Session-based Recommendation: Challenges and Recent Advances

Dietmar Jannach, AAU Klagenfurt, Austria

dietmar.jannach@aau.at
Recommender Systems

• A central part of our daily user experience
  – They help us locate potentially interesting things
  – They serve as filters in times of information overload
  – They have an impact user behavior and business
Recommendations everywhere
Recommendations everywhere
Recommendations everywhere
A field with a tradition

- **1970s**: Early roots in IR and what was called “Selective Dissemination of Information”
- **1990s**: A field develops, “content-based” approaches, Collaborative Filtering
- **2000s** and beyond: The Netflix Prize and its implications
- **Today and the future:**
  - Deep learning everywhere
  - But are we focusing on the most important problems?
The recommendation problem

• A very general definition:
  – “Find a good/optimal selection of items to place in the recommendation list(s) of users”

• The corresponding questions:
  – What determines a good/optimal selection?
    • Help users find something new?
    • Show the user alternatives to a certain item?
    • The diversity of the recommendations?
  – Good or optimal for whom?
    • The consumer, the platform or retailer, the manufacturer, all of them?

A common problem abstraction

- Recommendation as a matrix completion task

<table>
<thead>
<tr>
<th></th>
<th>Item1</th>
<th>Item2</th>
<th>Item3</th>
<th>Item4</th>
<th>Item5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>5</td>
<td>?</td>
<td>4</td>
<td>4</td>
<td>?</td>
</tr>
<tr>
<td>User1</td>
<td>3</td>
<td>1</td>
<td>?</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>User2</td>
<td>?</td>
<td>3</td>
<td>?</td>
<td>?</td>
<td>5</td>
</tr>
<tr>
<td>User3</td>
<td>3</td>
<td>?</td>
<td>1</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>User4</td>
<td>?</td>
<td>5</td>
<td>5</td>
<td>?</td>
<td>1</td>
</tr>
</tbody>
</table>

- Goal:
  - Learn/Optimize a prediction function from the data

- Quality assessment:
  - Prediction error on the test data
But, think again of about this one

- Past ratings do not play an obvious role
- There’s seemingly not even personalization
- Nonetheless, it is a key application example in the literature
Outline

• Characterizing the session-based recommendation problem

• Algorithmic approaches for “next event” predictions
  – Categorization
  – A performance comparison

• Session-aware recommendation in e-commerce
  – On short-term intents, reminders, trends and discounts
Outline

• Characterizing the session-based recommendation problem

• Algorithmic approaches for “next event” predictions
  – Categorization
  – A performance comparison

• Session-aware recommendation in e-commerce
  – On short-term intents, reminders, trends and discounts
Session-based Recommendation

• Instead of a rating matrix, we are given a sequentially ordered log of user interactions
  – e.g., item views, purchases, listening or viewing events, ...

• In many cases,
  – user cannot be identified (first-time users, users not logged in)
  – no longer-term preference information is available
  – user interest/intention/preferences must be assessed from a small set of interactions
Session-based Recommendation

• Guessing the intention can be difficult
Session-based Recommendation

- Also in online music recommendation
- Our user searched and listened to “Last Christmas” by Wham!
- Should we, ...
  - Play more songs by Wham!? 
  - More pop Christmas songs?
  - More popular songs from the 1980s?
  - Play more songs with controversial user feedback?
Session-aware Recommendation

- In some domains, past sessions of the current user are also known,
  - potential for personalization
  - possibility to remind users of objects
- We call “session-aware” recommendation
- Generally,
  - both session-based and session-aware recommendation as specific subtasks for sequence-aware recommenders

Quadrana, M., Cremonesi, P. and Jannach, D.: "Sequence-Aware Recommender Systems". ACM Computing Surveys. ACM, 2018
A Problem Abstraction

- Past sessions of the current user
- Past sessions of the user community
- Current session
Sequence-aware Recommendation

- High-level overview
  - specific types of inputs, specific computational tasks

Specific computational tasks:
- Context adaptation
- Trend detection
- Repeated recommendation
- Consideration of constraints
Outline

• Characterizing the session-based recommendation problem

• Algorithmic approaches for “next event” predictions
  – Categorization
  – A performance comparison

• Session-aware recommendation in e-commerce
  – On short-term intents, reminders, trends and discounts
Operationalizing the Research Problem

• **Background**
  – Intention of user is not known, not clear if we should recommend more similar items or, e.g., accessories

• **Computational task, simplified to:**
  – Predict subsequent user action(s), given
    • the last N actions by the user (e.g., in the current session)
    • other types of information (community behavior, metadata, ...)

