Toward Activity Discovery in the Personal Web

Tara Safavi, Adam Fourney, Robert Sim, Marcin Juraszek, Shane Williams, Ned Friend, Danai Koutra, Paul N. Bennett

Overview

Research problem

Identify high-level *activities* from low-level entities in individuals’ heterogeneous personal information collections (personal webs) in a private, unsupervised, online manner.

Why activity discovery?

• Task detection and reflection
• Entity search and recommendation
• Email prioritization and filtering

Graph-based representations

![Graph-based representations diagram](image)

**INTUITION**

• Model collection of personal information entities (emails, files, contacts, appts, web searches) as a graph
• Learn representations via graph propagation from a set of seeds

**CONTRIBUTIONS**

• Derive exact online updates of representations via outer product
• Update to entity representations
• Update strengths for each entity
• Update values from each attribute

\[
\Delta X = X_{new} - X = uv^T
\]

Linear in max(# of edges, # of new attributes)

Graph smoothness with Laplacian L

\[
\min \mathcal{L}(X) = \|X - X\|^2 + \lambda \mathcal{L}(X L X)
\]

Intrinsic evaluation

Data

• Participants: 10 interns, researchers, managers
• 2-7 days of data from local logging application
• Recent emails, appts, contacts, searches, files
• Around 100 to 1k entities per participant
• Extract noun phrases (NP) and topics (LSA) from text

Privacy-preserving task setup

• Task hosted locally on participants’ machines via USB
• Display pairs of personal information entities [1, 2]
• Participants rate the “activity relatedness” of pairs:
  - **Scope** (low-, mid-, high-level) and **grade** (0-4 points)
  - All feedback anonymized and aggregated

Extrinsic evaluation

Data + task setup

• Avocado dataset: 128 inboxes, 500 to 12k entities per inbox
• Learn models on training emails, predict last recipient on test emails
• Baselines: Email features [3] and node2vec [2]

<table>
<thead>
<tr>
<th>Participant</th>
<th># entities in C</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>Avg. grade (task)</th>
</tr>
</thead>
<tbody>
<tr>
<td>People Overlap</td>
<td>2.96(4)</td>
<td>2.47(3)</td>
<td>2.67(2)</td>
<td>2.87(2)</td>
<td>2.77(1)</td>
<td>2.66(2)</td>
<td>2.06(1)</td>
<td>2.45(3)</td>
<td>2.13(3)</td>
<td>2.22 ± 1.23(2)</td>
</tr>
<tr>
<td>node2vec</td>
<td>2.35(1)</td>
<td>1.86(2)</td>
<td>1.97(3)</td>
<td>2.10(2)</td>
<td>2.02(1)</td>
<td>1.96(2)</td>
<td>1.70(1)</td>
<td>2.13(3)</td>
<td>1.74(4)</td>
<td>1.82 ± 1.31(2)</td>
</tr>
<tr>
<td>Our-LP</td>
<td>2.17(6)</td>
<td>1.97(5)</td>
<td>2.33(1)</td>
<td>2.37(2)</td>
<td>2.17(4)</td>
<td>1.90(2)</td>
<td>2.03(1)</td>
<td>2.75(1)</td>
<td>1.72(4)</td>
<td>1.78 ± 1.31(2)</td>
</tr>
<tr>
<td>Our-LSA</td>
<td>2.17(6)</td>
<td>1.97(5)</td>
<td>2.33(1)</td>
<td>2.37(2)</td>
<td>2.17(4)</td>
<td>1.90(2)</td>
<td>2.03(1)</td>
<td>2.75(1)</td>
<td>1.72(4)</td>
<td>1.78 ± 1.31(2)</td>
</tr>
</tbody>
</table>

Entity representations: Rows of matrix after propagation

Our representations perform best at identifying “low-level” relationships among entities: Short-term tasks and goals

Scalability

Online updates 470x faster than offline

References