Aligning and Integrating Data in Karma

Craig Knoblock
University of Southern California
Data Integration Approaches
Data Integration Approaches

Data Warehousing
Data Integration Approaches

Data Warehousing

Virtual Integration
Domain Model

Describes the domain you have, e.g., people, events and their attributes
Key Ingredient: Source Mappings

Describe how data in sources relate to the domain model
Karma: A Data Integration Tool
Karma

Interactive tool for rapidly extracting, cleaning, transforming, integrating and publishing data

Tabular Sources
Hierarchical Sources
Services

Karma

RDF
Database
CSV
...

http://www.isi.edu/integration/karma
@KarmaSemWeb
Information Integration in Karma

Domain Model

Samples of Source Data

Karma

Karma semi-automatically generates Source Mappings

Source Mappings
Information Integration in Karma

Karma semi-automatically generates Source Mappings

Karma supports multiple integration regimes

Domain Model

Samples of Source Data
Secret Sauce: Karma Understands Your Data

Karma semi-automatically builds a semantic model of your data

Source Mappings

Semantic Model of the Data

Domain Model

Samples of Source Data
What is a Semantic Model?

Describe sources using classes & relationships in an ontology

<table>
<thead>
<tr>
<th>Source</th>
<th>name</th>
<th>date</th>
<th>city</th>
<th>state</th>
<th>workplace</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fred Collins</td>
<td>Oct 1959</td>
<td>Seattle</td>
<td>WA</td>
<td>Microsoft</td>
</tr>
<tr>
<td>2</td>
<td>Tina Peterson</td>
<td>May 1980</td>
<td>New York</td>
<td>NY</td>
<td>Google</td>
</tr>
</tbody>
</table>
### Semantic Types

<table>
<thead>
<tr>
<th>name</th>
<th>date</th>
<th>city</th>
<th>state</th>
<th>workplace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fred Collins</td>
<td>Oct 1959</td>
<td>Seattle</td>
<td>WA</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Tina Peterson</td>
<td>May 1980</td>
<td>New York</td>
<td>NY</td>
<td>Google</td>
</tr>
</tbody>
</table>
Relationships

```
<table>
<thead>
<tr>
<th>name</th>
<th>date</th>
<th>city</th>
<th>state</th>
<th>workplace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fred Collins</td>
<td>Oct 1959</td>
<td>Seattle</td>
<td>WA</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Tina Peterson</td>
<td>May 1980</td>
<td>New York</td>
<td>NY</td>
<td>Google</td>
</tr>
</tbody>
</table>
```
Semantic Model

Semantic models will be formalized as Source Mappings

Key ingredient to automate source discovery, data integration, and publishing semantic data (RDF triples)
so what?
Karma uses **semantic models** to create knowledge graphs.
Karma semi-automatically builds semantic models

Karma uses semantic models to create knowledge graphs
Karma semi-automatically builds semantic models... and provides a nice GUI to edit them.

Karma uses semantic models to create knowledge graphs.
Semi-automatically Building Semantic Models in Karma
Approach

[Knoblock et al, ESWC 2012]
Example

Source

<table>
<thead>
<tr>
<th>name</th>
<th>date</th>
<th>city</th>
<th>state</th>
<th>workplace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fred Collins</td>
<td>Oct 1959</td>
<td>Seattle</td>
<td>WA</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Tina Peterson</td>
<td>May 1980</td>
<td>New York</td>
<td>NY</td>
<td>Google</td>
</tr>
</tbody>
</table>

Domain Ontology

Find a semantic model for the source (map the source to the ontology)
Learning Semantic Types
[Krishnamurthy et al., ESWC 2015]

- class?
- property?

