DRN: A Deep Reinforcement Learning Framework for News Recommendation

Guanjie Zheng[†], Fuzheng Zhang[§], Zihan Zheng[§], Yang Xiang[§]

Nicholas Jing Yuan[§], Xing Xie[§], Zhenhui Li[†]

Pennsylvania State University[†], Microsoft Research Asia[§]
University Park, USA[†], Beijing, China[§]
gjz5038@ist.psu.edu,{fuzzhang,v-zihanzhe,yaxian,nicholas.yuan,xingx}@microsoft.com,jessieli@ist.psu.edu

Presented by: Mohammad Zangooei
CS885 - CS Dept. of UWaterloo

Introduction

Preliminaries

- Too many content to recommend!
- Traditional methods of personalized online content recommendation:
  - Content based
  - Collaborative filtering based
  - Hybrid
- Long live deep learning models!

[https://medium.com/swlh/news-recommendation-system-a8efde3cb233]

The Big Challenge: Dynamic Changes

- News become outdated very fast
- Users interest evolve during time

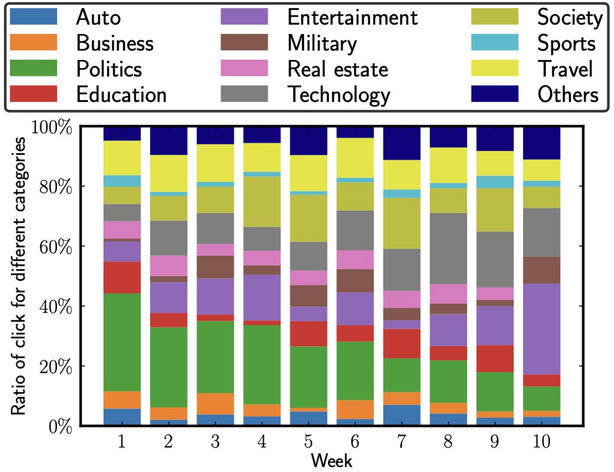

Figure 1: Distribution of clicked categories of an active user in ten weeks. User interest is evolving over time.

Room for Improvement

- Not overlooking long-term rewards!
    - Kobe Bryant Vs. thunderstorm alert
- “When the user will be back” as a feedback!
    - Click-through rate is not enough
- More effective exploration!
    - ε-greedy and UCB can be harmful

The Proposed Solution

- Deep Q-learning
  - Future reward
  - Scalable
- Activeness score as a user feedback
  - Better indication
- Dueling Bandit Gradient Descent (DBGD) method for exploration
  - Candidates in the neighborhood of the current recommender

Method

System Overview

Table 1: Notations

Notation	Meaning
G	Agent
u, U	User, User set
a	Action
s	State
r	Reward
i, I	News, Candidate news pool
L	List of news to recommend
B	List of feedback from users
Q	Deep Q-Network
W	Parameters of Deep Q-Network

