Evaluating Simulated User Interaction and Search Behaviour

Saber Zerhoudi[1][0000−0003−2259−0462], Michael Granitzer[1][0000−0003−3566−5507], Christin Seifert[2][0000−0002−6776−3868], and Joerg Schloetterer[2][0000−0002−3678−0390]

[1] University of Passau, Passau, Germany
{saber.zerhoudi,michael.granitzer}@uni-passau.de
[2] University of Duisburg-Essen, Duisburg, Germany
{christin.seifert,joerg.schloetterer}@uni-due.de

Abstract. Simulating user sessions in a way that comes closer to the original user interactions is key to generating user data at any desired volume and variety such that A/B-testing in domain-specific search engines becomes scalable. In recent years, research on evaluating Information Retrieval (IR) systems has mainly focused on simulation as means to improve users models and evaluation metrics about the performance of search engines using test collections and user studies. However, test collections contain no user interaction data and user studies are expensive to conduct. Thus there is a need in developing a methodology for evaluating simulated user sessions. In this paper, we propose evaluation metrics to assess the realism of simulated sessions and describe a pilot study to assess the capability of generating simulated search sequences representing an approximation of real behaviour. Our findings highlight the importance of investigating and utilising classification-based metrics besides the distribution-based ones in the evaluation process.

Keywords: Evaluation metrics · Simulating user session · Simulation evaluation · User search behaviour · User modelling.

1 Introduction

Developing evaluation methods to help improve the performance of Information Retrieval (IR) systems has been the focal point of researchers in IR community for many years [8, 14, 21]. Originally, the Cranfield evaluation methodology [12], which is so far the leading methodology for evaluating an IR system, is designed to evaluate the performance of a system using a test collection comprising a sample of queries, documents and a set of relevance judgments (indicating which documents are relevant/non-relevant to which queries) but lacks user interaction data (indicating the interaction sequences generated by users while expressing their information needs). Furthermore, users are represented in a highly abstracted form without considering the complexities of their interactions. As users' information needs become more complex in sophisticated IR

systems, assessing the performance of a system needs to be assessed over an entire interactive session. Such sophisticated IR systems have been so far evaluated mainly using controlled user studies [14] or using the history log of user interactions. However, the experimental results obtained in such a way would be expensive and hard to reproduce using the former method, or would require search logs that are mostly inaccessible due to users' privacy using the latter.

TREC Session Track [5] and Dynamic Test Collections [4] are two important attempts to evaluate IR system performance over an entire search session. The evaluation metrics they adopted (e.g. P@k and nDCG@k) are cheap and reusable, but they cannot cope with users' dynamic information need (i.e. query reformulation behaviors). Jiang et al. [11] explored the correlation between user models and metrics. They examined several evaluation metrics and showcased that session Rank-Biased Precision (sRBP) [15] and session-based DCG (sDCG) [10] have stronger correlations with user satisfaction compared with existing session-based metrics. Following their finding, Zhang et al. introduced Recency-aware Session-based Metrics (RSMs) [24] which characterise users' cognitive process in search sessions by incorporating the recency effect.

There has been growing interest in the generation of simulated interaction data, and in particular how to develop more realistic models of search [2,18], for multiple reasons: First, simulation offers a way to overcome the lack of experimental real-world data, especially when the acquisition of such data is costly or challenging. Second, simulation helps reducing the amount of collected user data while preserving the profiling efficiency and protecting the privacy and the confidentiality of users' personal information.

While most related studies focus on the browsing model (i.e., user browsing behavior when consulting a page of search results), querying model and document relevance model for search evaluation, few studies have investigated the utility of evaluating simulated user interactions. Carterette et al. [4] suggested a meta-evaluation methodology using session histories and evaluated the simulation model based on its effectiveness at predicting actual user interactions, using standard classification evaluation metrics (i.e., precision, recall, accuracy and AUC). However, as it will be discussed in this paper, the used classification-based metrics are difficult to be justified and lack an evaluation of distributional properties of the data. Inspired by Carterette et al. [4], we propose a method to evaluate simulated user sessions' realism in the context of a search session. Realism represents the level of authenticity that simulated sessions present compared to the real log data. We model users' browsing patterns using Markov models. Markov models have been widely used for discovering meaningful patterns in browsing data due to their good interpretability [16,23]. In particular, they capture sequences in search patterns using transitional probabilities between states and translate user sessions into Markov processes.

