Assessing Query Suggestions for Search Session Simulation

Sebastian Günther[1], Matthias Hagen[1]

[1]Martin-Luther-Universität Halle-Wittenberg, Halle (Saale), Germany

Abstract

Research on simulating search behavior has mainly dealt with result list interactions in the recent years. We instead focus on the querying process and describe a pilot study to assess the applicability of search engine query suggestions to simulate search sessions (i.e., sequences of topically related queries). In automatic and manual assessments, we evaluate to what extent a session detection approach considers the simulated query sequences as "authentic" and how humans perceive the quality in the sense of coherence, realism, and representativeness of the underlying topic. As for the actual suggestion-based simulation, we compare different approaches to select the next query in a sequence (always selecting the first suggestion, random sampling, or topic-informed selection) to the human TREC Session track sessions and a previously suggested simulation scheme. Our results show that while it is easy to create query logs that are authentic to both users and automated evaluation, keeping the sessions related to an underlying topic can be difficult when relying on given suggestions only.

Keywords

Simulating query sequences, Search session simulation, Query suggestion, TREC Session track, Task-based search

1. Introduction

Many studies on the simulation of search behavior focus on using simulated user behavior in system evaluations—while others cover aspects of user modeling in general. Using simulated interactions for evaluation purposes is usually motivated by retrieval setups with no or only few actual users whose behavior can be observed and used to improve the actual system (e.g., system variants in digital libraries or new (academic) search prototypes without an established user base). Such few-user systems could also be evaluated in lab studies. But lab studies are difficult to scale up and also consume a lot of time since actual users need to be hired, instructed, and observed. In such situations, simulation promises a way out but the extent to which simulated search interactions can actually authentically replace real users in specific scenarios is still an open question. In the recent years, mostly result clicks or stopping decisions have been the focus of user modeling and simulation studies while simulating querying behavior has received less attention.

In this paper, we describe a pilot study on query simulation that aims to assess the suitability of stitching together query suggestions to form "realistic" search sessions (i.e., sequences of queries on the same information need that some human might have submitted). The scenario we address is inspired by typical TREC style evaluation setups where search topics are given as a verbal description of some information need along with a title or first query. To simulate some search session with a couple of queries, we examine sequences of query suggestions provided by some suggestion approach—in our pilot experiments, we simply use the suggestions that the Google search engine returns, but any other suggestion approach could also be applied. Starting with the actual title or the first query of a TREC topic, the second query for the session is selected among the suggestions for the first query, the third query is selected from the suggestions for the second query, etc.

Our research question is how such suggestion-based simulated sessions compare to real user sessions in the sense of coherence, realism, and representativeness of the underlying topic. In our pilot study, we thus let a human annotator assess human sessions from the TREC Session track mixed with sessions generated from suggestion sequences and sessions generated by a previous more static query simulation scheme. The results show that suggestion-based sessions replicate patterns commonly seen in query logs. Both humans and a session detection framework were unable to differentiate the simulated sessions from real ones. However, keeping close to the given topic when using suggestions as simulated queries is rather difficult. Among other reasons, the limited terminology in the topic, query, and suggestions and most importantly the relatively small amount of suggestions provided by the Google Suggest Search API often cause the session to drift away from the given topic.

2. Related Work

Similar to recent developments in the field of recommenders [1], simulation in the context of information retrieval often aims to support experimental evaluation of retrieval systems (e.g., in scenarios with few user interactions like in digital libraries) in a cost-gain scenario [2]

Causality in Search and Recommendation (CSR) and Simulation of Information Retrieval Evaluation (Sim4IR) workshops at SIGIR, 2021

sebastian.guenther@informatik.uni-halle.de (S. Günther); matthias.hagen@informatik.uni-halle.de (M. Hagen)

(cost for retrieval system interactions, gain for retrieving good results). Different areas of user behavior have been addressed by simulation: scanning snippets / result pages, judging document relevance, clicking on results, reading result documents, deciding about stopping the search, and query (re-)formulation itself. Some simulation studies combine different of these areas but some also just focus on a particular one. In this paper, we focus on the domain of simulating query (re-)formulation behavior. While quite a few studies on user click models and stopping decisions have been published in the recent years, query formulation is still perceived as difficult to simulate [3] but also necessary to generate useful simulations for interactive retrieval evaluation [4].

