BioGraph

Connect, enrich and explore diverse biographical collections

Krishna Janakiraman
Sean Marimpietri

Advisor: Ray Larson
Find all cabinet officers associated with Abraham Lincoln who were involved with slavery.
How is Groucho Marx related to Cold War technology?
Historical archives and their descriptions are host to rich latent social networks.
Social Networks and Archival Context Project

Our primary collection of archival description records
We developed tools to make the SNAC collection more accessible
We created tools to link SNAC with online, crowdsourced resources.

We developed a new visual interface to discover latent social structures in the SNAC collection.
Why link?

Users of the linked collections can make use of the historic social network that SNAC offers.

Archival descriptions are primarily used as finding aids. Hence there is significant information gap in the biographical facts.
Missing facts in SNAC
Why new interface?

Existing faceted interface is not ideal for discovering latent social structures
Resources
Social Networks and Archival Context (SNAC)

Collection of biographical information derived from archive description records

Curated by professional archivists

Collection is encoded using the EAC-CPF standard
Collection derived from various sources: Wikipedia, MusicBrainz and many others.

Information is **crowdsourced**. Schema is editable.

Collection is encoded using Freebase’s quad store format.
Collection originated as lists kept by film enthusiasts.

All user submissions are reviewed. **Schema is fixed.**

Collection is encoded in plaintext.
Number of Person Entities

Freebase
- 1.9M+

SNAC
- 100k +

IMDb
- 200k +
Architecture

- Extract Records
- Match Schemas
- Merge Records
- Visual Interface
For each collection, we created an entity table and an assertion table
Match Schemas
Develop a tool that given a field in the SNAC collection can recommend fields from other collections that might be a match.
The schema matching problem
Our approach:

1. Represent each field as a vector of terms.
2. Terms are derived from the name of the field.
3. Terms are also taken from user-provided input.
4. We use standard IR techniques for matching field vectors of similar fields.
<table>
<thead>
<tr>
<th>Rank</th>
<th>Predicate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>/film/film_regional_release_date/release_date</td>
</tr>
<tr>
<td>2</td>
<td>/fictional_universe/fictional_date_time/other_date_time</td>
</tr>
<tr>
<td>3</td>
<td>/base/lostintime/probable_date/date</td>
</tr>
<tr>
<td>4</td>
<td>/fictional_universe/fictional_character/date_of_birth</td>
</tr>
<tr>
<td>5-7</td>
<td>....</td>
</tr>
<tr>
<td>8</td>
<td>/biology/organism/date_of_birth</td>
</tr>
<tr>
<td>9</td>
<td>/people/person/date_of_birth</td>
</tr>
<tr>
<td>10</td>
<td>/people/place_lived/end_date</td>
</tr>
</tbody>
</table>

Results from a query to match **birthDate** field from SNAC to freebase

Context words: 'type, birth, date, people, person'
Link Records
Develop techniques that given a SNAC entity can **find records to link/merge** in other collections

**Goal**
SNAC

Ritter, Tex 1905-1974
Ritter, Tex 1907-1974
Ritter, Tex
Ritter, Woodward Maurice 1905-1974
Ritter, Woodward Maurice

Freebase

Tex Ritter

IMDb

Tex Ritter (1905–1974)
Our approach

Train binary classifiers over varying names and existence dates

Perturb existing information to generate additional samples within specific error levels
**SNAC**

- **Features**
  - Names
  - Birth and Death dates

- **Names**
  - Shingle Language Model
  - String distance metrics

- **Features**
  - Train decision tree classifiers

- **Link Records**

**Train decision tree classifiers**

Freebase

IMDb
**Name**: Einstein Albert

**Shingle sequence**: ein, ins, nst, ste, tei, ein … , ert

Probability that the sequence (ins, nst, ste) follows ein is very high for the name einstein
Name 1: Einstein Albert

Name 2: Ainshtain Albert

Name 3: Albert Einstein

Shingle Language Model for names
Example Decision Tree For Von Neumann
<table>
<thead>
<tr>
<th>Name</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>TPR (%)</th>
<th>FPR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albert Einstein</td>
<td>78</td>
<td>11</td>
<td>25</td>
<td>145</td>
<td>75.7%</td>
<td>7%</td>
</tr>
<tr>
<td>George W Bush</td>
<td>39</td>
<td>9</td>
<td>6</td>
<td>60</td>
<td>86.6%</td>
<td>13%</td>
</tr>
<tr>
<td>Von Neumann</td>
<td>182</td>
<td>14</td>
<td>27</td>
<td>301</td>
<td>87%</td>
<td>4%</td>
</tr>
</tbody>
</table>

**Corpus Average**

TPR: 72.7%
FPR: 17%
15,300 records, thresh = 0.85

1100 records, thresh = 0.9

How many did we link?
Develop a visual interface for querying and discovering social structures in our collection.

Goal
Exploratory, faceted search interface
Pros

- Great for finding a specific person or persons under specific facets
- Gives a good overview of the distribution of facets in the entire collection

Cons

- Impossible to discover social structures that span the collection
- Requires multiple explore and filter steps, followed by an aggregation step
Alternate interfaces for discovering structures in social networks

Sophisticated query languages:
- hard to learn for the user

Interactive graph exploration tools:
- great for exploration, requires multiple steps to find the desired social structure pattern
Our Approach

Users express queries for discovering social structures as a *query graph*

A query by example interface

Techniques for visualizing the result graph
About

BioGraph
Discover latent social networks in biographical collections

Hint: Click on an entity type to place it on the grid. To draw an edge double-click on a vertex.
Questions?