Wrapper Learning

Craig Knoblock
University of Southern California

This presentation is based on slides prepared by Ion Muslea and Chun-Nan Hsu
GIVE ME:
Thai food
< $20
“A”-rated

Thai
< $20

“A”rated
Roadmap to Wrapper Building

• **Today:**
  • **Part 1:**
    • Wrapper Learning
  • **Part 2:**
    • Agent Builder
      • Extracting information from a page
      • Executing wrappers

• **Next Time:**
  • Automatic Wrapper Generation
  • Advanced Agent Builder
    • Navigating through a site
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
Example of Extraction Task

NAME: Casablanca Restaurant
STREET: 220 Lincoln Boulevard
CITY: Venice
PHONE: (310) 392-5751
In this part of the lecture …

- **Wrapper Induction Systems**
  - WIEN:
    - The rules
    - Learning WIEN rules
  - SoftMealy
- **The STALKER approach to wrapper induction**
  - The rules
  - The ECTs
  - Learning the rules
• Assumes items are always in fixed, known order
  ... Name: J. Doe; Address: 1 Main; Phone: 111-1111. <p>
  Name: E. Poe; Address: 10 Pico; Phone: 777-1111. <p> ...

• Introduces several types of wrappers

  • LR:
Rule Learning

- Machine learning:
  - Use past experiences to improve performance

- Rule learning:
  - INPUT:
    - Labeled examples: training & testing data
    - Admissible rules (hypotheses space)
    - Search strategy
  - Desired output:
    - Rule that performs well both on training and testing data
Learning LR extraction rules

<html> Name:<b> Kim’s </b> Phone:<b> (800) 757-1111 </b> ...

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

USC Information Sciences Institute
Learning LR extraction rules

- Admissible rules:
  - prefixes & suffixes of items of interest

- Search strategy:
  - start with shortest prefix & suffix, and expand until correct
Learning LR extraction rules

- Admissible rules:
  - prefixes & suffixes of items of interest

- Search strategy:
  - start with shortest prefix & suffix, and expand until correct

<table>
<thead>
<tr>
<th>Name</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim’s</td>
<td>(800) 757-1111</td>
</tr>
<tr>
<td>Joe’s</td>
<td>(888) 111-1111</td>
</tr>
</tbody>
</table>
Learning LR extraction rules

- Admissible rules:
  - prefixes & suffixes of items of interest

- Search strategy:
  - start with shortest prefix & suffix, and expand until correct
Learning LR extraction rules

- **Admissible rules:**
  - prefixes & suffixes of items of interest

- **Search strategy:**
  - start with shortest prefix & suffix, and expand until correct

<table>
<thead>
<tr>
<th>Name</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim’s</td>
<td>(800) 757-1111</td>
</tr>
<tr>
<td>Joe’s</td>
<td>(888) 111-1111</td>
</tr>
</tbody>
</table>
Learning LR extraction rules

- **Admissible rules:**
  - prefixes & suffixes of items of interest

- **Search strategy:**
  - start with shortest prefix & suffix, and expand until correct
Summary

- **Advantages:**
  - Fast to learn & extract

- **Drawbacks:**
  - Cannot handle permutations and missing items
  - Must label entire page
  - Requires large number of examples
In this part of the lecture …

- Wrapper Induction Systems
  - WIEN:
    - The rules
    - Learning WIEN rules
  - SoftMealy
- The STALKER approach to wrapper induction
  - The rules
  - The ECTs
  - Learning the rules
SoftMealy [Hsu & Dung, ‘98]

- Learns a transducer

![Diagram of SoftMealy model with nodes named Name, Addr, and Phone, and arrows connecting them with labels for Name, Addr, and Phone.]
SoftMealy --- extractor representation formalism

- Variation of finite state transducer (a.k.a. Mealy machine)
- Simple enough to be learnable from a small number of examples of extractions
  - fixed graph structure or strictly confined search space for graph structures
  - less edges, less outgoing edges
- Complex enough to handle irregular attribute permutations
  - missing attributes
  - multiple attribute values
  - variant attribute ordering
How **SoftMealy** extractors work

Mani Chandy, Professor of Computer Science and Executive Officer for Computer Science
Contextual rule

- Contextual rule looks like:

  TRANSFER FROM state N TO state N IF

  left context = capitalized string

  right context = HTML tag \( "<\/A>" \)

- When the "master" read head stops at the boundary between two tokens, the "secondary" read head scans the left and right context and matches what’s read with contextual rules.