To summarise, the main contributions of our work are twofold: (1) We model users' browsing patterns using a first-order Markov approach and a contextual Markov model that utilises user's browsing context based on common sense assumptions. We then conduct experiments on a real-world dataset and simulate

user search behaviour by the two approaches. (2) We propose a method to evaluate the realism of simulated user interactions in the context of a search session. We first utilise the Kolmogorov-Smirnov statistical test as an empirical validation to compare the similarity between data log and simulated sessions distribution, and then employ a classification-based evaluation technique to assess the quality of simulated search session.

2 Evaluation Methods

Our goal is to develop an evaluation methodology to evaluate to which extend simulated user models can replace or complement sample-based ones. The quality of simulated user search sessions is usually evaluated by comparing real log and simulated data. In fact, simulated data are expected to be similar to real data as we do not want them to be distinguishable. Our evaluation method assumes the following user models:

First-order Markov model: We propose investigating the use of Markov Chains to model the search dynamics. The theoretical model is based on first-order Markov models [22]. Let X_k be the random variable that models actions in a user search session. The transition probability is modelled using maximum likelihood estimation: $P(X_k = A_j | X_{k-1} = A_i) = \frac{N_{A_i,A_j}}{N_{A_i}}$, where N_{A_i} is the total amount of how many times the action A_i occurred in the training data and N_{A_i,A_j} is the amount of how many times the transition from action i to action j has been observed.

Contextual Markov model: During a search session, a user performs different search actions to find documents that fulfil their information needs. The technique that we propose here aims to categorise users into different groups based on their search behaviour. Search tasks are commonly divided into two major types of user's behaviour [1, 17]: i) *Exploratory*: where users are more likely to formulate more queries as they learn about the topic and explore the search result list exhaustively, ii) *Lookup*: where users only investigate the first few results and rephrase their queries quickly.

Kumaripaba et al. [1] extended the work of Marchionini [17] and provided a few simple indicators of information search behaviours (e.g. query length, maximum scroll depth, completion time) to categorise users into exploratory and lookup searchers. We utilise these indicators to split the training data into smaller portions. We build a first-order Markov model for each type (i.e. exploratory and lookup) and we compare them to the Markov model built from the whole data (i.e. first-order Markov model). This would allow us to evaluate the impact of context on the accuracy of simulated sessions.

2.1 Kolmogorov-Smirnov-based Evaluation

The two-sample Kolmogorov-Smirnov (KS-2) goodness-of-fit test [13] is one of the most useful and non-parametric methods for comparing two datasets. It is a convenient method for investigating whether two probability distributions can

be regarded as indistinguishable. Essentially, we test the null hypothesis that the two independent samples are drawn from the same distribution and proceed with calculating the absolute value of the distance between two data samples which we refer to as the test statistic d to compare their distribution for similarities.

We derive two separate simulation models for context-aware approaches (i.e., dividing the dataset into lookup and exploratory subsets and then for each subset we construct a Markov model to simulate user-type specific search sessions) and one global simulation model that is trained on whole dataset.

2.2 Classification-based Evaluation

Additionally, we define a classification-based evaluation to evaluate the simulation realism of our models. We first develop a set of features that represent the sequential nature of a user search session in the form of a feature vector. Then we train a classifier to distinguish simulated sessions from real log data sessions and report the results. Building upon previous work [9] about what kinds of engineered features are best suited to various machine learning model types, we developed a set of features that represent the sequentiality of the search session (i.e., typing a query; reformulating the query; clicking, viewing and exporting actions) and discarded those that only describe the user's overall search behaviour (e.g., tally of search actions, queries formulation and clicks). We used a binary vector to indicate the presence of a feature (i.e., (0) if present and (1) if not) and ordered features in the sequence (i.e., i_feature where i refer to the sequence order of the query in a session , e.g., 1_search, 2_view_record).

Each user session is converted to a feature vector, labelled and fed to a classifier. This process was repeated separately for each of the Markov approaches, i.e, first-order and contextual. We created an equal amount of simulated sessions as real log sessions for a balanced classification and evaluated three classifiers with 10-fold cross-validation. As per the classifier, we used the most popular algorithms in binary classification, namely, Support Vector Machine [6], Decision Trees [20] (XGBoost), Random Forests [3] and reported the average score. We also used automated machine learning (Auto-sklearn [7]) as it employs an ensemble of top performing models discovered during the optimisation process. Since we are interested in finding a classifier that is close to 100% Recall on the real log sessions (i.e., successful in detecting all real log sessions) and a high recall on the simulated sessions (i.e., good at detecting most of simulated sessions), we incorporate a bias in the classifier by weighting the class of real data ($w_{real} = 10^4$, $w_{simulated} = 1$) to penalise bad real log sessions predictions.

