HINSim: Meta-Path Guided Embedding for Similarity Search in Massive Heterogeneous Information Networks

Jingbo Shang, Meng Qu, Jialu Liu, Jian Peng, Jiawei Han
Feb 4, 2016
Presented by Jingbo Shang
Outline

- Motivation: Why Meta-Path is important?
- Related Work
- HINSim: Methodology
- Experimental Results
- Discussion and Future Work
Why Heterogenous?

- Heterogeneous v.s. Homogenous
  - The vertices and edges of a HIN are associated with different types
Why Meta-Path?

- The similarity between vertices
  - There are so many ways to define the similarity
  - Co-authorship: $\text{sim}(B,C) > \text{sim}(B,A)$
  - Research interests: $\text{sim}(B,A) > \text{sim}(B,C)$
Why Meta-Path?

- The representation of similarity
  - Meta-Path
  - Co-authorship: A-P-A
  - Research interests: A-P-V-P-A

- Meta-path guided similarity search
  - We leave the subjectivity to users by requesting a user-selected meta-path to indicate his/her preference
Why not Project Hetero to Homo?

- Loss of information
  - Vertices of other types may have useful information
- Scalability
  - The projected network might be dense
Outline

- Motivation: Why Meta-Path is important?
- Related Work
- HINSim: Methodology
- Experimental Results
- Discussion and Future Work
PathSim

- Use the normalized counts of the concrete paths belonging to the selected meta-path between any pairs of vertices.
- Can’t uncover the similarity if there are no concrete paths between two vertices.
Embed Projected Networks

- Project the original HIN to a bipartite or homogenous network and then apply network embedding techniques.

- Problems
  - Loss of information.
  - The projected network may be very dense. The computational costs of the projected networks are thus usually very high.
Outline

- Motivation: Why Meta-Path is important?
- Related Work
- HINSim: Methodology
- Experimental Results
- Discussion and Future Work
HINSim: Probability of a concrete meta-path

- The conditional probability of \( v \) generated by vertex \( u \) under type \( r \)

\[
Pr(v|u, r) = \frac{\exp(f(u, v, r))}{\sum_{v'} \exp(f(u, v', r))}
\]

- Function \( f \) is a scoring function modeling the relevance between \( u \) and \( v \) conditioned on edge type \( r \)

\[
f(u, v, r) = \mu_r + p_r^T x_u + q_r^T x_v + x_u^T S_r x_v
\]

- The probability of a concrete path

\[
Pr(\mathcal{P}_{e_1 \ldots e_L}|R) = Pr(u_1|R) \times Pr(\mathcal{P}_{e_1 \ldots e_L}|u_1, R) \\
\propto C(u_1, 1|R)^\gamma \times \prod_{l=1}^{L} Pr(v_l|u_l, r_l)
\]

- \( C(u, i|R) \) represents the count of paths constrained on \( R \) with the \( i \)-th vertex being \( u \)
HINSim: The framework

- Following the noise contrastive estimation (NCE), we reduce the problem to a binary classification

\[ \mathcal{L}_R \approx \sum_{P_{e_1 \rightarrow e_L} \in \mathcal{P}_R} \log \sigma(\sum_i f(u_i, v_i, r_i)) + \sum_{k=1}^{K} \mathbb{E}_{P_{e_1 \rightarrow e_L} \sim P_{r^{-}}} \left[ \log \left(1 - \sigma(\sum_i f(u_i^k, v_i^k, r_i^k))\right) \right] \]

- \( P_{r^{-}} \) is a noise distribution which is defined as (\( \gamma = 3/4 \) usually)

\[ P_{r^{-}}(P_{e_1 \rightarrow e_L} | R) \propto \prod_{i=1}^{L+1} C(u_i, i | R)^\gamma \]

- We also want to model the whole heterogeneous information network by viewing every edge type as a length-1 meta-path

\[ \mathcal{L} = \mathcal{L}_R + \lambda \sum_r \mathcal{L}_r \]
### HINSim

#### Algorithm 1: HINSim Training Framework

**Require:** Heterogenous information network $G = (V, E)$, user-specified meta-path $R$, sampling times $t$, and negative sampling ratio $K$