• **Evaluation**
  – Use standard IR measures (precision, recall, MRR, ...)
  – Some interaction log datasets are publicly available
    • But can be biased
A Problem Abstraction

Past sessions of the current user

Past sessions of the user community

Current session
A Problem Abstraction

Past sessions of the current user

Current session

Past sessions of the user community
Basic Algorithmic Approaches

- Item co-occurrences in individual sessions
  - “Customers who bought”, association rules of size 2
- Simple Markov Chains and “Sequential Rules”
  - Count how often appeared (immediately) after another in the training data
- Session-based nearest neighbors
  - Look for similar past sessions
  - Use a weighted prediction scheme
  - Different similarity functions possible
    - Sequence-agnostic and sequence-aware ones
Advanced Algorithmic Approaches

- Sequence-learning techniques
  - Frequent pattern mining
    - Frequent item sets, frequent sequential patterns
  - Sequence modeling
    - Markov Models, Recurrent Neural Networks
  - Distributed item representations
    - Distributional and Latent Markov embeddings
- Sequence-aware matrix factorization
- Hybrids
  - Factorized Markov Chains, others
Deep Learning

- Recurrent Neural Networks (RNN)
  - can learn from sequential data
  - a “natural choice” for the problem

- Recent algorithm: GRU4REC
  - Proposed by Hidasi et al. (2016)
    - Multiple improvements since then
  - Uses Gated Recurrent Units
  - Several technical innovations to ensure scalability

Performance Comparison

• Background
  – Today, no “standard” for benchmarking session-based algorithms exist
  – Researchers often use
    • different evaluation protocols and measures
    • different datasets
    • different baselines
  – This makes the assessment of the true value of new approaches difficult
Performance Comparison

• Recently conducted an extensive set of experiments, including
  – both simple and sophisticated algorithms,
  – different datasets (e-commerce, music, news), and
  – different performance measures
    • including computational complexity
  – single-split and multiple-split evaluation protocol

Scalability Issues

• Naïve nearest-neighbor methods do not scale
• Our approach
  – Use compact in-memory data structures
    • Supporting fast look-up of possible neighbors
  – Use data sampling
    • E.g., 1,000 out of several million past sessions
    • Focus on most recent events
    • Only small accuracy compromises in many cases

• Results
  – Prediction time, e.g., at about 30 ms per request
  – Immediate “model updates” possible
Scalability Issues

• Complex methods
  – can require substantial amount of resources for training
  – but are usually fast at prediction time

• Example
  – E-commerce dataset, about 7 million sessions
  – A few hours for training (GPU-based)
    • Challenge lies in parameter optimization
    • New data usually requires full re-training
  – Main memory requirements largely depend on catalog size
Performance Comparison

• Main outcomes
  – In almost all configurations, one of the simple or almost trivial methods outperformed the most recent sequence learning method GRU4REC (2.0)
  – Much room for improvement regarding the development of more complex methods
    • e.g., hybrids

Side observations / remarks

• Careful choice of baselines needed
• Finding a good baselines is a not trivial
  – Ranking of algorithms varies across datasets
• Neighborhood-based baselines should be included in future experiments as well
Outline

• Characterizing the session-based recommendation problem

• Algorithmic approaches for “next event” predictions
  – Categorization
  – A performance comparison

• Session-aware recommendation in e-commerce
  – On short-term intents, reminders, trends and discounts
Session-Aware Recommendation

• Investigated various aspects regarding the success of recommenders in e-commerce
  – based on data shared to us by a large retailer
  – partly based on field test (A/B study)
• Specific questions
  – How important are long-term models?
  – Should we remind users of already seen items?
  – Can we leverage community trends in the recommendation process?
  – Should we recommend discounted items?

Long-term and short-term models

- Being able to predict which kinds of things a certain user generally likes, is important.
- Here’s what the customer looked at or purchased during the last weeks:
  - Nike t-shirt
  - Blue jeans
  - Red t-shirt
  - Yellow t-shirt

- Now, he or she return to the shop and browse these items:
  - Converse sneakers
  - Digital watch
  - Gold watch
What to recommend?

- Some plausible options
  - Only shoes or only watches?
  - Mostly Nike shoes?
  - Maybe also some T-shirts?