To evaluate the realism of our models, we use metrics *Precision*, *Recall*, *F-score* and *Accuracy* common for objectively measuring the classifier's performance. In our case, we consider *True Positive* (TP) to be the scenario where the model classifies simulated sessions as simulated. A score of 0 means that the classifier cannot distinguish between simulated and real log sessions and therefore the simulated sessions are similar to real log data sessions, whereas with a score of 1, simulated sessions and log data are completely different. Since we can distinguish between real log and simulated sessions, reporting the accuracy alone

can obfuscate some of the performance that F-score would highlight. F-score tells how precise the classifier is (i.e., how many instances it classifies correctly), as well as how robust it is (i.e., does not miss a significant number of instances). In fact, if F-score showed low precision/recall along with a low accuracy, we can have better confidence in the results. Therefore, we utilise all four metrics to demonstrate relative performance and consistency of the results.

3 DataSet

We use Sowiport[3] *User Search Session Data Set (SUSS)*[4] [19] for our experiments, which includes 484,437 individual search sessions, 179,796 queries and around 8 million log entries that was collected over a period of one year (from April 2014 to April 2015). Sowiport describes users' search actions using a list of 58 different actions that covers all user's activities while interacting with the interface of the search engine (e.g., formulating a query, clicking on a document, viewing the full document's content, selecting a facet, using search filters). For each user interaction, a session id, date stamp, length of the action and other additional information are stored to describe user's path during the search process. From the 484,437 individual search sessions in the dataset, we filter sessions that do not contain a query (i.e., users having searched nothing) or have invalid query annotations and we sample 100,000 sessions which we refer to as SUSS$^-$.

4 Results

For this evaluation test, we derived two separate simulation models (i.e., exploratory and lookup) and one global simulation model (i.e., first-order) that is trained on whole SUSS$^-$ dataset. For each model, we utilise the transition probabilities between states which are drawn from the log sessions and the simulated sessions separately to generate two independent samples. By feeding these data points to KS-2 we obtain the test statistic value (i.e., 0.00417 first-order, 0.00381 and 0.00302 for exploratory and lookup respectively) and compare it to the critical value for the two samples (i.e., 0.00421 first-order, 0.00389 and 0.00356 for exploratory and lookup respectively).

Results show that the statistical value is smaller than the critical value across all models, hence we retain the null hypothesis. Therefore, we conclude that the simulated and the real log sessions belong to the same distribution.

Since the KS-2 critical values are all significant, it means that query change as context factor does not improve the simulation or at least it is hard to quantify the improvement using a KS-2 test. Therefore, we need to adopt a second evaluation method: we investigate whether we can train a classifier and try to distinguish between real log and simulated sessions through controlled scenarios. For each scenario, we simulate an equal amount of sessions as present in the log data to balance class distribution.

[3] http://www.sowiport.de

[4] The dataset is publicly available at http://dx.doi.org/10.7802/1380

Table 1: Classification of real log sessions vs simulated sessions using first-order and contextual Markov model (CMM) approaches. We report the accuracy, recall, precision and F-score across 10-CV folds (while (1) averaging over three classifiers defined in subsection 2.2 and (2) using Auto-sklearn (AS.). Bold indicates the best result in terms of the corresponding metric. Lowest results are the best as we aim to reduce the classifier's capability to distinguish between real log and simulated sessions.

Approach		**Size**	**Accuracy**		**Recall**		**Precision**		**F-score**	
			Avg.	***AS.***	***Avg.***	***AS.***	***Avg.***	***AS.***	***Avg.***	***AS.***
first-order Markov model		1	0.661	0.660	0.814	0.796	0.543	0.558	0.651	0.656
CMM	Exploratory	0.39	0.611	0.625	0.628	0.673	0.506	0.502	0.560	0.575
	Lookup	0.61	**0.572**	**0.577**	**0.612**	**0.624**	**0.452**	**0.463**	**0.519**	**0.531**

Table 1 shows that when using contextual Markov with the exploratory-lookup approach, the model did better while simulating sessions for *"Lookup"* with an F-score of 0.519 in comparison to *"Exploratory"* with a score of 0.560. One possible explanation for this is that lookup sessions are probably easier to simulate since there is less variation. The exploratory group of users generate longer sessions, thus higher total of state transitions which results a diverse number of simulated sessions. In addition, table 1 also reports low precision (i.e., 0.452 for *"Lookup"* and 0.506 for *"Exploratory"*)/recall (i.e., 0.612 for *"Lookup"* and 0.628 for *"Exploratory"*) values along with a low accuracy score (i.e., 0.572 for *"Lookup"* and 0.611 for *"Exploratory"*) when using the exploratory-lookup approach in comparison to global first-order model, which indicates that we can have better confidence in the results.