The existing approaches to query simulation can be divided into approaches that generate queries following rather static underlying schemes [5, 6, 7, 8, 9] and approaches that use language models constructed from the topic itself, from observed snippets, or from some result documents to generate queries of varying lengths [10, 11, 3, 12]. Not all, but most of the query simulations aim to simulate search sessions in the sense of query sequences that all have a similar intent [13, 14].

As for the static simulation schemes, many different ideas have been suggested. Jordan et al. [7] generate controlled sets of single-term, two-term, and multiple term queries for retrieval scenarios on the Reuters-21578 corpus by combining terms of selected specificity in the documents of the corpus (e.g., only highly discriminative terms to form very specific queries). Later studies have suggested to combine terms from manually generated query word pools and tested that on TREC topics. The respective querying strategies sample initial and subsequent query words from these pools and combine them to search sessions [5, 6, 8] following static schemes of for instance keeping the same two terms in every query but adding different third terms or for instance generating all possible three-permutations of three-term queries [6]. The suggested static schemes have been "idealized" from real searcher interactions [8] and have also been used in a later language modeling query simulator [12]. Similar to the mentioned keep-two-terms-but-vary-third-term query formulation strategy, Verberne et al. [9] create queries of n terms for the iSearch collection where $n-1$ terms are kept and the last term is varied to mimic academic information seeking behavior and to evaluate the cumulated gain over a simulated session.

One of the earliest more language model-based query simulators was suggested by Azzopardi et al. [10] in the domain of known-item search on the EuroGOV corpus (crawl of European government-related sites). Single queries for some given known-item document are generated from the term distribution within the document and some added "noise" to mimic imperfect human memory. The later InQuery system of Keskustalo et al. [15] used Bayesian inference networks to generate queries, Azzopardi [11] generated additional ad-hoc queries for existing TREC collections, while Carterette et al. [3] suggest a reformulation simulator to simulate whole sessions by also including the snippets from the seen result pages in the language model using TREC Session track data.

Some anchor text-based approaches to "simulate" complete query logs or to train query translation models also constitute a topic loosely related to ours [16, 17]. However, we aim to simulate shorter sequences of topically related queries instead of complete query logs. As for the simulation, we want to study in pilot experiments, whether and how well sequences of query suggestions stitched together may form search sessions. This idea is inspired by studies on query suggestions to support task-based search [18, 19] since more complicated tasks usually result in more interactions and queries from the respective users. Our research question thus is how "authentic" sessions can be that are formed from simply following suggestions up to some depth.

3. Query Log Generation

As described above, there are various types of datasets and models that have been suggested for query simulation. In this paper, we want to study a yet not covered source: query suggestions. Our reasoning is that query suggestions from large search engines are derived from their large query logs and thus represent "typical" user behavior. In our pilot experiments, we specifically focus on query suggestions provided by the Google Suggest Search API (that serves up to 10 suggestions at a time) but, in principle, any other suggestion approach could also be applied (e.g., suggestions from other large search engines or suggestion methods from the literature). Still, the characteristics of the suggestions may vary between different services such that the results of our pilot experiments should be tested in a more general setup with different suggestion approaches.

As our basis for simulated and real sessions, we use the TREC 2014 Session track dataset [20] containing 1021 sessions on 60 topics. Each topic is defined by an information need given as a short description. The respective sessions include (among other information) the queries some user formulated on the topic with timestamps, the shown snippets, and clicked results. We extract the first queries of the sessions as seed queries for the simulated sessions since the topics themselves do not have explicit titles that might be used as a first query. In addition to the TREC data we also sampled sessions from the Webis-SMC-12 dataset [21] that contains query sequences from the AOL log [22].

As suggestion-based session simulations, we consider the following three strategies in our pilot study.

First Suggestion. This strategy always selects the first suggestion provided by the Google Suggest Search API for the previous query of the session as input. A generated session will contain a maximum of four queries in addition to the original query (analyzing several query log datasets, the average sessions had up to five queries). A session might be terminated early if the API does not provide additional suggestions.

Random Suggestion. The random selection strategy randomly selects one of the suggestions provided by the Google Suggest Search API for the previous query of the session as input. Like with the first suggestion strategy, generated session contain up to four queries in addition to the original query. The same query can not appear back-to-back and a session might be terminated early if the API does not provide additional suggestions.

