FAst Multi-source Entity Resolution System
Alieh Saeedi, Eric Peukert, Erhard Rahm

www.scads.de
- Physical data integration
  - Knowledge graph: Store data from multiple sources in a graph-like structure
Automatic construction & maintenance of KG: **data quality**

Challenges for data quality

- **Entity Resolution**: The task of identifying and linking entities that refer to the same real-world entity

2 sources:
Binary linking

N sources:
Clustering
- **FAMER** (FAst Multi-source Entity Resolution system)
  - Scalable ER approaches for big data
    - Multiple data sources
    - Large volumes of data
  - Built on top of the distributed data flow framework Apache Flink and Gradoop
    - High scalability
    - Large amounts of data
    - Many machines
FAMER OVERVIEW

Input

Source A
Source B
Source C
Source D
Source E

Linking: Similarity Graph

Clustering

FAMER
<table>
<thead>
<tr>
<th>Id</th>
<th>Name</th>
<th>Surname</th>
<th>Suburb</th>
<th>Post code</th>
<th>Source Id</th>
</tr>
</thead>
<tbody>
<tr>
<td>a₀</td>
<td>George</td>
<td>Walker</td>
<td>winston salem</td>
<td>27106</td>
<td>Src A</td>
</tr>
<tr>
<td>b₀</td>
<td>George</td>
<td>Alker</td>
<td>winstom salem</td>
<td>27106</td>
<td>Src B</td>
</tr>
<tr>
<td>c₀</td>
<td>George</td>
<td>Walker</td>
<td>Winstons</td>
<td>27106</td>
<td>Src C</td>
</tr>
<tr>
<td>d₀</td>
<td>Geoahge</td>
<td>Waker</td>
<td>Winston</td>
<td>27100</td>
<td>Src D</td>
</tr>
<tr>
<td>a₁</td>
<td>Bernie</td>
<td>Davis</td>
<td>pink hill</td>
<td>28572</td>
<td>Src A</td>
</tr>
<tr>
<td>b₁</td>
<td>Bernie</td>
<td>Davis</td>
<td>Pinkeba</td>
<td>27872</td>
<td>Src B</td>
</tr>
<tr>
<td>c₁</td>
<td>Bernii</td>
<td>Davis</td>
<td>pink hill</td>
<td>28571</td>
<td>Src C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Id</th>
<th>Name</th>
<th>Surname</th>
<th>Suburb</th>
<th>Post code</th>
<th>Source Id</th>
</tr>
</thead>
<tbody>
<tr>
<td>a₂</td>
<td>Bertha</td>
<td>Summercille</td>
<td>Charlotte</td>
<td>28282</td>
<td>Src A</td>
</tr>
<tr>
<td>b₂</td>
<td>Bertha</td>
<td>Summeahville</td>
<td>Charlotte</td>
<td>2822</td>
<td>Src B</td>
</tr>
<tr>
<td>d₂</td>
<td>Brtha</td>
<td>Summersville</td>
<td>Charlotte</td>
<td>28222</td>
<td>Src D</td>
</tr>
<tr>
<td>c₃</td>
<td>Bereni</td>
<td>dan’lel</td>
<td>Pinkeba</td>
<td>27840</td>
<td>Src C</td>
</tr>
<tr>
<td>d₃</td>
<td>Bereni</td>
<td>Dasniel</td>
<td>Pinkeba</td>
<td>27880</td>
<td>Src D</td>
</tr>
</tbody>
</table>
FAMER OVERVIEW

Clustering
- ER Clustering
  - Connected Components
  - Correlation Clustering
- Center
- Merge Center
- Star-1
- Star-2

RLIP
- Overlap Resolution
- CLIP

CLIP
(Clustering based on Link Priority)

Sources
- Source 1
- Source 2
- Source k
FAMER CLUSTERING

Similarity graph

Source-inconsistency
overlap

ER Clustering
- Connected Components
- Correlation Clustering
- Center
- Merge Center
- Star-1
- Star-2
Prioritize links based on
- Link strength
  - Strong, Normal, Weak
- Link degree
- Similarity value

produces
- Source-consistent clusters
- No overlap
– Link Strength
  • Strong
  • Normal
  • Weak
- Geographical domain
- 4 sources
- F-Measure

CLIP QUALITY

Similarity Graph
RLIP QUALITY

- Average F-Measure
- 5 sources
- 5,000,000 entities
- Flink cluster of 16 workers
- CLIP
  - 69s (1 min) for 16 workers
  - 10 sources/
    10 million entities: 228 s

<table>
<thead>
<tr>
<th>clustering</th>
<th>#workers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td>ConCom</td>
<td>51</td>
</tr>
<tr>
<td>CCPiv</td>
<td>1530</td>
</tr>
<tr>
<td>Center</td>
<td>390</td>
</tr>
<tr>
<td>Merge Center</td>
<td>640</td>
</tr>
<tr>
<td>Star-1</td>
<td>288</td>
</tr>
<tr>
<td>Star-2</td>
<td>214</td>
</tr>
<tr>
<td>CLIP</td>
<td>190</td>
</tr>
</tbody>
</table>
Hackathon: June 30th- July 1st 2018
Summer School: July 2nd- July 6th 2018
Location: University of Leipzig
Bonus: LSWT attendants obtain early registration rates

www.scads.de/en/summerschool-2018
SUMMER SCHOOL 2018

- Topics & Highlights
  - Deep Learning
  - Data Mining in Stream Data
  - Machine Learning for the Sciences
  - Apache SystemML
  - Secure Cloud Databases
  - Visual Analytics for Big Data
  - Graph Analytics on Spark
  - Web-scale information extraction
  - Big Data Integration
  - Big Data and HPC

www.scads.de/en/summerschool-2018