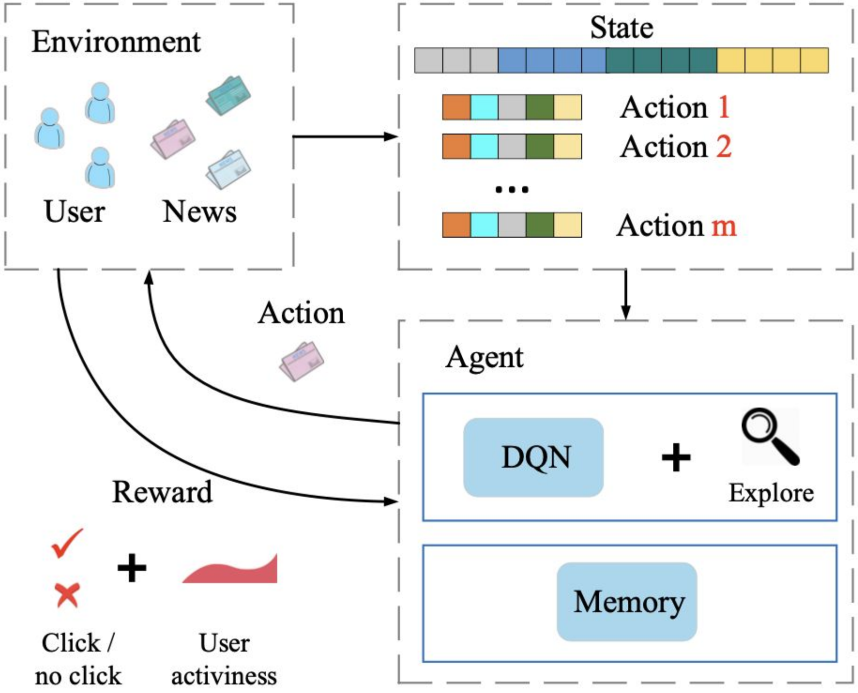

Figure 2: Deep Reinforcement Recommendation System

Model Framework

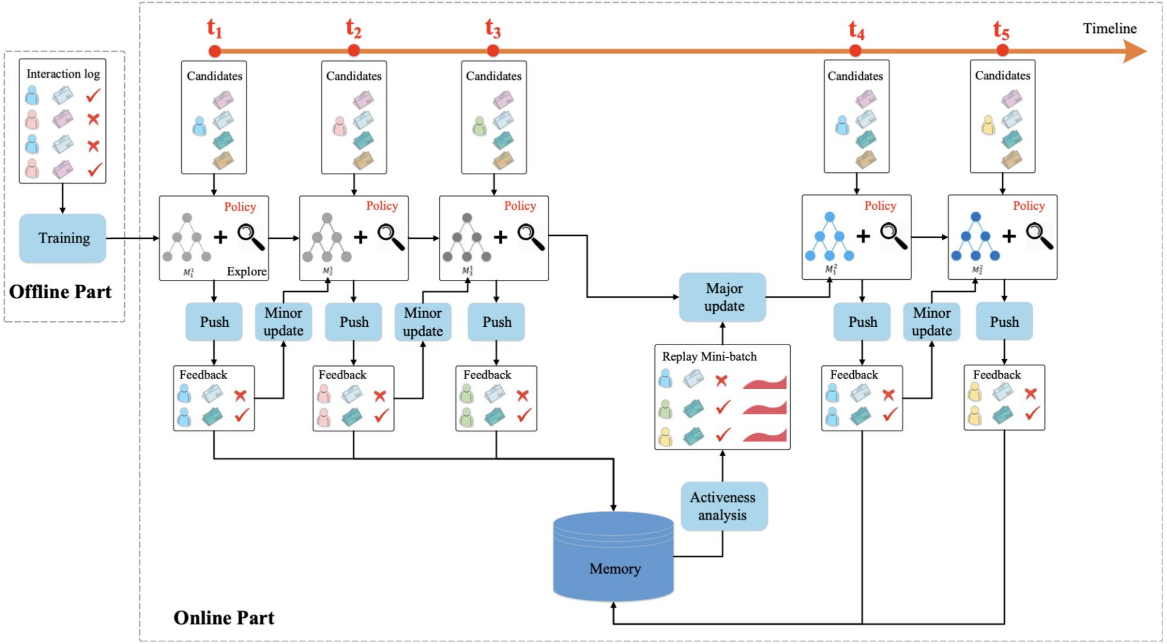

Figure 3: Model framework

- Offline
- Online
    - Push
    - Feedback
    - Minor update
    - Major update
    - Repeat!

Feature Construction

Action

- News features - 417 dim
  - Describes whether certain property appears in this piece of news, including headline, provider, ranking, entity name, category, topic category,
  - And click counts in last 1 hour, 6 hours, 24 hours, 1 week, and 1 year
- User news features - 25 dim
  - Describe the interaction between user and one certain piece of news

State

- Users features - 413 * 5 dim
  - Describes the features (i.e., headline, provider, ranking, entity name, category, and topic category) of the news that the user clicked in 1 hour, 6 hours, 24 hours, 1 week, and 1 year
  - Also a total click count for each time granularity.
- Context features - 32 dim
  - Describe the context when a news request happens, including time, weekday, and the freshness of the news

Deep Reinforcement Recommendation

- State: context features and user features
- Action: news features and user-news interaction features
- Reward:

$$y_{s,a} = Q(s, a) = r_{immediate} + \gamma r_{future}$$

- Predict the total reward regarding a specific action:

$$y_{s,a,t} = r_{a,t+1} + \gamma Q(s_{a,t+1}, \arg\max_{a'} Q(s_{a,t+1}, a'; W_t); W'_t)$$

