Cross-Domain Recommendation

--- Dataset and Case Study

Yong Li, Tsinghua University

Sept. 26th, 2017

fi.ee.tsinghua.edu.cn/~liyong/
Nowadays, **users**…

- Provide feedback for items of different types, e.g., in Amazon we can rate books, music, movies, ...
- Express their opinions/leave behavioral trace on different social media and different providers, e.g., Wechat, Facebook, Twitter, Amazon, Netflix, TripAdvisor, Dianping

Nowadays, **providers** wish to …

- Cross-sell products and services
- Provide recommendations to new users
Leverage all the available personal data (of all apps) provided in distinct domains to generate better recommendations!
Background

History and Problems

- **2002**: the term “cross-domain recommenders” appear for the first time in a patent:
  - Triplehop Technologies (now Oracle)
- **2005**: some papers suggest “cross-domain” as an interesting topic
  - Mark van Setten, Sean M. McNee, Joseph A. Konstan. 2005
  - Shlomo Berkovsky, Tsvi Kuflik, Francesco Ricci. 2005
- **2007**: first papers with contributions on “cross-domain”
  - Ronald Chung, David Sundaram, Ananth Srinivasan. 2007
  - Shlomo Berkovsky, Tsvi Kuflik, Francesco Ricci. 2007
  - Ronald Chung, David Sundaram, Ananth Srinivasan. 2007
New, real life cross-domain datasets are quite scarce and hard to reach in practice

• Gathered by big industry players, like BAT, Amazon, eBay, and Yelp, but rarely available to the broader research community;

• Moreover, cross large-domain data are very difficult to obtain, i.e., data from different service provider or apps.
Mobile Network Accessing Records Dataset

Device’s ID | Start time | End time | Location | HTTP logs

Large-scales
9800 BSs
6,500,000 users

Real-time
Generated day and night, accurate to second

Numerous-Apps
Most of the App behaviors are recorded

Cellular data access records from mobile operator is perfect to trace comprehensive user behaviors in different websites/apps/domains.
For a given app marketplace:

Crawl and download applications
1. Application webpage
2. Application executable (Free apps)

Operational result:

Application Profiles: web-pages and app-executables.
For an app with the profile:

Extract identifiers from
1. Application webpage
2. Application executable (Free apps)

Operational result:
Set of application identifiers grouped by type.
For an app in the repository:

Execute the application in an emulator to
1. Produce HTTP flows for the app
2. Store flows in labeled pcap files.

Operational result:

Labeled application flow headers.
For a pcap in the training set:

Identify lexical snippets in flows by
1. Looking for occurrence of identifiers
2. Extracting the general lexical context

Operational result:

Labeled lexical context training set.
Dataset

Raw data ➔ App & Activity & Content

For each lexical context:

1. Aggregating general contexts
2. Validating over test data

Operational result:

Conjunctive Rules.
Identify User Activities in the Apps: (wechat)
text, video call, voice call, location, picture, sight, moment ...
Identify the Accessed Content in the Apps: (weibo)
User ID, Clicked links, Posters, Viewed Video & Pages, etc.
Dataset

Structured personal dataset

Physical Domain: User ID || Time || BS ID || Location

Cyber Domain: User ID || Time || App || Activities || Content || Rela.
## Case Study 1: Cross-sites Video Recommendation

<table>
<thead>
<tr>
<th></th>
<th>YK</th>
<th>IQI</th>
<th>SH</th>
<th>KK</th>
<th>LE</th>
<th>TC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total user count</td>
<td>4479428</td>
<td>3156249</td>
<td>3156351</td>
<td>1351640</td>
<td>2190856</td>
<td>2781976</td>
</tr>
<tr>
<td>Total video count</td>
<td>1936643</td>
<td>169889</td>
<td>174229</td>
<td>59041</td>
<td>111250</td>
<td>130240</td>
</tr>
<tr>
<td>Total view count</td>
<td>90647660</td>
<td>24721556</td>
<td>39927427</td>
<td>8455611</td>
<td>24091297</td>
<td>17883308</td>
</tr>
</tbody>
</table>
Case Study 1: Cross-sites Video Recommendation

- Challenges for accurate user modeling
  - Data sparsity
  - Site-specific preferences
- Merging the multiple-site data can deal with the first challenges, but ignore the sites properties
- A generative model of **Multi-site Probabilistic Factorization (MPF)** to capture both the cross-site as well as site-specific preferences
Case Study 1  

Cross-sites Video Recommendation

Performance Gains

Cross-sites recommendation improves the F-measure by 12.96%, 8.24%, and 6.88% compared to SMF, CMF and MMF.