- Using the matrix completion formulation
  - One trains a model based only on past actions
  - The context of the user’s current shopping intent is considered only in “context-aware” recommenders
  - Without the context, the algorithm will probably most recommend mostly T-shirts and trousers. Is this what you expect?
On short-term intents

• Research question:
  – What is the relative importance of adapting recommendations to users’ short-term intents (shopping goals) when they visit the site?

• Measurement approach
  – “Hide-and-predict” simulation experiments on log data from a large online shop (Zalando)
  – Compare capability of session-aware and session-agnostic algorithms of predicting the purchased items in a given session

Contextualization Strategies

- Various comparably simple “real-time” strategies tested, e.g.,
  - CoOccur, i.e., “Customers who bought ... also bought”
- Feature Matching (FM)
  - Rank items up when they have features in common with those from the current session (e.g., same brand)
- Recently Viewed (RV)
  - Recommend recently viewed items in reverse chronological order
Technical approach

• Simple two-stage approach
  – Stage 1: Learn a long-term user profile, rank items
  – Stage 2: Filter or re-rank items based on the assumed short-term situation or intents

• Stage 1 can be done offline
  – Various algorithms were tested
    • Bayesian Personalized Ranking, Factorization Machines, Item-item-Nearest-Neighbors, Popularity-based and Random Baselines

• Stage 2 must be “real-time”
  – Can still be based on complex models, e.g., RNNs
Dataset

• Evaluations mostly based on an e-commerce log dataset
  – By Zalando, a major European online fashion retailer
  – Dataset contains sample of user activity logs
    • 1 million purchases
    • 20 million view events
    • 170,000 sessions
    • 800,000 users (many non-purchasers)
    • 150,000 different items

• Dataset is very sparse
  – Many users without any purchase
Empirical Results

• Observations for dense dataset (example)
  – Recall of best baseline method (BPR): 40%
  – Other:
    • Customers who bought ... : 49%
    • Just show me what I have seen : 64%
    • Show me similar things : 71%
    • Combining long- and short-term: 73%

[Chart showing comparison between methods: BPR, CoOccur, Recently Viewed, Feature Matching, Hybrid]
Insights

• Combination of various short-term and long-term signals as the most effective strategy

• Choice of baseline ranking method is relevant
  – Better baseline ranking in most cases leads to stronger overall results

• Importance of short-term adaptation
  – Contextualization-only methods often already better than the best long-term profile
  – Becomes more and more relevant, the more is known for the current session
  – Reminding is a very effective strategy
More on reminders

- Follow-up study
  - Deeper analysis of reminders
    - Using again the Zalando dataset
  - Development of more intelligent reminding strategies
  - Evaluation of reminding strategy in field test

A field study on the business value

• A/B-tested different strategies on an e-commerce site for electronic gadgets

• Competing strategies
  – BPR as a learning-to-rank model
  – Similarity-based recommendation (using a reference item)
  – A personalized similarity-based approach
  – Popularity-based baseline
  – Present recently viewed items
    • In reverse chronological order
Recommendations in A/B test

Einen großen Todesstern von etwa 6 cm könnt ihr mit diesem Gadget herstellen! Diese enorme Kühlpower führt also dazu, dass dieses Gadget nicht für ein einziges Glas geeignet ist – es will mehr! Als Material ist Silikon angegeben und wenn ihr nicht gerade gegen die Sonne kampft, dann lässt sich darin zusammen mit einer Feuerquelle auch Nahrung herstellen.

Hier geht's zum Gadget >>

 interesse Gadgets - Schon gesehen?

Für alle Mad scientists: Eislöwen in Gehirnform für 1.80€
Raaawrrr – Eisschweine im Halblosenlock ab 1.35€
Schwarzer Humor: Eisschnecken im Form der Titanic (mit passenden
Ja ich will: Die Kugel (Eischnecke)-Form für 1.39€
Field study outcomes

- “Success rate” as business measure
  - Click on recommendation and click on outgoing link to external retailer
  - Pure reminders led to best business value in this specific situation
Better reminders?

• General filtering strategy
  – Do not remind users of items in categories where recently a purchase was made

• Designed different “adaptive” strategies
  – **Recency-based baseline**: Use reverse chronological order
  – **Intensity-based ranking**: Rank reminder items based on the number of past clicks
  – **Item-similarity ranking**: Select reminder items based on their fit for the current session
  – **Session-similarity ranking**: Select reminders based on their occurrence in similar past sessions
Empirical Evaluation

- Baseline ranking method:
  - Session-based nearest neighbors
    - Configured to include reminders as well
- Results (hit rate, example, 2 evaluation variants)
  - v1 hides view event for target item, v2 reveals them

- Adaptive reminders better than simple reminders
On Trends and Discounts

• More general question
  – What are factors that determine the success of a recommendation?

