Wrap-up
Part 1
Web IE, Wrappers and Information Integration using Karma
Extracting Data from Semi-structured Sources

Casablanca Restaurant
220 Lincoln Boulevard, Venice, CA 90291
(310) 392-5751

Appears in the Category:
Restaurants

NAME: Casablanca Restaurant
STREET: 220 Lincoln Boulevard
CITY: Venice
PHONE: (310) 392-5751
Approaches to Wrapper Construction

- Manual Wrapper Construction
- Learning-based Wrapper Construction
- Automatic Wrapper Construction
  - Grammar learning using Roadrunner
  - Clustering and learning the structure of the clustered pages using the Inferlink tool
Information Integration in Karma

Domain Model

Karma

Samples of Source Data

Source Mappings

Karma semi-automatically generates Source Mappings

Karma supports multiple integration regimes
Karma semi-automatically builds semantic models.

Karma uses semantic models to create knowledge graphs.
Part 2

Information Extraction from ‘unstructured’ data
Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC’s Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

Frank Huymbrechts - COO
Mr. Huymbrechts has over 20 years of

<table>
<thead>
<tr>
<th>Name</th>
<th>Phone</th>
<th>Email</th>
<th>Department</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barto, Andrew G.</td>
<td>(413) 545-2109</td>
<td><a href="mailto:barto@cs.umass.edu">barto@cs.umass.edu</a></td>
<td>CS276</td>
</tr>
<tr>
<td>Professor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berger, Emery D.</td>
<td>(413) 577-4211</td>
<td><a href="mailto:emery@cs.umass.edu">emery@cs.umass.edu</a></td>
<td>CS344</td>
</tr>
<tr>
<td>Assistant Professor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brock, Oliver</td>
<td>(413) 577-0334</td>
<td><a href="mailto:oli@cs.umass.edu">oli@cs.umass.edu</a></td>
<td>CS246</td>
</tr>
<tr>
<td>Assistant Professor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clarke, Lori A.</td>
<td>(413) 545-1328</td>
<td><a href="mailto:clarke@cs.umass.edu">clarke@cs.umass.edu</a></td>
<td>CS304</td>
</tr>
<tr>
<td>Professor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohen, Paul R.</td>
<td>(413) 545-3638</td>
<td><a href="mailto:cohen@cs.umass.edu">cohen@cs.umass.edu</a></td>
<td>CS278</td>
</tr>
<tr>
<td>Professor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Charts**

**Tables**

**Charts**
Scope

Web site specific

Genre specific (e.g., forums)

Wide, non-specific
# Pattern Complexity

**E.g., word patterns**

<table>
<thead>
<tr>
<th><strong>Closed set</strong></th>
<th><strong>Regular set</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S. states</td>
<td>U.S. phone numbers</td>
</tr>
<tr>
<td>He was born in <strong>Alabama</strong>. . .</td>
<td><strong>Phone</strong>: (413) 545-1323</td>
</tr>
<tr>
<td>The big <strong>Wyoming</strong> sky . .</td>
<td>The CALD main office can be reached at <strong>412-268-1299</strong></td>
</tr>
</tbody>
</table>

**Complex pattern**

- U.S. postal addresses
  - University of Arkansas
    - P.O. Box 140
    - Hope, AR 71802
  - Headquarters:
    - 1128 Main Street, 4th Floor
    - Cincinnati, Ohio 45210

**Ambiguous patterns,**

- needing context and many sources of evidence
  - **Person names**
    - ...was among the six houses sold by **Hope Feldman** that
    - **Pawel Opalinski**, Software Engineer at WhizBang Labs.

**“YOU don’t wanna miss out on ME :) Perfect lil booty Green eyes Long curly black hair Im a Irish, Armenian and Filipino mixed princess :) ❤ Kim ❤ 7°7~7two7~7four77 ❤ HH 80 roses ❤ Hour 120 roses ❤ 15 mins 60 roses”**

**Courtesy of Andrew McCallum**
Practical Considerations

• How good (precision/recall) is necessary?
  • High precision when showing extractions to users
  • High recall when used for ranking results

• How long does it take to construct?
  • Minutes, hours, days, months

• What expertise do I need?
  • None (domain expertise), patience (annotation), simple scripting, machine learning guru

• What tools can I use?
  • Many ...
myDIG: A KG Construction Toolkit
Python, MIT license, https://github.com/usc-isci2/dig-etl-engine

• Enable end-users to construct domain-specific KGs
  • end users from 5 government orgs constructed KGs in less than one day

• Suite of extraction techniques
  • semi-structured HTML pages, glossaries, NLP rules, NER, tables (coming soon)

• KG includes provenance and confidences
  • enable research to improve extractions and KG quality

• Scalable
  • runs on laptop (~100K docs), cluster (> 100M docs)

• Robust
  • Deployed to many law enforcement agencies

• Easy to install
  • Docker deployment with single “docker compose up” installation
Part 3
Knowledge Graph Completion
What is knowledge graph completion?

