Offline Evaluation of Ranking Policies with Click Models

Shuai Li[1] Yasin Abbasi-Yadkori[2] Branislav Kveton[3] S. Muthukrishnan[4] Vishwa Vinay[2] Zheng Wen[2]

[1]The Chinese University of Hong Kong [2]Adobe Research [3]Google Research [4]Rutgers University

Motivation

- Recommendations happen everywhere, such as Amazon, Facebook, Adobe Stock, Google Play, Netflix
- Suppose the existing policy π

with the expected CTR $V(\pi)$

- Can we verify a new policy h

satisfies $V(h) \geq V(\pi)$ based on logged data under policy π?

Setting

- Ground set $E = \{1, \dots, L\}$ of L items
- A list is a K-permutation of E, which is an element of $\prod_K E = \{(a_1, \dots, a_K): a_1, \dots, a_K \in E; a_i \neq a_j, i \neq j\}$
- Context set X
- $\bar{w}(a, k|x)$: the expected CTR of putting item a in position k under context x
- A policy π is a conditional probability distribution of a list given context x: $\pi(\cdot\,|x)$
- The reward of list A

$$f(A, w) = \sum_{k=1}^{K} w(a_k, k)$$

- The value of a policy

$$V(\pi) = \mathbb{E}_x[\mathbb{E}_{A\sim\pi(\cdot|x)} f(A, \bar{w}(\cdot\,|x))]$$

- At each time t
  - the environment draws context x_t and click realizations w_t
  - The learner observes x_t and selects A_t according to policy π
  - The environment reveals $(w_t(a_k^t, k))_{k=1}^K$
- Logged dataset: $S = \{(x_t, A_t, w_t)\}_{t=1}^n$

- Objective
  - To design statistically efficient estimators based on logged dataset for any ranking policy

- Challenge
  - The number of different lists is exponential in K

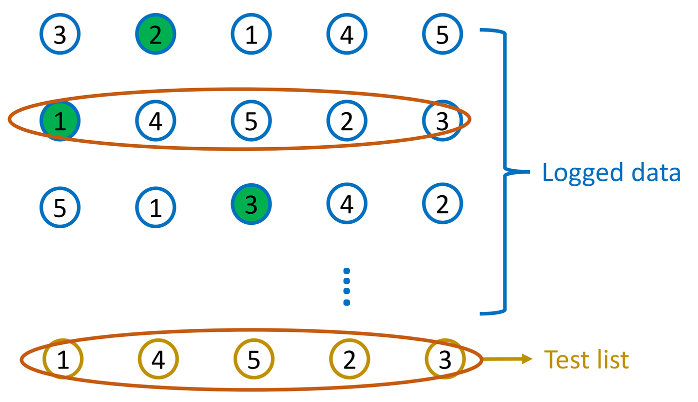
Click Models & Estimators

- List estimator [Strehl'2010]

$$\hat{V}_L(h) = \frac{1}{|S|} \sum_{(x,A,w)\in S} f(A, w) \min\left\{\frac{h(A|x)}{\hat{\pi}(A|x)}, M\right\}$$

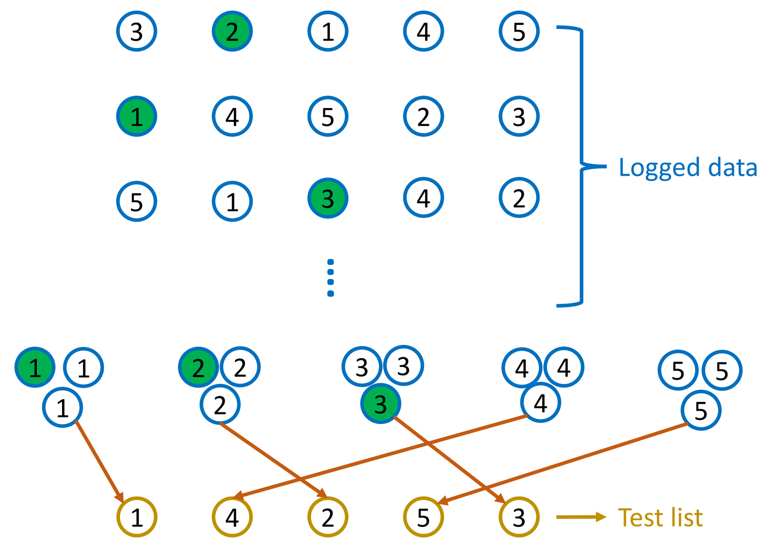
$\hat{\pi}$: estimates of the logging policy

