

# Beyond Conventional Metrics: Assessing User Simulators in Information Retrieval

Saber Zerhoudi, Michael Granitzer

University of Passau, Germany

## Abstract

Traditional evaluation methods for user simulators in Information Retrieval systems have limitations in assessing their reliability for comparative analysis. To address this, we apply the Fréchet Distance (FD) to measure similarity between real and simulated user search session distributions. Using the TREC Session 2014 Track dataset, we compare FD's performance against established metrics like session nDCG and Expected Global Utility. Our study explores FD's effectiveness in various scenarios, including those with minimal and extensive interaction data, and examines its sensitivity to different feature extraction methods. Results show that FD correlates strongly with existing metrics while offering unique insights into session similarity, particularly for complex, multi-query sessions. FD demonstrates robustness across feature extraction techniques and versatility in various evaluation scenarios. This research contributes to the field of Interactive Information Retrieval (IIR) by providing a more comprehensive framework for evaluating simulated search sessions.

## Keywords

Evaluation Metrics, Simulation Evaluation, Information Retrieval

## 1. Introduction

The evolution of Interactive Information Retrieval (IIR) systems has introduced new challenges in performance evaluation, particularly for simulated search sessions. Traditional metrics often fail to capture the complex, dynamic nature of user interactions in modern search environments, which involve query sequences, diverse user actions, and temporal elements. Conventional methods typically require extensive real user interaction data, which is costly and difficult to obtain, and may not adequately assess simulation fidelity for comparative analyses across different IIR systems.

To address these limitations, we propose the application of Fréchet Distance (FD) as a novel metric for evaluating the similarity between real and simulated user search sessions in IIR. This approach extends FD's successful application from fields like computer vision to information retrieval. FD offers a quantitative measure of how well simulated data replicates complex patterns of real user behavior, potentially serving as a standard for assessing user simulator performance and reliability in IIR systems. Our investigation into FD's efficacy for IIR systems is guided by four research questions: (1) How effectively does FD measure the quality of simulated search sessions with minimal interaction data? (2) Can FD accurately evaluate the quality of

---


*IIR2024: The 14th Italian Information Retrieval Workshop, 5th - 6th September 2024, Udine, Italy*

✉ [saber.zerhoudi@uni-passau.de](mailto:saber.zerhoudi@uni-passau.de) (S. Zerhoudi); [michael.granitzer@uni-passau.de](mailto:michael.granitzer@uni-passau.de) (M. Granitzer)

🆔 0000-0003-2259-0462 (S. Zerhoudi); 0000-0003-3566-5507 (M. Granitzer)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

simulated search sessions with extensive interaction data? (3) How correlated is the performance of IR systems in simulated search sessions, as measured by FD, compared to other metrics used to assess the similarity of simulated user sessions? (4) How sensitive is FD to different feature extraction methods when assessing the similarity of simulated search sessions?

Our experimental study utilizes the TREC Session 2014 Track dataset[1], which provides detailed logs of user interactions across multiple search sessions. This dataset is particularly suitable for our research due to its structured representation of query sequences and user actions over time. To compute FD, we employ various feature extraction methods to create vector representations of both simulated and real sessions, ranging from simple query-based embeddings to more sophisticated approaches like BERT-based Session Embedding and Time-aware Session Embedding.

We investigate the correlation between FD and established session similarity metrics such as session nDCG (sDCG) [2] and Expected Global Utility (EGU) [3], examine FD’s sensitivity to different feature extraction methods, and explore its performance across various session lengths and complexities. Additionally, we assess FD’s performance with both minimal and extensive interaction data, providing insights into its versatility as an evaluation metric. By integrating this sophisticated approach into IIR system evaluation, our research aims to demonstrate FD’s potential as a metric that not only compares the quality of simulated and real user sessions but also offers insights into the reliability and accuracy of different simulation methodologies. This study could significantly enhance the development and assessment of user simulators, leading to more effective and user-centric interactive information retrieval systems.

## 2. Related Work

The evaluation of interactive information retrieval (IIR) systems has evolved significantly, moving from single query-based metrics to more comprehensive session-based measures. This shift reflects the recognition of complex, multi-query search behaviors and the limitations of traditional evaluation methods. Early efforts to address this led to the development of session-based extensions of traditional metrics, such as Session nDCG (sDCG) by Järvelin et al. [2], which applies a discount to results from later queries in a session.

Building on this concept, Yang and Lad [3] proposed a framework for modeling user browsing behavior and computing Expected Global Utility (EGU) over a session, while Kanoulas et al. [4] introduced the concept of modeling a user’s browsing behavior as a “path.” Although these session-based metrics represented a significant advancement, they often did not model detailed browsing behaviors, such as clicking decisions. The development of click models [5] addressed this gap, providing insights into user interaction patterns. Fuhr [6] proposed the Interactive IR Probability Ranking Principle (IIR-PRP), which theoretically integrates a user’s clicking decision with a measure of overall utility of a ranked list. Zerhoudi et al. [7] used the two-sample Kolmogorov-Smirnov (KS-2) goodness-of-fit test and a classification-based evaluation to evaluate simulated user interactions in the context of a search session. However, most existing session-based evaluation measures and click models assume sequential browsing, which may not hold in modern search interfaces with complex layouts and interaction possibilities.

