Consumer Data Research Centre

An ESRC Data Investment

Session-Based Recommender Systems in E-Commerce

Sharon Liu[1], Emre Alper Yildirim[1] and Robert Franklin[2]
[1]University of Edinburgh, [2]Boots

Background and Motivation

Recommender systems (RS) are information retrieval systems that, in the context of e-commerce, take in variables such as item-item similarities, user preferences and user-item interactions to predict what would be the next best item to view or purchase. Due to their proven ability to increase consumer interaction and purchases, they are now a main staple for almost any internet-based businesses.

Boots' e-commerce website features an expansive catalogue spanning multiple categories with high traffic daily. Thus, a RS serves to not only recommend the most relevant items, but also prevent users from experiencing an overload of information and expose them to new items. Majority of web sessions occur on mobile platforms and are unidentifiable as users are typically not signed in. Furthermore, repeat visits from the same user are uncommon. As a result, there is limited user profile and user-specific historical session data to serve as model inputs. Thus, session-based recommender systems (SBRS) are ideal since recommendations are generated based solely only on the user's interactions in the current session.

As SBRS are trained only on session browsing sequences, recommendations are limited to items that appear during model training. This means that all other items cannot be considered for recommendation nor as inputs for prediction. This is known as the item cold start problem and notably, impacts newly launched items. As Boots is in the fast-moving consumer goods industry and releases on average about 100 new items daily, there is greater potential in addressing this problem.

Beyond accuracy metrics, diversity metrics are also critical for the long-term effectiveness of RS. This is due to the feedback loop between users and RS which can amplify popularity bias in the dataset and result in diminishing diversity and personalization in recommendations over time. In addition, as the RS works in real time serving recommendations to users, they also need to be highly scalable and adaptable to Boots' growth in catalogue size and web traffic in the long term.

Data and Methods

The main datasets were provided by Boots and consisted of users' timestamped product views, identified by their cookie ID. Conducting a survey of recent state-of-the-art SBRS, we shortlisted the following models based on their scalability and performance in past works:

- STAN – session-based nearest neighbour model that incorporates recency of items and sessions
- STAMP – neural network model that explicitly models session's current interests via the last click and applies attention mechanism to the session sequence
- GRU4Rec+ – recurrent neural network model that utilises Gated Recurrent Units (GRU) to model session sequences with novel ranking loss function
- SLIST – regression model that jointly optimizes two linear regression models, balancing between item-item similarity and sequential dependency of items

In addition, we propose a neighbourhood-based extension to SLIST, referred to as SLIST Ext, to incorporate new items for recommendation to alleviate the item cold start problem. This is done through augmenting the optimal item-item matrix by calculating the weighted average from the new items' nearest neighbours based on item features extracted from item metadata.

We investigate the performance of each model on Boots dataset in predicting the next item view after every time step of each session, among a recommendation list of 10 items. The models are evaluated on accuracy (hit rate, mean reciprocal rank), diversity (catalogue coverage, popularity bias) and scalability (training times, predicting times, memory usage). We also analyse each model's scalability by evaluating their performance across various training sizes. In alignment with evaluation performed in past works, we remove new items (i.e., items not present in the training set) from test sets for evaluation of base models, and perform a separate evaluation on the full test set with all items

retained for SLIST Ext and SLIST for comparison.

Key Findings

Overall, SLIST and GRU4Rec+ are the better performing models in terms of accuracy and diversity metrics. For scalability, GRU4Rec+ took 111x more training time while SLIST required 39x more memory. In addition, SLIST's computational needs scales with number of items, whereas GRU4Rec+ scales with number of sessions. As the scale of growth in web traffic is typically greater than the growth in catalogue size, SLIST would most likely be more sustainable in the long run. In addition, our proposed extension, SLIST Ext, displayed performance gains over the base model in terms of accuracy. Performance gains were marginal partially due to the low product views for new items, which may have stemmed from the lack of exposure and popularity bias perpetuated by the current RS employed.

Overall, SLIST is the optimal model choice as (i) it outperformed all other models based on accuracy – our primary evaluation metric, (ii) its short training time allows for more frequent model updates which could be critical for Boots in the fast-moving consumer goods industry, (iii) it is more feasible in the long run as its computational needs scales with catalogue size, and (iv) its model simplicity allows for simple model extensions to alleviate the item cold start problem.

Beyond deciding the optimal model choice for Boots, we also reported the strengths and weaknesses of each model specific to Boots' dataset. STAN's superior performance in hit rate but poor performance in mean reciprocal rank suggests that users' browsing behaviour demonstrate strong clustering behaviour which is easily exploited by the neighbourhood-based approach. However, such an approach also left it vulnerable to the popularity bias in the dataset and the inability to capture global patterns resulted in poorer ranking for recommendations. GRU4Rec+ outperformed in both diversity metrics as its negative-based sampling discounted the popularity bias to achieve consistently high performance in diversity metrics across various training sizes.

Note: These results were obtained on a small subset of Boots' dataset.

Value of the research

Our research serves as a starting point for Boots in identifying the optimal model. As there is often a gap in performance between online and offline evaluations, this work serves as the preliminary stage to shortlist models for online evaluations, as these can often be expensive and time-consuming to conduct.

We also revealed latent characteristics of users' browsing behaviour which interplay with the unique characteristics of each model to influence model performance. This research established the various strengths and weaknesses of each model specific to Boots, which could lead to further research such as building upon these models to capitalize on their strengths and address their weaknesses.

In addition, our proposed extension also demonstrated the value of simple model extensions for recommendation of new items, an area often overlooked by researchers. This could serve as motivation for further research into the online performance of recommender systems for the recommendation of newly launched items.

Model	Accuracy		Diversity		Scalability		
	Mean Reciprocal Rank	Hit Rate	Catalogue Coverage	Popularity Bias	Training (min)	Predicting (ms)	Memory (MB)
STAN	0.1417	**0.3986**	0.7130	0.0630	**0.01**	39.73	31.03
STAMP	0.1755	0.3385	0.7273	0.0543	50.81	9.95	109.08
GRU4Rec+	0.1732	0.3516	**0.8948**	**0.0266**	19.92	20.11	**11.33**
SLIST	**0.1816**	0.3583	0.8132	0.0559	0.18	**6.66**	444.92

*Table 1: Performance metrics of base models. The best performing model is marked in **bold**, and the second-best performing model is underlined.*

Model	Number of new items	Mean Reciprocal Rank	Hit Rate
SLIST	-	0.17988	0.35479
SLIST Ext	95	**0.17990**	**0.35492**

*Table 2: Accuracy metrics of SLIST and SLIST Ext on test set with new items retained. The better performing model is marked in **bold**.*