• An ‘intelligent’ way of doing data cleaning
  • Deduplicating entity nodes (entity resolution)
  • Collective reasoning (probabilistic soft logic)
  • Link prediction
  • Dealing with missing values
  • Anything that improves an existing knowledge graph!

• Also known as knowledge base identification
Some solutions we covered

• Entity Resolution (ER)
• Probabilistic Soft Logic (PSL)
• Knowledge Graph Embeddings (KGEs), with applications
Entity Resolution (ER)

- The algorithmic problem of grouping entities referring to the same underlying entity
Extraction Graph + Ontology + ER + PSL

**Uncertain Extractions:**
- 0.5: Lbl(Kyrgyzstan, bird)
- 0.7: Lbl(Kyrgyzstan, country)
- 0.9: Lbl(Kyrgyz Republic, country)
- 0.8: Rel(Kyrgyz Republic, Bishkek, hasCapital)

**Ontology:**
- Dom(hasCapital, country)
- Mut(country, bird)

**Entity Resolution:**
- SameEnt(Kyrgyz Republic, Kyrgyzstan)

(After Knowledge Graph Identification)
Knowledge graph embeddings

- Many ways to model the problem: entities are usually vectors, relations could be vectors or matrices
**Objective/loss/energy functions**

- What is an ‘optimal’ vector/matrix for an entity or relation?

<table>
<thead>
<tr>
<th>Model</th>
<th>Score function $f_r(h, t)$</th>
<th># Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>TransE (Bordes et al. 2013b)</td>
<td>$|h + r - t|<em>{\ell</em>{1/2}}, \ r \in \mathbb{R}^k$</td>
<td>$O(n_c k + n_r k)$</td>
</tr>
<tr>
<td>Unstructured (Bordes et al. 2012)</td>
<td>$|h - t|^2_2$</td>
<td>$O(n_c k)$</td>
</tr>
<tr>
<td>Distant (Bordes et al. 2011)</td>
<td>$|W_{rh}h - W_{rt}t|<em>1, \ W</em>{rh}, W_{rt} \in \mathbb{R}^{k \times k}$</td>
<td>$O(n_c k + 2n_r k^2)$</td>
</tr>
<tr>
<td>Bilinear (Jenatton et al. 2012)</td>
<td>$h^T W_r t, W_r \in \mathbb{R}^{k \times k}$</td>
<td>$O(n_c k + n_r k^2)$</td>
</tr>
<tr>
<td>Single Layer</td>
<td>$u_r^T f(W_{rh}h + W_{rt}t + b_r)$</td>
<td>$O(n_c k + n_r (sk + s))$</td>
</tr>
<tr>
<td></td>
<td>$u_r, b_r \in \mathbb{R}^s, W_{rh}, W_{rt} \in \mathbb{R}^{s \times k}$</td>
<td></td>
</tr>
<tr>
<td>NTN (Socher et al. 2013)</td>
<td>$u_r^T f(h^T W_r t + W_{rh}h + W_{rt}t + b_r)$</td>
<td>$O(n_c k + n_r (sk^2 + 2sk + 2s))$</td>
</tr>
<tr>
<td></td>
<td>$u_r, b_r \in \mathbb{R}^s, W_r \in \mathbb{R}^{k \times k \times s}, W_{rh}, W_{rt} \in \mathbb{R}^{s \times k}$</td>
<td></td>
</tr>
<tr>
<td>TransH</td>
<td>$(h - w_r^T hw_r) + d_r - (t - w_r^T tw_r)|_2^2$</td>
<td>$O(n_c k + 2n_r k)$</td>
</tr>
</tbody>
</table>
Applications

• Triples classification
• Link prediction
• Toponym Featurization
• Many more!
Hands-on activities