The Fréchet metric, a natural measure of similarity between two curves, has gained promi-

nence in various applications [8]. This metric can be intuitively understood by imagining a dog and its handler walking on separate curves. Both can control their speed but must move forward, with the Fréchet distance representing the minimal leash length required for them to traverse their respective paths from start to finish. Due to its effectiveness in comparing curve similarities, the Fréchet distance and its variants have found widespread use across diverse fields. These applications include dynamic time-warping [9], speech recognition [10], and matching of time series in databases [11]. The versatility of the Fréchet metric in these domains underscores its significance in analyzing and comparing complex, non-linear data patterns.

Inspired by the success of distribution-based metrics in other fields, such as the Fréchet Inception Distance (FID) in computer vision [12], our work introduces the Fréchet Distance (FD) as a novel metric for evaluating simulated search sessions in IIR. This approach addresses several limitations of existing methods by handling complex interactions, requiring minimal data, providing distribution-based comparisons, offering flexibility in feature representation, and investigating correlations with established metrics.

By adapting FD to the evaluation of simulated search sessions, our work bridges the gap between advanced evaluation techniques in other fields and the specific needs of IIR evaluation. This approach offers a promising direction for developing more accurate and comprehensive evaluation methods for modern IIR systems, particularly in scenarios where traditional metrics may fall short due to data sparsity or complex user interaction patterns.

Arabzadeh and Clarke [13]’s proposal to use the Fréchet Distance to measure the distance between the distributions of relevant judged items and retrieved results aligns with our approach, further validating its potential as a robust and flexible metric. Their research provides additional evidence for the effectiveness of distribution-based comparisons in scenarios with sparse data, complementing our exploration of FD for simulated search sessions.

### 3. Fréchet Distance for Evaluating Simulated Search Sessions

#### 3.1. Fréchet Distance

The Fréchet distance is a measure of dissimilarity between two curves or trajectories. It can be conceptualized as the minimum leash length required for a dog walking along one path while its owner walks along another, with both potentially moving at different speeds [14, 15].

Formally, given two curves  $A$  and  $B$  represented as sequences of points in a metric space, the Fréchet distance  $F(A, B)$  is computed as:

$$F(A, B) = \inf_{\alpha, \beta} \max_{t \in [0, 1]} d(A(\alpha(t)), B(\beta(t)))$$

where  $A$  and  $B$  are continuous maps from  $[0, 1]$  to a metric space, and  $\alpha$  and  $\beta$  are continuous, non-decreasing surjection functions representing reparameterizations of  $[0, 1]$ . This formulation ensures that neither the dog nor its owner can backtrack along their respective curves.

The Fréchet distance can also be applied to assess the disparity between probability distributions [12]. For two normal univariate distributions  $X$  and  $Y$ , the Fréchet Distance is given by:  $FD(X, Y) = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2$ , where  $\mu$  and  $\sigma$  represent the mean and standard deviation of the distributions, respectively. This versatility makes the Fréchet distance a powerful tool for comparing both geometric curves and statistical distributions in various fields of study.

### 3.2. Fréchet Distance for Evaluating Simulated Search Sessions

The evaluation of simulated search sessions using Fréchet Distance provides a robust method for assessing the quality of simulation models in Information Retrieval (IR). This approach considers both the semantic content and sequential nature of user actions within search sessions. Let  $S$  represent a set of  $n$  simulated search sessions, where each session  $s_i$  consists of a sequence of user actions  $A_{s_i}$ . These actions may include queries, clicks, scrolls, or other interactions with the search engine.  $R_S$  denotes the set of ideal or expected actions for the sessions in  $S$ . The function  $M(s_i)$  generates a sequence of actions for a given simulated session  $s_i$ , producing  $M_{s_i}$ . To apply Fréchet Distance, we map the actions to a suitable embedding space using function  $V$ , which transforms any action into a  $p$ -dimensional vector. This embedding captures both semantic and behavioral aspects of each action. The Fréchet Distance for Simulated Search Sessions ( $FD_S^M$ ) is then calculated as:

$$FD_S^M = FD(\mathbb{V}(R_S), \mathbb{V}(M(S))) \quad (5)$$

Here,  $FD$  measures the distance between the distribution of the set embeddings of the ideal actions  $V(R_S)$  and those of the simulated actions  $V(M(S))$ . A lower  $FD_S^M$  indicates higher similarity between simulated and ideal actions, suggesting better performance of the simulation model  $M$  on the session set  $S$ . To account for the sequential nature of search sessions, we extend this measure to consider the order of actions within each session:

$$FD_{S_{seq}}^M = \frac{1}{|S|} \sum_{s_i \in S} FD(\mathbb{V}(R_{s_i}), \mathbb{V}(M(s_i))) \quad (6)$$

