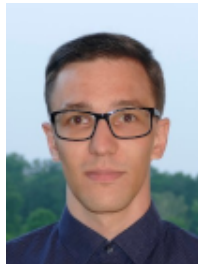


Assessing Query Suggestions for Search Session Simulation



Sebastian
Günther



Matthias
Hagen

Martin-Luther-Universität Halle-Wittenberg

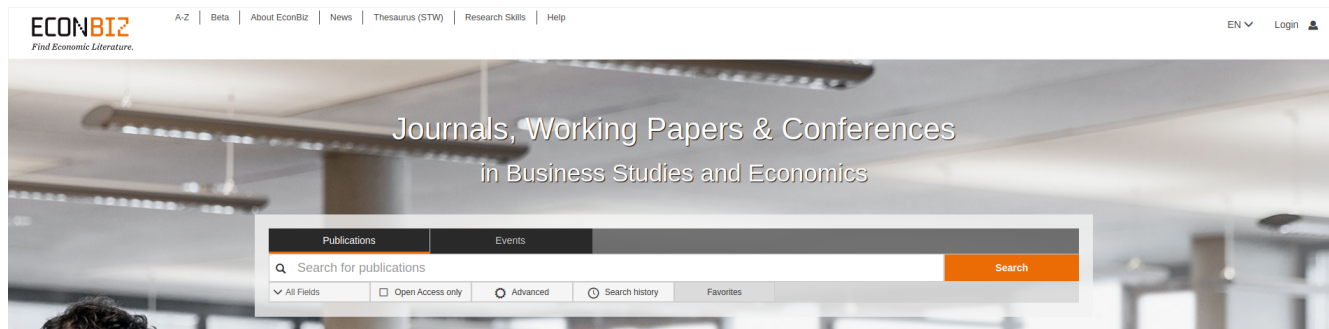
July 15, 2021

webis.de

Assessing Query Suggestions for Search Session Simulation

The SINIR Project

- ❑ SINIR: **S**imulating **I**nteractive **I**nformation **R**etrieval
- ❑ 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

Assessing Query Suggestions for Search Session Simulation

SINIR - Motivation (2)

- ❑ 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

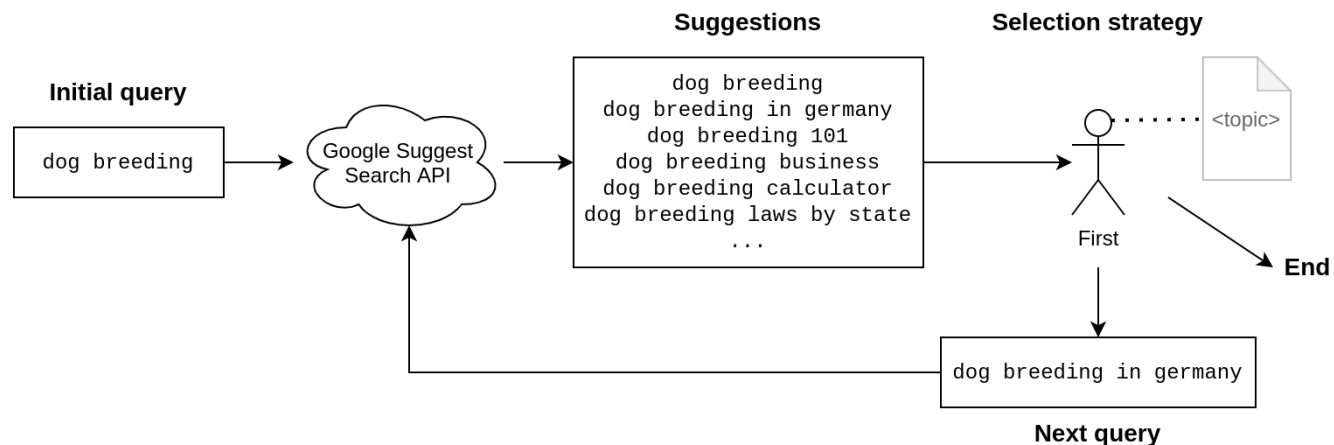
Can we create authentic sessions from suggestions only?

- ❑ 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)

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

(* 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)

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

- ❑ 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”

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

(* 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 misled

Assessing Query Suggestions for Search Session Simulation

Evaluation: Manual Assessment of Topicality

Strategy	Sessions	On Topic	(%)
First suggestion*	64	21	32.8
Random suggestion*	65	20	30.8
Three word queries	20	20	100.0

(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

Query String	Time
air conditioning alternatives	15:05:53
air conditioning alternatives car	15:10:22
no air conditioning in car alternatives	15:11:07
how can i keep my car cool without ac	15:15:28
ways to keep car cool without ac	15:21:16

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

Query String	Time
air conditioning alternatives	17:31:54
no air conditioning alternatives	17:32:27
what to use instead of ac	17:36:28
what to use instead of activator	17:45:42
what can i use instead of activator for nails	17:51:03
how to make nail activator	17:53:26

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`

Assessing Query Suggestions for Search Session Simulation

Conclusion

- ❑ 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

Assessing Query Suggestions for Search Session Simulation

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!