Evaluate Top-k Meta Path Queries on Large Heterogeneous Information Networks

Zichen Zhu
Background

- HIN: Directed graph $G = (V, E)$ with multiple node types and edge types
- Meta Path: a sequence of node and edge types (e.g. $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$)

(a) Network Schema

(b) Meta Paths
Problem Statement

• Top-k the most important meta paths in Heterogeneous Information Networks (HIN):
  ➢ A meta path can summarize high-level relationship patterns
  ➢ A longer meta path could be meaningless and even misleading [1], but it does not mean its length is the only criterion of importance
Example

• Found 13 paths of the same length, which can be summarized into just 2 meta paths:

\[ P_1: \quad \text{Author} \xrightarrow{\text{belong to}} \text{Affiliation} \xrightarrow{\text{belong}^{-1}} \text{Author} \quad 1 \]

\[ P_2: \quad \text{Author} \xrightarrow{\text{write}} \text{Paper} \xrightarrow{\text{write}^{-1}} \text{Author} \quad 12 \]

Which one is better? \( P_2 \) is better than \( P_1 \)

The shorter, the better? Not necessarily!

\[ P_3: \quad \text{Author} \xrightarrow{\text{write}} \text{Paper} \xrightarrow{\text{cite}} \text{Paper} \xrightarrow{\text{write}^{-1}} \text{Author} \quad 20 \]

\[ P_4: \quad \text{Author} \xrightarrow{\text{write}} \text{Paper} \xrightarrow{\text{cite}^{-1}} \text{Paper} \xrightarrow{\text{write}^{-1}} \text{Author} \quad 22 \]
Importance Function

For a meta path \( P = A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1} \) that connects \( s \) and \( t \), the importance function \( I(P) \) can be generalized as follows:

\[
I_{s,t}(P) = supp_{s,t}(P) \times rare_{s,t}(P, D) \times penalty(|P|)
\]

- Minimum Instances:
  - \( MNI_{s,t}(P) = \min_{1 < i < l+1} |\{p_i | p \in P_{s,t}\}| \)
  - \( MNI_{s,t}(P) \leq |\{p_2 | p \in P_{s,t}\}| \)
- Rarity: \( P \) is how rare among those similar pairs to \((s, t)\)
  - \( rare(P, D) = \log \frac{|D|}{|(u,v) | (u,v) \in D, P \in \mathbb{P}_{u \rightarrow v}|} \)
  - \( D = \{(s,v) | v \in V, L(v) = L(t)\} \cup \{(u,t) | u \in V, L(u) = L(s)\} \)
  - \( rare(P, D) \leq \log(|D|) \)
- \( supp_{s,t}(P) = (\prod_{1}^{l} \frac{1}{\min(o(A_i | R_i), l(A_{i+1} | R_i))}) \times MNI_{s,t}(P) \)
- \( Penalty(P) = \beta |P|, \beta \in (0, 1) \)
A* Searching

- MNI-based Importance function:
  \[ \overline{I(P)} = \prod_{i=1}^{l} \frac{1}{\min(o(A_i|R_i), I(A_{i+1}|R_i))} \min_{1<i<l+1} \{|p_i|p \in P\} \ast \log(|D|) \ast \beta^{|P|} \]
  \[ \overline{I(P \circ R)} \leq \overline{I(P)} \]

- Stop Condition: SQ.top < P_{kk}

SQ: Searching Priority Queue by \( \overline{I(P)} \)

RQ: Result Priority Queue by \( I_{s,t}(P) \)
Experiment

Case Study: Top-4 Meta Paths between Philip Yu and Jiawei Han

<table>
<thead>
<tr>
<th>Rank</th>
<th>Meta Path $\mathcal{P}$</th>
<th>$\mathcal{I}(\mathcal{P})$</th>
<th>MNI</th>
<th>Rarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$A \rightarrow P \leftarrow A$</td>
<td>0.154</td>
<td>12</td>
<td>5.8844</td>
</tr>
<tr>
<td>2</td>
<td>$A \rightarrow P \leftarrow P \leftarrow A$</td>
<td>$2.4 \times 10^{-3}$</td>
<td>14</td>
<td>4.3884</td>
</tr>
<tr>
<td>3</td>
<td>$A \rightarrow P \rightarrow P \leftarrow A$</td>
<td>$2.1 \times 10^{-3}$</td>
<td>13</td>
<td>4.1258</td>
</tr>
<tr>
<td>4</td>
<td>$A \rightarrow P \leftarrow A \rightarrow P \leftarrow A$</td>
<td>$3.65 \times 10^{-4}$</td>
<td>21</td>
<td>3.6735</td>
</tr>
</tbody>
</table>

Labeled-based Connectivity

MNIS and Shortest Paths

Different beta in MNIS


Backups

• Baseline 1: Shortest Meta Path (SP)
  ➢ \( supp_{s,t}(P) = 1 \)
  ➢ \( rare_{s,t}(P,D) = 1 \)
  ➢ \( penalty(|P|) = \frac{1}{|P|} \)

• Baseline 2: Weight Function in WsRel[2] (SLV1)
  ➢ \( supp_{s,t}(P) = \prod_1^l \frac{1}{\sqrt{o(A_i|R_i)\times I(A_{i+1}|R_i)}} \)
  ➢ \( rare_{s,t}(P,D) = 1 \)
  ➢ \( penalty(|P|) = \frac{1}{|P|} \)

• Baseline 3: Weight Function in HeteRecom[3] (SLV2)
  ➢ \( supp_{s,t}(P) = e^{\prod_1^l(o(A_i|R_i)\times I(A_{i+1}|R_i))^{\alpha}}, \alpha \in [0, 1] \)
  ➢ \( rare_{s,t}(P,D) = 1 \)
  ➢ \( penalty(|P|) = \frac{1}{e^{|P|}} \)
**Backups**

**Case Study: SLV1 and SLV2**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Meta Path $\mathcal{P}$</th>
<th>$\mathcal{I}(\mathcal{P})$ (SLV1)</th>
<th>$\mathcal{I}(\mathcal{P})$ (SLV2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$A \rightarrow P \leftarrow A$</td>
<td>$1.2 \times 10^{-2}$</td>
<td>0.139</td>
</tr>
<tr>
<td>2</td>
<td>$A \rightarrow F \leftarrow A$</td>
<td>$4.86 \times 10^{-4}$</td>
<td>0.135</td>
</tr>
<tr>
<td>3</td>
<td>$A \rightarrow P \rightarrow P \leftarrow A$</td>
<td>$4.82 \times 10^{-4}$</td>
<td>$4.99 \times 10^{-2}$</td>
</tr>
<tr>
<td>4</td>
<td>$A \rightarrow P \leftarrow P \leftarrow A$</td>
<td>$4.82 \times 10^{-4}$</td>
<td>$4.99 \times 10^{-2}$</td>
</tr>
</tbody>
</table>

**Connectivity Experiment**

![Graph showing Accuracy@k vs. k for different methods: MNIS(0.1), SP, SLV1, SLV2.](image)