Evaluating Simulated User Interaction and Search Behaviour

Saber Zerhoudi[1] , Michael Granitzer[1], Christin Seifert[2] and Jörg Schlötterer[2]

[1] University of Passau, Passau, Germany [2] University of Duisburg-Essen, Duisburg, Germany

Introduction

- **Research Question:** Can we develop an evaluation methodology to evaluate to which extend simulated user models can replace or complement sample-based ones?
- **Problem:** Information Retrieval (IR) system mainly use test collections *(contain no user interaction data)* and user studies *(expensive to conduct)* to evaluate its performance → Simulation
  - Most studies focus on the browsing, querying and document relevance models for search evaluation, but only few studies have investigated the utility of evaluating simulated user interactions [8]

→ Develop a methodology for evaluating simulated user sessions

Contribution

A pilot study to assess the capability of generating simulated search sequences representing an approximation of real behaviour.

- ❖ Conduct experiments on a real-world dataset and model users' browsing patterns using a first-order and a contextual Markov model approach that utilises user's browsing context based on common sense assumptions.
- ❖ Propose evaluation metrics to assess the realism of simulated user interactions using the *Kolmogorov-Smirnov statistical test* as an empirical validation and a *classification-based* evaluation technique to assess the quality of simulated search session.

Methods & Results

The quality of simulated user search sessions is usually evaluated by comparing real log and simulated data.

Our evaluation method utilizes the following user models:

1. ***First-Order Markov Model,*** where we propose to use of first- order Markov Chains to model the search dynamics [5].
2. ***Contextual Markov Model,*** where we categorise users into different groups based on their search behaviour [6]. 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.

Dataset

Sowiport User Search Session Data Set (SUSS)[1] [9] (~8 million log entries) describes users' search process using a list of actions that covers all user's activities while interacting with the interface of the search engine.

Kolmogorov-Smirnov-Based Evaluation

KS-two-sample goodness-of-fit test [7] help investigating whether two probability distributions can be regarded as indistinguishable.

Results show that the simulated and the real log sessions belong to the same distribution.

KS-2 critical values are all significant → It is hard to quantify the improvement using a KS-2 test.

Classification-Based Evaluation

1. Develop a set of features that represent the sequential nature of a user search session in the form of a feature vector.
2. Each user session is converted to a feature vector, labelled and fed to a classifier.
3. Train a classifier to distinguish simulated sessions from real log data sessions and report the results.
4. Report the Accuracy, Recall, Precision and F-score across tenfold cross-validation (while (1) averaging over the most popular algorithms in binary classification (i.e. Support Vector Machine [1], Decision Trees [2] (XGBoost), Random Forests [3]) and (2) using Auto-sklearn (AS.) [4].
5. Incorporate a bias in the classifier by weighting the class of real data to penalise bad real log sessions predictions. $(w_{real} = 10^4, w_{simulated} = 1)$

Table 1 shows that grouping user search sessions depending on their behavioural characteristics (*Exploratory vs. Lookup*) helps improving the simulation quality (i.e., reducing the accuracy of the classifier which is translated by lower F-score, recall and precision values).

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. Classification of real log sessions vs simulated sessions using first-order and contextual Markov model (CMM) approaches. 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.

Discussion

- Evaluating simulated user interactions can be used as economic alternatives of user studies.
- The proposed evaluation methods represents a theoretical foundation for experimental studies of sophisticated IR systems and opens up many new research directions.
- Classification-based methods can be used to derive potentially better metrics and open up many interesting opportunities to leverage search log data to generate various realistic user simulators for evaluating complicated search systems.

References

1. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. 20(3), 273–297 (1995)
2. Quinlan, J.R.: Simplifying decision trees. Int. J. Man Mach. Stud. 27(3), 221–234 (1987)
3. Belgiu, M., Dragu¸t, L.: Random forest in remote sensing: a review of applications and future directions. ISPRS J. Photogrammetry Remote Sens. 114, 24–31 (2016)
4. Feurer, M. et al. Auto-sklearn: efficient and robust automated machine learning (2019).
5. Shamshad, A. et al. First and second order Markov chain models for synthetic generation of wind speed time series. Energy 30, 693–708 (2005)
6. Marchionini, G.: Exploratory search: from finding to understanding. ACM 49(4), (2006)
7. Massey Jr., F.J.: The Kolmogorov-Smirnov test for goodness of fit. J. Am. Stat. (1951)
8. Carterette, B. et al.: Dynamic test collections for retrieval evaluation. (2015)
9. The dataset is publicly available at http://dx.doi.org/10.7802/1380.

This work has been partially carried out within the project "SINIR: Simulating INteractive Information Retrieval" funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 408022022.