- Disadvantages:
  - Have to match the exact lists. The number of lists is extremely large, thus $\hat{\pi}(A|x)$ is very small



- With click-model assumptions, we can build estimators that leverage structures of click feedback

- Document-Based Click Model (DCTR):
  - $\bar{w}(a, k|x)$ only depends on item a



$$\hat{V}_I(h) = \frac{1}{|S|} \sum_{(x,A,w)\in S} \sum_{k=1}^{K} w(a_k, k) \min\left\{\frac{h(a_k|x)}{\hat{\pi}(a_k|x)}, M\right\}$$

$$\pi(a|x) = \sum_A \pi(A|x)\, 1\{a \in A\}$$

- Item-Position Click Model (IP):
  - $\bar{w}(a, k|x)$ depends on both item a and position k

$$\hat{V}_{IP}(h) = \frac{1}{|S|} \sum_{(x,A,w)\in S} \sum_{k=1}^{K} w(a_k, k) \min\left\{\frac{h(a_k, k|x)}{\hat{\pi}(a_k, k|x)}, M\right\}$$

$$\pi(a, k|x) = \sum_A \pi(A|x)\, 1\{a_k = a\}$$

- Rank-Based Click Model (RCTR):
  - $\bar{w}(a, k|x)$ only depends on position k

$$\hat{V}_R(h) = \frac{1}{|S|} \sum_{(x,A,w)\in S} \sum_{k=1}^{K} w(a_k, k)$$

- Position-Based Click Model (PBM):
  - $\bar{w}(a, k|x) = \mu(a|x) p(k|x)$

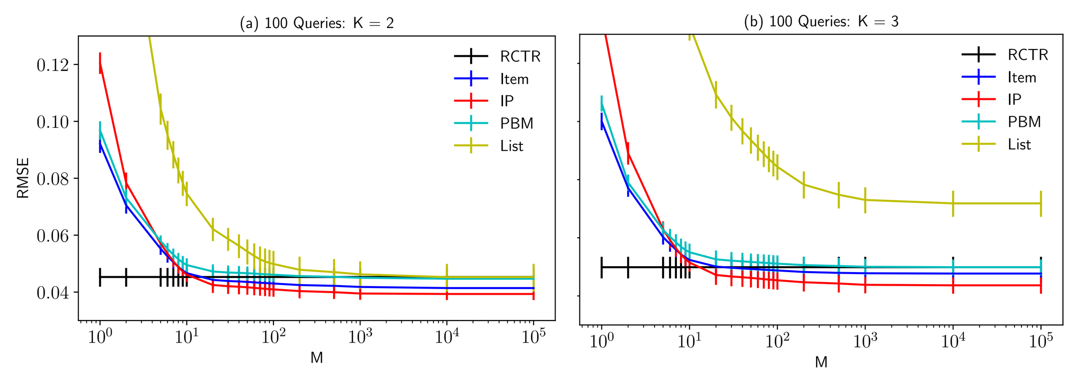
$$\hat{V}_{PBM}(h) = \frac{1}{|S|} \sum_{(x,A,w)\in S} \sum_{k=1}^{K} w(a_k, k) \min\left\{\frac{\langle p(\cdot\,|x), h(a_k, \cdot\,|x)\rangle}{\langle p(\cdot\,|x), \hat{\pi}(a_k, \cdot\,|x)\rangle}, M\right\}$$

