/its-good-to-be-on-the-first-page-of-google/)
Background: Examination Hypothesis

• The likelihood that a user will click on a search result is influenced by
  • Whether the user examined the search result
  • Whether the result is attractive/relevant

\[ C_i = 1 \rightarrow E_i = 1, R_i = 1 \]

• Examination: user has comprehended (part of) the result and made a decision on whether to click.

• How do users examine search results?
• How to estimate the probability of examination?

Result Examination: Click Models

• Examination Hypothesis
  \[ C_i = 1 \implies E_i = 1, R_i = 1 \]

• Estimating examination with search user behavior assumption

  • Cascade model:
    \[ P(E_{i+1} = 1|E_i = 1, C_i) = 1 - C_i \]

  • Dependent click model (DCM):
    \[ P(E_{i+1} = 1|E_i = 1, C_i = 0) = 1 \]
    \[ P(E_{i+1} = 1|E_i = 1, C_i = 1) = \lambda_i \]

  • User browsing model (UBM):
    \[ P(E_i = 1|C_1...i-1) = \lambda_{r_i,d_i} \]

  • Other models: DBM, DBN, CCM, ...
Result Examination: Eye-tracking

**Strong Eye-mind Hypothesis**

- There is no appreciable lag between what is fixated on and what is processed (Just et al., 1980).
- Most existing studies: Inferring Examination with a threshold in fixation (200-500ms)

![Graph 1: Percentage of times an abstract was viewed/clicked depending on the rank of the result.](image1)

![Graph 2: Mean time of arrival (in number of previous fixations) depending on the rank of the result.](image2)

Joachims et al., Eye-tracking analysis of user behavior in www search. SIGIR 2005
Result Examination: beyond Eye-tracking

• **Problems with Strong Eye-mind Hypothesis**

• While the duration of the gaze is closely related to the duration of cognitive processes, *the two durations are not necessarily identical*. (Just & Carpenter, 1980)
Research Questions

• RQ1: How do users examine results on SERPs
• RQ2: How do behavior biases happen in user’s examination process
• RQ3: How can we identify examination behavior
Collecting Examination Information

Commercial Search Engine

Fixed Queries

Search Tasks

Removal of Ads, verticals, ...

SERP Generation

Eye-tracker Calibration

Task completion and eye /mouse logging

Collecting user feedback on result comprehension

Search Users

Next Search Task
Collecting Examination Behavior on SERPs

• Search task details

• Data Collected: click-through, mouse movement, eye movement, explicit feedback on examination.

• 37 participants, 25 queries (INF:TRA:NAV = 2:2:1)
Examination Behavior Analysis

• Examination v.s. Fixation: *Eye fixation on a search result is a prerequisite for examining this result*

<table>
<thead>
<tr>
<th></th>
<th>Fixation=0</th>
<th>Fixation=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examine=0</td>
<td>31.61%</td>
<td>28.81%</td>
</tr>
<tr>
<td>Examine=1</td>
<td>5.49%</td>
<td>34.09%</td>
</tr>
</tbody>
</table>

Why don’t you annotate the fixed results as examined?

<table>
<thead>
<tr>
<th>Proportion</th>
<th>Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>48%</td>
<td>Take a glance at the result without thinking about it.</td>
</tr>
<tr>
<td>26%</td>
<td>Take a glance at the result and feel unattractive to read it</td>
</tr>
<tr>
<td>16%</td>
<td>Feel that the result is not relevant.</td>
</tr>
<tr>
<td>10%</td>
<td>Cannot tell clear reason.</td>
</tr>
</tbody>
</table>
Examination Behavior Analysis

• Examination v.s. Click: *Examining a search result is a prerequisite for clicking on the result.*

