Automating Data Integration with Machine Learning

Bringing Your Data Together

Natalia Rümmele | Data Scientist
September 2016

www.data61.csiro.au
Data Integration Problem

• Combine data from different sources
• Provide unified view of data
Data Integration Problem

• Combine data from different sources
• Provide unified view of data
  • Multiple datasets with common entities and content
  • Siloed systems
  • Sharing data across systems/schemas
  • Handling legacy systems
  • Handling different schemas
Data Integration Problem

• Combine data from different sources
• Provide unified view of data
  • Multiple datasets with common entities and content
  • Siloed systems
  • Sharing data across systems/schemas
  • Handling legacy systems
  • Handling different schemas
• Resource-intensive ETL...
Data Integration Problem

• Combine data from different sources
• Provide unified view of data
  • Multiple datasets with common entities and content
  • Siloed systems
  • Sharing data across systems/schemas
  • Handling legacy systems
  • Handling different schemas
• Resource-intensive ETL...
• Can machine learning help?
Areas of Investigation

• Schema matching & mapping
  • Connecting datasets by establishing a global data model

• Entity Resolution
  • Detecting entities across databases

• Privacy Preserving Analytics
  • Learning the global data model without sharing data

• Data Quality
  • Semantic and Syntactic scoring
Relational Schema Matching
**Schema Matching**

- **Goal:** Automatically connect datasets across relational schemas

<table>
<thead>
<tr>
<th>UserName</th>
<th>Names</th>
<th>People</th>
<th><em>USERS</em></th>
<th>IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe Blogs</td>
<td>Blogs, Joe</td>
<td>Blogs, J</td>
<td>Blogs_Joe</td>
<td>Joe Blogs</td>
</tr>
<tr>
<td>Jill Blogs</td>
<td>Blogs, Jill</td>
<td>Blogs, J</td>
<td>Blogs_Jill</td>
<td>Jill Blogs</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Schema Matching

• Goal: Automatically connect datasets across relational schemas

• Problem: Syntax vs Semantics
  • Requires humans to distinguish between similar columns
  • Machines get caught on syntax

 userName
 Joe Blogs
 Jill Blogs
 ...

 names
 Blogs, Joe
 Blogs, Jill
 ...

 people
 Blogs, J
 Blogs, J
 ...

 _users_
 Blogs_Joe
 Blogs_Jill
 ...

 ids
 Joe Blogs
 Jill Blogs
 ...

 Semantic type
Data Integration: Schema Matcher

• Can we automatically label columns with semantic types?

• Given multiple datasets and a set of semantic types, can we detect all columns of the same semantic type where:
  • Column names may be different
  • Common entries may not exist
  • Formatting issues may exist

<table>
<thead>
<tr>
<th>Column</th>
<th>Data</th>
</tr>
</thead>
</table>
| UserName | Joe Blogs, Jill Blogs, ...
| Names | Blogs, Joe Blogs, Jill Blogs, ...
| People | Blogs, J Blogs, J Blogs, Jill Blogs, Jill Blogs, ...
| _USERS_ | Blogs_Joe, Blogs_Jill, ...
| IDs | Joe Blogs, Jill Blogs, ...

Schema Matcher

Name  | Address  | Phone
--- | --- | ---
Name  | Name  | Name
UserName | Joe Blogs | Jill Blogs | ...
Names | Blogs, Joe | Blogs, Jill | ...
People | Blogs, J | Blogs, J | Blogs, J | Blogs, Jill | ...
_USERS_ | Blogs_Joe | Blogs_Jill | ...
IDs | Joe Blogs | Jill Blogs | ...

Diagram showing data integration with schema matching for Name, Address, and Phone columns.
Machine Learning: Feature Vector

- Multi-class classification problem
- Represent column as a vector of features
- Classify column as one of semantic types using a ML classifier
- Training data needed!

