BubbleRank: Safe Online Learning to Re-Rank via Implicit Click Feedback

Chang Li[1] Branislav Kveton[2] Tor Lattimore[3] Ilya Markov[1] Maarten de Rijke[1] Csaba Szepesvári[3,4] Masrour Zoghi[2]

[1]University of Amsterdam [2]Google Research [3]DeepMind [4]University of Alberta

Motivations

- Learning to rank: using machine learning to build ranking systems
- Offline approaches *lack exploration* and are limited to the provided training data
- Online approaches:
  - Balance exploration and exploitation
  - Learn from the user feedback, e.g., clicks
  - Tend to learn from scratch: *safety issue*

BubbleRank

A safe online learning to re-rank algorithm that combines the strength of both offline and online approaches.

Setup

- Item set $\mathcal{D} = [L]$ and $K \leq L$ positions
- Action set $\mathcal{R} \in \Pi_K(\mathcal{D})$: $\mathcal{R}(k)$ is the item at position k
- Click at any $k \in [K]$: $\boldsymbol{c}_t(k) = \boldsymbol{X}_t(\boldsymbol{\mathcal{R}}_t, k)\boldsymbol{A}_t(\boldsymbol{\mathcal{R}}_t(k))$
- Each time step t:
  - The agent chooses an action $\mathcal{R}_t \in \Pi_K(\mathcal{D})$
  - Observe the click feedback $\boldsymbol{c}_t \in \{0,1\}^K$
- Goal: maximize the expect number of clicks in top K positions
- Or minimize the expected cumulative regret in n steps:

$$R(n) = \sum_{t=1}^{n} \mathbb{E}\left[\max_{\mathcal{R}\in\Pi_K(\mathcal{D})} r(\mathcal{R}, \alpha, \chi) - r(\boldsymbol{\mathcal{R}}_t, \alpha, \chi)\right].$$

Assumptions

For any lists $\mathcal{R}, \mathcal{R}' \in \Pi_K(\mathcal{D})$ and positions $k, \ell \in [K]$ such that $k < \ell$:

A1. $r(\mathcal{R}, \alpha, \chi) \leq r(\mathcal{R}^*, \alpha, \chi)$, where $\mathcal{R}^* = (1 \dots K)$ is the optimal ranking;

A2. $\{\mathcal{R}(1), \dots, \mathcal{R}(k-1)\} = \{\mathcal{R}'(1), \dots, \mathcal{R}'(k-1)\}$ $\implies \chi(\mathcal{R}, k) = \chi(\mathcal{R}', k)$;

A3. $\chi(\mathcal{R}, k) \geq \chi(\mathcal{R}, \ell)$;

A4. If $\mathcal{R}$ and $\mathcal{R}'$ differ only in that the items at positions k and ℓ are exchanged, then $\alpha(\mathcal{R}(k)) \leq \alpha(\mathcal{R}(\ell)) \iff \chi(\mathcal{R}, \ell) \geq \chi(\mathcal{R}', \ell)$;

A5. $\chi(\mathcal{R}, k) \geq \chi(\mathcal{R}^*, k)$.

BubbleRank

Methodology: start with an *initial base list* $\mathcal{R}_0$ and improve it online by gradually exchanging higher-ranked less attractive items for lower-ranked more attractive items.

- At each step t:
  - $h \leftarrow t \bmod 2$
  - For $k \in [(K-h)/2]$: //*Building the display list*
    Randomly exchange items $\boldsymbol{\mathcal{R}}_t(2k-1+h)$ and $\boldsymbol{\mathcal{R}}_t(2k+h)$ in list $\boldsymbol{\mathcal{R}}_t$, if $\boldsymbol{s}_{t-1}(i,j) \leq 2\sqrt{\boldsymbol{n}_{t-1}(i,j)\log(1/\delta)}$
  - Display $\boldsymbol{\mathcal{R}}$ and observe clicks $\boldsymbol{c}_t \in \{0,1\}^K$
  - For $k \in [(K-h)/2]$ and $i \leftarrow \boldsymbol{\mathcal{R}}_t(2k-1+h)$, $j \leftarrow \boldsymbol{\mathcal{R}}_t(2k+h)$: // *Update stats*
    update $\boldsymbol{s}_t(i,j)$ and $\boldsymbol{n}_t(i,j)$ if $|\boldsymbol{c}_t(2k-1+h) - \boldsymbol{c}_t(2k+h)| = 1$
  - For $k \in [(K-h)/2]$ and $i \leftarrow \boldsymbol{\mathcal{R}}_t(2k-1+h)$, $j \leftarrow \boldsymbol{\mathcal{R}}_t(2k+h)$: // *Updating the base list*
    Permanently exchange $\boldsymbol{\mathcal{R}}_{t+1}(k)$ and $\boldsymbol{\mathcal{R}}_{t+1}(k+1)$ if $\boldsymbol{s}_t(j,i) > 2\sqrt{\boldsymbol{n}_t(j,i)\log(1/\delta)}$

