LISP – A Rich Interaction Dataset and Loggable Interactive Search Platform

Jana Isabelle Friese[1]✉[0009−0005−2483−0476], Andreas Konstantin Kruff[2][0009−0002−8350−154X], Philipp Schaer[2][0000−0002−8817−4632], Norbert Fuhr[1][0000−0002−0441−6949], and Nicola Ferro[3][0000−0001−9219−6239]

[1] University of Duisburg-Essen, Germany
{jana.friese,norbert.fuhr}@uni-due.de
[2] TH Köln - University of Applied Sciences, Germany
{andreas.kruff,philipp.schaer}@th-koeln.de
[3] University of Padua, Italy
nicola.ferro@unipd.it

Abstract. We present a reusable dataset and accompanying infrastructure for studying human search behavior in Interactive Information Retrieval (IIR). The dataset combines detailed interaction logs from 61 participants (122 sessions) with user characteristics, including perceptual speed, topic-specific interest, search expertise, and demographic information. To facilitate reproducibility and reuse, we provide a fully documented study setup, a web-based perceptual speed test, and a framework for conducting similar user studies. Our work allows researchers to investigate individual and contextual factors affecting search behavior, and to develop or validate user simulators that account for such variability. We illustrate the dataset's potential through an illustrative analysis and release all resources as open-access, supporting reproducible research and resource sharing in the IIR community.

Keywords: IIR · Interaction Dataset · Study Setup · User Aspects

1 Motivation

Resource sharing and re-use are essential for advancing research, ensuring transparency, and fostering community collaboration in Interactive Information Retrieval (IIR) [20]. While a number of IIR studies build on existing datasets, only a small fraction make their own data publicly available, and even fewer share complete experimental setups [9]. This mismatch between interest in reusing resources and their availability limits cumulative progress and highlights the need for more accessible, reusable datasets.

A larger pool of reusable resources would directly support IIR research. While user studies provide valuable insights into real-world scenarios beyond system performance [23], human search behavior varies widely, and effect sizes are often small. Still, most IIR studies rely on limited samples, undermining power and reproducibility. User simulation offers a scalable alternative [7], but click

models mostly rely on simplified assumptions about user behavior [15] and, even in advanced frameworks, lack systematic validation against real behavior [45]. Reliable baselines of authentic user behavior can enable reliable comparison and validation of such models, enhancing progress in IIR research.

However, in existing session interaction datasets some crucial information is often missing. Liu and Shah [27] note that participant variables, which may significantly affect results, are under-reported in IIR studies. Gäde et al. [20] identify three main resource types that should be documented and shared to enable effective reuse: (1) *research design*, (2) *infrastructure*, and (3) *data*. In practice, however, sharing typically focuses on the data alone, with limited attention to the other components. In particular, infrastructures——such as user interfaces and logging frameworks——are rarely reused but instead redeveloped from scratch, as they are difficult to adapt to new research scenarios [21].

Our work addresses this gap by introducing a comprehensive, reusable resource that integrates all three main resource types outlined by Gäde et al. The dataset includes detailed interaction logs enriched with additional participant information (i.e., the *data*), an extensively documented study setup (*design*), and the complete infrastructure for running and adapting the experiment (*infrastructure*). To ensure high reusability, we aligned our materials with the highest of Gäde et al.'s five levels of reusability standards, which require structured and openly documented archival of all three resource types. Because infrastructure often poses the greatest challenge for reuse, we also provide detailed instructions for adapting our setup to different scenarios.[4]

To build the resource, we conducted a user study focusing on perceptual speed and interest in an argument retrieval task, as prior work has shown that both cognitive abilities [2,3,11,6] and contextual factors [35,33,43,14] shape user behavior. We therefore selected perceptual speed and interest as representative dimensions of individual and situational factors. Including these factors makes the dataset suitable not only for studying human search behavior but also for developing and validating user simulators that account for such variability. Alongside the interaction logs, we provide each participant's perceptual speed scores, topic-specific interest ratings, search expertise, and demographic information.

Our study followed an exploratory design, enabling a wide range of research questions to be addressed. This paper includes an illustrative analysis of the effects of perceptual speed and interest on user behavior, demonstrating how the dataset can be used. The analysis both characterizes the collected data and highlights its potential for developing and evaluating simulators that adapt to different user types or contexts. In doing so, we also discuss the broader implications of these factors for simulation and user-centric evaluation in IIR.

Overall, the dataset and accompanying infrastructure offer the IIR community: (1) a robust baseline of human search behavior accounting for demographics, cognitive abilities, and situational factors; (2) a fully documented, adaptable study environment; and (3) a practical example of designing and sharing reusable resources that adhere to current best-practice standards.

[4] Resources accessible at: https://github.com/irgroup/LISP_Dataset_and_Platform

2 Related Work

To position our dataset within the current landscape, we conducted a literature review on existing session log resources and their characteristics. As noted by Reimer et al. [36], several large-scale query logs have been collected over the years; however, most focus on individual queries rather than full sessions and do not include click data. Many are also no longer accessible, underscoring the scarcity of suitable datasets and the challenge of long-term reproducibility.

