Learning Hierarchical Representation Model for Next Basket Recommendation

Pengfei Wang, Jiafeng Guo, Yanyan Lan, Jun Xu, Shengxian Wan, Xueqi Cheng
CAS Key Lab of Network Data Science and Technology
Institute of Computing Technology, Chinese Academy of Sciences
• Task
• Background
• Motivation
• Our model
  - structure of HRM
  - connections to previous methods
• Experiments
• Summary
Next basket recommendation: given a sequence of purchases, what items are purchased sequentially?
Background

- **Sequence models**
  - **MC** (Markov Chain) [Zimdars et al. UAI01, Chen et al. KDD12]

- **Collaborative Filtering**
  - **NMF** (Non negative Matrix Factorization) [Daniel D. Lee et al. NIPS01]

- **Hybrid methods**
  - **FPMC** (Factorized Personalized Markov Chain) [S Rendle et al. WWW10]
Weakness of Previous methods

Sequence models based on Markov Assumption: MC

Weakness of MC: Lack users' general interests

Transactions of \( u_1 \):
- \( a \rightarrow b \rightarrow d \)

Transactions of \( u_2 \):
- \( b \rightarrow a \rightarrow \{ \text{onions, potatoes} \} \rightarrow \{ \text{burger} \} \)
- \( c \rightarrow \{ \text{bread, butter} \} \rightarrow \{ \text{milk} \} \)
- \( \{ \text{beer} \} \rightarrow \{ \text{napkin} \} \)
- \( \{ \text{tennis shoes} \} \rightarrow \{ \text{tennis} \} \)

...
Weakness of Previous methods

Collaborative filtering matrix factorization: NMF

transactions of $u_1$

transactions of $u_2$

weakness of NMF: *Lack sequence behaviors*

![Diagram](diagram.png)
Weakness of Previous methods

Hybrid method: Factorized Personalized Markov Chain

[S Rendle et al. WWW10 best paper]

transactions of $u_1$

transactions of $u_2$

weakness of FPMC: **linear combination of different factors**

U represent users, V represent items
Weakness of Previous methods

Hybrid method: Factorized Personalized Markov Chain
[S Rendle et al. WWW10 best paper]

transactions of $u_1$

transactions of $u_2$

weakness of FPMC: **linear combination of different factors**

### General taste + Sequential behavior

\[
\hat{x}_{u,t,i} := \langle v_u^{U,I}, v_i^{I,U} \rangle + \frac{1}{|B_{t-1}^u|} \sum_{l \in B_{t-1}^u} \langle v_i^{I,L}, v_l^{L,I} \rangle
\]

**Independent Influence!**
Weakness of Previous methods

Is that linear combination enough for a good recommendation?

- Last transaction: pumpkin
- Last transaction: candy
- Next transaction: potato
- Next transaction: chocolate
- Next transaction: chips
- Next transaction: Halloween!
We need a model that is capable of incorporating more complicated interactions among multiple factors. This becomes the major motivation of our work.
The structure of HRM

![Diagram of hierarchical representation model](image)
Aggregation Method

• **Linear method**
  - average pooling
    \[ f_{avg}(V) = \frac{1}{|V|} \sum_{l=1}^{|V|} \bar{v}_l \]

• **Nonlinear method**
  - max pooling
  - other types of operators (top-k average pooling, k-max pooling, hadamard pooling)

(V is a set of input vectors to be aggregated)
The probability of purchasing one item next transaction:

\[
p(i \in T^u_t | u, T^u_{t-1}) = \frac{\exp(\langle \tilde{v}^I_i, \tilde{v}^{Hybrid}_{u,t-1} \rangle)}{\sum_{j=1}^{||I||} \exp(\langle \tilde{v}^I_j, \tilde{v}^{Hybrid}_{u,t-1} \rangle)}
\]

exponent of item and users` hybrid interest

sum of all items: too large!

Objective function

\[
\ell_{HRM} = \sum_{u \in U} \sum_{T^u_t \in T^u} \sum_{i \in T^u_t} \log p(i \in T^u_t | u, T^u_{t-1}) - \lambda \|\Theta\|^2_F
\]

all users  all trans  all items

Negative sampling

\[
\ell_{NEG} = \sum_{u \in U} \sum_{T^u_t \in T^u} \sum_{i \in T^u_t} \left( \log \sigma(\langle \tilde{v}^I_i, \tilde{v}^{Hybrid}_{u,t-1} \rangle) + k \cdot E_{i' \sim P_t} [\log \sigma(-\langle \tilde{v}^I_{i'}, \tilde{v}^{Hybrid}_{u,t-1} \rangle)] \right) - \lambda \|\Theta\|^2_F
\]

negative count  sample distribution
Connection to Previous methods

- **Degradation To MC**

\[
\ell_{CopyItem} = \sum_{u \in U} \sum_{T^u_i \in T^u} \sum_{i \in T^u_i} \left( \log \sigma(\vec{v}_i^l \cdot \vec{v}_s^l) \right) + k \cdot \mathbb{E}_{i' \sim P_I} \left[ \log \sigma(-\vec{v}_i^l \cdot \vec{v}_{s}'^l) \right] - \lambda \| \Theta \|^2_F
\]

\[
= \vec{v}_i^l \cdot \vec{v}_s^l = PMI(v_i^l, v_s^l) - \log k
\]