Main Results

Theorem 1 (Upper Bound). The expected n-step regret of BubbleRank is bounded as

$$R(n) \leq 180K\frac{\chi_{\max}}{\chi_{\min}}\frac{K-1+2|\mathcal{V}_0|}{\Delta_{\min}}\log(1/\delta) + \delta^{\frac{1}{2}}K^3n^2.$$

Lemma 2 (Safety). Let

$$\mathcal{V}(\mathcal{R}) = \{(i,j) \in [K]^2 : i < j, \mathcal{R}^{-1}(i) > \mathcal{R}^{-1}(j)\}$$

be the set of *incorrectly-ordered item pairs* in list $\mathcal{R}$. Then

$$|\mathcal{V}(\boldsymbol{\mathcal{R}}_t)| \leq |\mathcal{V}(\mathcal{R}_0)| + K/2$$

holds uniformly over time with probability of at least $1 - \delta^{\frac{1}{2}}K^2n$.

Experimental setup

- Yandex Click log: at least one query in each session with 10 ranked items and 30M search sessions in total
- We randomly choose 100 frequent search queries and learn their CMs, DCMs and PBMs
- $L = 10$ items with $K = 10$ positions
- *Goal*: place 5 most attractive items in the descending order of attractiveness at the 5 highest positions

Experimental results

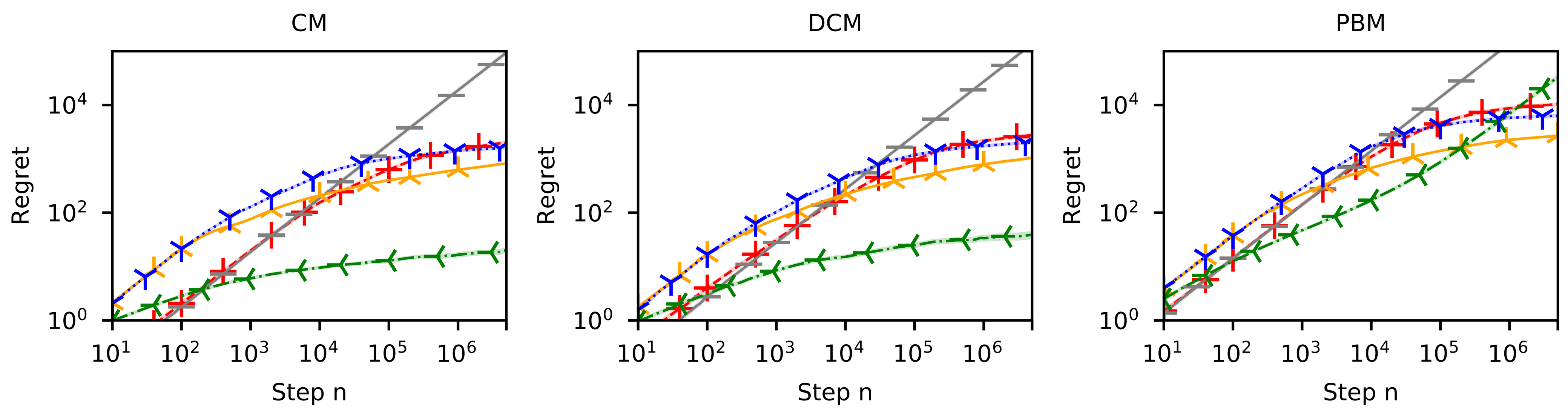

Figure: The n-step regret of BubbleRank (red), CascadeKL-UCB (green), BatchRank (blue), TopRank (orange), and Baseline (grey).

BubbleRank illustration

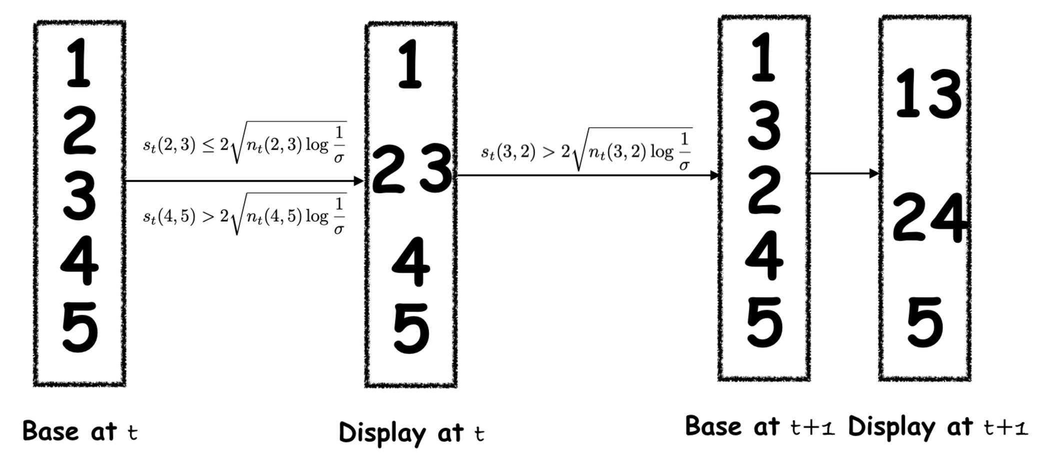

Figure: Illustration of BubbleRank with 5 items at step t.

