Wrapper Learning

Kristina Lerman
University of Southern California

This presentation is based in part on slides prepared by Ion Muslea
Web Wrappers

GIVE ME:
Thai food
< $20
“A”-rated

Thai
< $20

“A” rated
Example of Extraction Task

Wrapper is a procedure that translates a Web page to tuple(s)

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

Problem description:

• Web sources present data in *human-readable format*
  • take user query
  • apply it to data base
  • present results in “template” HTML page

• To integrate data from multiple sources, one must first *extract relevant information* from Web pages

• Task: learn extraction rules based on labeled examples
  • Hand-writing rules is tedious, error prone, and time consuming
In this lecture ...

- **Wrapper Induction System** [STALKER, Muslea et al, ’99, ’01,’06]
  - Input
    - Manually labeled Web pages
    - Data schema
  - Output
    - Extraction rules

- **Wrapper Maintenance** [Lerman et al., ’03]
  - Monitors wrappers to see if they are working correctly
Wrapper Learning and Maintenance

GUI

To be labeled

Wrapper Induction System

Web pages

Wrapper

Extracted data

Automatic Re-labeling

Wrapper Verification
STALKER [Muslea et al, ’98 ’99 ’01]

- Hierarchical wrapper induction
  - Decomposes a hard problem in several easier ones
  - Extracts items independently of each other
  - Each extraction rule is a finite automaton

- Benefits
  - Can handle pages with many different structures
    - Lists, embedded lists
  - Can efficiently learn wrappers from few labeled examples
STALKER: The Wrapper Architecture

Query ——— Data

Information Extractor

EC Tree

Extraction Rules
Extraction Rules

- Extraction rule: sequence of *landmarks*
- Landmarks are tokens that help locate information on the page
  
  E.g. to extract a Phone number from Web page

Name: Joel’s  <p> Phone: <i> (310) 777-1111 </i><p> Review: ...
More about Extraction Rules

Extraction rules can handle variability on pages

Name: Joel’s  Phone: <i>(310) 777-1111</i>  Review: ...

Name: Kim’s  Phone (toll free): <b>(800) 757-1111</b>  …

Name: Kim’s  Phone: <b>(888) 111-1111</b>  Review: ...

Start: EITHER  SkipTo( Phone: <i> )
OR  SkipTo( Phone ) SkipTo( : <b> )
The Embedded Catalog Tree (ECT)

ECT describes the structure of the page

**Name:** KFC  
**Cuisine:** Fast Food  
**Locations:**
- **Venice**  
  - (310) 123-4567,  
  - (800) 888-4412.  
- **L.A.**  
  - (213) 987-6543.  
- **Encino**  
  - (818) 999-4567,  
  - (888) 727-3131.
Learning the Extraction Rules

GUI

Labeled Pages

EC Tree

Inductive Learning System

Extraction Rules
Example of Rule Induction

Training Examples:

Name: Del Taco <p> Phone (toll free): <b>(800) 123-4567</b> <p>Cuisine ...

Name: Burger King <p> Phone: (310) 987-9876 <p>Cuisine: ...
Example of Rule Induction

Training Examples:

- Name: Del Taco <p> Phone (toll free) : <b>( 800 ) 123-4567 </b><p>Cuisine ...

- Name: Burger King <p> Phone : ( 310 ) 987-9876 <p>Cuisine: ...

Initial candidate:

SkipTo( )
Example of Rule Induction

**Training Examples:**

Name: Del Taco <p> Phone (toll free) : <b> ( 800 ) 123-4567 </b><p> Cuisine ...

Name: Burger King <p> Phone : (310) 987-9876 <p> Cuisine: ...

**Initial candidate:**

```
SkipTo( ()

SkipTo( <b> ( )

SkipTo(Phone) SkipTo( ()

... SkipTo(:) SkipTo()
```
Example of Rule Induction

Training Examples:

Name: Del Taco <p> Phone (toll free) : <b> (800) 123-4567 </b> <p>Cuisine ...

Name: Burger King <p> Phone : (310) 987-9876 <p>Cuisine: ...

Initial candidate:

SkipTo( ()

SkipTo( <b> () … SkipTo(Phone) SkipTo( () … SkipTo( :) SkipTo( ()

… SkipTo(Phone) SkipTo( :) SkipTo( () …

USC Information Sciences Institute /ISI
Active Learning [Muslea et al., 2006]

- Why stop at 95-99% accuracy?
  - Select most informative examples to get to 100% accuracy

- Active Learning
  - **Idea:** system selects most informative exs. to label
  - **Advantage:** fewer examples to reach same accuracy
Multi-view Learning

Two ways to find start of the phone number:

SkipTo( **Phone:** )

BackTo( **Number** )

Name: KFC  <p> Phone: (310) 111-1111 <p> Review: Fried chicken …
Which example should be labeled next?