- Feeding the feature into the network

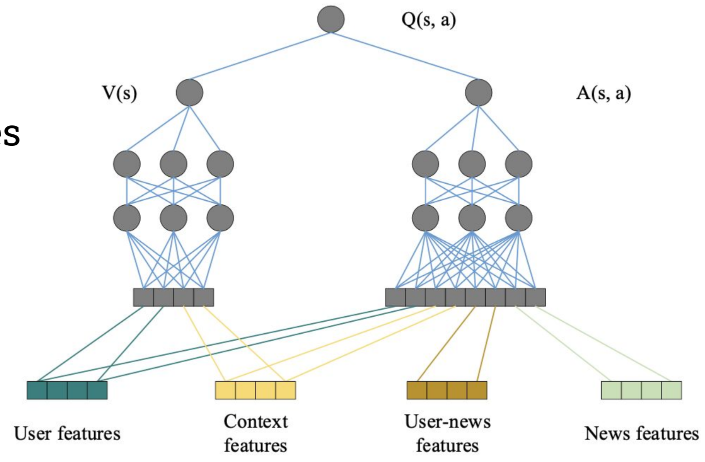

Figure 4: Q network

User Activeness

- Survival model
  - Starts from S_0
  - Constant rate of return λ_0
  - If return: add a constant value of S_a
  - Not exceeding 1

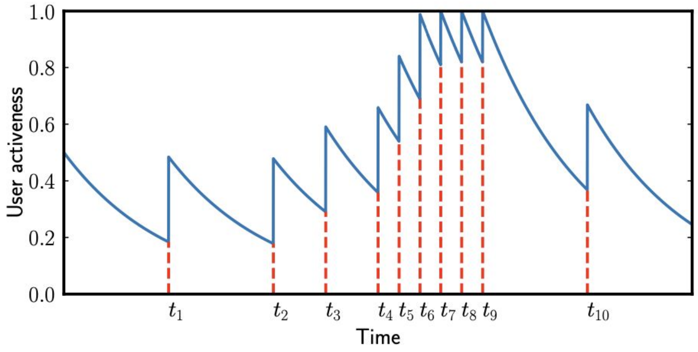

Figure 5: User activeness estimation

Explore

- The disturb to be added to Q parameters:

$$\Delta W = \alpha \cdot rand(-1, 1) \cdot W$$

- Update the target Q towards exploration network:

$$W' = W + \eta \tilde{W}.$$

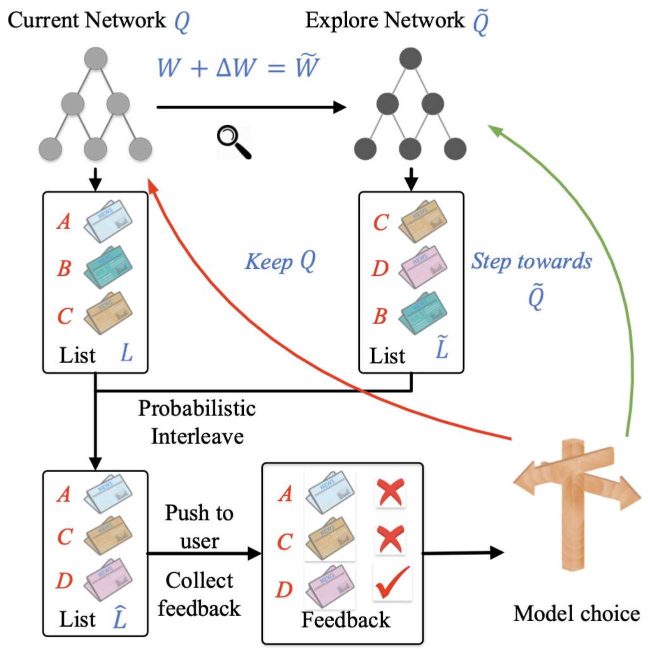

Figure 7: Exploration by *Dueling Bandit Gradient Descent*

Experiment

Evaluation Dataset and Metrics

- Two phases:
  - Offline
  - Online
- Metrics
  - Click through rate

$$CTR = \frac{\text{number of clicked items}}{\text{number of total items}}$$

  - Precision@k

$$Precision@k = \frac{\text{number of clicks in top-k recommended items}}{k}$$

  - nDCG

$$DCG(f) = \sum_{r=1}^{n} y_r^f D(r)$$

$$D(r) = \frac{1}{log(1+r)}$$

Compared Methods

- Variations of their model
  - DN: The basic model without future reward
  - DDQN: DN + future reward
  - DDQN + U: DDQN + user activeness
  - DDQN + EG: DDQN + ε-greedy
  - DDQN + DBGD: DDQN + Dueling Bandit Gradient Descent
- Baseline algorithms

