Assessing Query Suggestions for Search Session Simulation

Martin-Luther-Universität Halle-Wittenberg

July 15, 2021

webis.de
Assessing Query Suggestions for Search Session Simulation

The SINIR Project

- **SINIR:** Simulating Interactive Information Retrieval
- Funded by DFG
- Development of a simulation framework to:
  - Assess changes to the retrieval system and UI of digital libraries
  - Analyze the cost/gain impact of changes on (simulated) users
- Without having to implement changes to a live system
- In cooperation with ZBW (econbiz) and University of Passau
Assessing Query Suggestions for Search Session Simulation

SINIR - Motivation (1)

- Web search engines and digital libraries have a lot in common
  But: significant differences for development and evaluation

- **Evaluating changes to web search engines**
  Implement the changes
  A/B testing with part of the large user base

- **Evaluating changes to digital libraries**
  Fewer human resources for development
  No large user base to run A/B tests on

- **Solutions**
  Models of changed systems instead of implementations
  Increase user base by simulating users
If no sufficient user base is available, simulated users can be utilized to evaluate a search engine.

Simulated user has to:
- Solve a task (find information/document)
- Interact with an IR system
- Work on unseen data

Formulating queries is part of an authentic user model.
Assessing Query Suggestions for Search Session Simulation

Our Research Question in this Paper

- Most search engines will provide suggestions or autocompletion
- Users are influenced by those suggestions
- Search suggestions are built upon user queries
- Users use suggestions to formulate their queries

Can we create authentic sessions from suggestions only?

- For this talk: session = multiple queries on a given topic (no interaction, etc.)
Assessing Query Suggestions for Search Session Simulation

Approach

- Given an initial query, mine search suggestions from the Google Suggest Search API (up to 10)
- Using different strategies, choose one of the suggestions as next query or end the session
- Optional: utilize the topic description provided in the selection process
Assessing Query Suggestions for Search Session Simulation

Data

- Data sources: real sessions from
  - TREC 2014 Session track
  - Webis-SMC-12 dataset

- Inter-query times sampled from Webis-SMC-12 dataset
  - Will be used to generate timestamps in the session
  - Max. 20 minutes after removal of outliers
Assessing Query Suggestions for Search Session Simulation

Suggestion Selection Strategies

First suggestion

Random suggestion

Three word queries (based on “Session Strategy S3 [Keskustalo et al. 2009]

- Initial query forms the basis
- In each step, combined with one of the tf-idf-wise top-10 keywords from the topic description
- Check whether generated query exist in the suggestions

General rules

- Max. 5 queries per session
- Unique queries only
- Terminate, if no suggestions are available
Assessing Query Suggestions for Search Session Simulation Evaluation

Automatic session detection

- Run session detection approach [Hagen et al. 2013] to identify whether consecutive queries belong to the same information need

Human authenticity assessment

- All sessions in one pool
- Manually assess each session as “real” or “simulated”

Human topicality assessment

- Manually label a session as “on topic”, when at least one query addresses at least one information need from the topic description
### Assessing Query Suggestions for Search Session Simulation

#### Evaluation: Automatic Session Detection (1)

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Sessions</th>
<th>Splits</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First suggestion*</td>
<td>64</td>
<td>1</td>
<td>1.6</td>
</tr>
<tr>
<td>Random suggestion*</td>
<td>65</td>
<td>2</td>
<td>3.1</td>
</tr>
<tr>
<td>Three word queries</td>
<td>20</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>TREC 2014 Session Track</td>
<td>1257</td>
<td>142</td>
<td>11.3</td>
</tr>
<tr>
<td>Webis-SMC-12</td>
<td>2882</td>
<td>217</td>
<td>7.5</td>
</tr>
</tbody>
</table>

(* one-query sessions were removed)

- Number of simulated or real sessions actually split by the session detection
- More splits = more query pairs seem to be unrelated
- More splits in real sessions, as human users are less predictable
Assessing Query Suggestions for Search Session Simulation
Evaluation: Automatic Session Detection (2)

<table>
<thead>
<tr>
<th>Query String</th>
<th>Time</th>
<th>Split</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Good split</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>well im writing it down but i’ll scribble it out lyrics</td>
<td>19:40:23</td>
<td>True</td>
</tr>
<tr>
<td>pictures in exhibition lyrics</td>
<td>19:57:21</td>
<td>False</td>
</tr>
<tr>
<td><strong>Bad split</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIV CHARITIES IN AFRICA</td>
<td>04:39:35</td>
<td>True</td>
</tr>
<tr>
<td>hiv charities africa</td>
<td>04:42:26</td>
<td>False</td>
</tr>
</tbody>
</table>