<table>
<thead>
<tr>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>H: 3.5 in, W: 2.5 in</td>
</tr>
<tr>
<td>H: 64 in, W: 51.5 in, D: .75 in</td>
</tr>
<tr>
<td>L: 57 in, center back: 23 in</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>
Learning Semantic Types

CulturalHeritageObject

extent

Dimensions

<table>
<thead>
<tr>
<th>H: 3.5 in, W: 2.5 in</th>
</tr>
</thead>
<tbody>
<tr>
<td>H: 64 in, W: 51.5 in, D: .75 in</td>
</tr>
<tr>
<td>L: 57 in, center back: 23 in</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

1- User specifies

2- System learns
Learning Semantic Types

**CulturalHeritageObject**

- **extent**

### Dimensions
- H: 3.5 in, W: 2.5 in
- H: 64 in, W: 51.5 in, D: .75 in
- L: 57 in, center back: 23 in

### Extent
- 52.1 x 71.4 cm (20 1/2 x 28 1/8 in.)
- 9 3/4 x 7 9/16 in.
- H: 19 x W: 15 1/4 x D: 8 1/4 in.
Learning Semantic Types

CulturalHeritageObject

extent

Dimensions
H: 3.5 in, W: 2.5 in
H: 64 in, W: 51.5 in,
D: .75 in
L: 57 in, center
back: 23 in
...

CulturalHeritageObject

extent

Extent
52.1 x 71.4 cm (20
1/2 x 28 1/8 in.)
9 3/4 x 7 9/16 in.
H: 19 x W: 15 1/4 x
D: 8 1/4 in.
...

...
Requirements

• Learn from a small number of examples
• Work on both textual and numeric values
• Learn quickly and highly scalable to large number of semantic types
Approach for Textual Data

• **Document**: each column of data
• **Label**: each semantic type
• Use **Apache Lucene** to index the labeled documents
• Compute **TF/IDF vectors** for documents
• Compare documents using **Cosine Similarity** between TF/IDF vectors
Approach for Textual Data

**Dimensions**
- H: 3.5 in, W: 2.5 in
- H: 64 in, W: 51.5 in, D: .75 in
- L: 57 in, center back: 23 in

**Extent**
- 52.1 x 71.4 cm (20 1/2 x 28 1/8 in.)
- 9 3/4 x 7 9/16 in.
- H: 19 x W: 15 1/4 x D: 8 1/4 in.

**Term: TF-IDF score**
- h: 0.375
- w: 0.336
- in: 0.491
- centre: 0.241
- back: 0.301

**Term: TF-IDF score**
- h: 0.414
- w: 0.364
- d: 0.245
- cm: 0.354
- in: 0.395

**TF-IDF Cosine Similarity**

\[ tf(t, d) = \text{frequency}^{1/2} \]

\[ idf(t) = 1 + \log\left(\frac{\text{numDocs}}{\text{docFreq} + 1}\right) \]

\[ sim(q, d) = \frac{V(q) \cdot V(d)}{|V(q)| \cdot |V(d)|} \]
Approach for Numeric Data

• **Distribution** of values in different semantic types is different, e.g., temperature vs. population

• Use **Statistical Hypothesis Testing** to see which distribution fits best

• **Welch’s T-test**, **Mann-Whitney U-test** and **Kolmogorov-Smirnov Test**
Approach for Numeric Data
Similarity features

- Attribute names similarity
  - Jaccard
  - TF-IDF

- Value Similarity
  - Jaccard

- Distribution Similarity
  - Mann-Whitney test
  - Kolmogorov-Smirnov test

- Histogram Similarity
  - Mann-Whitney test
Training machine learning model
[Pham et al., ISWC 2016]
Predicting new attribute