Well-established log datasets, such as the TREC Session Track collections [12], remain publicly available and provide high-quality session-based interaction data from controlled lab studies. While being valuable for research, they are limited in size. In contrast, logs collected from real-world search applications enable the creation of much larger datasets. One example is TripClick [37], which offers extensive click and ranking data but lacks any information about users or their underlying information needs. User- and context-related factors, however, strongly influence search behavior. Yet, even in lab-based studies where such information is typically collected, these details are rarely published.

Detailed information about user characteristics and context is crucial not only for understanding search behavior but also for ensuring reproducibility. In recent years, there has been growing interest in how individual differences and motivational factors shape search behavior. Among the many traits studied, perceptual speed and task-specific interest have frequently been included, highlighting their relevance for modeling user interactions [2,11,42,4,6,24,17,28,22]. Without reporting these factors, valuable insights are lost, and the usefulness of datasets is limited. Additional constraints, such as language and domain coverage, further restrict the applicability of existing resources.

Table 1 provides an overview of publicly available session-based interaction log datasets, highlighting characteristics such as number of logged sessions, domain, language, and collection environment. The language column (Lang.) indicates the primary language of the dataset, as many logs—especially from real-world settings—contain queries in multiple languages.

Table 1. Summary of publicly available session-based interaction log datasets. The last row shows our own dataset.

Ref	User Profiles	Domain	Lab Setting	Year	# Logs	Lang.
Yandex [39]	✗	Mixed	✗	2011	797,867	en
TREC Session[12]	✗	Mixed	✓	2011–2014	1,564	en
TripClick [37]	✗	Medical	✗	2013–2020	1.6M	en
AOL [34]	✗	Mixed	✗	2006	283,207	en
Baidu-ULTR [46]	✗	Mixed	✗	2022	1.2B	ch
SoguoQ [41]	✗	Mixed	✗	2008	14.1M	ch
TianGong-ST [13]	✗	Mixed	✗	2015	147,155	ch
SUSS [30]	✗	Academic	✗	2014–2015	484,449	en/ger
Own	✓	Argument	✓	2025	122	en

While some datasets contain extensive click data, they reveal very little about the users: outside controlled lab settings, information about users' underlying information needs is missing, and no existing dataset provides comprehensive details on user or contextual factors.

This aligns with the observations reported by Crasswell et al. [16]: privacy concerns and the sensitivity of user-entered data often prevent researchers from sharing their logs, leaving few publicly available click datasets suitable for research. This further underscores the limited availability of resources that support meaningful comparison or reproduction of studies.

The scarcity of datasets is mirrored by the limited availability of shared frameworks for collecting and analyzing interaction data. Several logging frameworks have been proposed [29,38,25,8], and initial solutions for RAG systems have emerged [26]. However, fully integrated end-to-end frameworks remain rare, and most researchers rely on custom implementations. A notable exception is Podify [31], which provides standardized logging and reproducible experiments, but its focus on podcast streaming limits applicability to other document types or search tasks. This lack of accessible infrastructure aligns with bibliometric findings from Bogers et al. [9], who reported that only a small fraction of CHIIR authors shared research data or infrastructure components, reflecting a broader reproducibility gap in the field.

3 Resource Design and Development

To create the dataset, we conducted a user study and collected detailed interaction logs together with extensive participant profiles. We used a within-subject study design to investigate the effects of topic interest—i.e., a cognitive-emotional relationship between an individual and a topic [40]—on search behavior. Participants were asked to prepare for writing an opinion essay through an exploratory search, gathering arguments for both sides of a given debate. Each participant worked on two topics: one of high and one of no personal interest, making topic interest the within-subject variable. In addition, perceptual speed (PS)—defined as "the speed in finding figures, making comparisons, and carrying out other very simple tasks involving visual perception" [18]—was included as a between-subject variable to assess its impact on search behavior.

The study consisted of three parts: (1) Participants completed a pre-study questionnaire, including the Perceptual Speed Test, both administered remotely. (2) They then attended the main study session, (3) followed by a post-study questionnaire; these latter steps were conducted in person.

3.1 Dataset and System

For the experimental setup, we used the Conversational Argument Retrieval dataset from Touché 2020 [10], which is based on the args.me corpus of Debate.org threads. The corpus contains 387,606 arguments across 50 TREC-style topics, collected in mid-2019. We selected this corpus because its controversial

topics are well suited to elicit user interest, increasing the likelihood of finding topics with different levels of interest. In addition, the relatively short argumentative texts reduce reading time and encourage more frequent system interactions within the ten-minute session limit. Topics that were unlikely to be relatable to the participants (e.g., heavily US-centric issues) were excluded, as well as topics where one side of the debate was underrepresented, resulting in a set of 26 topics. From the corresponding arguments, we removed overly short or uninformative entries (<150 characters) as well as excessively long texts (>3,000 characters) that could distort reading time, leaving 226,468 documents in the collection.

To provide a familiar search experience, we implemented a simple search system using Terrier's BM25 ranking, allowing users to submit queries, view retrieved arguments, save them as supporting or opposing a stance, and review previously saved documents. Results were shown in pages of ten arguments, each presented with a title and a snippet, expandable to the full text on click. Snippets consisted of the first 200 characters of an argument. As the dataset does not include titles, we generated them using Llama3 (temperature = 0.3) based on the argument text and included them in the GitHub repository with the system. The current topic was displayed alongside the results at all times.