This sequential Fréchet Distance ( $FD_{S_{seq}}^M$ ) calculates the average Fréchet Distance between ideal and simulated action sequences for each session. This provides a more nuanced evaluation of the simulation model's ability to capture the temporal dynamics of user behavior in search sessions. By incorporating both semantic content and sequential information, this approach offers a comprehensive evaluation framework for simulated search sessions, enabling researchers to assess and improve the fidelity of their simulation models in IR experiments.

## 4. Experimental Setup

In this section, we describe the general settings of our experiments, including the dataset, traditional evaluation metrics, click models, simulation framework, and the embeddings used to represent user search sessions.

### 4.1. Dataset

This study employs the TREC Session 2014 Track dataset [1], which is designed for evaluating multi-query search behavior. The dataset comprises 1,257 sessions with 4,680 queries and 1,685 clicks, averaging 4.33 queries per session (median: 2). It includes real user queries, interaction logs, and ranked document lists with snippets, making it ideal for simulated search session evaluation. The dataset's diverse composition allows for a comprehensive analysis of user interactions and search strategies in multi-query scenarios, enabling robust conclusions about the effectiveness of various information retrieval techniques.

## 4.2. Click Models and Simulation Framework

This study employs a comprehensive approach to simulate user search sessions, utilizing both traditional probabilistic graphical models and a neural click model. The traditional models include the Position-Based Model (PBM) [16], User Browsing Model (UBM) [17], Dependent Click Model (DCM) [18], and Dynamic Bayesian Network Model (DBN) [19], implemented using the PyClick library. Additionally, we incorporate the neural click model NCM [20] to enhance simulation diversity. To create complete user search sessions, we use the SimIIR 2.0 framework [21], which simulates complex user behaviors including query formulation, result list examination, and click decisions. This integrated approach allows for a comprehensive assessment of FD’s ability to quantify the quality of simulated search sessions across a broad spectrum of user behaviors.

## 4.3. Embeddings for Search Session Representation

This study explores two approaches for embedding search sessions to apply the Fréchet Distance metric. The first approach uses action-based embeddings, employing a fine-tuned BERT model [22] to embed individual user actions and aggregating them through mean pooling or sequence modeling. The second approach adapts Doc2Vec [23] to create Session2Vec [24], learning fixed-length vector representations of entire search sessions. Both methods aim to capture semantic and behavioral aspects of user search sessions. For query and document representations, pre-trained word embeddings are used, with query embeddings computed as the average of query word embeddings and document embeddings derived from title and snippet words. These approaches provide insights into effectively capturing search behavior nuances and improving simulated search session evaluation using the Fréchet Distance metric.

## 4.4. Evaluation Process

Our evaluation process involves generating simulated search sessions using various click models and SimIIR 2.0, then embedding these sessions along with ground truth sessions from the TREC Session 2014 Track dataset. We compare simulation approaches by calculating the Fréchet Distance between simulated and ground truth session embedding distributions. To account for temporal dynamics, we also compute the sequential Fréchet Distance ( $FD_{Seq}$ ) as defined in equation (6). The effectiveness of Fréchet Distance as an evaluation metric is assessed by comparing FD scores with traditional metrics and analyzing results across different click models and simulation configurations. This approach provides insights into the strengths and limitations of various simulation methods in replicating realistic user search behavior.

## 5. Evaluating Search Sessions with Minimal Interaction Data

This section explores the effectiveness of the Fréchet Distance (FD) in assessing the quality of simulated search sessions with limited interaction data. We analyze the performance of various click models and the SimIIR 2.0 framework on the TREC Session 2014 Track dataset, comparing the FD metric with traditional session-based metrics.

Our study utilizes a subset of 200 sessions from the TREC Session 2014 Track, each containing an average of 4-5 queries. We generate simulated interactions using click models and the SimIIR 2.0 framework, then compute the FD between simulated and ground truth sessions using action-based embeddings and Session2Vec representations.

