Introduction

- **Research Question:** Can we develop an evaluation methodology 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.

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 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

Sowimport User Search Session Data Set (SUSS) [9] (~8 million log entries) describes users’ search 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.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Size</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Avg. AS</td>
<td>Avg. AS</td>
<td>Avg. AS</td>
<td>Avg. AS</td>
</tr>
<tr>
<td>First-order Markov model</td>
<td>1</td>
<td>0.661</td>
<td>0.660</td>
<td>0.814</td>
<td>0.796</td>
</tr>
<tr>
<td>CMM Exploratory</td>
<td>0.39</td>
<td>0.611</td>
<td>0.625</td>
<td>0.628</td>
<td>0.673</td>
</tr>
<tr>
<td>Lookup</td>
<td>0.61</td>
<td>0.572</td>
<td>0.577</td>
<td>0.612</td>
<td>0.624</td>
</tr>
</tbody>
</table>

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 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.

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

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).

References

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