3.2 Data Collection

Participant Recruitment For the study, we recruited information science students from a German university. Participation was fully voluntary, with informed consent obtained after participants were briefed on the study purpose, data handling, and their rights to reject the publication of their data. Students were compensated for their participation with course credits. All study steps were conducted under full anonymity, using pseudonyms to ensure that no personal data (as defined by the GDPR) was collected. Pseudonyms were further standardized during post-processing to strengthen data protection.

Questionnaires Before the search sessions, participants completed a pre-study questionnaire on demographics, socioeconomic background, online activity, and search engine experience (see 4.3 for further details). They also rated their interest in the 26 available topics.

At the end of the study, participants had to complete a post-study questionnaire. They were asked whether their stance on the topic had changed during the process, and if so, how it had changed. In addition, participants were asked to rate the difficulty of completing the tasks.

Perceptual Speed Test As part of the pre-study questionnaire, participants completed a perceptual speed test in their own environment to avoid potential lab effects. Several perceptual speed tests have been reviewed in the context of IR [19]. Following Azzopardi et al. [6], we used a modified Finding A's test, in which participants identified ε and ¥ characters, as this method has been shown to be a more accurate test of users actual perceptual speed [1].

After instructions, participants completed a 30-second practice phase. In the test, participants classified strings of 10–15 non-alphanumeric characters—presented one at a time—as containing both the symbols ε and ¥ (pressing j) or not (pressing n). The time limit was two minutes. All participants saw the same fixed order of strings, with a maximum of 500 provided to ensure comparability and prevent ceiling effects.

Study Procedure Based on the pre-study questionnaire, each participant was assigned two topics: one of high and one of no interest to them. To avoid ordering effects, half of the participants started with the high-interest and half with the no-interest topic. Topics were assigned using a heuristic pseudo-random approach to ensure a balanced coverage across participants, preventing any single topic from dominating either interest category. The task was identical for both topics. After an introduction explaining the procedure and task, the participants started the study on the provided computers. Each task had a ten-minute time limit. Upon completing the first task, they were automatically directed to the second. Before each task, participants were shown the full topic and task description. After completion, they could optionally explain why they ended their search before the time limit. These explanations are included in the dataset in their original form to avoid bias; for responses written in German, a machine-translated version is also provided (based on Google Translate API). Upon finishing both tasks, participants proceeded to the post-study questionnaire.

4 Resource Description

Building on the study design described in Section 3, this section documents the resources we release: (1) the interaction log dataset, (2) the loggable interactive search platform, and (3) the user profile data (demographics, perceptual speed, and search expertise). A more comprehensive documentation is available on the project website[5] to support reproducibility and adhere to the Level-5 reusability standard.

4.1 Interaction Log Dataset

The interaction dataset collected in the user study comprises 122 session log files, evenly divided between high-and no-interest topics. Each interaction contains an interaction type, a timestamp, and session identifier.

Events captured include query submissions and reformulations, document interactions (such as expanding or collapsing views, and selecting stances), navigating between result pages, and task completion actions. Each document entry is associated with its rank, retrieval score, and length. Stopping decisions were explicitly logged, and participants could provide a reason for ending their search early. Although each task was designed for a ten-minute duration, this limit was not hard-coded, so some sessions slightly exceeded this timeframe.

[5] https://irgroup.github.io/LISP_Dataset_and_Platform/

4.2 lisp - A Loggable Interactive Search Platform

For the study, we developed a loggable interactive search platform called lisp. While originally designed for argument retrieval tasks, the platform can be easily adapted to other search scenarios.

The interface provides a typical search experience: participants can enter queries, browse retrieved arguments, navigate between pages, expand snippets to view the full text, and save documents as supporting or opposing a stance. A side panel summarizes the task and topic, and tracks the number of labeled documents (Figure 1). At the end of each session, a pop-up window presents an overview of all saved documents with their assigned labels and titles, with stance indicated through color-coding (i.e., green for supporting, red for opposing).

For each session, all user interactions described in 4.1 are logged, and a logfile is created. Logging begins when a task is started and ends when the participant confirms completion. Logfiles are named using the participant ID, task number, and timestamp.

In the current workflow, participants first see a welcome page, then enter their username, and proceed through two successive search tasks before finishing on a page linking to the post-study questionnaire.

The platform is designed for easy adaptation: elements such as buttons, result presentation, logging output, and datasets can all be customized with minimal changes. Detailed instructions are provided in the repository.

The repository additionally includes the implementation of the Perceptual Speed Test. Running the test requires a MySQL database for storing results, and the GitHub README provides instructions on the necessary implementation changes. A demo version[6] of the Perceptual Speed Test is also provided, which allows for testing the application without setting up a database.

