Learning High Accuracy Rules for Object Identification

Sheila Tejada

Wednesday, December 12, 2001

Committee Chair: Craig A. Knoblock
Committee: Dr. George Bekey, Dr. Kevin Knight,
Dr. Steven Minton, Dr. Daniel O'Leary
Integrating Restaurant Sources

Zagat’s Restaurant Guide Source

Department of Health Restaurant Rating Source

ARIADNE Information Mediator

Question: What is the Review and Rating for the Restaurant “Art’s Deli”?
Extract web objects in the form of database records

### Zagat’s

<table>
<thead>
<tr>
<th>Name</th>
<th>Street</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art’s Deli</td>
<td>12224 Ventura Boulevard</td>
<td>818-756-4124</td>
</tr>
<tr>
<td>Teresa’s</td>
<td>80 Montague St.</td>
<td>718-520-2910</td>
</tr>
<tr>
<td>Steakhouse The</td>
<td>128 Fremont St.</td>
<td>702-382-1600</td>
</tr>
<tr>
<td>Les Celebrites</td>
<td>155 W. 58th St.</td>
<td>212-484-5113</td>
</tr>
</tbody>
</table>

### Dept of Health

<table>
<thead>
<tr>
<th>Name</th>
<th>Street</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art’s Delicatessen</td>
<td>12224 Ventura Blvd.</td>
<td>818/755-4100</td>
</tr>
<tr>
<td>Teresa’s</td>
<td>103 1st Ave. between 6th and 7th Sts.</td>
<td>212/228-0604</td>
</tr>
<tr>
<td>Binion’s Coffee Shop</td>
<td>128 Fremont St.</td>
<td>702/382-1600</td>
</tr>
<tr>
<td>Les Celebrites</td>
<td>5432 Sunset Blvd</td>
<td>212/484-5113</td>
</tr>
</tbody>
</table>
Multi-Source Inconsistency

Zagat’s Restaurant Guide Source

- Art’s Deli
- California Pizza Kitchen
- Campanile
- Citrus
- Grill, The
- Philippe The Original
- Spago

Department of Health Restaurant Source

- Art’s Delicatessen
- Ca’ Brea
- CPK
- The Grill
- Patina
- Philippe’s The Original
- The Tillerman

How can the same objects be identified when they are stored in inconsistent text formats?
Observations:

- Mapping objects can be application dependent
- Example:
  - The mapping is in the application, not the data
  - User input is needed to increase accuracy of the mapping

<table>
<thead>
<tr>
<th>Mapped?</th>
<th>Steakhouse The</th>
<th>128 Fremont Street</th>
<th>702-382-1600</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Binion's Coffee Shop</td>
<td>128 Fremont St.</td>
<td>702/382-1600</td>
</tr>
</tbody>
</table>

Binion's Coffee Shop
128 Fremont St.
702/382-1600
Key Ideas for Mapping Objects

- Learning important attributes for determining a mapping

<table>
<thead>
<tr>
<th>Name</th>
<th>Street</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zagat’s</td>
<td>Art’s Deli, 12224 Ventura Blvd</td>
<td>818-756-4124</td>
</tr>
<tr>
<td>Dept of Health</td>
<td>Art’s Delicatessen, 12224 Ventura Blvd</td>
<td>818/755-4100</td>
</tr>
</tbody>
</table>

- Learning general transformations to recognize objects

<table>
<thead>
<tr>
<th>Zagat’s</th>
<th>Transformations</th>
<th>Dept of Health</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art’s Deli</td>
<td>Prefix</td>
<td>Art’s Delicatessen</td>
</tr>
<tr>
<td>California Pizza</td>
<td>Acronym</td>
<td>CPK</td>
</tr>
<tr>
<td>Kitchen</td>
<td>Stemming</td>
<td>Philippe’s The Original</td>
</tr>
<tr>
<td>Philippe The</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Mapping Rules

## Zagat’s Restaurants

<table>
<thead>
<tr>
<th>Name</th>
<th>Street</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art’s Deli</td>
<td>12224 Ventura Boulevard</td>
<td>818-756-4124</td>
</tr>
<tr>
<td>Teresa’s</td>
<td>80 Montague St.</td>
<td>718-520-2910</td>
</tr>
<tr>
<td>Steakhouse The</td>
<td>128 Fremont St.</td>
<td>702-382-1600</td>
</tr>
<tr>
<td>Les Celebrites</td>
<td>155 W. 58th St.</td>
<td>212-484-5113</td>
</tr>
</tbody>
</table>