In summary, we report that grouping user search sessions depending on their behavioural characteristics helps improving the simulation quality (i.e., reducing the accuracy of the classifier which is translated by lower F-score, recall and precision values).

5 Conclusion

In this paper, we propose a method to evaluate simulated user interactions in the context of a search session, which can be used as economic alternatives of user studies. We performed experiments using a real-world academic dataset with contextual Markov models and provided empirical results showing that the context-aware models allow to account for finer context granularity, i.e., more specific models. The proposed evaluation methods represents a theoretical foundation for experimental studies of sophisticated IR systems and opens up many new research directions. For example, we can use the classification-based methods to derive potentially better metrics than the existing ones that we proposed. The evaluation methods also opens up many interesting opportunities to leverage search log data to generate various realistic user simulators for evaluating complicated search systems.

References

1. Kumaripaba Athukorala, Dorota Głowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. Is exploratory search different? a comparison of information search behavior for exploratory and lookup tasks. *J. Assoc. Inf. Sci. Technol.*, 67(11):2635–2651, November 2016.
2. Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. Time drives interaction: Simulating sessions in diverse searching environments. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 105–114, 2012.
3. Mariana Belgiu and Lucian Drăguţ. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:24–31, 04 2016.
4. Ben Carterette, Ashraf Bah, and Mustafa Zengin. Dynamic test collections for retrieval evaluation. In *Proceedings of the 2015 international conference on the theory of information retrieval*, pages 91–100, 2015.
5. Ben Carterette, Evangelos Kanoulas, Mark Hall, and Paul Clough. Overview of the trec 2014 session track. Technical report, Delaware Univ Newark Dept. Computer and Information Sciences, 2014.
6. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
7. Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter. *Auto-sklearn: Efficient and Robust Automated Machine Learning*, pages 113–134. Springer International Publishing, Cham, 2019.
8. Donna Harman. Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 3(2):1–119, 2011.
9. Jeff Heaton. An empirical analysis of feature engineering for predictive modeling. *SoutheastCon 2016*, Mar 2016.
10. Kalervo Järvelin, Susan L Price, Lois ML Delcambre, and Marianne Lykke Nielsen. Discounted cumulated gain based evaluation of multiple-query IR sessions. In *European Conference on Information Retrieval*, pages 4–15. Springer, 2008.
11. Jiepu Jiang and James Allan. Correlation between system and user metrics in a session. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 285–288, 2016.
12. Karen Sparck Jones and Peter Willett. *Readings in information retrieval*. Morgan Kaufmann, 1997.
13. Frank J. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.
14. Diane Kelly. *Methods for evaluating interactive information retrieval systems with users*. Now Publishers Inc, 2009.
15. Aldo Lipani, Ben Carterette, and Emine Yilmaz. From a user model for query sessions to session rank biased precision (sRBP). In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 109–116, 2019.
16. Eren Manavoglu, Dmitry Pavlov, and C Lee Giles. Probabilistic user behavior models. In *Third IEEE International Conference on Data Mining*, pages 203–210. IEEE, 2003.
17. Gary Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.

18. David Maxwell and Leif Azzopardi. Agents, simulated users and humans: An analysis of performance and behaviour. In *Proceedings of the 25th ACM International Conference on information and Knowledge Management*, pages 731–740, 2016.
19. Philipp Mayr. Sowiport User Search Sessions Data Set (SUSS) (Version: 1.0.0), 2016.
20. J.R. Quinlan. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27(3):221–234, 1987.
21. Mark Sanderson. *Test collection based evaluation of information retrieval systems*. Now Publishers Inc, 2010.
22. Ahmad Shamshad, M. A. Bawadi, Wan Muhd Aminuddin Wan Hussin, Taksiah A. Majid, and S. Ahmad Mohd. Sanusi. First and second order markov chain models for synthetic generation of wind speed time series. *Energy*, 30:693–708, 2005.
23. Vu Tran, David Maxwell, Norbert Fuhr, and Leif Azzopardi. Personalised search time prediction using Markov chains. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pages 237–240, 2017.
24. Fan Zhang, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Min Zhang, and Shaoping Ma. Cascade or recency: Constructing better evaluation metrics for session search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 389–398, 2020.