**Table 1:** Performance of simulation approaches on TREC Session 2014 Track subset

Model	nDCG@10	ERR@10	FD@1	FD@10
PBM	0.342	0.289	6.823	4.156
UBM	0.368	0.311	5.967	3.845
DCM	0.381	0.325	5.412	3.621
DBN	0.395	0.339	4.876	3.302
NCM	0.423	0.361	3.945	2.687
SimIIR 2.0	0.451	0.389	3.124	2.103

**Figure 1:** Bootstrap analysis of simulation approaches (ERR@10 vs. FD@10)

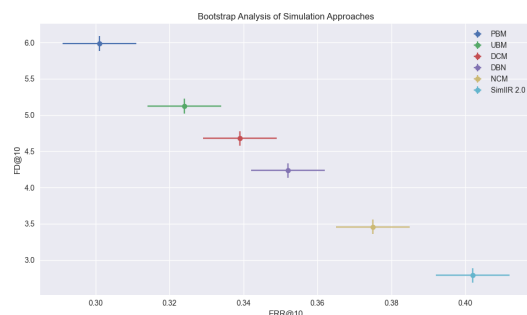


Table 1 presents the performance of different simulation approaches using traditional metrics (nDCG@10 and ERR@10) and Fréchet Distance metrics (FD@1 and FD@10). The results show that FD effectively quantifies the quality of simulated search sessions. The PBM model, being the simplest, shows the highest FD values, indicating the largest discrepancy from ground truth sessions. Conversely, SimIIR 2.0, incorporating more complex user behaviors, achieves the lowest FD values, suggesting simulations closest to the ground truth. FD aligns well with traditional evaluation metrics, consistently ranking simulation approaches. This indicates that FD can effectively capture simulation quality even with minimal interaction data. Bootstrap analysis (Figure 1) confirms these patterns across different samples, with narrow confidence intervals indicating stability. The FD metric demonstrates sensitivity to simulation model complexity, with more sophisticated models like NCM and SimIIR 2.0 achieving lower FD scores. However, its discriminative power decreases when comparing closely performing models. Overall, Fréchet Distance proves to be a promising metric for evaluating simulated search sessions, especially with minimal interaction data, complementing traditional evaluation metrics in interactive information retrieval.

## 6. Evaluating Search Sessions with Extensive Interaction Data

This section examines the effectiveness of the Fréchet Distance (FD) in assessing simulated search sessions with extensive interaction data. Using a subset of the TREC Session 2014 Track dataset, we focus on longer and more complex sessions to address whether FD can accurately evaluate the quality of simulated search sessions with extensive interaction data. Our experimental setup involved selecting 100 sessions from the dataset, each containing at least 10 queries and a rich set of user interactions. We generated simulated interactions using various click models and the SimIIR 2.0 framework, computing the FD between simulated and

ground truth sessions using both action-based embeddings and Session2Vec representations. To investigate the impact of interaction data quantity, we evaluated the simulations using full sessions, first 5 queries, and first 10 queries.

**Table 2**

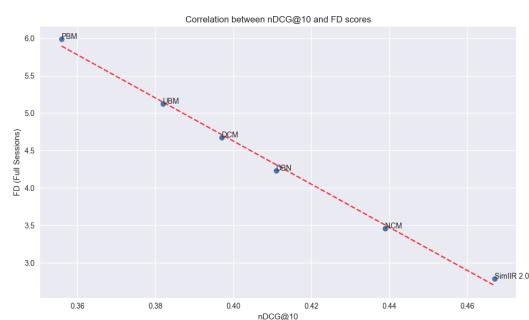
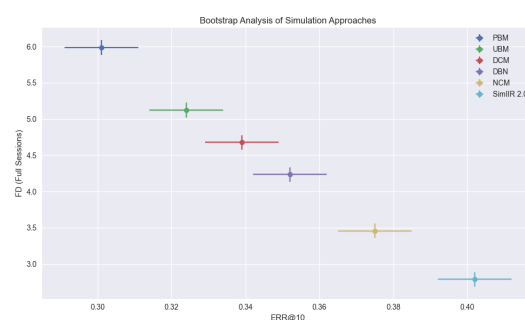
Performance of simulation approaches with varying amounts of interaction data

Model	nDCG@10	ERR@10	FD (Full)	FD (5Q)	FD (10Q)
PBM	0.356	0.301	5.987	6.234	6.102
UBM	0.382	0.324	5.123	5.456	5.289
DCM	0.397	0.339	4.678	4.912	4.795
DBN	0.411	0.352	4.234	4.567	4.401
NCM	0.439	0.375	3.456	3.789	3.623
SimIIR 2.0	0.467	0.402	2.789	3.123	2.956

**Table 3**

Kendall's  $\tau$  correlation between FD scores for full and partial sessions

Comparison	Kendall's $\tau$
FD (Full) vs FD (5Q)	0.923
FD (Full) vs FD (10Q)	0.956

**Figure 2:** Correlation between nDCG@10 and FD scores for full sessions**Figure 3:** Bootstrap analysis of simulation approaches (ERR@10 vs. FD for full sessions)

Results in Table 2 demonstrate that Fréchet Distance (FD) effectively quantifies the quality of simulated search sessions across varying amounts of interaction data. FD scores for full sessions are generally lower than those for partial sessions, indicating improved simulation accuracy with more interaction data. Figure 2 shows a strong negative correlation between nDCG@10 and FD scores for full sessions suggests alignment with traditional evaluation metrics.

