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

(Work done while interning at Google)

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

April 23, 2020
Outline

1. Introduction

2. EDAM: 
   End-to-end Deep Attentive Model for Personalized Item Retrieval

3. Experiments

4. Conclusions
Online content-sharing platforms are popular in nowadays.
Search is one of the most essential functions for platforms.

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

An example of YouTube search.
Item Retrieval is Hard.

Can we conduct item retrieval without using noisy descriptive information?
Users can have different intents for a query.

“Tomorrow”
Silverchair (1985)
Chris Young (2011)

Personalization is important for the retrieval performance.
## Comparisons between Different Retrieval Tasks

<table>
<thead>
<tr>
<th>Personalized Task</th>
<th>Descriptive Information</th>
<th>Meta Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad-hoc Search</td>
<td>✓ (documents)</td>
<td>✗</td>
</tr>
<tr>
<td>Web Search</td>
<td>✓ (web pages)</td>
<td>✗</td>
</tr>
<tr>
<td>Microblog Search</td>
<td>✓ (tweets)</td>
<td>✓ (hashtag)</td>
</tr>
<tr>
<td>Product Search</td>
<td>✓ (product reviews)</td>
<td>✓ (categories)</td>
</tr>
<tr>
<td>Item Search</td>
<td>✗</td>
<td>✓ (content providers)</td>
</tr>
</tbody>
</table>
Our Proposed Framework

- **Search Results**
  - Fully-connected Layer
  - Historical Provider Representation
  - Provider Attention
    - Provider Embedding
    - History Content Providers: $c_1, c_2, c_{|H^C(u)|}$
    - Query Tokens: $t_1, \ldots, t_{|q|}$
  - History Items: $v_1, v_2, v_{|H^V(u)|}$

- **Training**
  - Auxiliary Ranker with Item Key Softmax
  - Item Key Embedding

- **Serving**
  - Classification as Ranking
  - Softmax

- **Label Item**
  - Locality Preservation
  - Historical Item Representation
  - Item Attention
  - Item Embedding
  - Item Key Embedding
  - Fully-connected Layer

- **Softmax**
  - as softmax weights
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:

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

- $k_v$ is the external key embedding.
- The historical item representation $h_v$ can be re-written as

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

Ultimate Features

- Representations of historical items, providers, and the query.
Item key embeddings can be considered as ranking weights for regularization.
Locality Preservation

Context information in sessions can be leveraged.

\[ \sum_{v} v \]

\( \mathcal{V}_i \)

Softmax Classifier

Hidden Layer

Historical Items

Time

\( \mathcal{V}_1 \)

\( \mathcal{V}_i - \frac{L}{2} \)

\( \mathcal{V}_i - 1 \)

\( \mathcal{V}_i + 1 \)

\( \mathcal{V}_i + \frac{L}{2} \)

\( \mathcal{V}_|HV(u)| \)
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)

![Diagram of user history, item history, query, and label]

J.-Y. Jiang et al. (UCLA & Google)  End-to-End Deep Attentive Personalized Item Retrieval  April 23, 2019  11 / 16
Experimental results

![Graph showing success rate (SR) for different models at various metrics SR@1, SR@3, SR@5, SR@10. The models compared are QEM, HEM, ACNN, ARNN, AEM, ZAM, and EDAM. The graph illustrates the performance improvement of EDAM over other models in terms of success rate.]
Performance with different lengths of historical items

![Bar chart showing success rate at top-1 (SR@1) for different models and historical item lengths.]

- QEM
- HEM
- ACNN
- ARNN
- AEM
- ZAM
- EDAM

Number of Historical Items:
- 0~50
- 51~100
- 101~200
- 201~400

Success Rate at Top-1 (SR@1):
- 26%
- 28%
- 30%
- 32%
- 34%
- 36%
- 38%
- 40%
- 42%
## Ablation Study on Model Components

<table>
<thead>
<tr>
<th>Method</th>
<th>Length of User History</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td></td>
</tr>
<tr>
<td>[0, 50]</td>
<td>0.4097</td>
</tr>
<tr>
<td>[51, 100]</td>
<td>0.3297</td>
</tr>
<tr>
<td>[101, 200]</td>
<td>0.3718</td>
</tr>
<tr>
<td>[201, 400]</td>
<td>0.3822</td>
</tr>
<tr>
<td>EDAM</td>
<td>0.4196</td>
</tr>
<tr>
<td>without Auxiliary Ranking</td>
<td></td>
</tr>
<tr>
<td>without Locality Preservation</td>
<td></td>
</tr>
<tr>
<td>0.3973</td>
<td>0.3031</td>
</tr>
<tr>
<td>0.3522</td>
<td>0.3696</td>
</tr>
<tr>
<td>0.4089</td>
<td></td>
</tr>
<tr>
<td>0.4039</td>
<td>0.3143</td>
</tr>
<tr>
<td>0.3591</td>
<td>0.3729</td>
</tr>
<tr>
<td>0.4155</td>
<td></td>
</tr>
</tbody>
</table>
## Performance on Key Embeddings

<table>
<thead>
<tr>
<th>Method</th>
<th>Length of User History</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall</td>
</tr>
<tr>
<td>ZAM</td>
<td>0.3957</td>
</tr>
<tr>
<td>EDAM (Item)</td>
<td>0.4097</td>
</tr>
<tr>
<td>EDAM (Provider)</td>
<td>0.3808</td>
</tr>
</tbody>
</table>
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/