Three Word Queries (adapted). This strategy is based on the idea of the Session Strategy S3 described by Keskustalo et al. [8] which is also implemented in the SimIIR framework[1] as `TriTermQueryGenerator`. The original idea uses two terms as the basis extended by a third term selected from a topic description. We adapt this strategy with a few modifications. Initially we start with the original query from the real session without any additions. We then extract the 10 keywords from the topic's description with highest $tf \cdot idf$ scores (idf computed on the English Wikipedia). In each round, we calculate the cosine similarity of each suggestion and each original query–keyword pair. We select the suggestion that is closest to one of the query–keyword pairs. We limit the sessions to a maximum of four queries in addition to the original query. We also employ a dynamic threshold for the cosine similarity that stops accepting suggestions when the similarity falls below a certain threshold. Due to the varying length and specificity of the descriptions and the ambiguity of the topics, the threshold has to be manually adjusted for each topic. In our evaluation, we note that choosing an important term from the topic description provides an advantage to this strategy over the previous two with respect to the topic representativeness of the generated sessions.

For the three approaches, we generate 100, 100, and 20 sessions, respectively (in case of the *three word* strategy, the strict selection process and the small pool of suggestions often results in very short sessions such that we could only include 20 in the evaluation). While we mostly focus on the textual aspect of the queries in this paper, user session logs often come with additional information like user agent, user identification, IP address, date and time of the interaction. Each of our sessions consists of at least one query with a fixed user assigned to it. To run automatic session detection, we also simulate timestamps for each query submission.

Inter-Query Time. To simulate the time gap between query submissions, we have extracted the timings from user sessions from the Webis-SMC-12 dataset [21]. Our analysis shows that 25% of the time gap are shorter than 41 seconds, while half of the gaps is no longer than 137 seconds. The distribution of timings shows a peak at 8 seconds and a long tail with the highest values in the multi-hour range. To account for logging and annotation errors, we have removed outliers by deleting 10% of the longest gaps, which limits the simulated time between query submissions to no longer than 20 minutes. We use this remaining pool of time gaps to accurately reproduce the timing distribution for our generated sessions by randomly drawing values from it—which naturally then favors shorter time spans since they are more frequent.

Limits when using Suggestions. While working on our pilot study, we experimented with various combinations of suggestion selection strategies and session lengths. We identified issues in our strategies that are a direct result of the nature of search engine suggestions.

The first suggestion strategy is particularly prone to loops, when two queries are the top-ranked suggestions for each other—causing the generated session to alternate between two query strings; also observed for singular–plural pairs or categories (i.e., file formats, programming languages). To counter the looping issue, we use a unique query approach, which ensures that queries are not repeated in loops within a session. Additionally, another policy ensures a minimum dissimilarity between consecutive queries that helps to avoid plurals as top suggestions. However, while unique / dissimilar queries mitigate looping, we find that especially longer sessions (say, ten queries) narrow down to very specific topics. A possible reason is that today's search engine query suggestions do not only show related queries, but often offer more specific autocompletions. Further details on the evaluation are provided in Section 4.

4. Evaluation

In the evaluation, we compare the sessions generated by our three approaches to sessions from both the Webis-SMC-12 dataset and the TREC 2014 Session track. As a

[1] https://github.com/leifos/simiir

Strategy	Sessions	Splits
First suggestion*	64	1
Random suggestion*	65	2
Three word queries	20	0
TREC 2014 Session Track	1257	142
Webis-SMC-12	2882	217

Table 1
Number of within-session splits the automatic session detection introduced for simulated and real sessions (more splits mean more query pairs seem to be unrelated; * indicates that one-query sessions were removed).

Strategy	Sessions	Real	Simulated
First suggestion*	64	62	2
Random suggestion*	65	62	3
Three word queries	20	17	3
TREC 2014 Session Track	50	49	1
Webis-SMC-12	50	50	0

Table 2
Manual judgments for all sessions whether they are simulated or "real" (* indicates that one-query sessions were removed). "Real" in the upper group and "simulated" in the lower group indicate cases where the judge was mislead.

first step, we perform an automated evaluation by running the sessions through the session detection approach of Hagen et al. [21]. Ideally, the simulated sessions should not be split by the session detection in order to count as "authentic". In a second step, a human assessor looked at the simulated sessions as well as original sessions and had to judge whether a session seems to be simulated or of human origin. In a third step, a human assessor judged whether a session actually covers the intended information need given by the topic description.

