Attacking Black-box Recommendations via Copying Cross-domain User Profiles

Wenqi Fan, Tyler Derr, Xiangyu Zhao, Yao Ma, Hui Liu, Jianping Wang, Jiliang Tang, and Qing Li

19-22 April 2021
37th IEEE International Conference on Data Engineering (ICDE)
Recommender systems

- Goal: suggest items that best fit users’ preferences
Recommender systems

• Security (Attacking) in Recommender Systems
  • Data poisoning attacks: promote/demote a set of items
Attacking in recommender systems

- **Challenges in existing attacking methods:**
  - Less "realistic" user profiles (easily detected)
Attacking in recommender systems

- **Cross-domain Information**
  - Share a lot of items
  - Users from these platforms with **similar functionalities** also share similar behavior patterns/preferences.
Attacking in recommender systems

- Challenges in existing attacking methods:
  - Less "realistic" user profiles (easily detected)
  - Copy cross-domain users with real profiles from other domains
Attacking in recommender systems

• Challenges in existing attacking methods:
  • Less "realistic" user profiles (easily detected)
  • Cross-domain Information

• White-box setting (i.e., model architecture and parameters, and datasets)
  → impossible and unrealistic (privacy and security)

• Black-box setting
  ➢ Reinforcement Learning (RL) -- Query Feedback (Reward)
CopyAttack

- **Problem Statement**
  - **Target RecSys A**
    - Users: $\mathcal{U}^A = \{u_1^A, u_2^A, ..., u_n^A\}$
    - User profile: $P_{u_i}^A = \{v_1 \rightarrow ... \rightarrow v_j \rightarrow ... \rightarrow v_l\}$
    - Item profile: $P_{v_j}^A = \{u_1^A \rightarrow ... \rightarrow u_i^A \rightarrow ... \rightarrow u_o^A\}$
  - **Source RecSys B**
    - Users: $\mathcal{U}^B$  
    - User Profile: $P_{u}^B = \{v_1 \rightarrow ... \rightarrow v_j \rightarrow ... \rightarrow v_l\}$
  - Overlapping items: $\mathcal{V} = \mathcal{V}^A \cap \mathcal{V}^B$
  - **Goal**: $\mathcal{U}^{A'} = \mathcal{U}^A \cup \mathcal{U}^B \rightarrow A$

\[ y_i^{A, > k} = \{v[1], v[2], ..., v[k]\} = \text{Rec}(P_{u_i}^A, P_{v}^A) \]
CopyAttack

• **Attacking RL Environment**
  - Action A: user profiles in source domain B

  - Reward R (Hitting Ratio, HR):
    - **Spy users**
      
      \[
      r(s_t, a_t) = \frac{1}{|U^A_*|} \sum_{i=1}^{|U^A_*|} HR(u^A_{i*}, v_*, k)
      \]

      \[
      HR(u^A_{i*}, v_*, k) = \begin{cases} 
      1, & v_* \in y_{u^*}, >k, \\
      0, & v_* \notin y_{u^*}, >k 
      \end{cases}
      \]

  - Terminal: reach the budget or successfully satisfy the promotion task
CopyAttack

User Profile Selection in Source Domain B

User Profile Crafting in Source Domain B

Target RecSys in Target Domain A

Users:

Items:

Feedback (Queries)

Top-k List

Reward (+/-)

Reward (+/-)

Path for User Profile Selection

Non-leaf Node

Mask (Stop Sign)

Target Item to be Attacked

Users with Target Item

Users without Target Item

Raw User Profile

Crafting User Profile

PN-* Policy Network

Sharing Items in Two Domains

Spy Users

Real Users
CopyAttack

• **User Profile Selection**
  • Construct hierarchical clustering tree
  • **Masking Mechanism** - specific target items
  • Hierarchical-structure Policy Gradient

\[ a_t^u = \{a_{[t,1]}^u, a_{[t,2]}^u, \ldots, a_{[t,d]}^u\} \]

\[ p^u(a_t^u|s_t^u) = \prod_{d} p_d^u(a_{[t,d]}^u|s_t^u) \]

\[ = p_d^u(a_{[t,d]}^u|s_t^u) \cdot p_{d-1}^u(a_{[t,d-1]}^u|s_t^u) \cdots p_1^u(a_{[t,1]}^u|s_t^u) \]

\[ x_{\psi^*} = \text{RNN}(U_t^B \rightarrow A), \]

\[ p_i^u(\cdot|s_t^u) = \text{softmax}(\text{MLP}([q_{\psi^*}^B \oplus x_{\psi^*}^u]|\theta_i^u)) \]

Time Complexity: \( \mathcal{O}(|U^B|) \rightarrow \mathcal{O}(d \times |U^B|^{1/d}) \)
CopyAttack

- **User Profile Crafting**
  - Clipping operation to craft the raw user profiles
    
    \[ W = \{10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%, 100\%\} \]
  
  - Sequential patterns (forward/backward)

Example:

- \[ p^B_{u_i} = \{v_1 \rightarrow v_2 \rightarrow v_3 \rightarrow v_4 \rightarrow v_{5*} \rightarrow v_6 \rightarrow v_7 \rightarrow v_8 \rightarrow v_9 \rightarrow v_{10}\} \]
- \[ \hat{p}^B_{u_i} = \{v_3 \rightarrow v_4 \rightarrow v_{5*} \rightarrow v_6 \rightarrow v_7\} \]

\[ p^l(\cdot | s^l_t) = \text{softmax}(\text{MLP}([p^B_i \oplus q^B_{v*}] | \theta^l)) \]
CopyAttack

User Profile Selection in Source Domain B

User Profile Crafting in Source Domain B

Target RecSys in Target Domain A

PN-1
PN-2
PN-3

Users:

Items:

Reward (+/-)

Top-k List

Feedback (Queries)

PN-L

Users with Target Item

Users without Target Item

Raw User Profile

Crafting User Profile

PN-* Policy Network

Sharing Items in Two Domains

Spy Users

Real Users

Reward (+/-)

26x469
CopyAttack
12
Thank You

Wenqi Fan, wenqifan03@gmail.com

Please see my homepage for more details: https://wenqifan03.github.io
Attacking Black-box Recommendations via Copying Cross-domain User Profiles

Wenqi Fan, Tyler Derr, Xiangyu Zhao, Yao Ma, Hui Liu, Jianping Wang, Jiliang Tang, and Qing Li

19-22 April 2021
37th IEEE International Conference on Data Engineering (ICDE)