# Serving Low-Latency Session-Based Recommendations at bol.com

Barrie Kersbergen[1,2], Olivier Sprangers[1], Sebastian Schelter[1]
[1]*AIRLab, University of Amsterdam* [2]*bol.com*
bkersbergen@bol.com [o.sprangers, s.schelter]@uva.nl

Session-based recommendation targets a core scenario in e-commerce and online browsing. Given a sequence of interactions of a visitor with a selection of items, we want to recommend to the user the next item(s) of interest to interact with [1]–[3]. This machine learning problem is crucial for e-commerce platforms, which aim to recommend interesting items to buy to users browsing the site.

**Challenges in scaling session-based recommendation**. Scaling session-based recommender systems is a difficult undertaking, because the input space (sequences of item interactions) for the recommender system is exponentially large, which renders it impractical to precompute recommendations offline and serve them from a data store. Instead, session-based recommenders have to maintain state in order to react to online changes in the evolving user sessions, and compute next item recommendations with low latency [3], [4] in real-time. Recent research indicates that nearest-neighbor methods provide state-of-the-art performance for session-based recommendation, and even outperform complex neural network-based approaches in offline evaluations [2], [3]. It is however unclear whether this superior offline performance also translates to increased user engagement in real-world recommender systems. Furthermore, it is unclear whether the academic nearest-neighbor approaches scale to industrial use cases, where they have to efficiently search through hundreds of millions of historical clicks while adhering to strict service-level-agreements for response latency.

**A novel recommender system for bol.com**. We created a scalable adaptation of the state-of-the-art session-based recommendation algorithm VS-kNN [2]. Our approach minimises intermediate results, controls the memory usage and prunes the search space with early stopping. As a consequence, this approach drastically outperforms VS-kNN in terms of prediction latency, while still providing the desired prediction quality advantages over neural network-based approaches. Furthermore, we designed and implemented a real-world system around this algorithm, which is deployed in production at bol.com. Figure 1 illustrates the high-level architecture of our system. In order to tackle the scalability challenge, we leverage an offline data-parallel Spark job that generates a session similarity index. We replicate our index to all recommendation servers, and colocate the session storage with the update and recommendation requests, so that we only have to use machine-local reads and writes for maintaining sessions and computing recommendations. Our system currently computes recommendations on the product detail pages, e.g., the “others also viewed” recommendations on https://go.bol.com/p/9200000055087295.

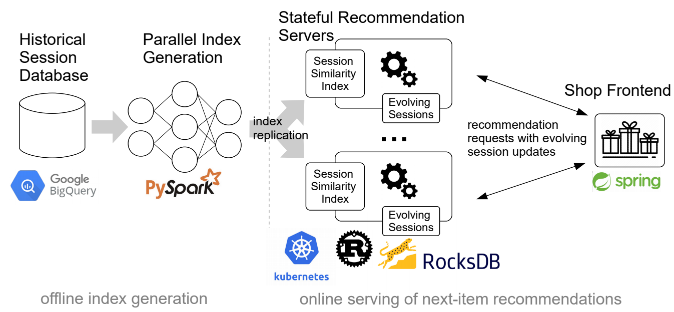


Fig. 1. High level architecture of our recommendation system. The offline component (left) generates a session similarity index 1 from several billion historical click events via a parallel Spark job in regular intervals. The online serving machines (right) maintain state about the evolving user sessions 2, and leverage the session similarity index to compute next item recommendations in response to recommendation requests from the shopping frontend 3.

**Evaluation**. We ran load tests on our system with 6.5 million distinct items in its index and find that it gracefully handles more than 1,000 requests per second and responds within less than 7 milliseconds at the 90th percentile while using only two vCPU’s in total. Our system easily handles up to 600 requests per second during an A/B test on the e-commerce platform at bol.com with very low response latencies at the 90th percentile of around 5 milliseconds. The session recommendations produced by our system significantly increase customer engagement by 2.85% compared to classical item-to-item recommendations (as produced by our legacy system).

*This work was supported by Ahold Delhaize. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their employers.*

## References

[1] Q. Liu *et al.*, “Stamp: short-term attention/memory priority model for session-based recommendation,” *KDD*, 2018.
[2] M. Ludewig, N. Mauro, S. Latifi, and D. Jannach, “Performance comparison of neural and non-neural approaches to session-based recommendation,” in *RECSYS*, 2019.
[3] B. Kersbergen and S. Schelter, “Learnings from a retail recommendation system on billions of interactions at bol.com,” *ICDE*, 2021.
[4] I. Arapakis, X. Bai, and B. B. Cambazoglu, “Impact of response latency on user behavior in web search,” in *SIGIR*, 2014.