Recommender System for Publisher of Technical News

Yawen Sun, Yue Li, Yuzi Luo
Mentor: Rok Sosic, Jeffrey Ullman
Our Problem

Digital Trends (https://www.digitaltrends.com)
- Technology news, lifestyle, and information website.

Dataset
- Information about 170,000 News articles collected by content management system.
- Event data from real-time data collection platform over the past 15 months.

Goal
- Provide personalized content to users.
Methodology - Feature Extraction

News Articles Feature Extraction
- Timestamp of Publish Date
- 32-D Vector Representing Keywords using Word2Vec
- Category using K-means Clustering

User Feature Extraction
- Location
- Number of Clicks from Each Cluster
- Timestamp of Each Click Event
Posts > Word Vectors > Clusters

- clustering of posts by topic is our critical issue
- our clusters are generated based on tags and title words
- examination of clusters by either titles or tags show a lot of consistency
Clusters Matter

- consistency of clusters: \( \text{distortion} \propto \frac{1}{k} \)
- size of clusters: \( \text{size} \propto \frac{1}{k} \)

\[
\begin{align*}
\text{The Elbow Method showing the optimal } k \text{ from 1 to 10000} \\
\text{consistency always goes up}
\end{align*}
\]

\[
\begin{align*}
\text{Clusters Size Distribution (50, 100, 200, 400)} \\
\text{small min cluster size}
\end{align*}
\]

trade off between inter-cluster semantic consistency & cluster size distribution & computational resource
Clusters Optimization

- used ‘elbow method’ to pick the optimal number of clusters (consistency vs. computation)

- customized stop words according to posts’ frequent but less meaningful words (everything you need to know about... the best...)

- re-clustered the above-average-size clusters (deviation between cluster sizes)
Experiment Setup

Training & Test Data
- Select 7,000 users frequently visiting the website over 15 months.
- First 9 months’ viewing history for training
- Last 6 months for val & test

Learning Model
- Similarity Learning \( f_W(x, z) = x^T W z \).
- Wide and Deep combination of a linear model and a neural network

Evaluation
- Mean reciprocal rank (MRR)
  - by highest rank of exact posts
  - by highest rank of similar (cosine similarity > 0.95) posts from the same cluster

\[
MRR = \frac{1}{|Q|} \sum_{i=1}^{\left|Q\right|} \frac{1}{\text{rank}_i}.
\]
Results

cluster number = 50

exact posts:

SL:
val MRR: 0.0012
test MRR: 0.0012

W&D:
val MRR: 0.023
test MRR: 0.021

similar posts:

SL:
val MRR: 0.0328
test MRR: 0.0356

W&D:
val MRR: 0.397
test MRR: 0.392
# Results - evaluate by similarity

num of test users = 3500, threshold = 0.95

<table>
<thead>
<tr>
<th>num_cluster K</th>
<th>mrr</th>
<th>num_rank &lt; 10</th>
<th>training time</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>similarity learning</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.0356</td>
<td>126</td>
<td>2h15min</td>
</tr>
<tr>
<td>100</td>
<td>0.1111</td>
<td>1246</td>
<td>2h12min</td>
</tr>
<tr>
<td>200</td>
<td>0.0743</td>
<td>1186</td>
<td>2h16min</td>
</tr>
<tr>
<td>400</td>
<td>0.2068</td>
<td>1243</td>
<td>2h32min</td>
</tr>
<tr>
<td><strong>wide and deep</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.3924</td>
<td>2112</td>
<td>7h03min</td>
</tr>
<tr>
<td>100</td>
<td>0.4731</td>
<td>2261</td>
<td>8h38min</td>
</tr>
<tr>
<td>200</td>
<td>0.4980</td>
<td>2225</td>
<td>12h02min</td>
</tr>
<tr>
<td>400</td>
<td>0.2932</td>
<td>1356</td>
<td>19h37min</td>
</tr>
</tbody>
</table>
Limitations & Future Work

- Lack of Article Content
- Lack of Negative Data Points
- More Factors in News Recommendation to Consider
  - Short Time Big News Event
  - Cold Start Problem / Model Updating
  - Novelty Exploration