4.1. Automatic Session Detection

The goal of a session detection system is to identify consecutive queries as belonging to the same information need or not. When a consecutive pair is detected that seems to belong to two different information needs, a split is introduced. Later some of these sessions might be run through a mission detection to identify non-consecutive sessions that belong to the same search task, etc.

As an automatic evaluation of the the simulated sessions' authenticity, we individually run each simulated session and the individual sessions from the TREC and Webis-SMC-12 data through the session detection approach of Hagen et al. [21]. A simulated or original session "passes" the automatic authenticity test iff the detection approach does not introduce a split. The results are shown in Table 1 (sessions with only one query were removed since they will never be split).

Altogether, the simulated sessions are hardly split by the automatic detection. The one wrong split for the first suggestion strategy and one wrong split for the random suggestion strategy are likely due to the first query being uppercased while the subsequent suggestions are lowercased, while the second "wrong" split for the random suggestions strategy is likely caused by a reformulation with abbreviation and no term overlap ("no air conditioning alternatives" to "what to use instead of ac"). These examples serve as a good demonstration for the limitations of a fully automatic authenticity evaluation such that we also manually assess the simulated sessions.

4.2. Human Authenticity Assessment

An automated session detection system only "assesses" whether the consecutive queries seem to belong together based on factors like lexical or semantic similarity and time gaps. However, we want to complement this purely automatic relatedness detection by a manual assessment of how "authentic" the simulated sessions are perceived by humans, i.e., whether a human can distinguish simulated from real sessions.

Procedure. All simulated sessions and a sample of original sessions are combined into one session pool. The sessions are then presented to the judge as kind of log excerpts with user ID, timestamps, and queries. The judge has no accurate knowledge about the amount of queries for each approach and there is no obvious way to determine the source of a session. The judge then labels each session as *real* (sampled from actual query logs) or *simulated* (by one of the three approaches). The results in Table 2 indicate that the simulated sessions are perceived as real even though the assessor was told that some sessions actually are simulated.

During the assessment, the assessor took notes of which features of a session or query determine the judgment. This helps us in understanding how humans and algorithms may come up with different verdicts. The primary criteria for the relatedness of two queries are their term composition and length. Similarities in those aspects are perceived as patterns. This is also true for small editing actions (adding or replacing single words) which naturally comes with the specialization towards a topic. The opposite effect is perceived for rapid topic changes. When multiple closely related tasks have to be fulfilled within one session, there may be large changes from query to query. This is also true for replacing words by synonyms or abbreviations. While a human judge will usually be able to infer context to those rapid changes, an automatic process is more likely to detect a new session. Another discrepancy between human and algorithmic evaluation becomes apparent when we consider outlier

Strategy	Sessions	On Topic
First suggestion*	64	21
Random suggestion*	65	20
Three word queries	20	20

Table 3
Number of simulated sessions judged as "on topic" with respect to the TREC topic description (* indicates that one-query sessions were removed).

behavior like text formatting (e.g., all-uppercase) that a human might be able to judge as a simple typing error while a detection approach without lowercasing preprocessing might be mislead.

In a nutshell, while both humans and algorithms look for patterns in the sessions and queries, the human judge does so more selective by looking for mistakes. If found, the type of a mistake usually heavily influences the assessment of a session. Finally, note that due to the nature of the three word query strategy there might be a chance for an informed human to guess the sessions origin.

4.3. Human Topicality Assessment

So far, we have shown that the authenticity of a session is largely influenced by its term composition and appearance. However, to serve as a replacement for humans, a session generator not only has to provide sessions that a detection approach or some human would assess as authentic, but also has to simulate sessions that follow the topic given as part of the evaluation study.

Procedure. Determining if a session or query is on topic is a non-trivial task. While a query like "car" overlaps with the topic "find information on used car prices", it does not address the information need formulated in the topic description. We therefore set the following criteria to evaluate if a session is "on topic": A session is "on topic", if the last query addresses at least one information need formulated in the topic description or shows clear signs that the session is headed in that direction—such that very short sessions are more likely to be on topic. A session is also "on topic", if any query of the session addresses at least one information need formulated in the topic description—necessary condition to account for topics with multiple subtasks.