[6] https://andykruff.github.io/demo-ps-test/

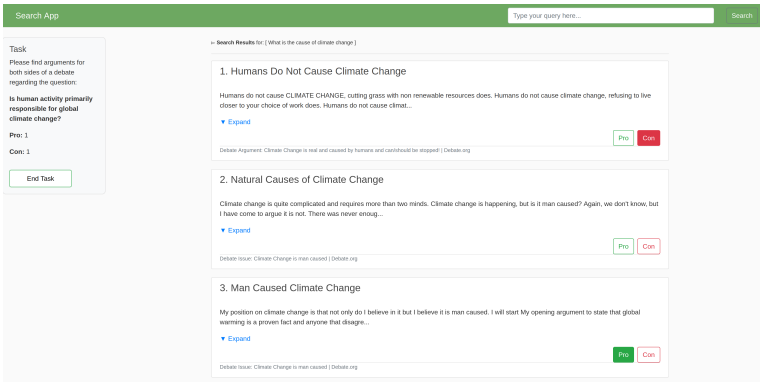

Fig. 1. Screenshot of the search interface of lisp

4.3 User Profile Data

In addition to interaction logs, the dataset includes user profiles with demographics, search experience, and perceptual speed scores. Figure 2 summarizes the demographic and experience data. As all participants were recruited from the same university course, the sample is relatively homogeneous.

Perceptual speed results (Figure 3) show an average score of 101.66 with a median of 103 and a standard deviation of 28.5, indicating substantial variability within the group. The dataset also reports detailed response patterns: the number of j- and n-key presses (indicating both symbols were perceived or not) and the correctness of these responses. Failure rates were low, with 1.72% false positives and 1.32% false negatives, and an overall failure rate of 1.46%.

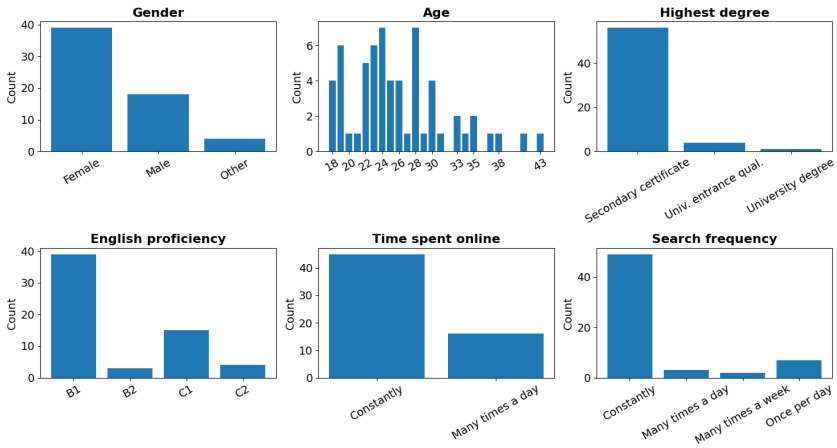

Fig. 2. Demographic and experience profile of the user sample (N = 61). Answer options that were not selected are not displayed.

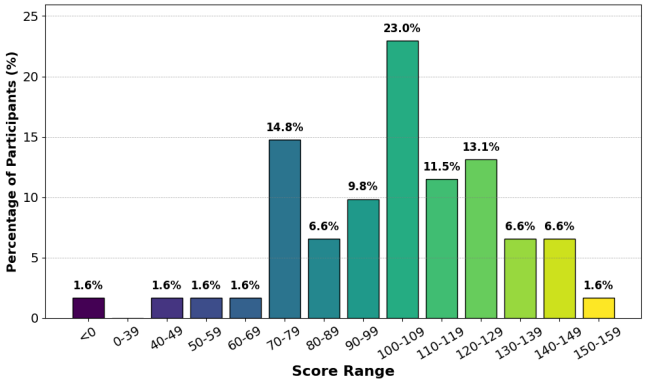

Fig. 3. Distribution of perceptual speed scores among participants (N = 61)

5 Influence of Perceptual Speed and Interest on Search Behaviors

To illustrate the practical value of the dataset collected through our framework, we present an initial case study on the questions how perceptual speed and the level of interest influences the observed search behaviour.

While the main contribution of this paper is the dataset itself, we provide this case study as an illustrative analysis to demonstrate the potential use cases of our research resource. We address the following research questions:

RQ1: How does perceptual speed influence the observed search behaviors of users?

RQ2: How does their level of interest in a topic influence the observed search behaviors of users?

5.1 Outcome Measures

To comprehensively assess search behavior, we considered interactions and times. Most interactions were extracted directly from the logfiles, while interaction durations were estimated from timestamps (e.g., document viewing time was measured from click to marking or the next document click). Interactions considered were queries issued, documents clicked or marked, pages viewed, and snippets viewed, both per session and per query; times included session duration, time per query formulation, per query, per document, and per snippet.

Since common measures do not capture the full search process, we also built Markov models from the interaction sequences to compare entire search processes across groups. Figure 4 shows the resulting Markov model based on all sessions.

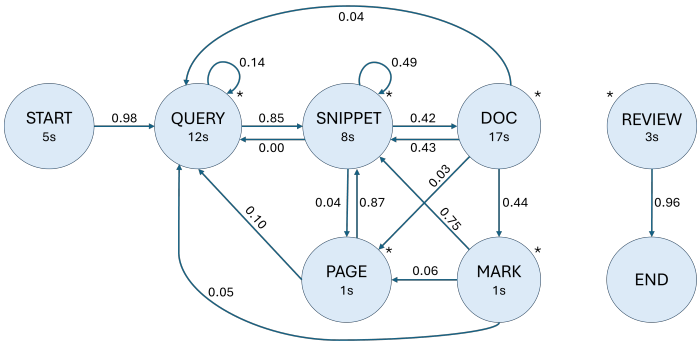

Fig. 4. Markov model of user interactions. Transitions from states marked with * to REVIEW and from REVIEW back to those states were omitted for clarity. Transition probabilities indicated at the arrows and mean interaction times per state are averaged across all sessions. As MARK and PAGE have no actual duration, they are assigned a nominal value of 1 second for completeness.

