Inventor, Assignee, and Location Disambiguation through PatentsView

WIPO STANDARDS WORKSHOP
ON NAME STANDARDIZATION
USPTO current practice

• No system-wide approach for operational application

• Disambiguation (entity resolution) for research and data analytics purposes through PatentsView web-tool
“...thanks for making this invaluable tool freely available to the public. As an academic researcher I deeply appreciate and strongly believe that public access to good quality data is a powerful accelerator of scientific and technological progress.”

PatentsView user, Massachusetts Institute of Technology
PatentsView data process

Collection & Extraction of New Data
- XML Parsing
- USPC, WIPO, CPC Classification Parsing
- Withdrawn Patents

Intermediate Processing
- Government Interest Patents: NER Processing & Extraction
- Inventor Gender Processing

Creation of Raw Database (Old Data + New Data)
- Copy Old Database
- Add in New Data
- Comparison/QC of new database & old database

Disambiguation
- Inventor
- Assignee
- Location

Creation of Reporting Database
- Create reporting database from raw database
- Comparison/QC of reporting database & raw database

Move Updated Data to Production
- Create Bulk Download Files
- Update Website
- Update QueryTool/API

uspto
2015 inventor disambiguation workshop

- Improve on original disambiguation algorithm by UC Berkeley based on Jaro-Winkler distance (Li et al., 2015)
- Seven teams from United States, Europe, Australia, and China
- Human-labeled training data from Pierre Azoulay (MIT), Erica Fuchs (Carnegie Mellon), Ivan Png (National University of Singapore), and Manuel Trajtenberg (Tel Aviv University)
- Winner: UMass Amherst with hierarchical co-referencing algorithm
- www.patentsview.org/workshop
Overview of disambiguation process

- Records
  - 15M inventors
  - 5M assignees
  - 24M locations

Canopies
- Canopy A
- Canopy B

Clusters
- Cluster 1
- Cluster 2
- Cluster 1
- Cluster 2
Canopies

• Inventors – four layers of canopies
  – First, last, and middle name exact match
  – Share last name and first name
  – Share last name and first five characters of first name
  – Share last name and first three characters of first name

• Assignees
  – Exact four-character prefix overlap for any word
Inventor canopies

Canopies

- John Smithson
- John Jacob Smith
- John Jacob Smith
- John Jacob Smith
- Kathryn Jones
- Kathryn P. Jones
- Katheryn Jones
- John Jacob Smith
- John Jacob Smith
- Kathryn Jones
- Kathryn P. Jones
- Katheryn Jones

Records

- John Jacob Smith
- John Jacob Smith
- John Smithson
- Kathryn Jones
- Kathryn P. Jones
- Katheryn Jones

- Exact Match
- Share first and last name
- Share last name and first 5 letters of first name
- Share last name and first 3 letters of first name
Assignee canopies

4 character prefix overlap

General Electric
General Electirc
Motorola
Motorola Electronics

Gene
Elec
Moto

Gene canopy
- General Electric
- General Electirc

Elec canopy
- General Electric
- General Electirc

Moto canopy
- Motorola
- Motorola Electronics

Canopies

Records

Motorola Electronics
Motorola
Elec
Gene
Clustering based on similarity measures

- Define rules to produce a numerical score for the similarity of two records within a canopy
- Cluster based on scores
Inventor similarity measures

- First name
- Middle name
- Assignee(s) name(s)
- Co-inventor(s) name(s)
- Lawyer(s) name(s)
- Location (concatenated city, state, country)
- Patent classification (CPC, USPC, IPC)
- Patent title word embedding
Assignee similarity measures

• Name string-based metrics: exact match, acronym match, prefix/suffix match, and Jaccard similarity
• Inventor(s) name(s)
• Name and location pairs
• Patent classification (CPC, USPC, IPC)
Hierarchical clustering algorithm for inventors

1. Beginning with smallest canopies (exact match), compute similarity score for random sample of records within each canopy

2. Join together into a cluster any records with similarity score above threshold determined through experimentation

3. Continue to add records to clusters or form new clusters when joining ‘improves’ the group

4. Move to larger canopies (less precise match) and repeat steps 1-3

5. Assign complete clusters most common name as the canonical name
‘Greedy’ agglomerative clustering for assignees

1. Compute similarity score for each pair of records within canopy
2. Group together the two most similar records to form cluster
3. Repeat comparison between all records and cluster from step 2
4. Add into a cluster any records with similarity score above threshold determined through experimentation
Agglomerative clustering for assignees

**Clusters**

- **Cluster 1**
  - Toyota Motor Engineering & Manufacturing North America
  - Toyota Motor Engineering & Manufacturing
  - Toyota Motor Engineering & Manufacturing N.A. (TEMA)

- **Cluster 2**
  - Toyota Motor Engineering

- **Cluster 3**
  - Toyota Motor Manufacturing

- **Cluster 4**
  - Toyota Foods

**Final Clusters**

- **Cluster 1**
  - Toyota Motor Engineering & Manufacturing North America
  - Toyota Motor Engineering & Manufacturing
  - Toyota Motor Engineering & Manufacturing N.A. (TEMA)

- **Cluster 2**
  - Toyota Motor Engineering

- **Cluster 4**
  - Toyota Foods
Evaluation metrics

• Standard precision, recall, and F1 metrics
  – Number of records correctly and incorrectly classified
  – Number of distinct entity names or ‘name variation’ evaluation

<table>
<thead>
<tr>
<th>Metric</th>
<th>Prec. PV</th>
<th>New PV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prec</td>
<td>0.9675</td>
<td>0.99</td>
</tr>
<tr>
<td>Rec</td>
<td>0.975</td>
<td>0.9975</td>
</tr>
<tr>
<td>F1</td>
<td>0.9825</td>
<td>0.9975</td>
</tr>
</tbody>
</table>
Next steps

- Extension to pre-grant published applications
- Second workshop
- New labeled data
- Operational applications
THANK YOU
ADDITIONAL SLIDES
PatentsView usage

107K API queries per day

38.8M total API queries

204K total files downloaded

170 hits per day to patent visualization & search interface

Source: PatentsView user statistics for 2018
Location canopies

Inventor: Fred Smith

Assignee: IBM

Locations
- Greensville, NC
- Greenville, NC
- Iowa City, IA

Canopies
- Greensville, NC
- Greenville, NC
- Iowa City, IA
- New York, NY
- New York City, NY
Location similarity measures

• Exact name match
• Non-existent location match
• Relaxed name match
• City name Jaro-Winkler similarity
• Inventor or assignee
• Disproportionate records match