Hypothesis: The *first* and *random* approach do not take the topic into account. Both strategies simply converge to anything the search suggestion API provides for the initial query. Instead, the *three word* approach makes informed decisions when choosing suggestions and should therefore be able to stay more "on topic".

Query String	Time
First suggestion	
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
Random suggestion	
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
Random suggestion	
Philadelphia	03:31:29
philadelphia cheese	03:34:50
philadelphia cheese recipes	03:35:05
philadelphia cheese recipes salmon pasta	03:53:17

Table 4
Example sessions with unusual editing patterns.

Results: We have manually judged all generated sessions. The results are shown in Table 3 show that even the uninformed strategies stay "on topic" on about one third of the sessions. This can largely be attributed to the nature of the TREC Session track topics that often contain several subtasks. Sessions generated by the *three word* strategy stay "on topic" even more.

4.4. Notable Examples

As part of the judgment process, we have also taken note of simulated sessions which contain conspicuous editing steps or queries. The examples in Table 4 include a positive and a negative example with respect to authenticity.

The first example was judged as "real" based on the usage of an abbreviation for air conditioning in the fourth query. The replacement of terms or groups of terms with a common abbreviation might be seen as a typical step for a human user after gaining more insight into a topic. The second example includes an issue that was caused by the autocomplete feature of the Google suggestions: the abbreviation 'ac' was falsely extended to the term 'activator', which ultimately changed the subject of the session. The third example shows a very common issue of ambiguous first queries. For the first and random suggestion strategies, there is no way to determine that a city is referenced in this example such that the session quickly diverges to the food domain.

4.5. Long Sessions

The simulated sessions up to this point had parameters like session length and inter-query time been set to values that deemed appropriate in some initial experiments on our end in order to generate "close to real" sessions. We also did not include navigational queries or known-item searches, which often could result in either very short or very long sessions. To investigate the applicability of our approaches to such outlier behavior we have also further assessed some sessions with up to 20 queries.

In many of the cases without imposing any limits on the generation process, the sessions still were often terminated early due to a lack of suggestions. This was mostly caused by two reasons: either the query became too specific to still yield additional suggestions or the pool of unique and dissimilar queries was used up. In cases where long sessions could actually be generated, the session usually quickly was rather specific and diverged substantially from the actual given topic towards the end of the session.

Using a different set of more technically oriented topics, we were able to generate longer sessions more frequently. For this to work, we had to limit the dissimilarity filter, as abbreviations within the query were more frequent and therefore editing distances were smaller. We also observed that queries from this field were mostly comprised of categorical keywords stitched together compared to the more natural looking sessions from standard query logs.

Those observations, while helping to shape our pilot study, show that parameters and strategies for authentic session generation are a very dynamic and potentially also topic-specific issue.

5. Conclusion

In this paper, we have investigated how well authentic sessions can be simulated using web search engine query suggestions. By employing different strategies of selecting and combining the suggestions, we showcased the potential but also the limits of the overall usefulness of suggestion-based session simulation. Our evaluation showed that both humans and a session detection framework are unable to distinguish suggestion-based sessions from sampled real sessions. While some kind of authenticity can thus be attributed to the simulated sessions, staying on topic proved to be rather difficult. Addressing the outlined shortcomings is an interesting direction for future work. We plan to continue investigating query simulation as follows.

Data Independence. Relying on suggestions as query candidates limits the flexibility and applicability of the simulated sessions. We will work on query modifications that include "knowledge" from language models or predefined editing rules.

Influence on the Topic. For accurate session simulation, it is necessary to influence the topic that the queries follow. We will evaluate how and where those decisions have to be made to create an effective user model.

User Types and Editing. Since query modifications often follow well-known patterns, we will also investigate ways to replicate editing patterns in simulated queries that are typical for specific user groups or tasks.

Acknowledgments

This work has been partially supported by the DFG through the project "SINIR: Simulating INteractive Information Retrieval" (grant HA 5851/3-1).

References

[1] S. Zhang, K. Balog, Evaluating conversational recommender systems via user simulation, in: R. Gupta, Y. Liu, J. Tang, B. A. Prakash (Eds.), KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020, ACM, 2020, pp. 1512–1520. URL: https://doi.org/10.1145/3394486.3403202. doi:10.1145/3394486.3403202.