Training Examples

Name: Joel’s  <p> Phone: (310) 777-1111 <p>Review: The chef...
Name: Kim’s  <p> Phone: (213) 757-1111 <p>Review: Korean ...

Unlabeled Examples

Name: Chez Jean  <p> Phone: (310) 666-1111 <p> Review: ...
Name: Burger King  <p> Phone: (818) 789-1211 <p> Review: ...
Name: Café del Rey  <p> Phone: (310) 111-1111 <p> Review: ...
Name: KFC  <p> Phone: <b>(800) 111-7171</b> <p> Review: ...

USC Information Sciences Institute
ISI
<table>
<thead>
<tr>
<th>Name</th>
<th>Phone</th>
<th>Review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joel’s</td>
<td>(310) 777-1111</td>
<td>Review: ...</td>
</tr>
<tr>
<td>Kim’s</td>
<td>(213) 757-1111</td>
<td>Review: ...</td>
</tr>
<tr>
<td>Chez Jean</td>
<td>(310) 666-1111</td>
<td>Review: ...</td>
</tr>
<tr>
<td>Burger King</td>
<td>(818) 789-1211</td>
<td>Review: ...</td>
</tr>
<tr>
<td>Café del Rey</td>
<td>(310) 111-1111</td>
<td>Review: ...</td>
</tr>
<tr>
<td>KFC</td>
<td>(800) 111-7171</td>
<td>Review: ...</td>
</tr>
</tbody>
</table>
Discussion

• Wrapper Learning = basic problem is to learn how to extract the data from a page

• STALKER
  • Input: ECT (schema), labeled examples
  • Rules to extract data from pages
  • Active Learning can reduce the number of examples required to learn

• But approach
  • Requires labeled examples
  • Sensitive to changes to sources
Wrapper Maintenance

Problem

- Landmark-based extraction rules are fast and efficient…but they rely on stable Web Page layout.
  - If the page layout changes, the wrapper fails!
  - Unfortunately, the average site on the Web changes layout more than twice a year.
- Requirement: Need to detect changes and automatically re-induce extraction rules when layout changes
## Phone Search Results

**Showing 1 - 2 of 2**

First | Prev | Next | Last  | **Search Again**

<table>
<thead>
<tr>
<th><strong>Name</strong></th>
<th><strong>Address</strong></th>
<th><strong>Phone</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrew Philpot</td>
<td>Mar Vista Calif</td>
<td>(310)822-9994</td>
</tr>
<tr>
<td></td>
<td>Los Angeles, CA 90066</td>
<td></td>
</tr>
<tr>
<td>Andrew Philpot</td>
<td>600 S Curson Ave</td>
<td>(323)936-5549</td>
</tr>
<tr>
<td></td>
<td>Los Angeles, CA 90036-3666</td>
<td></td>
</tr>
</tbody>
</table>

First | Prev | Next | Last  | **Search Again**
### Phone Search Results

**Showing 1 – 2 of 2**

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>Phone (click to call)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrew Philpot</td>
<td>Mar Vista Calif, Los Angeles, CA 90066</td>
<td>(310)822-9994</td>
</tr>
<tr>
<td>Andrew Philpot</td>
<td>600 S Curson Ave, Los Angeles, CA 90036-3666</td>
<td>(323)936-5549</td>
</tr>
</tbody>
</table>

### Table

<table>
<thead>
<tr>
<th>NAME</th>
<th>ADDRESS</th>
<th>CITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrew Philpot</td>
<td>Mar Vista Calif</td>
<td>Los Angeles</td>
</tr>
<tr>
<td>Andrew Philpot</td>
<td>600 S Curson Ave</td>
<td>Los Angeles</td>
</tr>
</tbody>
</table>
## Phone Search Results

### Showing 1 - 2 of 2

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>Phone (click to call)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrew Philpot</td>
<td>Mar Vista Calif</td>
<td>(310)822-9994</td>
</tr>
<tr>
<td></td>
<td>Los Angeles, CA</td>
<td></td>
</tr>
<tr>
<td>Andrew Philpot</td>
<td>600 S Curson Ave</td>
<td>(323)936-5549</td>
</tr>
<tr>
<td></td>
<td>Los Angeles, CA</td>
<td></td>
</tr>
</tbody>
</table>