Bootstrap analysis, using 1000 subsets of 50 sessions each, revealed narrow confidence intervals for both ERR@10 and FD scores, indicating stability across different session subsets. High correlations between FD scores for full and partial sessions (5Q and 10Q) suggest that FD maintains discriminative power even when evaluating partial sessions (i.e., Table 3). FD demonstrates several advantages, including consistency across different amounts of interaction data, robustness in assessments, sensitivity to simulation complexity, and alignment with traditional metrics. However, computational complexity may increase with larger datasets, warranting future exploration of efficient approximation methods.

In conclusion, Fréchet Distance proves to be an effective and robust metric for evaluating simulated search sessions, particularly with extensive interaction data. Its ability to capture distributional similarities between simulated and ground truth sessions makes it valuable for assessing and improving search session simulation models.

## 7. Correlation Analysis with Session Similarity Metrics

This section examines the correlation between the Fréchet Distance (FD) and other established metrics used to evaluate the similarity of simulated user sessions in Information Retrieval (IR) systems. The study aims to understand how FD's performance in measuring simulated search sessions correlates with other session similarity metrics.

We employ the TREC Session 2014 dataset, selecting 200 sessions of varying lengths and complexities. Simulated sessions are generated using click models and the SimIIR 2.0 framework. FD is compared with Session nDCG (sDCG), Expected Global Utility (EGU), Path-based Session Evaluation (PSE), and Interactive IR Probability Ranking Principle (IIR-PRP) based metric.

Results show strong negative correlations between FD and all other metrics (i.e., Table 4), with the strongest correlation observed with PSE. As FD decreases, indicating better simulation quality, other metrics tend to increase.

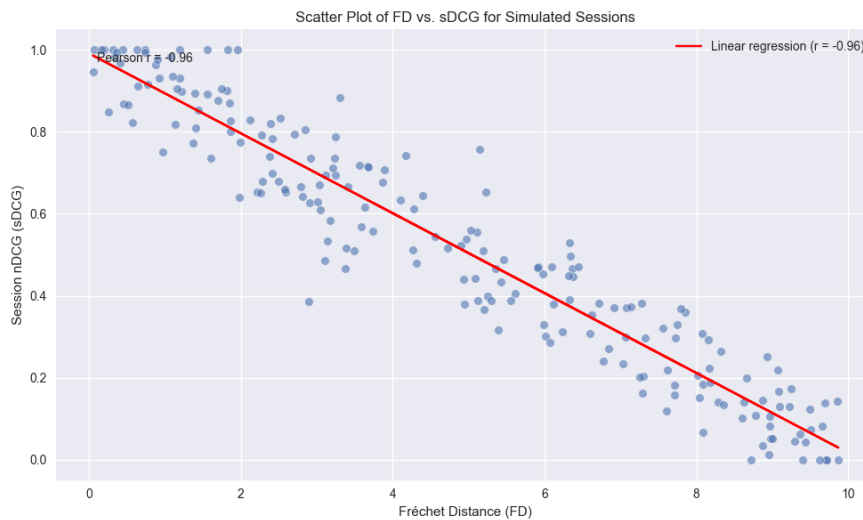
**Table 4:** Correlation coefficients between FD and other metrics

Metric	Pearson Correlation	Spearman's Rank Correlation
sDCG	-0.823 (-0.856, -0.789)	-0.841 (-0.872, -0.809)
EGU	-0.791 (-0.827, -0.752)	-0.805 (-0.839, -0.768)
PSE	-0.836 (-0.867, -0.803)	-0.852 (-0.881, -0.821)
IIR-PRP	-0.779 (-0.817, -0.738)	-0.795 (-0.831, -0.756)

**Table 5:** Correlation coefficients between FD and sDCG for different session lengths

Session Length	Pearson Correlation	Spearman's Rank Correlation
Short (1-3 queries)	-0.789 (-0.834, -0.739)	-0.803 (-0.846, -0.755)
Medium (4-7 queries)	-0.835 (-0.872, -0.794)	-0.851 (-0.886, -0.812)
Long (8+ queries)	-0.867 (-0.901, -0.829)	-0.882 (-0.914, -0.846)

**Figure 4:** Scatter plot of FD vs. sDCG for simulated sessions



Correlation analysis for different session lengths reveals that the relationship between FD and other metrics strengthens as session length increases as shown in Table 5, suggesting FD's effectiveness in capturing complex interaction patterns in longer sessions. The study supports FD as a valid and effective measure for evaluating simulated search sessions, demonstrating consistency with established metrics and sensitivity to session complexity.