<table>
<thead>
<tr>
<th></th>
<th>Examine=0</th>
<th>Examine=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Click=0</td>
<td>59.24%</td>
<td>17.57%</td>
</tr>
<tr>
<td>Click=1</td>
<td>1.18%</td>
<td>22.01%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Fixation=0</th>
<th>Fixation=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Click=0</td>
<td>34.96%</td>
<td>41.85%</td>
</tr>
<tr>
<td>Click=1</td>
<td>2.15%</td>
<td>21.04%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Fixation=0</th>
<th>Fixation=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examine=0</td>
<td>31.61%</td>
<td>28.81%</td>
</tr>
<tr>
<td>Examine=1</td>
<td>5.49%</td>
<td>34.09%</td>
</tr>
</tbody>
</table>
A Two-Stage Examination Model

Stage 1: Examination
- Eye Fixation (Skimming)
- Judgment

- Not Relevant
- Potential Relevant

Stage 2: Examination
- Explicit Feedback (Reading)
- Behavior Biases (Position, Attractiveness, Domain, ...)
- Behavior Biases (Position, ...), Snippet content (title/abstract)

- Not Relevant
- Relevant

Skip the result
Click the result
Landing Page Reading
A Two-Stage Examination Model

• Answer to RQ1: Users examine results with a two-stage model
  • Stage1: skimming process, careful reading or not
  • Stage2: reading process, clicking or not

• Relationship with information triage
  • the process of determining the priority of processing

• Relationship with selective attention
  • the process whereby the brain selectively filters out large amounts of sensory information to focus
Research Questions

• RQ1: How do users examine results on SERPs

• RQ2: How do behavior biases happen in user’s examination process

• RQ3: How can we identify examination behavior
Behavior Biases in Two-Stage Model

- Behavior biases in Web search environment
  - **Position bias:** Higher-ranked results receive more user attention (Craswell et al. 2008)
  - **Attractiveness bias:** Attractiveness in result titles and abstracts affects user judgment (Bar-Ilan et al. 2009), multimedia vertical results draws much user attentions (Wang et al. 2013)
  - **Trust bias:** Results from trust-worthy Web domains are preferred by users (Ieong et al. 2012)
Position Bias

• User judgments (for relevant results) in two stages are both affected by positions
Attractiveness Bias

• Attractive results draws significantly more attention in Stage 1 while doesn’t affect the judgment in Stage 2.

• Attractive results: Results with the longest title and abstract exact match on SERPs

<table>
<thead>
<tr>
<th></th>
<th>Attractive results</th>
<th>Other results</th>
</tr>
</thead>
<tbody>
<tr>
<td>**P(E</td>
<td>F)**</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>0.637301</td>
<td>0.484615</td>
</tr>
<tr>
<td>Variance</td>
<td>0.058769</td>
<td>0.066037</td>
</tr>
<tr>
<td>p-value</td>
<td></td>
<td>0.005788</td>
</tr>
<tr>
<td>**P(C</td>
<td>E)**</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>0.57775</td>
<td>0.472463</td>
</tr>
<tr>
<td>Variance</td>
<td>0.122599</td>
<td>0.082748</td>
</tr>
<tr>
<td>p-value</td>
<td></td>
<td>0.158477</td>
</tr>
</tbody>
</table>
Attractiveness Bias v.s. Position Bias

- Attractiveness bias happens in all result positions for judgments in Stage 1.
Trust Bias

• Reputable results draws significantly more attention in Stage 1 while doesn’t affect the judgment in Stage 2.
• Reputable results: results from Alexa.com top 100 popular sites in China

<table>
<thead>
<tr>
<th></th>
<th>Average</th>
<th>Variance</th>
<th>p-value</th>
<th>Average</th>
<th>Variance</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>**P(E</td>
<td>F)**</td>
<td>0.613371</td>
<td>0.065817</td>
<td>0.000656</td>
<td>0.519443</td>
<td>0.079853</td>
</tr>
<tr>
<td>**P(C</td>
<td>E)**</td>
<td>0.470799</td>
<td>0.063693</td>
<td>0.311937</td>
<td>0.473674</td>
<td>0.089271</td>
</tr>
</tbody>
</table>
Trust Bias v.s. Position Bias

- Trust bias happens in relatively lower result positions for judgments in Stage 1.
Effectiveness of Judgments in Two Stages