- It is not necessary that both left context and right context are used in a contextual rule.

- A contextual rule may have disjunctions.
Summary

• **Advantages:**
  - Also learns order of items
  - Allows item permutations & missing items
  - Uses wildcards (eg, Number, AllCaps, etc)

• **Drawback:**
  - Must “see” all possible permutations
In this part of the lecture …

- **Wrapper Induction Systems**
  - WIEN:
    - The rules
    - Learning WIEN rules
  - SoftMealy

- **The STALKER approach to wrapper induction**
  - The rules
  - The ECTs
  - Learning the rules
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 rule is a finite automaton
STALKER: The Wrapper Architecture

Query → Information Extractor → Data

- EC Tree
- Extraction Rules
Extraction Rules

Extraction rule: sequence of landmarks

SkipTo(Phone) SkipTo(<i>) SkipTo(</i>)

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

Name: Joel’s  Phone: (310) 777-1111  Review: ...

Name: Kim’s  Phone (toll free): (800) 757-1111  …

Name: Kim’s  Phone: (888) 111-1111  Review: …

Start:  EITHER  SkipTo( Phone : )  OR  SkipTo( Phone ) SkipTo( : )
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  Phone (toll free) :  (800) 123-4567
Cuisine ...

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

Initial candidate:

SkipTo( )

SkipTo( )  SkipTo(Phone)  SkipTo( )  ...  SkipTo(:)  SkipTo()

...  SkipTo(Phone)  SkipTo(:)  SkipTo( )  ...
Active Learning & Information Agents

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

- **Information Agents**
  - One agent may use hundreds of extraction rules
    - Small reduction of *examples per rule* => big impact on user
  - Why stop at 95-99% accuracy?
    - Select most informative examples to get to 100% accuracy
Which example should be labeled next?

<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>The chef...</td>
</tr>
<tr>
<td>Kim’s</td>
<td>(213) 757-1111</td>
<td>Korean...</td>
</tr>
<tr>
<td>Chez Jean</td>
<td>(310) 666-1111</td>
<td></td>
</tr>
<tr>
<td>Burger King</td>
<td>(818) 789-1211</td>
<td></td>
</tr>
<tr>
<td>Café del Rey</td>
<td>(310) 111-1111</td>
<td></td>
</tr>
<tr>
<td>KFC</td>
<td>(800) 111-7171</td>
<td></td>
</tr>
</tbody>
</table>
Two ways to find start of the phone number:

SkipTo( **Phone:** )  

BackTo( **Number** )

**Name:** KFC  
**Phone:** (310) 111-1111  
**Review:** Fried chicken …
Co-Testing

RULE 1

RULE 2

Labeled data

Unlabeled data
<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>...</td>
</tr>
<tr>
<td>Kim’s</td>
<td>(213) 757-1111</td>
<td>...</td>
</tr>
<tr>
<td>Chez Jean</td>
<td>(310) 666-1111</td>
<td>...</td>
</tr>
<tr>
<td>Burger King</td>
<td>(818) 789-1211</td>
<td>...</td>
</tr>
<tr>
<td>Café del Rey</td>
<td>(310) 111-1111</td>
<td>...</td>
</tr>
<tr>
<td>KFC</td>
<td>(800) 111-7171</td>
<td>...</td>
</tr>
</tbody>
</table>
Not all queries are equally informative
Weak Views

• Learn “content description” for item to be extracted

  • Too general for extraction
    • (Nmb) Nmb–Nmb can’t tell a phone number from a fax number

  • Useful at discriminating among query candidates

• Learned field description
  • Starts with: (Nmb)
  • Ends with: Nmb–Nmb
  • Contains: Nmb Punct
  • Length: [6,6]
Naïve & Aggressive Co-Testing

- **Naïve Co-Testing:**
  - Query: randomly chosen contention point
  - Output: rule with fewest mistakes on queries