**Column Header + Table Name**
- Nearest neighbours to class training labels
- Edit distance metrics
- WordNet distance

**Content**
- Character Frequencies
- Missing Values
- Number repeated values
- Entropy
### Machine Learning: Cost Matrix

- Class imbalance: unknown class over-represented (unlabeled columns)
- Class resampling strategies: oversampling/undersampling to mean, etc.
- Cost-sensitive learning by introducing asymmetric costs of misclassifications

<table>
<thead>
<tr>
<th>Semantic Type</th>
<th>Predicted</th>
<th>Unknown</th>
<th>Name</th>
<th>Address</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Name</td>
<td>40</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Address</td>
<td>40</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Phone</td>
<td>40</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Machine Learning: Bagging

• Number of columns usually much smaller than dataset
• Can use bagging to build many smaller samples by subsampling
• Also addresses class imbalance
Data Integration: Schema Matcher

- Schema Matcher is trained on examples and user feedback
- System improves as predictions are corrected
Data Integration: Schema Matcher

• User supplies semantic types and datasets
• User labels some example sets
Data Integration: Schema Matcher

- User supplies semantic types and datasets
- User labels some example sets
- Predictions are generated
Data Integration: Schema Matcher

- User supplies types and datasets
- User labels some example sets
- Predictions are generated, user corrects
Data Integration: Schema Matcher

- User supplies types and datasets
- User labels some example sets
- Predictions are generated, user corrects
- User adds more data
**Data Integration: Schema Matcher**

- User supplies types and datasets
- User labels some example sets
- Predictions are generated, user corrects
- User adds more data
- Repeat
Applications

• Linking and connecting data
• Class-wide transforms
• Relabelling columns to unified naming convention
• Labelling no-header datasets
• Merging tables by semantic type
• A component for further semantic modelling
Semantic Modelling
Implicit Semantics

- The semantic meaning of a dataset is more than a column label

<table>
<thead>
<tr>
<th>Name</th>
<th>BirthDate</th>
<th>City</th>
<th>State</th>
<th>Workplace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe Blogs</td>
<td>21-05-1986</td>
<td>Perth</td>
<td>WA</td>
<td>Data61</td>
</tr>
<tr>
<td>Jill Blogs</td>
<td>97-12-1990</td>
<td>Adelaide</td>
<td>SA</td>
<td>CSIRO</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Making Implicit Explicit

• The semantic meaning of a dataset is more than a column label

• There are usually relationships implied between the columns
Semantic Model

- The semantic meaning of a dataset is more than a column label.
- There are usually relationships implied between the columns.

```
<table>
<thead>
<tr>
<th>UserName</th>
<th>Location</th>
<th>Contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe Blogs</td>
<td>Melbourne</td>
<td>99110002</td>
</tr>
<tr>
<td>Jill Blogs</td>
<td>Perth</td>
<td>45878723</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
```
Semantic Modelling

• For this we need an ontology or graph schema

• Can come from:
  • Built up iteratively from definitions
  • Pre-defined domain ontologies
  • Downloaded ontologies from Semantic Web

Ontology

Class Node: Abstract Concept
Ontology

Class Node: Abstract Concept

Data Node: Properties
Ontology

Class Node: Abstract Concept

Data Node: Properties

Relationship: Relationship between Concepts
RDB2RDF Schema Matching

• Given an ontology and a set of known semantic models, can we generate a semantic model for a new dataset?
Constructing Semantic Model

• Map all possible semantic types matched for columns onto the Ontology
• Need to find most likely semantic model – subgraph which covers matched semantic types
• Minimum Cost Steiner Tree Problem (approximate)
Advantages

Not just better understanding of your data....

- Better entity resolution
- Easier integration to graph databases or merging between relational
- Enables more sophisticated and accurate merging
- More powerful search

[Diagram showing relationships between Person, City, Phone with attributes like Name, Location, Contact, _USERS_, Number]
Advanced Search

- The ontology allows transitive and subclass relationships
- Searches can associate new columns e.g. City -> State -> Country
- A search for something in a country can also proceed to the subclasses
Graph Database

- The ontology can act as the intermediary between graph databases and relational databases
- Functions as a graph schema and global (unified) schema of integrated datasets
Summary
Data Integration

- **Schema Matcher**
  - Relational schema
  - Find semantically similar columns across data sets

- **Semantic Modelling**
  - Graph schema
  - Find semantically similar columns + relationships between them
Open Questions

• Can we learn column transformations?
• Complex column matches
  • One-to-many matches
  • Many-to-one matches
• Applications in entity resolution
• Applications in search
Thank you

Data Platforms Team
Engineering and Design

e  natalia.ruemmele@data61.csiro.au