Experimental results cont'd

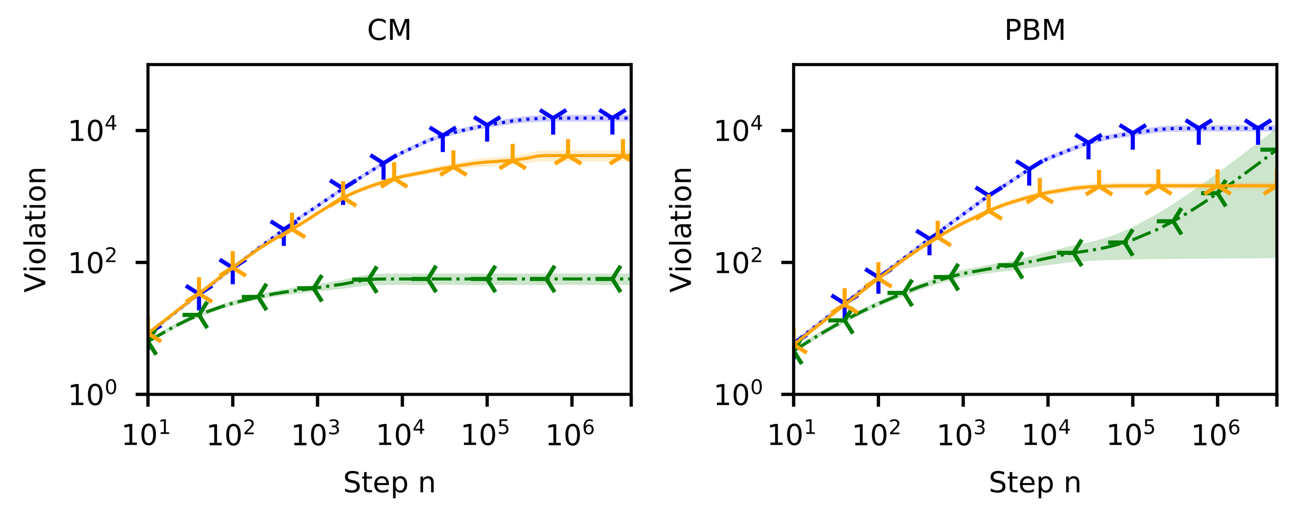

Figure: The n-step violation of the safety constraint.

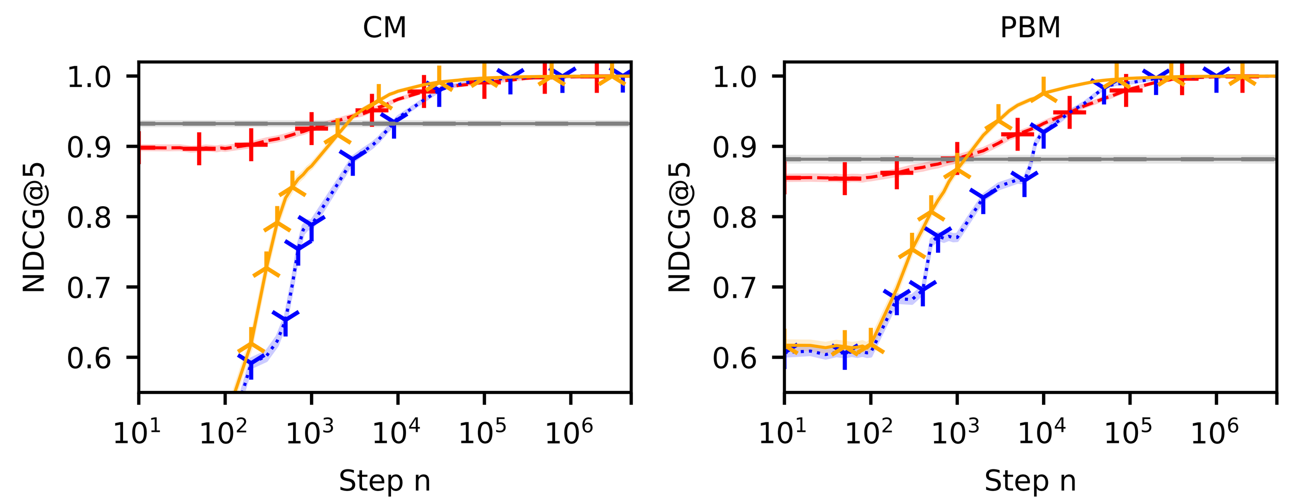

Figure: The per-step NDCG@5.

Conclusion

- BubbleRank fills the gap between online and offline LTR approaches in literature
- BubbleRank explores under a safety constraint
- BubbleRank learns slower than TopRank but can learn the optimal ranking eventually
- Future work: further theoretical and experimental analysis on BubbleRank in the online learning to rank setup

University of Amsterdam

c.li@uva.nl, bkveton@google.com, tor.lattimore@gmail.com, i.markov@uva.nl, derijke@uva.nl, szepesva@cs.ualberta.ca, masrour@zoghi.org