## Dept. of Health

<table>
<thead>
<tr>
<th>Name</th>
<th>Street</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art’s Delicatessen</td>
<td>12224 Ventura Blvd.</td>
<td>818/755-4100</td>
</tr>
<tr>
<td>Teresa’s</td>
<td>103 1st Ave. between 6th and 7th Sts.</td>
<td>212/228-0604</td>
</tr>
<tr>
<td>Binion’s Coffee Shop</td>
<td>128 Fremont St.</td>
<td>702/382-1600</td>
</tr>
<tr>
<td>Les Celebrites</td>
<td>160 Central Park S</td>
<td>212/484-5113</td>
</tr>
</tbody>
</table>

**Mapping rules:**

- Name > .9 & Street > .87 => mapped
- Name > .95 & Phone > .96 => mapped
Transformation Weights

- Transformations can be more appropriate for a specific application domain
  - Restaurants, Companies or Airports

- Or for different attributes within an application domain
  - Acronym more appropriate for the attribute Restaurant Name than for the Phone attribute

- Learn likelihood that if transformation is applied then the objects are mapped

Transformation Weight = P(mapped | transformation)
Thesis Statement

By simultaneously learning to tailor mapping rules and transformation weights to a specific domain, an object identification system can achieve high accuracy without sacrificing domain independence.
Contributions

• Approach to learning mapping rules that achieve high accuracy mapping while minimizing user involvement

• Only approach developed to tailor a general set of transformations to a specific domain application

• Novel method to combine both forms of learning to create a robust object identification system
Overview

• Approach
  – Computing textual similarity
  – Learning important attributes for mapping
    • Mapping rule learning
  – Learning transformation weights

• Experimental Results

• Related Work on Object Identification

• Conclusions & Future Work
Learning Object Mappings

- **Candidate Generator:**
  - Judge textual similarity of mappings
  - Reduce number of mappings considered for classification
- **Mapping Learner:**
  - Active learning technique to learn mapping rules and transformation weights
  - Minimize the amount of user interaction

---

![Diagram of the Active Atlas process]

- **Active Atlas**
- **Candidate Generator**
- **Mapping Learner**
- **User Input**
- **Set of Mapped Objects**

---

- **Source 1**
- **Source 2**
Computing Textual Similarity

Zagat’s Restaurant
Objects

Department of Health
Objects

Name | Street | Phone
---|---|---
Z1, | | |
Z2, | | |
Z3 | | |

* Candidate Generator returns sets of similarity scores

Name | Street | Phone
---|---|---
| | |
| | |
| | |
| | |

S_{name} | S_{street} | S_{phone}
---|---|---
.9 | .79 | .4
.17 | .3 | .74
...
Types of Transformations

Type I Transformations
- Equality (Exact match)
- Stemming
- Soundex (e.g. “Celebrities” => “C453”)
- Abbreviation (e.g. “3rd” => “third”)

Type II Transformations
- Initial
- Prefix (e.g. “Deli” & “Delicatessen”)
- Suffix
- Substring
- Acronym (e.g. “California Pizza Kitchen” & “CPK”)
- Drop Word
Applying Type I Transformations

- Employs Information Retrieval Techniques
- One set of attribute values broken into words or tokens
  - “Art” “s” “Delicatessen”
- Apply Type I transformations to tokens
  - “Art” “A630” “s” “S000” “Delicatessen” “D423”
- Enter tokens into inverted index
- Tokens from second set used to query the index
  - Transformed query set: “Art” “A630” “s” “S000” “Deli” “Del” “D400”
Applying Type II Transformations

- Type II transformations improve measurement of similarity
Attribute Similarity Function

- Transformations determine similarity of attribute values
- Each attribute value is represented as a vector

\[ \langle 2 \ 4 \ 3 \ 0 \ 5 \ 6 \ 6 \ 0 \ 0 \ 0 \ 0 \ 0 \ 5 \ 0 \ 0 \ 0 \ 0 \ ... \rangle \]

- Attribute Similarity Function:
  - Cosine Measure with a TFIDF

\[
\text{Similarity} (A, B) = \frac{\sum_{i=1}^{t} (w_{ia} \times w_{ij})}{\sqrt{\sum_{i=1}^{t} (w_{ia})^2 \times \sum_{i=1}^{t} (w_{ij})^2}}
\]

\[
w_{ia} = (0.5 + 0.5 \text{freq}_{ia}) \times \text{IDF}_i
\]

\[
w_{ij} = \text{freq}_{ij} \times \text{IDF}_i
\]

\[\text{freq}_{ia} = \text{frequency of term } i \text{ for attribute value } a\]

\[\text{IDF}_i = \text{IDF} \text{ of term } i \text{ in the entire collection}\]

\[\text{freq}_{ij} = \text{frequency of term } i \text{ in attribute value } j\]
Total Object Similarity Scores