<table>
<thead>
<tr>
<th></th>
<th>F-measure</th>
<th>SMF</th>
<th>MMF</th>
<th>CMF</th>
<th>MPF</th>
</tr>
</thead>
<tbody>
<tr>
<td>YK</td>
<td>0.880</td>
<td>0.888</td>
<td>0.881</td>
<td></td>
<td>0.911</td>
</tr>
<tr>
<td>IQI</td>
<td>0.699</td>
<td>0.779</td>
<td>0.761</td>
<td></td>
<td>0.829</td>
</tr>
<tr>
<td>SH</td>
<td>0.771</td>
<td>0.808</td>
<td>0.806</td>
<td></td>
<td>0.868</td>
</tr>
<tr>
<td>KK</td>
<td>0.694</td>
<td>0.717</td>
<td>0.706</td>
<td></td>
<td>0.788</td>
</tr>
<tr>
<td>LE</td>
<td>0.800</td>
<td>0.837</td>
<td>0.833</td>
<td></td>
<td>0.892</td>
</tr>
<tr>
<td>TC</td>
<td>0.692</td>
<td>0.769</td>
<td>0.770</td>
<td></td>
<td>0.834</td>
</tr>
<tr>
<td>Overall</td>
<td>0.756</td>
<td>0.799</td>
<td>0.789</td>
<td></td>
<td>0.854</td>
</tr>
</tbody>
</table>

Site-site Improvement

1. Site-site correlation of multi-homed users’ latent factors

2. Data sparsity of each site is 0.716

C. Yang, H. Yan, D. Yu, Yong Li, D. Chiu, Multi-site User Behavior Modeling and its Application in Video Recommendation, in *ACM SIGIR 2017.*
Case Study 2  Cross-domain App Usage Prediction

Given a place, can we know the (popular) apps that are being used by the people around?

Establish relationship between the characteristics of a physical location and the apps usage?
Case Study 2

Cross-domain App Usage Prediction

Insights of Leveraging POIs for app prediction
Case Study 2  Cross-domain App Usage Prediction

Cold-start Problem:
Only use POI

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Model</th>
<th>CMF</th>
<th>MLR</th>
<th>AOP</th>
<th>SMF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top5 hitrate</td>
<td>0.84</td>
<td>0.77</td>
<td>0.74</td>
<td>0.57</td>
<td>0.04</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.148</td>
<td>0.167</td>
<td>0.172</td>
<td>-</td>
<td>0.319</td>
</tr>
</tbody>
</table>

D. Y, Yong Li, F. Xu, P. Zhang, V. Kostakos. Smartphone App Usage Prediction Using Points of Interest, Accepted by ACM IMWUT (UbiComp)
Summary

• A nice dataset for cross-domain human behavior understanding, modeling and utilization

• Before opening it, we are still dealing with the privacy issues to avoid re-identification attacks (but can be shared in closed community)

• Future directions: cross-domains of social and information recommendation, joint cyber-physical cross-domain modeling, etc.
Thank You!

Yong Li, Tsinghua University

[Link to Yong Li's page](fi.ee.tsinghua.edu.cn/~liyong/)
Evaluation

Experiments and Baselines Setting

- Top-N hit rate
- Top-N prediction accuracy
- Root Mean Square Error (RMSE)

- **AOP**: Based on the total usage amount of all locations
- **MLR**: Multiple logistic regression
- **SMF**: Single matrix factorization
- **CMF**: Collective matrix factorization
- **Our Temporal CMF Model**