**Search Again**

---

## Phone Search Results

### Showing 1 - 1 of 1

<table>
<thead>
<tr>
<th>Name</th>
<th>Phone (click to call)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrew Philpot</td>
<td>(323)936-5549</td>
</tr>
<tr>
<td>600 S Curson Ave</td>
<td>Los Angeles, CA</td>
</tr>
</tbody>
</table>

**Search Again**

---

**USC Information Sciences Institute**
Wrapper Applied to Changed Source

Phone Search Results
Showing 1 - 1 of 1

<table>
<thead>
<tr>
<th>Name</th>
<th>Phone (click to call)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrew Philpot</td>
<td>(323)936-5549</td>
</tr>
<tr>
<td>600 S Curson Ave</td>
<td>Los Angeles, CA</td>
</tr>
</tbody>
</table>

NAME item
Begin_Rule
__ST__  * *
End_Rule
__ST__  </td> <td nowrap >
ADDRESS item
Begin_Rule
__ST__  </td> <td nowrap >
End_Rule
__ST__  <br>
...

<table>
<thead>
<tr>
<th>NAME</th>
<th>ADDRESS</th>
<th>CITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIL</td>
<td>NIL</td>
<td>600 S Curson Ave&lt;BR&gt; Los Angeles</td>
</tr>
</tbody>
</table>
Wrapper Verification

- **Wrapper Verification**
  - Automatically detect when the wrapper is no longer correctly extracting data from an information source
    - [Kushmerick 1999]
    - [Lerman et al., 2003]

- **Wrapper Reinduction**
  - Rebuild the wrapper automatically if it is not extracting data correctly from new pages
    - [Lerman et al., 2003]
Learning Global Properties for Wrapper Verification [Kushmerick, 1999]

- Each data field described by global numeric features
  - Word count, average word length, HTML density, alphabetic density
- Computationally efficient learning
- HTML density alone could account for almost all changes on test set
- Large number of false negatives on real changes to web sources [Lerman et al. 2003]
Learning Data Structure [Lerman et al., ’00,’03]

- Approach to learning the structure of data
- Token level syntactic description
  - descriptive but compact
  - computationally efficient
- Structure is described by a sequence (pattern) of general and specific tokens.
- Data prototype = starting & ending patterns

STREET_ADDRESS
220 Lincoln Blvd
420 S Fairview Ave
2040 Sawtelle Blvd

start with:
_NUM _CAPS
_end with:
_CAPS Blvd
_CAPS _CAPS
Token Syntactic Hierarchy

- Tokens = words
- Syntactic types
  - e.g., NUMBER, ALPHA
- Hierarchy of types
  - allows generalization
- Extensible
  - new types
  - domain-specific information
Prototype Learning Algorithm

- No explicit negative examples
- Learn from positive examples of data
- Find patterns that
  - describe many of the positive examples of data
  - highly unlikely to describe a random token sequence (implicit negative examples)
- are statistically significant patterns at $\alpha=0.05$ significance level
- **DataPro** – efficient (greedy) algorithm
DataPro Algorithm

- Process examples
- Seed patterns
- Specialize patterns loop
  - Extend the pattern
    - find a more specific description
    - is the longer pattern significant given the shorter pattern?
  - Prune generalizations
    - is the pattern ending with general type significant given the patterns ending with specific tokens

Examples:

220 Lincoln Blvd
420 S Fairview Ave
2040 Sawtelle Blvd
Examples: PHONE

- starting patterns:
  ( _NUM ) _NUM - _NUM

- ending patterns:
  ( _NUM ) _NUM - _NUM
Wrapper Verification

Data prototypes can be used for web wrapper maintenance applications.

- Automatically detect when the wrapper is no longer correctly extracting data from an information source
  - (Kushmerick 1999)
Prototypes can be used for web wrapper verification. Given

- Set of correct old examples of data
- Set of new examples
- Do the patterns describe the same proportions of new examples as old examples?
Results

- Monitored 27 wrappers (23 distinct sources)
- There were 37 changes over ~ 1 year
- Algorithm discovered 35/37 changes with 15 mistakes
  - 13 false positives
- Overall:
  - Average precision = 73%
  - Average recall = 95%
  - Average accuracy = 97%
Discussion

- Flexible data representation scheme
- Algorithm to learn description of data fields
- Used in wrapper maintenance applications

Limitations:
- Needs to be extended to lists and tables
- Excellent recall, but lower recall will precision in many false positives