Meta-Path Guided Embedding for Heterogeneous Information Network Consolidation

CS512 Course Project

- Yuning Mao (yuningm2), Xiaotao Gu (xiaotao2), Sha Li (shal2), Jiaming Shen (js2)
- May 7, 2018
01

Motivation
What’s HIN Consolidation?

- **Structured Data**
  - *Venue*
  - *Paper* (Limited)
  - *Research Interests*
  - *Affiliation* (Natural, but limited)

- **Unstructured Data**
  - *Terms*
  - *Missing attribute*

- *DBLP*
Rich Information from Unstructured Data

position

Jiawei Han
Abel Bliss Professor, Department of Computer Science
Univ. of Illinois at Urbana-Champaign
Rm 2132, Siebel Center for Computer Science
201 N. Goodwin Avenue
Urbana, IL 61801, USA
E-mail: hanj[at]cs.uiuc.edu
Ph.D. (1985), Computer Science, Univ. Wisconsin-Madison

Knowledge Discovery and Data Mining, Database Systems

affiliation

Data Mining Research Group
Data and Information Systems Research Laboratory
UIUC Calendar: (17-18) (Cites: Exchange) (CS)
Office: (217) 333-6903
Fax: (217) 265-6494
Web: hanj.cs.illinois.edu
Schedule: Meetings and Appointments

interests
Rich Information from Unstructured Data

tags

- pytorch
- Tensors and Dynamic neural networks in Python with strong GPU acceleration
- http://pytorch.org
- neural-network
- autograd
- gpu
- numpy
- deep-learning
- tensor
- python

- 11,020 commits
- 47 branches
- 16 releases
- 612 contributors

Latest commit 4ad03a4 5 minutes ago
- bddpq Add dependency from caffe2_gpu to ATen in CMake (#7117)
- _github_ updated the environment collection script URL to the raw version on G...
- _jenkins_ Update C++ API tests to use Catch2 (#7108)
Methodology
Methodology

- Meta-path Guided Network Embedding
  - User-guided
  - Different meta-paths have different semantics
  - Different feature representations for different attributes

- A link prediction model that utilizes the learned embeddings
  - Learning to Match
  - Matching between objects in two heterogeneous domains
Methodology

Existing HIN

Unstructured data associated with nodes

Representation Learning for HIN

Learning to Match

Consolidated HIN

Other affiliations:
- Database group at CMU
- Machine Learning Department
- PDL (Parallel Data Lab)

Past affiliations:
- Informedia (1997–2007)
- Computational Biology Department (2003–2015)

Research Interests:
- Data Mining for graphs and streams
- Fractals, self-similarity and power laws
- Indexing and data mining for video, biological and medical databases
- Database performance evaluation (data placement, workload characterization)
Meta-path Guided Network Embedding

- Learn embeddings for nodes in the HIN

Probability of node $v$ being connected to node $u$ under metapath $\mathcal{M}$:

$$Pr(v|u, \mathcal{M}) = \frac{\exp(f(u, v, \mathcal{M}))}{\sum_{v' \in V} \exp(f(u, v', \mathcal{M}))}$$

Metapath scoring function $f$:

$$f(u, v, \mathcal{M}) = \mu_{\mathcal{M}} + \mathbf{p}_{\mathcal{M}}^T \mathbf{x}_u + \mathbf{q}_{\mathcal{M}}^T \mathbf{x}_v + \mathbf{x}_u^T \mathbf{x}_v$$

Node embeddings
Learning to Match

• Map nodes and their attributes into latent space
• Exploit the correlations between labels
• Semantics preserved vs. Multi-label Classification
Experiments
Results on DBLP

- DBLP dataset (1.03M authors, 2.7M nodes, 103M edges)
- 4K authors with research interests
- 30 unique phrases that represent research interests
- 3K training, 1K test
- Precision@1/3/5

<table>
<thead>
<tr>
<th>Methods</th>
<th>P @ 1</th>
<th>P @ 3</th>
<th>P @ 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.816</td>
<td>0.577</td>
<td>0.465</td>
</tr>
<tr>
<td>GuidedHeteEmbedding (WSDM’17)</td>
<td>0.597</td>
<td>0.507</td>
<td>0.435</td>
</tr>
<tr>
<td>Ours (A-P-A)</td>
<td>0.786</td>
<td>0.579</td>
<td>0.481</td>
</tr>
<tr>
<td>Ours (A-P-V-P-A)</td>
<td><strong>0.866</strong></td>
<td><strong>0.625</strong></td>
<td><strong>0.510</strong></td>
</tr>
</tbody>
</table>
Case Study (test set)

Jiawei Han
- database
- data_mining
- information_retrieval
- machine_learning
- artificial_intelligence

Wei Wang
- data_mining
- machine_learning
- neural_network
- bioinformatics
- database

Xin Luna Dong
- database
- data_mining
- information_retrieval
- distributed_system
- information_system

Michael I. Jordan
- machine_learning
- data_mining
- artificial_intelligence
- bioinformatics
- neural_network

Ting Liu
- machine_learning
- information_retrieval
- natural_language_processing
- data_mining
- text_mining

Christos Faloutsos
- database
- data_mining
- information_retrieval
- machine_learning
- world_wide_web
Results on GitHub

- GitHub dataset (user-repo bipartite graph)
- 372,817 repos with tags
- 393 unique tags
- 80% training, 20% test
- Precision@1/3/5

<table>
<thead>
<tr>
<th>Methods</th>
<th>$P @ 1$</th>
<th>$P @ 3$</th>
<th>$P @ 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.654</td>
<td>0.398</td>
<td>0.290</td>
</tr>
<tr>
<td>GuidedHeteEmbedding (WSDM'17)</td>
<td>0.732</td>
<td>0.428</td>
<td>0.302</td>
</tr>
<tr>
<td>Ours (R-U-R)</td>
<td>0.713</td>
<td>0.422</td>
<td>0.301</td>
</tr>
</tbody>
</table>
Thank you for your attention!

Yuning Mao, Xiaotao Gu, Sha Li, Jiaming Shen