Experiments

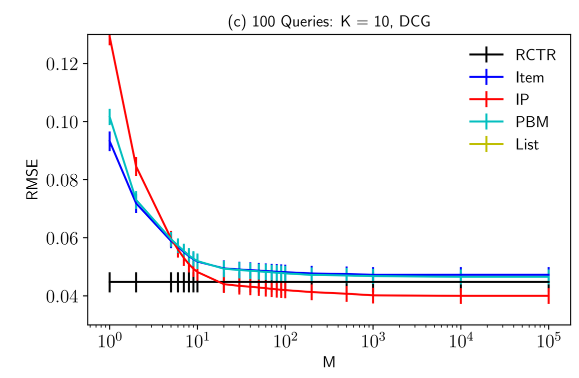
- Yandex dataset
- The dataset is recorded over 27 days
- Each record contains
  - a query ID
  - the day when the query occurs
  - 10 displayed items as a response to the query
  - the corresponding click indicators of each displayed items

- Logged dataset S
  - any records except day d
  - $\hat{\pi}$ is the empirical distribution over S
- Evaluation policy h
  - Take the records of day d
  - h is the empirical distribution of these records
  - The value $V(h)$ is the average CTR for these records

- Prediction errors on 100 most frequent queries as a function of clipping parameter M
  - Records of $K = 2$ or 3 positions



  - Records of $K = 10$ positions with DCG value



  - The performance of list estimator deteriorates fast with more positions
  - The IP estimator performs best

Analysis

Proposition 1.[Unbiased in a larger class of policies] Let $\mathcal{H}_Y$ contains all policies such that $\hat{V}_Y$ is unbiased, for any $Y \in \{L, IP, I, PBM\}$. Then $\mathcal{H}_L \subseteq \mathcal{H}_{IP} \subseteq \mathcal{H}_I / \mathcal{H}_{PBM}$.

Proposition 2.[Lower bias in estimating policy]

$$\mathbb{E}_S[\hat{V}_L] \leq \mathbb{E}_S[\hat{V}_{IP}] \leq \mathbb{E}_S[\hat{V}_I] / \mathbb{E}_S[\hat{V}_{PBM}] \leq V(h)$$

Proposition 3.[Policy optimization] Suppose $\tilde{h}_Y$ is the best policy under $\hat{V}_Y$, for any $Y \in \{L, IP, I, PBM\}$. Then the lower bound on $\tilde{h}_Y$ is at least as high as that on $\tilde{h}_L$.

Conclusions

- We propose various estimators for the expected number of clicks on lists generated by ranking policies that leverage the structure of click models
- We prove that our estimators are better than the unstructured list estimators, in the sense that they are less biased and have better guarantees for policy optimization
- Our estimators consistently outperform the list estimator in our experiments

Contact

Shuai Li
Email: shuaili@cse.cuhk.edu.hk

Yasin Abbasi-Yadkori
Email: abbasiya@adobe.com

Branislav Kveton
Email:bkveton@google.com

S. Muthukrishnan
Email: muthu@cs.rutgers.edu

Vishwa Vinay
Email: vinay@adobe.com

Zheng Wen
Email: zwen@adobe.com

Full Paper

References

1. Strehl, Alex, John Langford, Lihong Li, and Sham M. Kakade. "Learning from logged implicit exploration data." In Advances in Neural Information Processing Systems, pp. 2217-2225. 2010.
2. Swaminathan, Adith, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. "Off-policy evaluation for slate recommendation." In Advances in Neural Information Processing Systems, pp. 3632-3642. 2017.
3. Joachims, Thorsten, Adith Swaminathan, and Tobias Schnabel. "Unbiased learning-to-rank with biased feedback." In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, pp. 781-789. ACM, 2017.