Seq2Bubbles: Region-Based Embedding Learning for User Behaviors in Sequential Recomenders

Qitian Wu, Chenxiao Yang, Shuodian Yu, Xiaofeng Gao, Guihai Chen
Shanghai Jiao Tong University
Background for Recommendation

- Predict the **next item** based on historically clicked items of the user
- Most existing sequential recommendation models:
  1. **Embedding**: transform the item sequence into a sequence of vectors
  2. **Encoding**: encode the sequence to estimate user interests
  3. **Decoding**: compute similarity between the user state and a target item

![Diagram of Seq2Bubbles](image)

Squash a high-dimensional sequence into a single point
Motivation

- User interests often distribute over items of different aspects
  - Distribution of user interest tends to be multi-modal
- User interests for different items have distinct concentration levels
  - user’s concentration: variance of user’s clicked items in a specific aspect
  - more (less) diverse items in the aspect with stronger (weaker) concentration

Traditional point embedding fails to capture such distinct concentration levels!
Our Solutions: Region-based Embedding

☐ Basic idea: embed a sequence into a set of bubbles
  • a hyper-ellipsoid in vector space
  • bubble center: clicked item embedding
  • bubble radius: embody concentration of user interests
  • a union of bubble embedding for sequence reflect user interests

\[ \bigcup_{k=1}^{m} \{ x : \| (x - c_k) \odot \frac{1}{r_k} \|_2 \leq 1 \} \]

☐ Advantages:
  • Superior Expressiveness
  • Enough Flexibility
  • Interpretability

Key insight: regions enclosed by bubbles represent multi-modal interest and user intent
Proposed Model Overview
Encode item embedding sequence to extract useful information:

- Filter out *noise* existing in behavior sequences
- Mine *temporal dependency* and user’s interests evolution
- Distinguish the *importance* of different historical behaviors
Self-attentive architecture:

- Lower-level sequential unit $\Phi_A(\cdot)$ to aggregate historical items

$$z_k = \sum_{j=1}^{k} \alpha_{jk} q_{ij}, \quad \text{where} \quad \alpha_{jk} = \sigma \left( \frac{(W_{k}^1 q_{ik})^T (W_{Q}^1 q_{ij})}{\sqrt{d}} \right)$$

$$h_k = \text{Dropout}(\text{PReLU}(W_{N}^1 z_k + b_{N}^1))$$

- High-level readout unit $\Phi_R(\cdot)$ to estimate radius of bubbles

$$z_k = \sum_{j=1}^{m} \beta_{jk} \cdot h_{j}, \quad \text{where} \quad \beta_{jk} = \sigma \left( \frac{(W_{k}^2 h_{k})^T (W_{Q}^2 h_{j})}{\sqrt{d}} \right)$$

$$r_k = \text{Softplus}(W_{N}^2 z_k + b_{N}^2), \quad k = 1, \cdots, m$$
Model: Decoding Layer

- Compute the similarity between bubble embedding and target item
  → the distance from a point to the surface of a hyper-ellipsoid?

- Approximation:
  - Consider a circumscribed hyper-cube outside the hyper-ellipsoid region

\[
b = \{c, r\}: [c_1 - r_1, c_1 + r_1] \times \cdots \times [c_d - r_d, c_d + r_d]
\]

\[
D(b, q) := \min_{e \in \{-1,1\}^d} d(c + e \odot r, q)
\]

\[
D(\mathcal{B}^m, q_i) := \min_{1 \leq k \leq m} D(b_k, q_i),
\]

\[
= \min_{1 \leq k \leq m} d(c_k + \delta(q_i - c_k) \odot r_k, q_i)
\]

\[
S(\mathcal{B}^m, q_i) = \max_{1 \leq k \leq m} s(c_k + \delta(q_i - c_k) \odot r_k, q_i)
\]
Maximum operation only selects one bubble
- The gradient only update one item
- Ignore effects from different feature dimensions