[2] M. McGregor, L. Azzopardi, M. Halvey, Untangling cost, effort, and load in information seeking and retrieval, in: F. Scholer, P. Thomas, D. Elsweiler, H. Joho, N. Kando, C. Smith (Eds.), CHIIR '21: ACM SIGIR Conference on Human Information Interaction and Retrieval, Canberra, ACT, Australia, March 14-19, 2021, ACM, 2021, pp. 151–161. URL: https://doi.org/10.1145/3406522.3446026. doi:10.1145/3406522.3446026.

[3] B. Carterette, A. Bah, M. Zengin, Dynamic test collections for retrieval evaluation, in: J. Allan, W. B. Croft, A. P. de Vries, C. Zhai (Eds.), Proceedings of the 2015 International Conference on The Theory of Information Retrieval, ICTIR 2015, Northampton, Massachusetts, USA, September 27-30, 2015, ACM, 2015, pp. 91–100. URL: https://doi.org/10.1145/2808194.2809470. doi:10.1145/2808194.2809470.

[4] B. Carterette, E. Kanoulas, E. Yilmaz, Simulating simple user behavior for system effectiveness evaluation, in: C. Macdonald, I. Ounis, I. Ruthven (Eds.), Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM

2011, Glasgow, United Kingdom, October 24-28, 2011, ACM, 2011, pp. 611–620. URL: https://doi.org/10.1145/2063576.2063668. doi:10.1145/2063576.2063668.

[5] F. Baskaya, H. Keskustalo, K. Järvelin, Time drives interaction: Simulating sessions in diverse searching environments, in: W. R. Hersh, J. Callan, Y. Maarek, M. Sanderson (Eds.), The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012, ACM, 2012, pp. 105–114. URL: https://doi.org/10.1145/2348283.2348301. doi:10.1145/2348283.2348301.

[6] F. Baskaya, H. Keskustalo, K. Järvelin, Modeling behavioral factors ininteractive information retrieval, in: Q. He, A. Iyengar, W. Nejdl, J. Pei, R. Rastogi (Eds.), 22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013, ACM, 2013, pp. 2297–2302. URL: https://doi.org/10.1145/2505515.2505660. doi:10.1145/2505515.2505660.

[7] C. Jordan, C. R. Watters, Q. Gao, Using controlled query generation to evaluate blind relevance feedback algorithms, in: G. Marchionini, M. L. Nelson, C. C. Marshall (Eds.), ACM/IEEE Joint Conference on Digital Libraries, JCDL 2006, Chapel Hill, NC, USA, June 11-15, 2006, Proceedings, ACM, 2006, pp. 286–295. URL: https://doi.org/10.1145/1141753.1141818. doi:10.1145/1141753.1141818.

[8] H. Keskustalo, K. Järvelin, A. Pirkola, T. Sharma, M. Lykke, Test collection-based IR evaluation needs extension toward sessions - A case of extremely short queries, in: G. G. Lee, D. Song, C. Lin, A. N. Aizawa, K. Kuriyama, M. Yoshioka, T. Sakai (Eds.), Information Retrieval Technology, 5th Asia Information Retrieval Symposium, AIRS 2009, Sapporo, Japan, October 21-23, 2009. Proceedings, volume 5839 of *Lecture Notes in Computer Science*, Springer, 2009, pp. 63–74. URL: https://doi.org/10.1007/978-3-642-04769-5_6. doi:10.1007/978-3-642-04769-5_6.

[9] S. Verberne, M. Sappelli, K. Järvelin, W. Kraaij, User simulations for interactive search: Evaluating personalized query suggestion, in: A. Hanbury, G. Kazai, A. Rauber, N. Fuhr (Eds.), Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015, Proceedings, volume 9022 of *Lecture Notes in Computer Science*, 2015, pp. 678–690. URL: https://doi.org/10.1007/978-3-319-16354-3_75. doi:10.1007/978-3-319-16354-3_75.

[10] L. Azzopardi, M. de Rijke, K. Balog, Building simulated queries for known-item topics: An analysis using six european languages, in: W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, N. Kando (Eds.), SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007, ACM, 2007, pp. 455–462. URL: https://doi.org/10.1145/1277741.1277820. doi:10.1145/1277741.1277820.

[11] L. Azzopardi, Query side evaluation: An empirical analysis of effectiveness and effort, in: J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, J. Zobel (Eds.), Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009, ACM, 2009, pp. 556–563. URL: https://doi.org/10.1145/1571941.1572037. doi:10.1145/1571941.1572037.