• Our dataset includes additional information:
  – For each view event, the three recommendations displayed on the item detail page
  – Click events on the recommended items
  – Information about discounts (visible to customers) at the time of recommendation
  – (Purchase information)
Research Goals

• Goals
  – Analyze which recommendations are successful, i.e., lead to a purchase event later on
  – Operationalize these insights in new recommendation algorithms
Analyzing the effectiveness of the recommendations on the site

• Reminders:
  – Only 10% of the recommendations seen before
  – But 44% of the successful ones were already known

• Short-term intents
  – Recommendations are more likely to be successful when from the same brand, category etc.

• Trends
  – Success rate of four times higher when the recommended item is trending on that day

• Discounts
  – Recommending on-sale items boosts the success rate
A systematic feature analysis

- Engineered about 90 features to predict the success of a recommendation
  - Framed as a classification problem
  - Systematically determined feature importance values

<table>
<thead>
<tr>
<th>Feature</th>
<th>Gain Ratio</th>
<th>Chi Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discount level</td>
<td>0.439</td>
<td>0.556</td>
</tr>
<tr>
<td>Current popularity (day)</td>
<td>0.371</td>
<td>1.000</td>
</tr>
<tr>
<td>Discount flag</td>
<td>0.325</td>
<td>0.556</td>
</tr>
<tr>
<td>Viewed before</td>
<td>0.286</td>
<td>0.435</td>
</tr>
<tr>
<td>Current popularity (week)</td>
<td>0.242</td>
<td>0.785</td>
</tr>
<tr>
<td>Distance to first item view (in days)</td>
<td>0.232</td>
<td>0.428</td>
</tr>
<tr>
<td>Distance to last item view (in days)</td>
<td>0.217</td>
<td>0.441</td>
</tr>
<tr>
<td>Distance to first item view (in sessions)</td>
<td>0.214</td>
<td>0.428</td>
</tr>
<tr>
<td>Distance to last item view (in sessions)</td>
<td>0.210</td>
<td>0.443</td>
</tr>
<tr>
<td>Current popularity (month)</td>
<td>0.201</td>
<td>0.563</td>
</tr>
</tbody>
</table>
 Operationalization in algorithms

• Approach 1: A weighted two-phase approach
  – Create a candidate list of, e.g., 200 items using a baseline technique, in our case nearest-neighbors
  – Re-Rank items based on a weighted scoring scheme
  – Scoring functions
    • Content-wise similarity
    • Appearance in recent history (reminders)
    • Recent popularity on site
    • Level of discount
  – Importance weights have to be fine-tuned
Operationalization in algorithms

• Approach 2: A classification based approach
  – Framed scoring problem as a classification problem
  – Engineered 32 features in a similar way as was done for the feature importance analysis
    • Slightly different problem formulation
    • We are not interested analyzing success factors in general, but to predict the purchase probability for the given session
  – Used a deep neural-network approach for classification (H2O.ai library)
    • Outperformed Random Forests and manually tuned weights
Results

• Deep Learning based method led to best results
  – Independent of the chosen baseline ranking technique

• Random Forests were not better than the manually tuned weighted hybrid

<table>
<thead>
<tr>
<th>Baseline Metric@10</th>
<th>C-KNN HR</th>
<th>MRR</th>
<th>C-CoOcc HR</th>
<th>MRR</th>
<th>BPR HR</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>No post-processing</td>
<td>0.268</td>
<td>0.091</td>
<td>0.123</td>
<td>0.046</td>
<td>0.062</td>
<td>0.021</td>
</tr>
<tr>
<td>FM</td>
<td>0.281</td>
<td>0.093</td>
<td>0.145</td>
<td>0.052</td>
<td>0.119</td>
<td>0.046</td>
</tr>
<tr>
<td>IRec-FM</td>
<td>0.306</td>
<td>0.097</td>
<td>0.266</td>
<td>0.096</td>
<td>0.262</td>
<td>0.111</td>
</tr>
<tr>
<td>DR-FM</td>
<td>0.316</td>
<td>0.177</td>
<td>0.242</td>
<td>0.120</td>
<td>0.168</td>
<td>0.094</td>
</tr>
<tr>
<td>RPOP-FM</td>
<td>0.361</td>
<td>0.187</td>
<td>0.233</td>
<td>0.103</td>
<td>0.216</td>
<td>0.096</td>
</tr>
<tr>
<td>RFPredict</td>
<td>0.381</td>
<td>0.248</td>
<td>0.274</td>
<td>0.150</td>
<td>0.241</td>
<td>0.119</td>
</tr>
<tr>
<td>WR(RPOP,DR,0.5)-FM</td>
<td>0.382</td>
<td>0.220</td>
<td>0.262</td>
<td>0.121</td>
<td>0.225</td>
<td>0.100</td>
</tr>
<tr>
<td>DeepPredict</td>
<td>0.405</td>
<td>0.284</td>
<td>0.322</td>
<td>0.205</td>
<td>0.301</td>
<td>0.188</td>
</tr>
</tbody>
</table>
General insights