A generalized version:
- max-pooling to select dominant bubbles in each feature dimension
\[
p_k = [c_k + \delta(q_i - c_k) \odot r_k] \odot q_i, \quad k = 1, \ldots, m,
\]
\[
a_m = \text{MaxPooling}([p_1, p_2, \cdots, p_m])
\]
\[
S(B_m, q_i) = s(a_m, q_i).
\]
\[
\hat{y}_{ui}^{m} = (q_i)^	op a_m
\]
Model: Context-Aware Representation

- **Context-aware bubble**
  - incorporate information of clicked items related to the target item

\[
\tilde{c}_m = \sum_{k=1}^{m} \gamma_{km} q_{ik}, \quad \text{where} \quad \gamma_{km} = \sigma\left(\frac{(W_k^3 h_k)^	op (W_Q^3 q_i)}{\sqrt{d}}\right) \quad \tilde{r}_u^m = \text{Softplus}(W_N^3 [\tilde{c}_m || q_i] + b_N^3)
\]

- **Estimate with bubble embedding and context-aware state**
  - inherent interests from observed sequence
  - relations between historical behaviors and target items

\[
\tilde{p}_m = \tilde{c}_m + \delta(q_i - \tilde{c}_m) \odot \tilde{r}_m
\]
\[
\hat{y}_{ui} = (q_{it})^\top a_m + (q_{it})^\top \tilde{p}_m
\]
Model Optimization: Supervised Learning

- The model estimates the probability with the relevance score
  \[ P(i|T_u^m) = \sigma(\hat{y}_{ui}^m) \]

- Adopt Bayesian Personalized Ranking as objective
  \[ L = \sum_{u \in U} \sum_{m=1}^{n_u-1} \log P(i_{m+1}^u = \hat{i}_{m+1}^u | T_u^m) \]

- For the mini-batch data \( \{T_u\}_{u \in U_b} \)
  \[ L_{sup} = \sum_{u \in U_b} \sum_{m=1}^{n_u-1} \log \sigma(\hat{y}_{u,i_{m+1}^u}^m - \hat{y}_{u,i_{m+1}^u}^m) \]
Model Optimization: Contrastive Regularization

- Directly optimize the loss function lead to over-fitting
  - Radius vectors of bubbles tend to be updated radically

- Inspired by contrastive learning
  - Enforce self-consistency within a user sequence
  - Enlarge the mutual information between estimated bubble embedding and historical items
  - Guide the model to ‘look back’

\[
L_{reg} = - \sum_{u \in U_b} \sum_{m=t+1}^{n_u} \log \frac{\exp(S(\overline{B}_u^m, q_{i_{m-t}}^u))}{\sum_{u' \in U_b} \exp(S(\overline{B}_u^m, q_{i_{m-t}}^{u'}))}
\]
# Experiments: Overall Results