*Select copy*: copy item when constructing the transaction representation from item vectors, the operation randomly selects one item vector and copies it.
Connection to Previous methods

- Degradation To MF

\[ \ell_{CopyUser} = \sum_{u \in U} \sum_{T_t^u \in T^u} \sum_{i \in T_t^u} \left( \log \sigma(\tilde{v}_u^I \cdot \tilde{v}_u^U) \right) + \sum_{i \sim P_t} [\log \sigma(-\tilde{v}_u^I \cdot \tilde{v}_u^U)] - \lambda \|\Theta\|_F^2 \]

\[ = \tilde{v}_u^U \cdot \tilde{v}_i = PMI(v_u^U, v_i^I) - \log k \]

**Select copy**: always select and copy user vector in the second layer, ignoring the sequential information.
Degradation To FPMC

 Avg Pooling is used, each instance corresponds to 1 negative sample
## Data sets

<table>
<thead>
<tr>
<th></th>
<th>retails</th>
<th>ecommerce</th>
</tr>
</thead>
<tbody>
<tr>
<td>#transactions</td>
<td>67964</td>
<td>91294</td>
</tr>
<tr>
<td>#items</td>
<td>7982</td>
<td>5845</td>
</tr>
<tr>
<td>#users</td>
<td>9238</td>
<td>9321</td>
</tr>
<tr>
<td>#avg.transaction size</td>
<td>5.9</td>
<td>5.8</td>
</tr>
<tr>
<td>#avg.transaction per user</td>
<td>7.4</td>
<td>9.7</td>
</tr>
</tbody>
</table>
● Evaluation Metric

● F1-score:

\[
\text{Precision}(T_{t,u}^u, R(u)) = \frac{|T_{t,u}^u \cap R(u)|}{|R(u)|}
\]

\[
\text{Recall}(T_{t,u}^u, R(u)) = \frac{|T_{t,u}^u \cap R(u)|}{|T_{t,u}^u|}
\]

\[
\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

harmonic mean of precision and recall

● Hit-ratio:

\[
\text{Hit-Ratio} = \frac{\sum_{u \in U} I(T_{t,u}^u \cap R(u) \neq \phi)}{|U|}
\]

coverage

● NDCG:

\[
\text{NDCG@k} = \frac{1}{N_k} \sum_{j=1}^{k} \frac{2I(R_j(u) \in T_{t,u}^u)}{\log_2(j + 1)} - 1
\]

a ranking measure
Experiments

• Comparison among Different HRMs

<table>
<thead>
<tr>
<th>Models</th>
<th>F1-score</th>
<th>Hit-ratio</th>
<th>NDCG</th>
<th>F1-score</th>
<th>Hit-ratio</th>
<th>NDCG</th>
<th>F1-score</th>
<th>Hit-ratio</th>
<th>NDCG</th>
<th>F1-score</th>
<th>Hit-ratio</th>
<th>NDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRM _Avg_Avg</td>
<td>0.051</td>
<td>0.240</td>
<td>0.073</td>
<td>0.060</td>
<td>0.276</td>
<td>0.082</td>
<td>0.063</td>
<td>0.283</td>
<td>0.080</td>
<td>0.063</td>
<td>0.286</td>
<td>0.086</td>
</tr>
<tr>
<td>HRM _Max_Avg</td>
<td>0.059</td>
<td>0.275</td>
<td>0.080</td>
<td>0.064</td>
<td>0.279</td>
<td>0.087</td>
<td>0.065</td>
<td>0.290</td>
<td>0.083</td>
<td>0.067</td>
<td>0.298</td>
<td>0.086</td>
</tr>
<tr>
<td>HRM _Avg_Max</td>
<td>0.057</td>
<td>0.262</td>
<td>0.080</td>
<td>0.064</td>
<td>0.288</td>
<td>0.085</td>
<td>0.065</td>
<td>0.289</td>
<td>0.082</td>
<td>0.068</td>
<td>0.293</td>
<td>0.090</td>
</tr>
<tr>
<td>HRM _Max_Max</td>
<td><strong>0.062</strong></td>
<td><strong>0.282</strong></td>
<td><strong>0.089</strong></td>
<td><strong>0.065</strong></td>
<td><strong>0.293</strong></td>
<td><strong>0.088</strong></td>
<td><strong>0.068</strong></td>
<td><strong>0.298</strong></td>
<td><strong>0.085</strong></td>
<td><strong>0.070</strong></td>
<td><strong>0.312</strong></td>
<td><strong>0.093</strong></td>
</tr>
</tbody>
</table>

