End-to-End Deep Attentive Personalized Item Retrieval for Online Content-sharing Platforms

Jyun-Yu Jiang, Tao Wu, Georgios Roumpos, Heng-Tze Cheng, Xinyang Yi, Ed Chi, Harish Ganapathy, Nitin Jindal, Pei Cao and Wei Wang

(Work done while interning at Google)

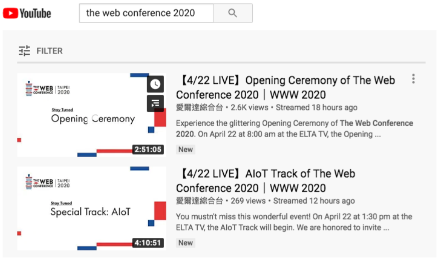
University of California, Los Angeles (UCLA)
Google Inc.

April 23, 2020

Outline

Online content-sharing platforms are popular in nowadays.

Search is one of the most essential functions for platforms

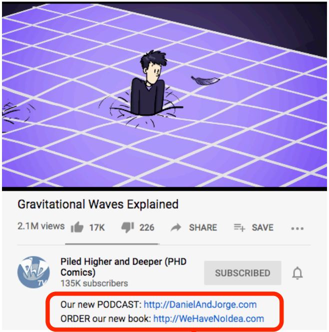


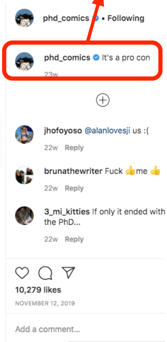
An example of YouTube search.

- Billions of videos on YouTube.
- Millions of musics on Spotify.
- Billions of photos on Instagram.
- . . . and more.

Item Retrieval is Hard.

Short or no description

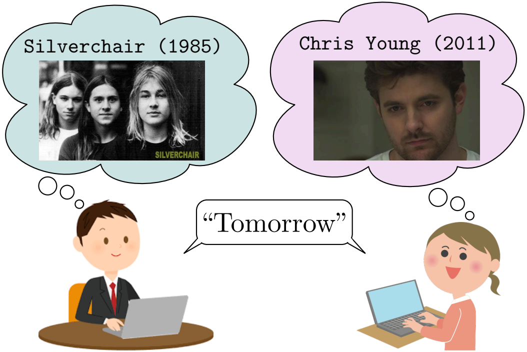




Noisy and advertising messages

Can we conduct item retrieval without using noisy descriptive information?

Users can have different intents for a query.

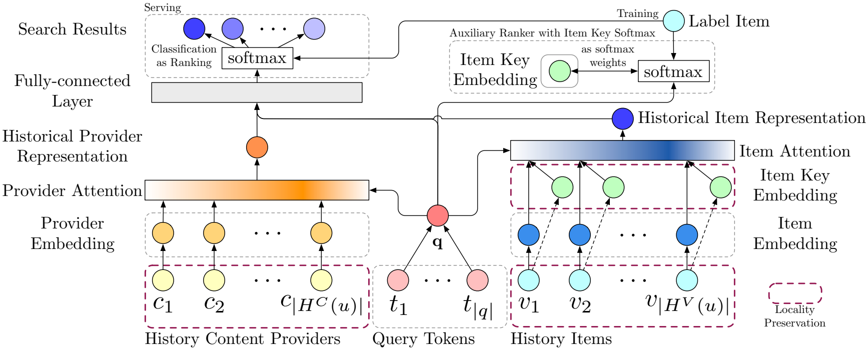


Personalization is important for the retrieval performance.

Comparisons between Different Retrieval Tasks

Personalized Task	Descriptive Information	Meta Information
Ad-hoc Search	✓ (documents)	✗
Web Search	✓ (web pages)	✗
Microblog Search	✓ (tweets)	✓ (hashtags)
Product Search	✓ (product reviews)	✓ (categories)
Item Search	✗	✓ (content providers)

Our Proposed Framework



Query-aware Attention with External Key Memory

Query embedding space can be much different from the item one.

- Given a query q, the attention weight of a historical item v can be:

$$\alpha_k(v, q) = \frac{\exp\left(\boldsymbol{q}^T \boldsymbol{k_v} / \sqrt{d}\right)}{\sum_{v' \in H^V(u)} \exp\left(\boldsymbol{q}^T \boldsymbol{k_{v'}} / \sqrt{d}\right)}.$$

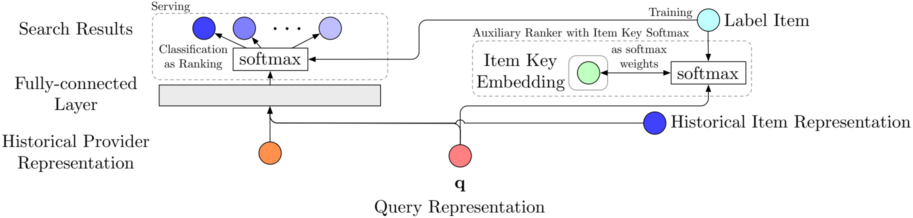
- $\boldsymbol{k_v}$ is the external key embedding.
- The historical item representation $\boldsymbol{h_V}$ can be re-written as

$$\boldsymbol{h_V} = \sum_{v \in H^V(u)} \alpha_k(v, q) \cdot \boldsymbol{v}.$$

Ultimate Features

Representations of historical items, providers, and the query.