<table>
<thead>
<tr>
<th>Name</th>
<th>Street</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zagat’s</td>
<td>Art’s Deli 12224 Ventura Boulevard</td>
<td>818-756-4124</td>
</tr>
<tr>
<td>Dept of Health</td>
<td>Art’s Delicatessen 12224 Ventura Blvd.</td>
<td>818/755-4100</td>
</tr>
</tbody>
</table>

Candidate Mapping Similarity Scores:

<table>
<thead>
<tr>
<th>Name</th>
<th>Street</th>
<th>Phone</th>
<th>Total Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>.967</td>
<td>.973</td>
<td>.3</td>
<td>2.034</td>
</tr>
<tr>
<td>.17</td>
<td>.3</td>
<td>.74</td>
<td>1.182</td>
</tr>
<tr>
<td>.8</td>
<td>.5</td>
<td>.49</td>
<td>1.749</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Learning Object Mappings

Active Atlas

Source 1
Candidate Generator
Mapping Learner
Set of Mapped Objects
Source 2
User Input
Learning Object Mappings

- The goal is to classify with high accuracy the proposed mappings while minimizing user input
  - Active learning technique
  - System chooses most informative example for the user to label
## Mapping Rules

<table>
<thead>
<tr>
<th>Set of Similarity Scores</th>
<th>Mapping Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Street</td>
</tr>
<tr>
<td>.967</td>
<td>.973</td>
</tr>
<tr>
<td>.17</td>
<td>.3</td>
</tr>
<tr>
<td>.8</td>
<td>.542</td>
</tr>
<tr>
<td>.95</td>
<td>.97</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Mapping Rule Learner

Choose initial examples

Generate committee of learners

Learn Rules
Classify Examples
Votes

Learn Rules
Classify Examples
Votes

Learn Rules
Classify Examples
Votes

Choose Example

Set of Mapped Objects

USER

Label
Committee Disagreement

- Chooses an example based on the disagreement of the query committee

<table>
<thead>
<tr>
<th>Examples</th>
<th>Committee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art’s Deli, Art’s Delicatessen</td>
<td>Yes</td>
</tr>
<tr>
<td>CPK, California Pizza Kitchen</td>
<td>Yes</td>
</tr>
<tr>
<td>Ca’Brea, La Brea Bakery</td>
<td>No</td>
</tr>
</tbody>
</table>

- In this case CPK, California Pizza Kitchen is the most informative example based on disagreement
Choosing Next Example

- The user labels the example, and the system updates the committee
- Mapping Rule Learner outputs classified examples
Set of Mappings between the Objects

((A_3, B_2) mapped)
((A_{45}, B_{12}) not mapped)
((A_5, B_2) mapped)
((A_{98}, B_{23}) mapped)

(Object pairs, Similarity Scores, Total Score, Transformations)

((A_3, B_2, (s_1, s_2, s_k), W_{32}, ((T_1, T_o), (T_3, T_2, T_o), (T_4)))
((A_{45}, B_{12}, (s_1, s_2, s_k), W_{45, 12}, ((T_2, T_o), (T_3, T_2), (T_1, T_4)))...)

Mapping Learner

Mapping Rule Learner

Transformation Weight Learner

Set of Mappings between the Objects

((A_3, B_2) mapped)
((A_{45}, B_{12}) not mapped)
((A_5, B_2) mapped)
((A_{98}, B_{23}) mapped)

Label

USER
Transformation Weight Learner

- Calculate Transformation Weights
- Compute Attribute Similarity Scores

Set of Similarity Scores
Calculate Transformation Weights

\[
P(\text{mapped} \mid \text{transformation}) = \frac{P(\text{transformation} \mid \text{mapped}) \cdot P(\text{mapped})}{P(\text{transformation})}
\]

<table>
<thead>
<tr>
<th>Examples</th>
<th>Classification</th>
<th>Labeled by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art’s Deli, Art’s Delicatessen</td>
<td>Mapped</td>
<td>Learner</td>
</tr>
<tr>
<td>CPK, California Pizza Kitchen</td>
<td>Mapped</td>
<td>User</td>
</tr>
<tr>
<td>Ca’Brea, La Brea Bakery</td>
<td>Not Mapped</td>
<td>Learner</td>
</tr>
</tbody>
</table>
## Recalculating Similarity Scores