[12] D. Maxwell, L. Azzopardi, K. Järvelin, H. Keskustalo, Searching and stopping: An analysis of stopping rules and strategies, in: J. Bailey, A. Moffat, C. C. Aggarwal, M. de Rijke, R. Kumar, V. Murdock, T. K. Sellis, J. X. Yu (Eds.), Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015, ACM, 2015, pp. 313–322. URL: https://doi.org/10.1145/2806416.2806476. doi:10.1145/2806416.2806476.

[13] R. Jones, K. L. Klinkner, Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs, in: J. G. Shanahan, S. Amer-Yahia, I. Manolescu, Y. Zhang, D. A. Evans, A. Kolcz, K. Choi, A. Chowdhury (Eds.), Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008, ACM, 2008, pp. 699–708. URL: https://doi.org/10.1145/1458082.1458176. doi:10.1145/1458082.1458176.

[14] A. H. Awadallah, X. Shi, N. Craswell, B. Ramsey, Beyond clicks: query reformulation as a predictor of search satisfaction, in: Q. He, A. Iyengar, W. Nejdl, J. Pei, R. Rastogi (Eds.), 22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013, ACM, 2013, pp. 2019–2028. URL: https://doi.org/10.1145/2505515.2505682. doi:10.1145/2505515.2505682.

[15] H. Keskustalo, K. Järvelin, A. Pirkola, Evaluating the effectiveness of relevance feedback based on a user simulation model: Effects of a user scenario on cumulated gain value, Inf. Retr. 11 (2008) 209–228. URL: https://doi.org/10.1007/s10791-007-9043-7. doi:10.1007/s10791-007-9043-7.

[16] V. Dang, W. B. Croft, Query reformulation using anchor text, in: B. D. Davison, T. Suel, N. Craswell, B. Liu (Eds.), Proceedings of the Third International Conference on Web Search and Web Data Mining,

WSDM 2010, New York, NY, USA, February 4-6, 2010, ACM, 2010, pp. 41–50. URL: https://doi.org/10.1145/1718487.1718493. doi:10.1145/1718487.1718493.

[17] N. Craswell, B. Billerbeck, D. Fetterly, M. Najork, Robust query rewriting using anchor data, in: S. Leonardi, A. Panconesi, P. Ferragina, A. Gionis (Eds.), Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013, ACM, 2013, pp. 335–344. URL: https://doi.org/10.1145/2433396.2433440. doi:10.1145/2433396.2433440.

[18] D. Garigliotti, K. Balog, Generating query suggestions to support task-based search, in: N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries, R. W. White (Eds.), Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017, ACM, 2017, pp. 1153–1156. URL: https://doi.org/10.1145/3077136.3080745. doi:10.1145/3077136.3080745.

[19] H. Ding, S. Zhang, D. Garigliotti, K. Balog, Generating high-quality query suggestion candidates for task-based search, in: G. Pasi, B. Piwowarski, L. Azzopardi, A. Hanbury (Eds.), Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings, volume 10772 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 625–631. URL: https://doi.org/10.1007/978-3-319-76941-7_54. doi:10.1007/978-3-319-76941-7_54.

[20] B. Carterette, E. Kanoulas, M. M. Hall, P. D. Clough, Overview of the TREC 2014 session track, in: E. M. Voorhees, A. Ellis (Eds.), Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014, volume 500-308 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2014. URL: http://trec.nist.gov/pubs/trec23/papers/overview-session.pdf.

[21] M. Hagen, J. Gomoll, A. Beyer, B. Stein, From search session detection to search mission detection, in: J. Ferreira, J. Magalhães, P. Calado (Eds.), Open research Areas in Information Retrieval, OAIR '13, Lisbon, Portugal, May 15-17, 2013, ACM, 2013, pp. 85–92. URL: http://dl.acm.org/citation.cfm?id=2491769.

[22] G. Pass, A. Chowdhury, C. Torgeson, A picture of search, in: X. Jia (Ed.), Proceedings of the 1st International Conference on Scalable Information Systems, Infoscale 2006, Hong Kong, May 30-June 1, 2006, volume 152 of *ACM International Conference Proceeding Series*, ACM, 2006, p. 1. URL: https://doi.org/10.1145/1146847.1146848. doi:10.1145/1146847.1146848.