5.2 Analysis

We used the Wilcoxon signed-rank test as a non-parametric alternative to the paired t-test to examine differences between conditions of the within-subject variable (topic interest), since the data were not normally distributed. Differences in the between-subject variable (PS) were analyzed using the Mann–Whitney U test, the non-parametric equivalent of the independent-samples t-test. Participants were divided into high- and low-PS groups based on the sample median, following previous work [4,42,6,2,11]; the median value was included in the higher group to ensure more balanced group sizes. To control for multiple testing, we applied the Benjamini–Hochberg correction.

Regarding the Markov models, we used the Frobenius norm to assess overall model similarity, while using the Jensen–Shannon divergence and the Kolmogorov–Smirnov test to assess differences at the state level.

5.3 Results

Perceptual Speed Comparing users with high versus low perceptual speed revealed only minor differences (see Table 2). On average, high-PS users spent 2.94 seconds longer on each document and examined 9.89 fewer snippets per session. However, none of these differences reached statistical significance.

The Markov models also showed minimal variation between groups (Frobenius norm = 0.18), while state-level comparisons yielded very low Jensen–Shannon divergences (0.00–0.13). Likewise, Kolmogorov–Smirnov tests indicated no significant differences (all $p > 0.98$) between transition probability distributions.

Interest Aggregated over entire sessions, the only significant difference for interest was in the number of queries: participants submitted more queries for topics of personal interest. At a finer level, however, clear differences emerged: users spent less time per query, clicked and marked fewer documents, and viewed fewer pages and snippets, with time per query and the numbers of viewed snippets and marked documents reaching significance (see Table 2).

The Markov models revealed no notable differences across conditions: overall, a Frobenius norm of 0.10 indicates highly similar behavior, while state-level comparisons showed minimal differences (Jensen–Shannon divergence = 0.02–0.11) and Kolmogorov–Smirnov tests confirmed no significant effects (all $p > 0.98$).

Post-Study questionnaire We analyzed post-questionnaire responses to examine how topic interest affected stance changes, certainty, and perceived task difficulty. Users rarely changed stance (5 for high-interest vs. 7 for no-interest topics), and interest did not affect task difficulty (2.66 vs. 2.87). Nearly one-third of participants reported changes in certainty: for no-interest topics, increases and decreases were balanced (9 vs. 8), whereas for high-interest topics, more participants became more certain (13) than less certain (3).

Table 2. Mean values and statistical results for all behavioral measures by Interest (*high int.* vs. *no int.*) and Perceptual Speed (*high PS* vs. *low PS*). For Interest, the mean within-participant differences (MD) between high- and no-interest topics are additionally reported. Statistically significant results ($p < 0.05$) are indicated by an asterisk (*).

		Interest				Perceptual Speed		
	Measure	high int.	no int.	MD	p (corr.)	high PS	low PS	p (corr.)
#.../session	Queries	3.82	2.95	0.87	0.037*	3.35	3.42	0.990
	Documents	23.02	21.52	1.49	0.458	22.29	22.25	0.823
	Marked docs	12.36	14.13	-1.77	0.110	12.82	13.68	0.208
	Snippets	51.93	47.85	4.08	0.458	45.03	54.92	0.140
	Pages	7.16	6.10	1.07	0.110	6.00	7.28	0.188
#.../query	Documents	10.01	12.24	-2.23	0.091	11.16	11.08	0.847
	Marked docs	5.16	7.61	-2.44	0.028*	5.78	7.01	0.723
	Snippets	18.44	25.32	-6.88	0.042*	19.37	24.48	0.268
	Pages	2.28	2.78	-0.50	0.090	2.23	2.83	0.182
time per...	Session	662.88	653.12	9.76	0.458	665.98	649.75	0.188
	Query	276.78	360.32	-83.54	0.037*	320.39	316.64	0.823
	Q.Formulation	11.62	12.36	-0.74	0.580	11.93	12.04	0.823
	Snippet	8.17	7.90	0.27	0.580	7.63	8.45	0.939
	Document	16.62	18.23	-1.61	0.860	18.87	15.93	0.140

5.4 Discussion

Perceptual Speed None of the observed trends for perceptual speed reached significance, likely due to the exploratory nature of our study and stricter corrections for multiple comparisons. Nevertheless, the trends—longer document viewing times and fewer viewed snippets for high-PS users—align with Azzopardi et al. [6], despite contradicting the expectation that higher perceptual speed leads to shorter viewing times. Overall, the results suggest that perceptual speed does influence search behavior, even if not statistically conclusive here.