• First approach in academia to “reconstruct” success factors of recommendations from log data
• Could successfully operationalize the insights in a new prediction method
• Feature engineering is important
• Domain-dependent aspects should be considered
  – Reminding or not
  – Recommending discounted items or not
  – Recommending trending items or not
Summary

- Session-based recommendation as a highly relevant problem in practice
- Recently increased interest
  - public datasets
  - success of RNNs (non-success at matrix completion)
- Agreed upon benchmark setup still needed
  - protocols, measures, baselines
- Domain-specific characteristics can be important (e.g., short-term community trends)
Outlook

- Better situation-dependent recommendations
  - Try to guess the customer’s decision and consumption situation
    - e.g., trying to learn what the options are, forming a choice set
    - e.g. listening to favorite artists, willing to explore new
  - Appropriate evaluation procedures needed, e.g., based on user studies

- More sophisticated algorithms for next-event prediction

• Thank you for your attention
• dietmar.jannach@aau.at

• Thanks to:
  – Paolo Cremonesi, Michael Jugovac, Iman Kamehkhosh, Lukas Lerche, Malte Ludewig, Massimo Quadrana
Sequence-aware Recommendation

- High-level overview
  - specific types of inputs, specific computational tasks
Evaluation method

- “Hide-and-predict” simulation experiments on log data from a large online shop (Zalando)
- Proposed parameterized evaluation protocol
  - Parameters, e.g., How many interactions of the current session are revealed
  - Allows us to assess how quickly algorithms adapt, how the importance of short-term intents increases
  - Recall and MRR as measures
“Reco-minders” in practice

• Log data contains recommendation list for the view events
  – Every 10th recommendation was a reminder
  – More than 40% of the successful recommendations were already known items
    • This also means that recommending unknown items is also very important, and helps users discover things
  – Users inspect an item multiple times before making a purchase
  – During one session, users inspect items of a small set of categories
    • Reminders as navigation shortcuts?
# Top features (recommendation)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Popularity</td>
<td>Normalized popularity of the item in the same day, week, or month</td>
</tr>
<tr>
<td>Viewed before</td>
<td>True, if the item was viewed before by the user</td>
</tr>
<tr>
<td>Views count</td>
<td>Number of previous views of the item by the user</td>
</tr>
<tr>
<td>Distance to first view</td>
<td>Distance to the first item view by the user in days or sessions</td>
</tr>
<tr>
<td>Distance to last view</td>
<td>Distance to the last view of the item in days or sessions</td>
</tr>
<tr>
<td>Brand ratio</td>
<td>Fraction of items of the same brand in the last 1, 2, and 3 sessions</td>
</tr>
<tr>
<td>Brand popularity</td>
<td>Overall popularity of the brand for the same day, week, or month</td>
</tr>
<tr>
<td>Color ratio</td>
<td>Fraction of items with the same color in the last 1, 2, and 3 sessions</td>
</tr>
<tr>
<td>Color popularity</td>
<td>Overall popularity of the color for the same day, week, or month</td>
</tr>
<tr>
<td>Category ratio</td>
<td>Fraction of items of the same category in the last 1, 2, and 3 sessions</td>
</tr>
<tr>
<td>Category popularity</td>
<td>Overall popularity of the category for the same day, week, or month</td>
</tr>
<tr>
<td>Price level ratio</td>
<td>Fraction of items of the same price level in the last 1, 2, and 3 sessions</td>
</tr>
<tr>
<td>Discount granted?</td>
<td>True, if the item is discounted</td>
</tr>
<tr>
<td>Discount level</td>
<td>Level of discount from 0 (no discount) to 3 (high discount)</td>
</tr>
</tbody>
</table>
Customer are often well-focused

- Impact of item features on success of recommendations visualized
- About 1% success rate in general