**Return:** Vertex Embedding Vector Representations $x_u, \forall u$

initialize parameters $x, S, p, q$

```plaintext
while not converge do
    // for network edges
    for each edge type $r$ do
        for $i = 1$ to $t$ do
            $e^+$ ← a sampled positive type-$r$ edge
            Optimize for an edge $e^+$ with label 1.
            for $j = 1$ to $K$ do
                $e^-$ ← a sampled negative type-$r$ edge
                Optimize for an edge $e^-$ with label 0.
        
    // for meta-path
    for $i = 1$ to $t$ do
        $p^+$ ← a sampled positive concrete $R$ meta-path
        Optimize for a concrete path $p^+$ with label 1.
        for $j = 1$ to $K$ do
            $p^-$ ← a sampled negative concrete $R$ meta-path
            Optimize for a concrete path $p^-$ with label 0.

return $x$.
```

Update for edges

Update for paths
Outline

- Motivation: Why Meta-Path is important?
- Related Work
- HINSim: Methodology
- Experimental Results
- Discussion and Future Work
Experiments: Settings

- Datasets:
  - DBLP
    - A star schema network and includes paper (P), author (A), venue (V) as vertices
    - Selected Meta-Paths: A-P-A  A-P-V-P-A
  - Yelp
    - A star schema network and includes review (R), business (B), user (U), and words (W) as vertices
Experiments: Settings

- Evaluation
  - Groundtruth: group labels (e.g., 4 research areas in DBLP and business types in Yelp)
  - Rank nodes based on cosine similarity
  - Metric: AUC
    - The probability that the nodes from the same group have higher ranks than the nodes from different groups
Experiments: Quantitative Results

- The effect of the scoring function
- The effect of the meta-paths
- Different meta-paths lead to different semantic meaning

<table>
<thead>
<tr>
<th>Table 2: AUC Evaluation in the DBLP dataset.</th>
<th>Table 3: AUC Evaluation in the Yelp dataset.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Model</td>
</tr>
<tr>
<td>HINSim-APVPA 83.22% 86.25%</td>
<td>HINSim-BRWRB 77.97% 75.17%</td>
</tr>
<tr>
<td>HINSim-APA 54.01% 88.17%</td>
<td>HINSim-BRURB 56.17% 55.29%</td>
</tr>
<tr>
<td>LINE-1st 78.73 82.51</td>
<td>LINE-1st 74.41% 70.16%</td>
</tr>
<tr>
<td>LINE-2nd 74.83 81.21</td>
<td>LINE-2nd 75.06% 69.41%</td>
</tr>
<tr>
<td>PathSim-APVPA 79.23% 77.50%</td>
<td>PathSim-BRWRB 73.57% 69.93%</td>
</tr>
<tr>
<td>PathSim-APA 50.05% 56.38%</td>
<td>PathSim-BRURB 52.16% 52.57%</td>
</tr>
<tr>
<td>HINSim-NoPath 80.15 83.14</td>
<td>HINSim-NoPath 75.75% 71.65%</td>
</tr>
<tr>
<td>LINE-APV 80.19% 85.01%</td>
<td>LINE-BRW 77.07% 74.06%</td>
</tr>
<tr>
<td>LINE-AP 54.45% 87.35%</td>
<td>LINE-BRU 51.30% 51.82%</td>
</tr>
</tbody>
</table>
Experiments: Visualization

Figure 3: Visualization of embedding vector representations of 10% random sampled authors in DBLP Research Area dataset. Colors correspond to different research areas, including blue: Database, orange: Data Mining, magenta: Machine Learning, and green: Information Retrieval. The projection to a 2-D space is done by the t-SNE package [30]. Note that the projected points in different figures with similar positions may refer to different vertices in the network.
Experiments: Visualization

Figure 4: Visualization of embedding vector representations of all authors in DBLP Research Group dataset. Colors correspond to different research groups, including blue: Jiawei Han, orange: Christos Faloutsos, magenta: Dan Roth, and green: Michael I. Jordan. The projection to a 2-D space is done by the t-SNE package [30]. Note that the projected points in different figures with similar positions may refer to different vertices in the network.
Experiments: Efficiency

- The time complexity is quadratic to the number of vertices $|V|$
- The speed-up ratio is quite close to linear

**Figure 5: Efficiency evaluation.**
Outline

- Motivation: Why Meta-Path is important?
- Related Work
- HINSim: Methodology
- Experimental Results
- Discussion and Future Work
Future Work

- Generate Meta-Paths automatically, supervised by some specific tasks.