Player

birthName

height

? -> player-birthName

Classifier

P_{True} = 0.8

P_{True} = 0.1

<table>
<thead>
<tr>
<th>player</th>
<th>height</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hazard, Eden</td>
<td>172</td>
</tr>
<tr>
<td>Cahill, Gary</td>
<td>191</td>
</tr>
<tr>
<td>Rooney, Wayne</td>
<td>176</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>player</th>
<th>P_{True}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wayne Smith</td>
<td>0.4</td>
</tr>
<tr>
<td>Tim Cahill</td>
<td>0.5</td>
</tr>
<tr>
<td>Juan Mata</td>
<td>0.05</td>
</tr>
</tbody>
</table>

jaccard | cosine
---|---
0.4 | 0.5

jaccard | cosine
---|---
0.05 | 0.1
Construct a Graph

Construct a graph from semantic types and ontology

<table>
<thead>
<tr>
<th>name</th>
<th>date</th>
<th>city</th>
<th>state</th>
<th>workplace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fred Collins</td>
<td>Oct 1959</td>
<td>Seattle</td>
<td>WA</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Tina Peterson</td>
<td>May 1980</td>
<td>New York</td>
<td>NY</td>
<td>Google</td>
</tr>
</tbody>
</table>
Construct a Graph

Construct a graph from semantic types and ontology
Inferring the Relationships

• Search for minimal explanation

• Steiner tree connecting semantic types over ontology graph
  • Given graph $G=(V,E)$, nodes $S \subseteq V$, cost $c: E \rightarrow \mathbb{R}$
  • Find a tree of $G$ that spans $S$ with minimal total cost
  • Unfortunately, NP-complete

• Approximation Algorithm [Kou et al., 1981]
  • Worst-case time complexity: $O(|V|^2|S|)$
  • Approximation Ratio: less than 2
Inferring the Relationships
Select minimal tree that connects all semantic types

• A customized Steiner tree algorithm
Result in Karma

<table>
<thead>
<tr>
<th>name</th>
<th>birthDate</th>
<th>city</th>
<th>state</th>
<th>workplace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fred Collins</td>
<td>Oct 1959</td>
<td>Seattle</td>
<td>WA</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Tina Peterson</td>
<td>May 1980</td>
<td>New York</td>
<td>NY</td>
<td>Google</td>
</tr>
<tr>
<td>Richard Smith</td>
<td>Feb 1975</td>
<td>Los Angeles</td>
<td>CA</td>
<td>Apple</td>
</tr>
</tbody>
</table>
Refining the Model

Impose constraints on Steiner Tree Algorithm

– Change weight of selected links to $\varepsilon$
– Add source and target of selected link to Steiner nodes

Correct Model
Karma Learns the Source Models
Taheriyani et al., ISWC 2013, ICSC 2014

Sample Data
Domain Ontology
Known Semantic Models
Learn Semantic Types
Construct a Graph
Generate Candidate Models
Rank Results
Karma Use Cases
Source Mapping Phase

Mapping Phase
Source Mapping and Query Time

Mapping Phase

Query Phase

Domain Model

Samples of Source Data

Karma

Domain Expert

Source Mappings

Karma Runtime System

Data Warehousing

Virtual Integration

Query

Analyst

Pedro Szekely and Craig Knoblock

University of Southern California
VIVO

- **VIVO** is a system to build researcher networks across institutions

- Used Karma to map the data about USC faculty to VIVO ontology and publish it as RDF

- VIVO ingest the RDF data

- **Video**
Smithsonian American Art Museum

- Used Karma to convert data of 44000 museum objects to Linked Open Data
- Modeled according to Europeana Data Model (EDM)
- Linked the generated RDF to DBpedia, ULAN, NY Times Linked Data
- News: USC press, Viterbi
- Video
• **DIG**: Domain-specific Insight Graphs

• Building and using knowledge graphs to combat human trafficking

• Used Karma to map extracted data and structured sources to shared domain ontology

• News: [Forbes](https://www.forbes.com), [Wired.co.uk](https://www.wired.co.uk)
Demo
Using Karma to map museum data to the CIDOC CRM ontology

https://www.youtube.com/watch?v=h3_yiBhAJIc
Discussion

• Automatically build rich semantic descriptions of data sources

• Exploit the background knowledge from (i) the domain ontology, and (ii) the known source models

• Semantic descriptions are the key ingredients to automate many tasks, e.g.,
  • Source Discovery
  • Data Integration
  • Service Composition