• Observation
  Avg pooling perform worst
  When apply max pooling on any layer, the performance improved a little
  When apply max pooling on all layers, HRM performed best
Experiments

- Comparison with baselines

  top popular  sequential behavior  general interest  hybrid method

- Observation

  Top method performed worst
  NMF and MC performed better than top method
  FPMC performed better than NMF and MC
  HRM performed best
Experiments

● Comparison over groups

<table>
<thead>
<tr>
<th>user activeness</th>
<th>method</th>
<th>F1-score</th>
<th>Hit-Ratio</th>
<th>NDCG@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inactive</td>
<td>Top</td>
<td>0.036</td>
<td>0.181</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>MC</td>
<td>0.042</td>
<td>0.206</td>
<td>0.058</td>
</tr>
<tr>
<td></td>
<td>NMF</td>
<td>0.037</td>
<td>0.198</td>
<td>0.046</td>
</tr>
<tr>
<td></td>
<td>FPMC</td>
<td>0.043</td>
<td>0.216</td>
<td>0.060</td>
</tr>
<tr>
<td></td>
<td>HRM\text{MaxMax}</td>
<td>0.048</td>
<td>0.236</td>
<td>0.062</td>
</tr>
<tr>
<td>Medium</td>
<td>Top</td>
<td>0.051</td>
<td>0.230</td>
<td>0.084</td>
</tr>
<tr>
<td></td>
<td>MC</td>
<td>0.059</td>
<td>0.262</td>
<td>0.088</td>
</tr>
<tr>
<td></td>
<td>NMF</td>
<td>0.052</td>
<td>0.234</td>
<td>0.072</td>
</tr>
<tr>
<td></td>
<td>FPMC</td>
<td>0.059</td>
<td>0.263</td>
<td>0.087</td>
</tr>
<tr>
<td></td>
<td>HRM\text{MaxMax}</td>
<td>0.068</td>
<td>0.299</td>
<td>0.097</td>
</tr>
<tr>
<td>Active</td>
<td>Top</td>
<td>0.045</td>
<td>0.207</td>
<td>0.074</td>
</tr>
<tr>
<td></td>
<td>MC</td>
<td>0.050</td>
<td>0.212</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>NMF</td>
<td>0.056</td>
<td>0.223</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>FPMC</td>
<td>0.054</td>
<td>0.224</td>
<td>0.080</td>
</tr>
<tr>
<td></td>
<td>HRM\text{MaxMax}</td>
<td>0.062</td>
<td>0.246</td>
<td>0.087</td>
</tr>
</tbody>
</table>

● observation

NMF perform better than MC on active group, while MC performs better than NMF on inactive group
HRM performed best
The Impact of Negative Sampling

More negative count we choose, the more F1-score we obtain. The sampling number $k$ increases, the performance gain between two consecutive trials decreases.

**Observation**

More negative count we choose, the more F1-score we obtain. The sampling number $k$ increases, the performance gain between two consecutive trials decreases.
A next basket recommendation task

A Hierarchical Representation Model

- model both sequential behavior and users’ general taste
- Aggregation operators to connect two level factors.
- HRM can produce multiple recommendation models by introducing different aggregation operations

Future works

- More aggregations operations will be analyzed
- Integrate other types of information, e.g. timestamp of transaction
Thank You!

EMAIL: wangpengfei@software.ict.ac.cn
Pointwise mutual information (PMI) is a widely used word similarity measure.

\[ I(x', y') = \log_2 \frac{P(x', y')} {P(x')P(y')} \]

\[ = \log_2 \frac{P(x' | y')} {P(x')} \]

\[ = \log_2 \frac{P(y' | x')} {P(y')} \]
Weakness of Previous methods

purchase history

<table>
<thead>
<tr>
<th>The Matrix</th>
<th>X Men</th>
<th>Dark night</th>
<th>Inception</th>
<th>The Avengers</th>
<th>Lucy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.95</td>
<td>0.8</td>
<td>0.5</td>
<td>0.2</td>
</tr>
</tbody>
</table>

last purchase

<table>
<thead>
<tr>
<th>Match point</th>
<th>Lost in translation</th>
<th>Girl with a Pearl Earring</th>
<th>Lucy</th>
<th>The Avengers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.9</td>
<td>0.85</td>
<td>0.3</td>
<td>0.2</td>
</tr>
</tbody>
</table>

the sum of score in general recommend and score of sequential recommend

linear combination

Dark night | Lost in translation | The Avengers | Lucy |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.95</td>
<td>0.9</td>
<td>0.7</td>
<td>0.5</td>
</tr>
</tbody>
</table>