Interest Users formulated more queries for topics of interest but spent less time per query, clicked fewer documents, and marked fewer arguments. Assuming greater prior knowledge, they may have approached searches with higher expectations, which can influence behavior [44]. However, this contrasts with Edwards et al. [17], who found that task interest affected engagement but not observable behavior. A likely explanation for this discrepancy is the task design: Edwards et al. examined information-seeking tasks requiring the identification and evaluation of multiple options from neutral sources, whereas our study focused on gathering subjective arguments for controversial topics. As topics rated as interesting likely coincide with strong pre-existing opinions, cognitive biases like the confirmation bias were likely more present [32,5]. Though they were instructed to consider both sides, interested users may have actively sought evidence confirming their beliefs, leading to more focused and less exploratory behavior. In contrast, the tasks in Edwards et al. required an inherently exploratory approach.

These results indicate that the impact of interest on search behavior depends strongly on task type and cognitive framing. Distinguishing between cognitive and emotional involvement may therefore be crucial for understanding interest effects in interactive IR.

5.5 Limitations

Although the dataset provides rich and detailed information, it is limited in size, which means that some results are not statistically conclusive, particularly for subtle effects with small effect sizes. Consequently, the findings may not generalize beyond the specific setting examined. Moreover, the focus on a single domain further restricts the scope of applicability of the results.

5.6 Conclusion

Overall, the findings suggest that user characteristics, as well as the context of a search influence search behavior and should be accounted for in user models and simulators. At the same time, the Markov model results illustrate the limits of aggregated process representations: Although interest effects were more pronounced than those of perceptual speed, the Markov models for high- and no-interest sessions were more similar, since the models average over entire sessions while the interest effects primarily occurred at the query-level. Because querying is relatively infrequent compared to other interactions, the observed differences in query behavior had little impact on the overall transition probabilities. In addition, opposing tendencies within a session may balance out. This highlights the problem of oversimplifying user behavior for simulations: without considering search context—such as the current query, document rank, or evolving task state—important dynamics are lost. For user models and simulators to be realistic, they need to capture not only aggregate session patterns but also contextual dependencies that shape search as it unfolds.

6 Outlook

Beyond potential insights into how interest and perceptual speed shape search behavior, the presented resource opens up multiple avenues for future research. The dataset can support both empirical and modeling work—for instance, in validating and refining user simulators that account for cognitive traits and situational factors, or in developing models that incorporate retrieval scores and document characteristics to explain user interactions. The detailed log files also allow for analyses such as inter-rater agreement on marked arguments across participants and topics, offering new perspectives on subjectivity and consistency in argument relevance.

From a practical perspective, the data can also inform the design of user-adaptive search systems. By revealing how users with different cognitive profiles or levels of interest interact with search results, the dataset can be used to study

personalization strategies, interface adaptations, or ranking approaches aimed at supporting engagement and efficient information exploration in argumentative or opinion-driven search scenarios.

In addition, the adaptable study setup and open-source infrastructure provide a foundation for conducting new user studies with different contexts. Together, these resources enable cumulative IIR research that bridges human behavior, simulation, and system evaluation.

Acknowledgments. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 509543643.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ackerman, P.L., Cianciolo, A.T.: Cognitive, perceptual-speed, and psychomotor determinants of individual differences during skill acquisition. Journal of Experimental Psychology: Applied **6**(4), 259–290 (Dec 2000). https://doi.org/10.1037//1076-898x.6.4.259
2. Al-Maskari, A., Sanderson, M.: The effect of user characteristics on search effectiveness in information retrieval. Information Processing & Management **47**(5), 719–729 (2011). https://doi.org/https://doi.org/10.1016/j.ipm.2011.03.002, https://www.sciencedirect.com/science/article/pii/S030645731100029X, managing and Mining Multilingual Documents
3. Allen, B.: Cognitive differences in end user searching of a cd-rom index. In: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 298–309. SIGIR '92, Association for Computing Machinery, New York, NY, USA (1992). https://doi.org/10.1145/133160.133212, https://doi.org/10.1145/133160.133212
4. Arguello, J., Choi, B.: The effects of working memory, perceptual speed, and inhibition in aggregated search. ACM Trans. Inf. Syst. **37**(3) (May 2019). https://doi.org/10.1145/3322128, https://doi.org/10.1145/3322128
5. Azzopardi, L.: Cognitive biases in search: A review and reflection of cognitive biases in information retrieval. In: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval. p. 27–37. CHIIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3406522.3446023, https://doi.org/10.1145/3406522.3446023
6. Azzopardi, L., Maxwell, D., Halvey, M., Hauff, C.: Driven to distraction: Examining the influence of distractors on search behaviours, performance and experience. In: Proceedings of the 2023 Conference on Human Information Interaction and Retrieval. p. 83–94. CHIIR '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3576840.3578298, https://doi.org/10.1145/3576840.3578298
7. Balog, K., Zhai, C.: User simulation for evaluating information access systems. Foundations and Trends® in Information Retrieval **18**(1-2), 1–261 (2024). https://doi.org/10.1561/1500000098, http://dx.doi.org/10.1561/1500000098