- Split marks the last query in a session
- Should be triggered by topic changes or timing
- False triggers: rephrasing, syntactic changes
Assessing Query Suggestions for Search Session Simulation

Evaluation: Manual Assessment of “Realism”

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Sessions</th>
<th>Real</th>
<th>Simulated</th>
<th>“Real” (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First suggestion*</td>
<td>64</td>
<td>62</td>
<td>2</td>
<td>96.9</td>
</tr>
<tr>
<td>Random suggestion*</td>
<td>65</td>
<td>62</td>
<td>3</td>
<td>95.4</td>
</tr>
<tr>
<td>Three word queries</td>
<td>20</td>
<td>17</td>
<td>3</td>
<td>85.0</td>
</tr>
<tr>
<td>TREC 2014 Session Track</td>
<td>50</td>
<td>49</td>
<td>1</td>
<td>98.0</td>
</tr>
<tr>
<td>Webis-SMC-12</td>
<td>50</td>
<td>50</td>
<td>0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

(* one-query sessions were removed)

- Manual judgments for all sessions whether they are simulated or real
- Majority judged as “real”, even in the simulated group
- “Real” in the upper group and “simulated” in the lower group indicate cases where the judge was mislead
Assessing Query Suggestions for Search Session Simulation
Evaluation: Manual Assessment of Topicality

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Sessions</th>
<th>On Topic</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First suggestion*</td>
<td>64</td>
<td>21</td>
<td>32.8</td>
</tr>
<tr>
<td>Random suggestion*</td>
<td>65</td>
<td>20</td>
<td>30.8</td>
</tr>
<tr>
<td>Three word queries</td>
<td>20</td>
<td>20</td>
<td>100.0</td>
</tr>
</tbody>
</table>

(* one-query sessions were removed)

- Number of simulated sessions judged as “on topic” with respect to the TREC topic description
Assessing Query Suggestions for Search Session Simulation
Evaluation: Example of a Well-formed Simulated Session

<table>
<thead>
<tr>
<th>Query String</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>air conditioning alternatives</td>
<td>15:05:53</td>
</tr>
<tr>
<td>air conditioning alternatives car</td>
<td>15:10:22</td>
</tr>
<tr>
<td>no <strong>air conditioning</strong> in car alternatives</td>
<td>15:11:07</td>
</tr>
<tr>
<td>how can i keep my car cool without <strong>ac</strong></td>
<td>15:15:28</td>
</tr>
<tr>
<td>ways to keep car cool without ac</td>
<td>15:21:16</td>
</tr>
</tbody>
</table>

**Strategy:** First suggestion

- Does look convincing to a human eye
  - Original query rephrased
  - **ac** abbreviation used
Assessing Query Suggestions for Search Session Simulation
Evaluation: Example of a Rather Odd Simulated Session

<table>
<thead>
<tr>
<th>Query String</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>air conditioning alternatives</td>
<td>17:31:54</td>
</tr>
<tr>
<td>no <strong>air conditioning</strong> alternatives</td>
<td>17:32:27</td>
</tr>
<tr>
<td>what to use instead of <strong>ac</strong></td>
<td>17:36:28</td>
</tr>
<tr>
<td>what to use instead of <strong>activator</strong></td>
<td>17:45:42</td>
</tr>
<tr>
<td>what can i use instead of <strong>activator for nails</strong></td>
<td>17:51:03</td>
</tr>
<tr>
<td>how to make nail activator</td>
<td>17:53:26</td>
</tr>
</tbody>
</table>

Strategy: Random suggestion

- Does look odd to a human eye
  - Quick topic drift from air conditioning to nail care
  - Cause: completion of **ac** to **activator**
We evaluated how authentic sessions simulated via query suggestions are

Both human assessment and automatic session detection are unable to distinguish suggestion-based sessions from real ones

Authenticity does not appear to be a main issue

Issues for suggestions-based session simulation:
- Small number of suggestions for each topic
- Staying on topic
Future Work

Data independence

- (Few) suggestions limit the flexibility and applicability
- Instead: query modifications that use language models

Influence on the topic

- Necessary for accurate session simulation
- Which part of the model makes these decisions?

User types and editing

- Follow known editing patterns
- Replicate queries that are typical for specific user groups or tasks
Assessing Query Suggestions for Search Session Simulation

Future Work

Data independence

- Suggestions limit the flexibility and applicability
- Instead: query modifications that use language models

Influence on the topic

- Necessary for accurate session simulation
- Which part of the model makes these decisions?

User types and editing

- Follow known editing patterns
- Replicate queries that are typical for specific user groups or tasks

Thank you!