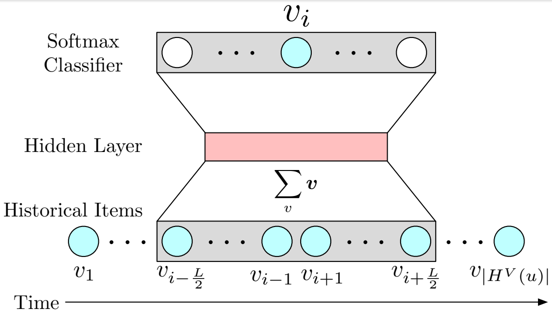
Classification as Ranking and Auxiliary Ranker



- Item key embeddings can be considered as ranking weights for regularization.

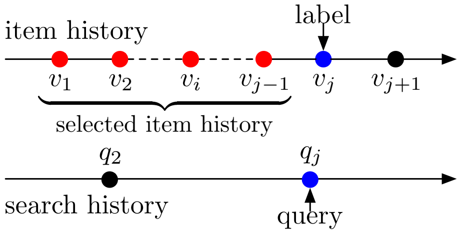
Locality Preservation

Context information in sessions can be leveraged.

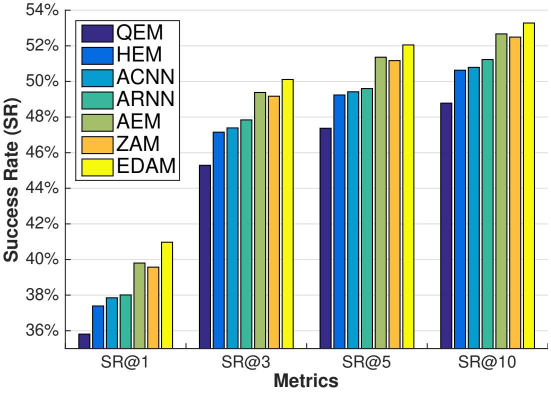


Experimental Datasets and Protocol

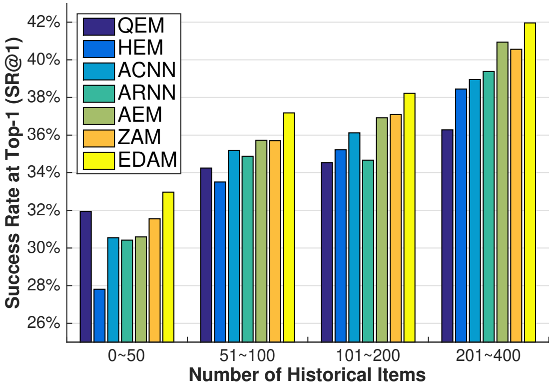
- User logs from a large video platform at Google
  - Videos as items and channels as content providers.
  - 400 most recent accessed items for 184M users.
- 90% of users are randomly sampled as training users.
  - The remaining 10% of users are considered testing users for evaluation.
- Evaluate the performance with Success Rate at K (SR@K)



Experimental results



Performance with different lengths of historical items



Ablation Study on Model Components

Method	Length of User History				
	Overall	[0, 50]	[51, 100]	[101, 200]	[201, 400]
EDAM	**0.4097**	**0.3297**	**0.3718**	**0.3822**	**0.4196**
without Auxiliary Ranking	0.3973	0.3031	0.3522	0.3696	0.4089
without Locaility Preservation	0.4039	0.3143	0.3591	0.3729	0.4155

Performance on Key Embeddings

Method	Length of User History				
	Overall	[0, 50]	[51, 100]	[101, 200]	[201, 400]
ZAM	0.3957	0.3155	0.3570	0.3709	0.4056
EDAM (Item)	**0.4097**	**0.3297**	**0.3718**	**0.3822**	**0.4196**
EDAM (Provider)	0.3808	0.3106	0.3513	0.3608	0.3892

Conclusions

- We propose a novel approach of personalized item retrieval.
- Independent item key embeddings improve the attention quality.
- Auxiliary ranker further sharpens the item key embeddings.
- Locality preservation as regularization also benefit performance.
- Significant improvements on a large-scale commercial dataset.
- Analysis shows the effectiveness and robustness of our approach.

Questions?
Or ask me by email: jyunyu@cs.ucla.edu
Personal Site: https://jyunyu.csie.org/