8. Bhattacharya, N., Gwizdka, J.: Yasbil: Yet another search behaviour (and) interaction logger. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2585–2589. SIGIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3462800, https://doi.org/10.1145/3404835.3462800
9. Bogers, T., Gäde, M., Hall, M.M., Koolen, M., Petras, V., Larsen, B.: How we work, share, and re-use at chiir. In: Proceedings of the 2023 Conference on Human Information Interaction and Retrieval. p. 351–356. CHIIR '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3576840.3578305, https://doi.org/10.1145/3576840.3578305
10. Bondarenko, A., Fröbe, M., Beloucif, M., Gienapp, L., Ajjour, Y., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of touché 2020: Argument retrieval. In: Conference and Labs of the Evaluation Forum (2020), https://api.semanticscholar.org/CorpusID:225073856
11. Brennan, K., Kelly, D., Arguello, J.: The effect of cognitive abilities on information search for tasks of varying levels of complexity. In: Proceedings of the 5th Information Interaction in Context Symposium. p. 165–174. IIiX '14, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2637002.2637022, https://doi.org/10.1145/2637002.2637022
12. Carterette, B., Kanoulas, E., Hall, M.M., Clough, P.D.: Overview of the trec 2014 session track. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of The Twenty-Third Text Retrieval Conference (TREC 2014). vol. 500-308, pp. 1–12. National Institute of Standards and Technology (NIST) (2014), http://trec.nist.gov/pubs/trec23/papers/overview-session.pdf
13. Chen, J., Mao, J., Liu, Y., Zhang, M., Ma, S.: Tiangong-st: A new dataset with large-scale refined real-world web search sessions. In: Zhu, W., Tao, D., Cheng, X., Cui, P., Rundensteiner, E.A., Carmel, D., He, Q., Yu, J.X. (eds.) Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019. pp. 2485–2488. ACM (2019). https://doi.org/10.1145/3357384.3358158, https://doi.org/10.1145/3357384.3358158
14. Choi, Y.: Effects of contextual factors on image searching on the web. Journal of the American Society for Information Science and Technology **61**(10), 2011–2028 (2010). https://doi.org/https://doi.org/10.1002/asi.21386, https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21386
15. Chuklin, A., Markov, I., de Rijke, M.: Click Models for Web Search. Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers (2015)
16. Craswell, N., Campos, D., Mitra, B., Yilmaz, E., Billerbeck, B.: Orcas: 18 million clicked query-document pairs for analyzing search. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. p. 2983–2989. CIKM '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3340531.3412779, https://doi.org/10.1145/3340531.3412779
17. Edwards, A., Kelly, D.: How does interest in a work task impact search behavior and engagement? In: Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval. p. 249–252. CHIIR '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2854946.2855000, https://doi.org/10.1145/2854946.2855000

18. Ekstrom, R.B., French, J.W., Harman, H.H.: Manual for kit of factor-referenced cognitive tests (1976), https://api.semanticscholar.org/CorpusID:141329865
19. Foulds, O., Azzopardi, L., Halvey, M.: Reflecting upon perceptual speed tests in information retrieval: Limitations, challenges, and recommendations. In: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval. p. 234–242. CHIIR '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3343413.3377982, https://doi.org/10.1145/3343413.3377982
20. Gäde, M., Koolen, M., Hall, M., Bogers, T., Petras, V.: A manifesto on resource reuse in interactive information retrieval. In: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval. p. 141–149. CHIIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3406522.3446056, https://doi.org/10.1145/3406522.3446056
21. Hall, M.M.: To re-use is to re-write: Experiences with re-using iir experiment software. In: CEUR Workshop Proceedings. vol. 2337, pp. 19–23 (2019)
22. Huang, K., Guo, Q., Zhou, C., Hao, X.: How Do Emotional Tasks Influence Information Seeking Behavior? Association for Computing Machinery, New York, NY, USA (2025), https://doi.org/10.1145/3677389.3703185
23. Kelly, D.: Methods for evaluating interactive information retrieval systems with users. Found. Trends Inf. Retr. **3**(1—2), 1–224 (Jan 2009). https://doi.org/10.1561/1500000012, https://doi.org/10.1561/1500000012
24. Kelly, D., Arguello, J., Edwards, A., Wu, W.c.: Development and evaluation of search tasks for iir experiments using a cognitive complexity framework. In: Proceedings of the 2015 International Conference on The Theory of Information Retrieval. p. 101–110. ICTIR '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2808194.2809465, https://doi.org/10.1145/2808194.2809465
25. Li, H., Lu, H., Huang, S., Ma, W., Zhang, M., Liu, Y., Ma, S.: Privacy-aware remote information retrieval user experiments logging tool. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2615–2619. SIGIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3462793, https://doi.org/10.1145/3404835.3462793
26. Liang, Y., Wu, Z., He, Y., Liang, F., Liu, K., Mao, J.: A flexible user study platform for generative information retrieval. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 4066–4070. SIGIR '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3726302.3730140, https://doi.org/10.1145/3726302.3730140
27. Liu, J., Shah, C.: Interactive IR User Study: Design, Evaluation, and Reporting. Morgan & Claypool Publishers (2019)
28. Liu, J., Jung, Y.J.: Interest development, knowledge learning, and interactive ir: Toward a state-based approach to search as learning. In: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval. p. 239–248. CHIIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3406522.3446015, https://doi.org/10.1145/3406522.3446015
29. Maxwell, D., Hauff, C.: Logui: Contemporary logging infrastructure for web-based experiments. In: Hiemstra, D., Moens, M.F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) Advances in Information Retrieval. pp. 525–530. Springer International Publishing, Cham (2021)