Table 1: Comparative results for different methods

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Metric</th>
<th>POP</th>
<th>BPR-MF</th>
<th>NCF</th>
<th>FPMC</th>
<th>GRURec</th>
<th>GRURec+</th>
<th>Caser</th>
<th>SASRec</th>
<th>TiSASRec</th>
<th>BERT4Rec</th>
<th>DisenRec</th>
<th>Seq2Bubbles</th>
<th>Improv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beauty</td>
<td>N@5</td>
<td>0.0241</td>
<td>0.0803</td>
<td>0.0844</td>
<td>0.0921</td>
<td>0.0821</td>
<td>0.1186</td>
<td>0.1054</td>
<td>0.1439</td>
<td>0.1310</td>
<td>0.1585</td>
<td>0.2404</td>
<td>0.2767</td>
<td>+13.1%</td>
</tr>
<tr>
<td></td>
<td>H@5</td>
<td>0.0396</td>
<td>0.1219</td>
<td>0.1304</td>
<td>0.1372</td>
<td>0.1321</td>
<td>0.1791</td>
<td>0.1613</td>
<td>0.1929</td>
<td>0.1804</td>
<td>0.2201</td>
<td>0.3225</td>
<td>0.3508</td>
<td>+8.0%</td>
</tr>
<tr>
<td></td>
<td>N@10</td>
<td>0.0337</td>
<td>0.1059</td>
<td>0.1132</td>
<td>0.1215</td>
<td>0.1064</td>
<td>0.1448</td>
<td>0.1361</td>
<td>0.1636</td>
<td>0.1566</td>
<td>0.1856</td>
<td>0.2709</td>
<td>0.2959</td>
<td>+8.4%</td>
</tr>
<tr>
<td></td>
<td>H@10</td>
<td>0.0755</td>
<td>0.1998</td>
<td>0.2146</td>
<td>0.2415</td>
<td>0.2347</td>
<td>0.2646</td>
<td>0.2593</td>
<td>0.2656</td>
<td>0.2581</td>
<td>0.3029</td>
<td>0.4171</td>
<td>0.4503</td>
<td>+7.3%</td>
</tr>
<tr>
<td>Steam</td>
<td>N@5</td>
<td>0.0477</td>
<td>0.0744</td>
<td>0.0717</td>
<td>0.0945</td>
<td>0.1370</td>
<td>0.1613</td>
<td>0.1131</td>
<td>0.1727</td>
<td>0.3252</td>
<td>0.1842</td>
<td>0.2863</td>
<td>0.3566</td>
<td>+9.7%</td>
</tr>
<tr>
<td></td>
<td>H@5</td>
<td>0.0805</td>
<td>0.1177</td>
<td>0.1203</td>
<td>0.1517</td>
<td>0.2171</td>
<td>0.2391</td>
<td>0.176</td>
<td>0.2559</td>
<td>0.4155</td>
<td>0.2710</td>
<td>0.3986</td>
<td>0.4384</td>
<td>+5.5%</td>
</tr>
<tr>
<td></td>
<td>N@10</td>
<td>0.0665</td>
<td>0.1005</td>
<td>0.1026</td>
<td>0.1026</td>
<td>0.1283</td>
<td>0.1802</td>
<td>0.1484</td>
<td>0.2147</td>
<td>0.3557</td>
<td>0.2261</td>
<td>0.3332</td>
<td>0.3875</td>
<td>+8.9%</td>
</tr>
<tr>
<td></td>
<td>H@10</td>
<td>0.1389</td>
<td>0.1993</td>
<td>0.2169</td>
<td>0.2551</td>
<td>0.3313</td>
<td>0.3594</td>
<td>0.2870</td>
<td>0.3783</td>
<td>0.5239</td>
<td>0.4013</td>
<td>0.5437</td>
<td>0.5661</td>
<td>+4.1%</td>
</tr>
<tr>
<td>ML-1m</td>
<td>N@5</td>
<td>0.0416</td>
<td>0.1903</td>
<td>0.1146</td>
<td>0.2885</td>
<td>0.3196</td>
<td>0.3705</td>
<td>0.3832</td>
<td>0.3980</td>
<td>0.4243</td>
<td>0.4454</td>
<td>0.4615</td>
<td>0.5035</td>
<td>+9.1%</td>
</tr>
<tr>
<td></td>
<td>H@5</td>
<td>0.0715</td>
<td>0.2866</td>
<td>0.1932</td>
<td>0.4297</td>
<td>0.4673</td>
<td>0.5103</td>
<td>0.5353</td>
<td>0.5434</td>
<td>0.5755</td>
<td>0.5876</td>
<td>0.6025</td>
<td>0.6351</td>
<td>+5.4%</td>
</tr>
<tr>
<td></td>
<td>N@10</td>
<td>0.0621</td>
<td>0.2365</td>
<td>0.1640</td>
<td>0.3439</td>
<td>0.3627</td>
<td>0.4064</td>
<td>0.4268</td>
<td>0.4368</td>
<td>0.4641</td>
<td>0.4818</td>
<td>0.5003</td>
<td>0.5447</td>
<td>+8.8%</td>
</tr>
<tr>
<td></td>
<td>H@10</td>
<td>0.1358</td>
<td>0.4301</td>
<td>0.3477</td>
<td>0.5946</td>
<td>0.6207</td>
<td>0.6351</td>
<td>0.6692</td>
<td>0.6629</td>
<td>0.7008</td>
<td>0.6970</td>
<td>0.7219</td>
<td>0.7422</td>
<td>+2.8%</td>
</tr>
<tr>
<td>ML-20m</td>
<td>N@5</td>
<td>0.0511</td>
<td>0.1332</td>
<td>0.0771</td>
<td>0.2239</td>
<td>0.3090</td>
<td>0.3630</td>
<td>0.2538</td>
<td>0.4208</td>
<td>0.5134</td>
<td>0.4967</td>
<td>0.5058</td>
<td>0.5666</td>
<td>+10.3%</td>
</tr>
<tr>
<td></td>
<td>H@5</td>
<td>0.0805</td>
<td>0.2128</td>
<td>0.1358</td>
<td>0.3601</td>
<td>0.4657</td>
<td>0.5118</td>
<td>0.3804</td>
<td>0.5727</td>
<td>0.6499</td>
<td>0.6323</td>
<td>0.6528</td>
<td>0.6931</td>
<td>+6.1%</td>
</tr>
<tr>
<td></td>
<td>N@10</td>
<td>0.0695</td>
<td>0.1786</td>
<td>0.1271</td>
<td>0.2895</td>
<td>0.3637</td>
<td>0.4087</td>
<td>0.3062</td>
<td>0.4665</td>
<td>0.5440</td>
<td>0.5340</td>
<td>0.5398</td>
<td>0.6189</td>
<td>+13.7%</td>
</tr>
<tr>
<td></td>
<td>H@10</td>
<td>0.1378</td>
<td>0.3538</td>
<td>0.2922</td>
<td>0.5201</td>
<td>0.5844</td>
<td>0.6524</td>
<td>0.5427</td>
<td>0.7136</td>
<td>0.7606</td>
<td>0.7473</td>
<td>0.7579</td>
<td>0.8015</td>
<td>+5.3%</td>
</tr>
</tbody>
</table>