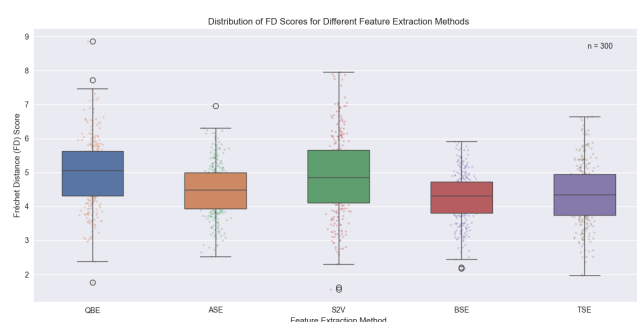


However, limitations include dataset specificity, simulation model dependency, and the need for further investigation into metric assumptions and computational efficiency. Future work should validate findings on other datasets, explore a wider range of simulation approaches, and correlate metric-based assessments with human judgments of session similarity. In conclusion, this study supports the use of FD as a valuable tool in evaluating IR systems, particularly for complex, multi-query sessions, potentially offering complementary insights to existing metrics and enhancing the assessment and improvement of IR systems in interactive, session-based search contexts.

## 8. Sensitivity to Feature Extraction in Simulated Sessions

This section examines the impact of various feature extraction methods on the Fréchet Distance when evaluating simulated search sessions. The study utilizes the TREC Session 2014 dataset, selecting 300 sessions of varying lengths and complexities. Five feature extraction methods are investigated: Query-based Embedding (QBE), Action Sequence Embedding (ASE), Session2Vec (S2V), BERT-based Session Embedding (BSE), and Time-aware Session Embedding (TSE).

**Figure 5:** Box plot of FD scores for different feature extraction methods.



**Table 6:** Percentage of rank changes in simulation model performance.

Method Change	% Rank Changes
QBE to ASE	18.3%
QBE to S2V	23.7%
QBE to BSE	15.9%
QBE to TSE	20.1%
ASE to S2V	12.6%
ASE to BSE	9.8%
ASE to TSE	11.2%
S2V to BSE	14.5%
S2V to TSE	17.8%
BSE to TSE	8.7%

**Table 7:** Spearman’s rank correlation between FD scores from different feature extraction methods.

Method	QBE	ASE	S2V	BSE	TSE
QBE	1.000	0.782	0.715	0.801	0.743
ASE	0.782	1.000	0.856	0.889	0.872
S2V	0.715	0.856	1.000	0.834	0.791
BSE	0.801	0.889	0.834	1.000	0.905
TSE	0.743	0.872	0.791	0.905	1.000

**Table 8:** Spearman’s rank correlation between FD scores for different session lengths (BSE method).

Session Length	Short	Medium	Long
Short	1.000	0.823	0.756
Medium	0.823	1.000	0.891
Long	0.756	0.891	1.000

Table 7 presents the Spearman’s rank correlation coefficients between FD scores obtained using different feature extraction methods, showing moderate to high correlations between all methods. The distribution of FD scores for each feature extraction method across all simulated sessions is illustrated in Figure 5.

Table 6 shows the percentage of rank changes in simulation model performance when switching between feature extraction methods. These changes range from 8.7% to 23.7%, indicating that the choice of feature extraction method can significantly impact the evaluation of simulated sessions. Table 8 displays the Spearman’s rank correlation between FD scores for different session lengths using the BSE method, revealing that correlations are generally higher between adjacent length categories.

The analysis of Fréchet Distance’s behavior across different feature extraction methods reveals moderate to high correlations between methods, but also method-specific variations. The choice of method can affect simulation model performance rankings, with sensitivity to session length observed. Contextual and time-aware approaches demonstrate greater robustness. Based on these findings, recommendations for using Fréchet Distance in evaluating simulated search sessions include: careful selection of feature extraction methods, preference for contextual approaches, maintaining consistency in comparisons, considering session length effects, validating using multiple methods, and ensuring transparency in reporting. While limitations exist, such as dataset specificity and the limited range of methods examined, the Fréchet Distance proves to be a robust measure for evaluating simulated search sessions, albeit with some sensitivity to feature extraction method choice.

## 9. Conclusion and Future Work

This paper explores the application of Fréchet Distance (FD) as a novel metric for evaluating simulated search sessions in interactive information retrieval systems. Our experiments demonstrate FD’s effectiveness and robustness in assessing the quality of simulated user interactions across various scenarios.

FD shows strong correlations with established metrics like session nDCG and Expected Global Utility, capturing similar aspects of session quality while offering additional insights due to its distributional nature. It proves particularly effective in evaluating longer, more complex sessions and demonstrates effectiveness even with minimal interaction data. These findings have significant implications for interactive information retrieval, potentially leading to more accurate and realistic simulation models. Future research directions include investigating FD’s performance with multi-modal session representations, extending this work to larger datasets, correlating FD assessments with human judgments, and exploring FD in evaluating generative models for search session simulation.

While our study offers valuable insights, it’s important to acknowledge certain limitations. FD requires sets of sessions for evaluation and assumes multivariate normal distributions, which may not always hold for all types of session data. Additionally, as an unbounded metric, the interpretation of FD scores may vary depending on dataset characteristics and sample sizes.

