Counterfactual Explanations for Neural Recommenders

Khanh Hiep Tran, Azin Ghazimatin, Rishiraj Saha Roy
Max Planck Institute for Informatics, Germany

Motivation

• Understanding recommendations increases trust and satisfaction of users
• Neural recommenders are state-of-the-art but too complex
• Existing explanation methods for neural models are limited:
  - Relies on attention
  - Not tangible actionable for end-users
  - Have privacy concerns
  - Need external information

% Need counterfactual explanations: A set of the user’s own actions, when removed, produces a different recommendation.

Approach

• Extend the basic idea in PRINCE [2] to neural models
• Estimate influence of actions on item scores using Fast Influence Analysis (FIA) [1]
• Extend influence on one item to influence on pairs of items
• Iteratively closing the gap between the recommendation and the replacement item with most influential actions

% ACCENT: Action-based Counterfactual Explanations for Neural Recommenders for Tangibility.

Method

For each replacement item \(rec^*\):
  - For each interaction \(z\) of user \(u\):
    • Calculate the influence of \(z\) on scores of \(rec\) and \(rec^*\) (using FIA)
    • Sort interactions by influence on the score gap
  - Add interactions to the result set until the gap is filled
  - Update the smallest result set found so far

Counterfactual explanation

You were recommended "The Godfather II" because:
• You liked "Goodfellas", and
• You liked "The Godfather".
Otherwise, the recommendation would have been: "Apt Pupil".

<table>
<thead>
<tr>
<th>Candidate top-k set of replacement items</th>
<th>k = 5</th>
<th>k = 10</th>
<th>k = 20</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recommendation</strong></td>
<td><strong>Explanation</strong></td>
<td><strong>Replacement</strong></td>
<td></td>
</tr>
<tr>
<td>The Silence of the Lambs</td>
<td>Contact</td>
<td>Fargo</td>
<td></td>
</tr>
<tr>
<td>Titanic</td>
<td>True Romance</td>
<td>The Basketball Diaries</td>
<td>East Of Eden</td>
</tr>
<tr>
<td>The Devil's Advocate</td>
<td>Speed</td>
<td>Eraser</td>
<td></td>
</tr>
<tr>
<td>It's A Wonderful Life</td>
<td>My Fair Lady</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forrester Gump</td>
<td>Nell</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Das Boot</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Counterfactual sets generated by ACCENT.

<table>
<thead>
<tr>
<th>Recommendation</th>
<th>ACCENT Explanation</th>
<th>Replacement</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Silence of the Lambs</td>
<td>Contact</td>
<td>Fargo</td>
</tr>
<tr>
<td>Titanic</td>
<td>True Romance</td>
<td>The Basketball Diaries</td>
</tr>
<tr>
<td>The Devil's Advocate</td>
<td>Speed</td>
<td>Eraser</td>
</tr>
<tr>
<td>It's A Wonderful Life</td>
<td>My Fair Lady</td>
<td></td>
</tr>
<tr>
<td>Forrester Gump</td>
<td>Nell</td>
<td></td>
</tr>
<tr>
<td>Das Boot</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results

• ACCENT can find concise counterfactual explanations
• Current methods based on attention and influence fall short of ACCENT in percentage and size of CF sets
• ACCENT is applicable to a broad class of neural models

References

5. Xin Xin, Xiangnan He, Yongfeng Zhang, Yongdong Zhang, and Ioanniswofis. 2019. Relational Collaborative Filtering: Modeling Multiple Item Relations for Recommendation. In SIGIR.

Contact:
Data and Code: [https://github.com/hieptk/ACCENT](https://github.com/hieptk/ACCENT)