30. Mayr, P., Kacem, A.: A complete year of user retrieval sessions in a social sciences academic search engine. In: Kamps, J., Tsakonas, G., Manolopoulos, Y., Iliadis, L.S., Karydis, I. (eds.) Research and Advanced Technology for Digital Libraries - 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings. Lecture Notes in Computer Science, vol. 10450, pp. 560–565. Springer (2017). https://doi.org/10.1007/978-3-319-67008-9_46, https://doi.org/10.1007/978-3-319-67008-9_46
31. Meggetto, F., Moshfeghi, Y.: Podify: A podcast streaming platform with automatic logging of user behaviour for academic research. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 3215–3219. SIGIR '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3539618.3591824, https://doi.org/10.1145/3539618.3591824
32. Nickerson, R.S.: Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology **2**(2), 175–220 (1998). https://doi.org/10.1037/1089-2680.2.2.175, https://doi.org/10.1037/1089-2680.2.2.175
33. O'Brien, H.L., Arguello, J., Capra, R.: An empirical study of interest, task complexity, and search behaviour on user engagement. Information Processing & Management **57**(3), 102226 (2020). https://doi.org/https://doi.org/10.1016/j.ipm.2020.102226, https://www.sciencedirect.com/science/article/pii/S0306457319301591
34. Pass, G., Chowdhury, A., Torgeson, C.: A picture of search. In: Jia, X. (ed.) Proceedings of the 1st International Conference on Scalable Information Systems, Infoscale 2006, Hong Kong, May 30-June 1, 2006. ACM International Conference Proceeding Series, vol. 152, p. 1. ACM (2006). https://doi.org/10.1145/1146847.1146848, https://doi.org/10.1145/1146847.1146848
35. Qu, P., Liu, C., Lai, M.: The effect of task type and topic familiarity on information search behaviors. In: Proceedings of the Third Symposium on Information Interaction in Context. p. 371–376. IIiX '10, Association for Computing Machinery, New York, NY, USA (2010). https://doi.org/10.1145/1840784.1840841, https://doi.org/10.1145/1840784.1840841
36. Reimer, J.H., Schmidt, S., Fröbe, M., Gienapp, L., Scells, H., Stein, B., Hagen, M., Potthast, M.: The archive query log: Mining millions of search result pages of hundreds of search engines from 25 years of web archives. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2848–2860. SIGIR '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3539618.3591890, https://doi.org/10.1145/3539618.3591890
37. Rekabsaz, N., Lesota, O., Schedl, M., Brassey, J., Eickhoff, C.: Tripclick: The log files of a large health web search engine. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2507–2513. SIGIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3463242, https://doi.org/10.1145/3404835.3463242
38. Scells, H., Jimmy, Zuccon, G.: Big brother: A drop-in website interaction logging service. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2590–2594. SIGIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3462781, https://doi.org/10.1145/3404835.3462781

39. Serdyukov, P., Dupret, G., Craswell, N.: Log-based personalization: the 4th web search click data (wscd) workshop. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining. p. 685–686. WSDM '14, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2556195.2556207, https://doi.org/10.1145/2556195.2556207
40. Sinnamon, L., Tamim, L., Dodson, S., O'Brien, H.L.: Rethinking interest in studies of interactive information retrieval. In: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval. p. 39–49. CHIIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3406522.3446031, https://doi.org/10.1145/3406522.3446031
41. Song, R., Zhang, M., Luo, C., Sakai, T., Liu, Y., Dou, Z.: SogouQ: The First Large-Scale Test Collection with Click Streams Used in a Shared-Task Evaluation, pp. 143–150. Springer Singapore, Singapore (2021). https://doi.org/10.1007/978-981-15-5554-1_10, https://doi.org/10.1007/978-981-15-5554-1_10
42. Turpin, L., Kelly, D., Arguello, J.: To blend or not to blend? perceptual speed, visual memory and aggregated search. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1021–1024. SIGIR '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2911451.2914739, https://doi.org/10.1145/2911451.2914739
43. Vuong, T., Saastamoinen, M., Jacucci, G., Ruotsalo, T.: Understanding user behavior in naturalistic information search tasks. Journal of the Association for Information Science and Technology **70**(11), 1248–1261 (2019). https://doi.org/https://doi.org/10.1002/asi.24201, https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24201
44. Wang, B., Liu, J.: Investigating the role of in-situ user expectations in web search. Information Processing & Management **60**(3), 103300 (2023). https://doi.org/https://doi.org/10.1016/j.ipm.2023.103300, https://www.sciencedirect.com/science/article/pii/S0306457323000377
45. Zerhoudi, S., Granitzer, M.: Beyond conventional metrics: Assessing user simulators in information retrieval. In: Proceedings of the 14th Italian Information Retrieval Workshop. vol. 3802, pp. 3–12 (2024)
46. Zou, L., Mao, H., Chu, X., Tang, J., Ye, W., Wang, S., Yin, D.: A large scale search dataset for unbiased learning to rank (2022), https://arxiv.org/abs/2207.03051