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Mapped</th>
<th>Not Mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>(EQUAL &quot;Art&quot; &quot;Art&quot;)</td>
<td>.8</td>
<td>.2</td>
</tr>
<tr>
<td>(EQUAL &quot;s&quot; &quot;s&quot;)</td>
<td>.8</td>
<td>.2</td>
</tr>
<tr>
<td>(PREFIX &quot;Deli&quot; &quot;Delicatessen&quot;)</td>
<td>.1</td>
<td>.9</td>
</tr>
</tbody>
</table>

Total mapped score \( m = .064 \)

Total not mapped score \( n = .004 \)

Normalized Attribute Similarity Score \( = \frac{m}{m + n} \)

\[ = \frac{.064}{(.064 + .004)} \]

Attribute Similarity Score \( = .941 \)
Set of Mappings between the Objects

((A₁ B₂ mapped)
(A₄₅ B₁₂ not mapped)
(A₅ B₂ mapped)
(A₉₈ B₂₃ mapped)

(Object pairs, Similarity Scores, Total Score, Transformations)

((A₁ B₂, (s₁ s₂ s₃), W₂₂₂, ((T₁₄,T₁₃),(T₃₄),(T₄)))
(A₄₅ B₁₂, (s₁ s₂ s₃), W₄₅₁₂,(T₂₅),(T₃₅),(T₄₅)))...
Enforcing One-to-One Relationship

- Viewed as weighted bipartite matching problem

**Zagat’s**

(Name, Street, City)
(Art’s Deli, 1745 Ventura Boulevard, Encino)
(Citrus, 267 Citrus Ave., LA)
(Spago, 456 Sunset Bl. LA)
(Z1, Z2, Z3)

.. (not in source)

**Dept of Health**

(Name, Street, City)
(Art’s Delicatessen, 1745 Ventura Blvd, Encino)
(Ca’ Brea, 6743 La Brea Ave., LA)
(Patina, 342 Melrose Ave., LA)
(D1, D2, D3)

.. (not in source)

Given weights W, matching method determines mostly likely Matching Assignment
Experimental Results

- Three domains: Restaurant, Company, Airport
- Three types of experiments
  - Active Atlas (Mapping Learner)
  - Passive Atlas (Decision tree learner)
  - Candidate Generator (Baseline) (only Stemming)
- Three Variations of Active Atlas
  - Without Transformation weight learning
  - Without using Dissimilarity for choosing queries
  - Without enforcing One-to-One Relationship
- Learned Weights and Rules
## Restaurant Domain

### Zagat’s Restaurants

<table>
<thead>
<tr>
<th>Name</th>
<th>Street</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art’s Deli</td>
<td>12224 Ventura Boulevard</td>
<td>818-756-4124</td>
</tr>
<tr>
<td>Teresa's</td>
<td>80 Montague St.</td>
<td>718-520-2910</td>
</tr>
<tr>
<td>Steakhouse The</td>
<td>128 Fremont St.</td>
<td>702-382-1600</td>
</tr>
<tr>
<td>Les Celebrites</td>
<td>155 W. 58th St.</td>
<td>212-484-5113</td>
</tr>
</tbody>
</table>

### Dept. of Health

<table>
<thead>
<tr>
<th>Name</th>
<th>Street</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art’s Delicatessen</td>
<td>12224 Ventura Blvd.</td>
<td>818/755-4100</td>
</tr>
<tr>
<td>Teresa’s</td>
<td>103 1st Ave. between 6th and 7th Sts.</td>
<td>212/228-0604</td>
</tr>
<tr>
<td>Binion’s Coffee Shop</td>
<td>128 Fremont St.</td>
<td>702/382-1600</td>
</tr>
<tr>
<td>Les Celebrites</td>
<td>160 Central Park S</td>
<td>212/484-5113</td>
</tr>
</tbody>
</table>

112 mapped objects / 3310 mappings proposed
Restaurant Results

[Graph showing accuracy over number of examples for Baseline, Passive Atlas, and Active Atlas]
Active Atlas Results
HooversWeb

<table>
<thead>
<tr>
<th>Name</th>
<th>Url</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soundworks</td>
<td><a href="http://www.sdw.com">www.sdw.com</a></td>
<td>Stereos</td>
</tr>
<tr>
<td>Cheyenne Software</td>
<td><a href="http://www.chey.com">www.chey.com</a></td>
<td>Software</td>
</tr>
<tr>
<td>Alpharel</td>
<td><a href="http://www.alpharel.com">www.alpharel.com</a></td>
<td>Computers</td>
</tr>
</tbody>
</table>