Despite these limitations, we believe that FD offers a powerful and flexible approach to evaluating simulated search sessions, complementing existing metrics and potentially driving improvements in both simulation models and real-world information retrieval systems.

## References

- [1] 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>.
- [2] K. Järvelin, S. L. Price, L. M. L. Delcambre, M. L. Nielsen, Discounted cumulated gain based evaluation of multiple-query IR sessions, in: C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, R. W. White (Eds.), *Advances in Information Retrieval*, 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. *Proceedings*, volume 4956 of *Lecture Notes in Computer Science*, Springer, 2008, pp. 4–15. URL: [https://doi.org/10.1007/978-3-540-78646-7\\_4](https://doi.org/10.1007/978-3-540-78646-7_4). doi:10.1007/978-3-540-78646-7\_4.
- [3] Y. Yang, A. Lad, Modeling expected utility of multi-session information distillation, in: L. Azzopardi, G. Kazai, S. E. Robertson, S. M. Rüger, M. Shokouhi, D. Song, E. Yilmaz (Eds.), *Advances in Information Retrieval Theory, Second International Conference on the Theory of Information Retrieval, ICTIR 2009*, Cambridge, UK, September 10-12, 2009, *Proceedings*, volume 5766 of *Lecture Notes in Computer Science*, Springer, 2009, pp. 164–175. URL: [https://doi.org/10.1007/978-3-642-04417-5\\_15](https://doi.org/10.1007/978-3-642-04417-5_15). doi:10.1007/978-3-642-04417-5\_15.
- [4] E. Kanoulas, B. Carterette, P. D. Clough, M. Sanderson, Evaluating multi-query sessions, in: W. Ma, J. Nie, R. Baeza-Yates, T. Chua, W. B. Croft (Eds.), *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011*, Beijing, China, July 25-29, 2011, ACM, 2011, pp. 1053–1062. URL: <https://doi.org/10.1145/2009916.2010056>. doi:10.1145/2009916.2010056.
- [5] A. Chuklin, I. Markov, M. de Rijke, *Click Models for Web Search, Synthesis Lectures on Information Concepts, Retrieval, and Services*, Morgan & Claypool Publishers, 2015. URL: <https://doi.org/10.2200/S00654ED1V01Y201507ICR043>. doi:10.2200/S00654ED1V01Y201507ICR043.
- [6] N. Fuhr, A probability ranking principle for interactive information retrieval, *Inf. Retr.* 11 (2008) 251–265. URL: <https://doi.org/10.1007/s10791-008-9045-0>. doi:10.1007/s10791-008-9045-0.
- [7] S. Zerhoudi, M. Granitzer, C. Seifert, J. Schloetterer, Evaluating simulated user interaction and search behaviour, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørnvåg, V. Setty (Eds.), *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022*, Stavanger, Norway, April 10-14, 2022, *Proceedings, Part II*, volume 13186 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 240–247. URL: [https://doi.org/10.1007/978-3-030-99739-7\\_28](https://doi.org/10.1007/978-3-030-99739-7_28). doi:10.1007/978-3-030-99739-7\_28.
- [8] A. Efrat, L. J. Guibas, S. Har-Peled, J. S. B. Mitchell, T. M. Murali, New similarity measures between polylines with applications to morphing and polygon sweeping, *Discret. Comput. Geom.* 28 (2002) 535–569. URL: <https://doi.org/10.1007/s00454-002-2886-1>. doi:10.1007/s00454-002-2886-1.
- [9] E. J. Keogh, M. J. Pazzani, Scaling up dynamic time warping to massive dataset, in: J. M. Zytzkow, J. Rauch (Eds.), *Principles of Data Mining and Knowledge Discovery, Third European Conference, PKDD '99*, Prague, Czech Republic, September 15-18, 1999, Pro-

