Information Mining in large-scale Github Network

Yang Liu, Du Su, Ailing Zhang
{liu301,dusu3,azhang41}@illinois.edu
University of Illinois Urbana-Champaign
May 9, 2016
Github Archive

- All github events available at: https://www.githubarchive.org/
- Server log includes:
  - Totally 20+ Events
  - PushEvent: Pushed code to remote repository.
  - ForkEvent: Copy repository.
  - PullRequestEvent: User apply to become a Collaborator
  - StarEvent: User stared repository
Our dataset

• Crawled four types of events from 2015/01/01 to 2015/06/30.
  ○ 1,515,548 users, 4,192,273 repos
  ○ 102,572,900 events

• Heterogeneous network.
  ○ Entity types: user, repository
  ○ Link types: commit, fork, pull request, star
  ○ Rich information: other potentially useful information such as commit messages and release notes.
What could we mine?
Commit Distribution

- Commit: Comments before pushing the code
- Commit Distribution: User vs Repo
User Commit Distribution

• How about commit distribution for one user?

• Detect user’s peak: Dynamic Gaussian Mixture Model
User-Repo Peak Distribution

- Peak Number Distribution

![Statistics of User-Peak](image1)

![Statistics of Repo-Peak](image2)
Demo
Link Prediction

- With multiple link types and entity types
- Large-scale one, millions of nodes/edges
- Embedding:
  - TransE
    - Translations operating on the low-dimensional embeddings of the entities.
  - LINE
    - 1st order similarity: nodes with strong ties turn to be similar
    - 2nd order similarity: nodes share many neighbors turn to be similar
\[
\sum \log P(v_1, v_2)
\]
\[
\sum \log P(v_1 | v_2)
\]
• $\sum \log P(v_1, v_2) + \sum \log P(w, v_2)$
• $\sum \log P(v_1|v_2) + \sum \log P(w|v_2)$
Our Result

• Experiment Setup
  ○ Training set: 2015 Jan -> 2015 May, 123,569,237 Triplets
  ○ Testing set: 2015 Jun, 761,295 Triplets
  ○ Method: Firstly embedding, then logistic regression with liblinear.

• Result

<table>
<thead>
<tr>
<th>Algr.</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>TransE</td>
<td>0.749</td>
</tr>
<tr>
<td>LINE-1st</td>
<td>0.676</td>
</tr>
<tr>
<td>LINE-2nd</td>
<td>0.842</td>
</tr>
<tr>
<td>SemLINE-1st</td>
<td>0.818</td>
</tr>
<tr>
<td>SemLINE-2nd</td>
<td>0.853</td>
</tr>
</tbody>
</table>
Summary

• Github network provided rich information
  ◦ Multiple link Type: Pull, Push, Fork, Star etc.
  ◦ Rich Semantic Information with every link

• Observation:
  ◦ Most users contribute periodly
  ◦ Word frequency could indicate user’s expression

• Link Prediction:
  ◦ Incorporate semantic information on each link into embedding method
  ◦ Achieve better performance on link prediction
Thank You!