Clik Probability

  - LR: Logistics Regression
  - FM: Factorization Machines
  - W&D: Wide & Deep

Potential Reward

  - LinUCB: Linear Upper Confidence Bound
  - HLinUCB: Hidden Linear Upper Confidence Bound

Offline Evaluation

Table 4: Offline recommendation accuracy

Method	CTR	nDCG
LR	0.1262	0.3659
FM	0.1489	0.4338
W&D	0.1554	0.4534
LinUCB	0.1447	0.4173
HLinUCB	0.1194	0.3491
DN	0.1587	0.4671
DDQN	**0.1662**	**0.4877**
DDQN + U	**0.1662**	**0.4878**
DDQN + U + EG	0.1609	0.4723
DDQN + U + DBGD	**0.1663**	**0.4854**

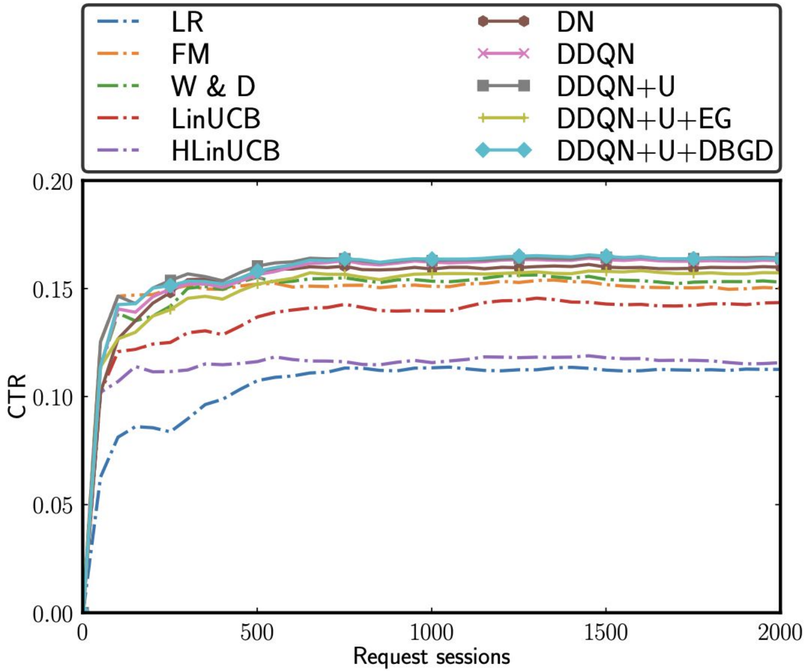

Figure 9: Offline cumulative CTR of different methods

Online Evaluation

Table 5: Online recommendation accuracy

Method	CTR	Precision@5	nDCG
LR	0.0059	0.0082	0.0326
FM	0.0072	0.0078	0.0353
W&D	0.0052	0.0067	0.0258
LinUCB	0.0075	0.0091	0.0383
HLinUCB	0.0085	0.0128	0.0449
DN	0.0100	0.0135	0.0474
DDQN	0.0111	0.0139	0.0477
DDQN + U	0.0089	0.0110	0.0425
DDQN + U + EG	0.0083	0.0100	0.03391
DDQN + U + DBGD	**0.0113**	**0.0149**	**0.0492**

$$ILS(\mathsf{L}) = \frac{\sum_{b_i \in \mathsf{L}} \sum_{b_j \in \mathsf{L}, b_j \neq b_i} S(b_i, b_j)}{\sum_{b_i \in \mathsf{L}} \sum_{b_j \in \mathsf{L}, b_j \neq b_i} 1}$$

Table 6: Diversity of user clicked news in the online experiment. Smaller *ILS* indicates better diversity. Similarity between news is measured by the cosine similarity between the bag-of-words vectors of news.

Method	ILS
LR	0.1833
FM	0.2014
W&D	0.1647
LinUCB	0.2636
HLinUCB	0.1323
DN	0.1546
DDQN	0.1935
DDQN + U	0.1713
DDQN + U + EG	0.1907
DDQN + U + DBGD	**0.1216**

Conclusion

Contributions and Future work

- Framework features
  - Effectively model the dynamic news features and user preferences and plan for future explicitly
  - User return pattern as a supplement to click / no click label
  - Effective exploration strategy to improve the recommendation diversity

- Future directions:
  - Clustering users and developing models for each group