Higher H (HR) and N (NDCG) are better
Experiments: Ablation Study

Table 2: Ablation analysis

<table>
<thead>
<tr>
<th>Variants</th>
<th>ML-1M</th>
<th>Beauty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HR@10</td>
<td>NDCG@10</td>
</tr>
<tr>
<td>w/o Contextual</td>
<td>0.731 (-1.4%)</td>
<td>0.536 (-1.5%)</td>
</tr>
<tr>
<td>w/o Regularization</td>
<td>0.730 (-1.6%)</td>
<td>0.537 (-1.3%)</td>
</tr>
<tr>
<td>w/o Self-Attention</td>
<td>0.621 (-16.3%)</td>
<td>0.483 (-11.2%)</td>
</tr>
<tr>
<td>w/o Max Pooling</td>
<td>0.611 (-17.6%)</td>
<td>0.503 (-7.5%)</td>
</tr>
<tr>
<td>Default</td>
<td><strong>0.742</strong></td>
<td><strong>0.544</strong></td>
</tr>
</tbody>
</table>

- Comparison with the simplified version that replace the bubble embedding by point embedding

![Graphs](image-url)
### Experiments: Robustness and Scalability

#### Further discussions:
- The regularization term helps to **alleviate over-fitting**
- The training time scales linearly w.r.t. sequence length and hidden size
Conclusions

Our contributions:
- Methodology: propose a new representation model for distributions of user interests with multi-modality and heterogeneous concentration
- Techniques: design an efficient distance computing scheme of new embedding and devise a self-supervised contrastive to enhance training
- Evaluation: achieve state-of-the-art performance on several benchmarks and conduct ablation studies to thoroughly dissect the effectiveness

Thanks for listening!