IonTech

<table>
<thead>
<tr>
<th>Name</th>
<th>Url</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soundworks</td>
<td><a href="http://www.sdw.com">www.sdw.com</a></td>
<td>AV Equipment</td>
</tr>
<tr>
<td>Cheyenne Software</td>
<td><a href="http://www.cheyenne.com">www.cheyenne.com</a></td>
<td>Software</td>
</tr>
<tr>
<td>Altris Software</td>
<td><a href="http://www.alpharel.com">www.alpharel.com</a></td>
<td>Software</td>
</tr>
</tbody>
</table>

294 mapped objects / 14303 mappings proposed
Company Results

![Graph showing accuracy over number of examples for Baseline, Passive Atlas, and Active Atlas. The graph indicates that Active Atlas consistently outperforms Baseline and Passive Atlas, reaching near-perfect accuracy with fewer examples.]
Active Atlas Results

![Graph showing Active Atlas Results](image-url)

- Accuracy values: 0.985, 0.988, 0.991, 0.994, 0.997
- Y-axis: Accuracy
- X-axis: Number of Examples
- Graph lines represent:
  - No Transformation Learning
  - No Dissimilarity
  - No 1-to-1
  - Active Atlas
### Airport/Weather Domain

<table>
<thead>
<tr>
<th>Code</th>
<th>Location</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>PADQ</td>
<td>KODIAK, AK</td>
<td>ADQ, Kodiak, AK USA</td>
</tr>
<tr>
<td>KIGC</td>
<td>CHARLESTON AFB, VA</td>
<td>CHS, Charleston, VA USA</td>
</tr>
<tr>
<td>KCHS</td>
<td>CHARLETON, VA</td>
<td>CHS, Charleston, VA USA</td>
</tr>
</tbody>
</table>

418 mapped objects / 17120 mappings proposed
Airport/Weather Results

![Graph showing accuracy over number of examples for Baseline, Passive Atlas, and Active Atlas. The graph indicates that Active Atlas consistently outperforms the other two methods, achieving a higher accuracy across all numbers of examples.](image-url)
Active Atlas Results

![Graph showing accuracy vs. number of examples for different methods: No Transformation Learning, No Dissimilarity, No 1-to-1, and Active Atlas. The Active Atlas method consistently achieves higher accuracy.]
## Applying Learned Weights & Rules

<table>
<thead>
<tr>
<th>Application Domain</th>
<th>Total Number of Examples</th>
<th>Total Number of Test Examples</th>
<th>Average Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restaurant</td>
<td>3310</td>
<td>662</td>
<td>.9989</td>
</tr>
<tr>
<td>Company</td>
<td>14303</td>
<td>2861</td>
<td>.9995</td>
</tr>
<tr>
<td>Airport</td>
<td>17120</td>
<td>3624</td>
<td>.9960</td>
</tr>
</tbody>
</table>
Related Work

• Key characteristics
  – Manual methods to customize rules for each domain
  – User-applied fixed threshold to match objects
  – No transformation weight learning

• Related work areas
  – Database Community (Ganesh et al, Monge&Elkan)
  – Information Retrieval (Cohen)
  – Sensor Fusion (Huang & Russell)
  – Record Linkage (Jaro, Winkler)
Database Community

- Removing duplicate records
  - Hernandez & Stolfo, Monge & Elkan
  - User-defined transformations
  - Manual generated mapping rules

- Data Mining
  - Work conducted by Pinheiro & Sun
    - User-defined transformations
    - Learned attribute model (supervised learning)
  - Work by Ganesh et al
    - Learned mapping rules (decision tree learner)
Information Retrieval

• Whirl Information Retrieval System (Cohen)
• Stemming is the only transformation
  – “CPK” would not match “California Pizza Kitchen.”
• The user reviews ranked set of objects to determine the threshold of the match
Sensor Fusion & Record Linkage

• Appearance Model (Huang & Russell)
  – Appearance probabilities will not be helpful for an attribute with a unique set of instances
  – (“Art’s Deli”, “Art’s Delicatessen”)

• Record Linkage community (Jaro, Winkler)
  – Hand tailored domain specific transformations
  – The EM algorithm is applied to classify the data into three classes:
    • Matched
    • Not matched
    • To be reviewed
  – Unsupervised learning technique
Conclusions

• Novel approach combines both mapping rule learning and transformation weight learning to create a robust object identification system
• Learns to classify examples with 100% accuracy
• Requires less user involvement than other baseline techniques (Passive Atlas & Information Retrieval)
Future Work

- Noise in User Labels
- Learning Specific Transformations Weights
- Learning to Generate New Transformations
- Scaling: Approach currently applied to sets of examples on the order of 10,000. What are the issues for millions?
- Reconciling textual differences