- ceedings, volume 1704 of *Lecture Notes in Computer Science*, Springer, 1999, pp. 1–11. URL: [https://doi.org/10.1007/978-3-540-48247-5\\_1](https://doi.org/10.1007/978-3-540-48247-5_1). doi:10.1007/978-3-540-48247-5\_1.
- [10] S. Kwong, Q. He, K. Man, C. Chau, K. Tang, Parallel genetic-based hybrid pattern matching algorithm for isolated word recognition, *Int. J. Pattern Recognit. Artif. Intell.* 12 (1998) 573–594. URL: <https://doi.org/10.1142/S0218001498000348>. doi:10.1142/S0218001498000348.
- [11] M. Kim, S. Kim, M. Shin, Optimization of subsequence matching under time warping in time-series databases, in: H. Haddad, L. M. Liebrock, A. Omicini, R. L. Wainwright (Eds.), *Proceedings of the 2005 ACM Symposium on Applied Computing (SAC)*, Santa Fe, New Mexico, USA, March 13-17, 2005, ACM, 2005, pp. 581–586. URL: <https://doi.org/10.1145/1066677.1066814>. doi:10.1145/1066677.1066814.
- [12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 6626–6637. URL: <https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fef65871369074926d-Abstract.html>.
- [13] N. Arabzadeh, C. L. A. Clarke, Fréchet distance for offline evaluation of information retrieval systems with sparse labels, in: Y. Graham, M. Purver (Eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers*, St. Julian’s, Malta, March 17-22, 2024, Association for Computational Linguistics, 2024, pp. 420–431. URL: <https://aclanthology.org/2024.eacl-long.26>.
- [14] T. Eiter, H. Mannila, Computing discrete fréchet distance (1994).
- [15] H. Alt, The computational geometry of comparing shapes, in: S. Albers, H. Alt, S. Näher (Eds.), *Efficient Algorithms, Essays Dedicated to Kurt Mehlhorn on the Occasion of His 60th Birthday*, volume 5760 of *Lecture Notes in Computer Science*, Springer, 2009, pp. 235–248. URL: [https://doi.org/10.1007/978-3-642-03456-5\\_16](https://doi.org/10.1007/978-3-642-03456-5_16). doi:10.1007/978-3-642-03456-5\_16.
- [16] N. Craswell, O. Zoeter, M. J. Taylor, B. Ramsey, An experimental comparison of click position-bias models, in: M. Najork, A. Z. Broder, S. Chakrabarti (Eds.), *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008*, Palo Alto, California, USA, February 11-12, 2008, ACM, 2008, pp. 87–94. URL: <https://doi.org/10.1145/1341531.1341545>. doi:10.1145/1341531.1341545.
- [17] G. Dupret, B. Piwowarski, A user browsing model to predict search engine click data from past observations, in: S. Myaeng, D. W. Oard, F. Sebastiani, T. Chua, M. Leong (Eds.), *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008*, Singapore, July 20-24, 2008, ACM, 2008, pp. 331–338. URL: <https://doi.org/10.1145/1390334.1390392>. doi:10.1145/1390334.1390392.
- [18] F. Guo, C. Liu, Y. M. Wang, Efficient multiple-click models in web search, in: R. Baeza-Yates, P. Boldi, B. A. Ribeiro-Neto, B. B. Cambazoglu (Eds.), *Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009*, Barcelona,

- Spain, February 9-11, 2009, ACM, 2009, pp. 124–131. URL: <https://doi.org/10.1145/1498759.1498818>. doi:10.1145/1498759.1498818.
- [19] O. Chapelle, Y. Zhang, A dynamic bayesian network click model for web search ranking, in: J. Quemada, G. León, Y. S. Maarek, W. Nejdl (Eds.), Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009, ACM, 2009, pp. 1–10. URL: <https://doi.org/10.1145/1526709.1526711>. doi:10.1145/1526709.1526711.
- [20] A. Borisov, I. Markov, M. de Rijke, P. Serdyukov, A neural click model for web search, in: J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks, B. Y. Zhao (Eds.), Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016, ACM, 2016, pp. 531–541. URL: <https://doi.org/10.1145/2872427.2883033>. doi:10.1145/2872427.2883033.
- [21] S. Zerhoudi, S. Günther, K. Plassmeier, T. Borst, C. Seifert, M. Hagen, M. Granitzer, The simiir 2.0 framework: User types, markov model-based interaction simulation, and advanced query generation, in: M. A. Hasan, L. Xiong (Eds.), Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022, ACM, 2022, pp. 4661–4666. URL: <https://doi.org/10.1145/3511808.3557711>. doi:10.1145/3511808.3557711.
- [22] F. A. Acheampong, H. Nunoo-Mensah, W. Chen, Transformer models for text-based emotion detection: a review of bert-based approaches, *Artif. Intell. Rev.* 54 (2021) 5789–5829. URL: <https://doi.org/10.1007/s10462-021-09958-2>. doi:10.1007/s10462-021-09958-2.
- [23] Q. V. Le, T. Mikolov, Distributed representations of sentences and documents, in: Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, volume 32 of *JMLR Workshop and Conference Proceedings*, JMLR.org, 2014, pp. 1188–1196. URL: <http://proceedings.mlr.press/v32/le14.html>.
- [24] L. Bing, Z. Niu, W. Lam, H. Wang, Learning a semantic space of web search via session data, in: S. Ma, J. Wen, Y. Liu, Z. Dou, M. Zhang, Y. Chang, W. X. Zhao (Eds.), Information Retrieval Technology - 12th Asia Information Retrieval Societies Conference, AIRS 2016, Beijing, China, November 30 - December 2, 2016, Proceedings, volume 9994 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 83–97. URL: [https://doi.org/10.1007/978-3-319-48051-0\\_7](https://doi.org/10.1007/978-3-319-48051-0_7). doi:10.1007/978-3-319-48051-0\_7.