- User examines more results in Stage 1, but the effectiveness of judgments in Stage 2 is higher
  - Relevance judgment in Stage 1: entering Stage 2
  - Relevance judgment in Stage 2: result clicking

<table>
<thead>
<tr>
<th></th>
<th>Stage 1</th>
<th>Stage 2</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of examined results</td>
<td>5598/8900</td>
<td>3034/5598</td>
<td>-45.80%</td>
</tr>
<tr>
<td>Number of results judged as relevant</td>
<td>3034/5598</td>
<td>1779/3034</td>
<td>-38.27%</td>
</tr>
<tr>
<td>Precision</td>
<td>0.5968</td>
<td>0.6738</td>
<td>+11.43%</td>
</tr>
<tr>
<td>Recall</td>
<td>0.6040</td>
<td>0.6755</td>
<td>+10.58%</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.6004</td>
<td>0.6747</td>
<td>+11.01%</td>
</tr>
<tr>
<td>AUC/ROC</td>
<td>0.6523</td>
<td>0.7169</td>
<td>+9.011%</td>
</tr>
</tbody>
</table>
Research Questions

• RQ1: How do users examine results on SERPs
• RQ2: How do behavior biases happen in user’s examination process
• RQ3: How can we identify examination behavior
Identifying Examination Behavior

• Existing solution: identification with eye fixation
  • Cognitive studies: McConkie 1975; Just et al. 1980;
  • Web search: Cutrell et al, 2007; Buscher et al, 2012;

• Problems:
  • Equipment is too expensive
  • Users are required to calibrate
  • Fixed threshold setting is not reasonable
Identifying Examination Behavior

• Alternative solution: Mouse movement behavior
  - Mouse movement information could be collected at large scale without interrupting users

• Existing studies on fixation prediction
  - Eye-mouse coordination: Rodden, 2008; Huang, 2012
  - Fixation prediction: Guo, 2010; Huang, 2012

• Problem:
  - Previous target: predict the whole fixation sequence
  - New target: predict whether a result is examined or not
Mouse Movement Features

• Distance features:
  • User’s total leftwards/rightwards/upwards/downwards movement distances in the result zone

• Position features:
  • The leftmost/rightmost/upmost/bottommost position cursor ever reaches in the result zone

• Duration features
  • Total mouse dwell time on a result/SERP/search task
Examination Prediction and Relevance Estimation

- Actual v.s. Predicted user behavior

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>KAPPA</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>E0</td>
</tr>
<tr>
<td>GBRT</td>
<td>0.6393</td>
<td>0.4519</td>
<td>0.7531</td>
</tr>
<tr>
<td>LogisticRegression</td>
<td>0.6310</td>
<td>0.4389</td>
<td>0.7517</td>
</tr>
<tr>
<td>SVM</td>
<td>0.6191</td>
<td>0.4163</td>
<td>0.7369</td>
</tr>
<tr>
<td>RandomForest</td>
<td>0.6151</td>
<td>0.4167</td>
<td>0.7332</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.6056</td>
<td>0.3972</td>
<td>0.7236</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>KAPPA</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>E0</td>
</tr>
<tr>
<td>Two-stage model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single-stage model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.6440</td>
<td>0.5760</td>
<td>0.6400</td>
</tr>
<tr>
<td>Precision</td>
<td>0.6910</td>
<td>0.8221</td>
<td>0.6872</td>
</tr>
<tr>
<td>Recall</td>
<td>0.6970</td>
<td>0.3356</td>
<td>0.6941</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.6865</td>
<td>0.4747</td>
<td>0.6799</td>
</tr>
</tbody>
</table>
Take-Home Messages

• RQ1: How do users examine results on SERPs
  • Two-stage examination: from skimming to reading
  • Information triage / selective attention in Web search

• RQ2: How do behavior biases happen in user’s examination process
  • Users rely on different signals in different stages

• RQ3: How can we identify examination behavior
  • Supervised learning with mouse movement features
Thank you

Welcome to visit my homepage

http://www.thuir.cn/group/~YQLiu/