- **Aggressive Co-Testing:**
  - Query: contention point that most violates weak view
  - Output: committee vote (2 rules + weak view)
Empirical Results: 33 Difficult Tasks

- 33 *most difficult* of the 140 extraction tasks
  - Each view: > 7 labeled examples for best accuracy
  - At least 100 examples for task
Results in 33 Difficult Domains

Extraction Tasks

Examples to 100% accuracy

Random sampling

Examples to 100% accuracy
Results in 33 Difficult Domains

Extraction Tasks

Examples to 100% accuracy

Naïve Co-Testing
Random sampling
Results in 33 Difficult Domains

- Aggressive Co-Testing
- Naïve Co-Testing
- Random sampling

Extraction Tasks

Examples to 100% accuracy

USC Information Sciences Institute
Summary

- **Advantages:**
  - Powerful extraction language (e.g., embedded list)
  - One hard-to-extract item does not affect others

- **Disadvantage:**
  - Does not exploit item order (sometimes may help)
Discussion

• Basic problem is to learn how to extract the data from a page

• Range of techniques that vary in the
  • Learning approach
  • Rules that can be learned
  • Efficiency of the learning
  • Number of examples required to learn

• Regardless, all approaches
  • Require labeled examples
  • Are sensitive to changes to sources
Wrapper Validation and Maintenance

Craig Knoblock
USC Information Sciences Institute
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
Learning Regular Expressions
[Goan, Benson, & Etzioni, 1996]

- Character level description of extracted data
- Based on ALERGIA [Carrasco and Oncina, 1994]
  - Stochastic grammar induction algorithm
  - Merges too many states resulting in over-general grammar
- WIL reduced faulty merges by imposing syntactic categories:
  - Number, lower upper, and delim
- Only merges when nodes contain the same syntactic category
- Requires large number of examples to learn
- Computationally expensive
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, Knoblock, Minton, 2002]
Learning Data Prototypes
[Lerman & Minton, 2000]

- 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

(310) 577 - 8182
(310) 652 - 9770
(310) 396 - 1179
(310) 477 - 7242
(626) 792 - 9779
(310) 823 - 4446
(323) 870 - 2872
(310) 855 - 9380
(310) 578 - 2293
(310) 392 - 5751
(805) 683 - 8864
(310) 301 - 1004
(626) 793 - 8123
(310) 822 - 1511

• starting patterns:
  ( _NUM ) _NUM - _NUM

• ending patterns:
  ( _NUM ) _NUM - _NUM
Example: STREET_ADDRESS

- starting patterns:
  - _NUM S _CAPS Blvd
  - _NUM _CAPS Ave
  - _NUM _CAPS

- ending patterns:
  - _NUM _CAPS _CAPS
  - _NUM S _CAPS Blvd
  - _NUM _CAPS Ave
  - _NUM _CAPS Blvd

13455 Maxella Ave
903 N La Cienega Blvd
110 Navy St
2040 Sawtelle Blvd
87 E Colorado Blvd
4325 Glencoe Ave
2525 S Robertson Blvd
998 S Robertson Blvd
523 Washington Blvd
220 Lincoln Blvd
420 S Fairview Ave
13490 Maxella Ave
363 S Fair Oaks Ave
4676 Admiralty Way
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)
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?
Wrapper Verification

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%
Wrapper Reinduction

- Rebuild the wrapper automatically if it is not extracting data correctly from new pages
- Data extraction step
  Identify correct examples of data on new pages
- Wrapper induction step
  Feed the examples, along with the new pages, to the wrapper induction algorithm to learn new extraction rules
The Lifecycle of A Wrapper

- GUI
- Wrapper Induction System
- To be labeled
- Web pages
- Extracted data
- Automatic Re-labeling
- Wrapper Verification
- Wrapper
## 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></td>
<td>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</td>
<td></td>
</tr>
<tr>
<td></td>
<td>90036-3666</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>First</th>
<th>Prev</th>
<th>Next</th>
<th>Last</th>
</tr>
</thead>
</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>

<table>
<thead>
<tr>
<th>First</th>
<th>Prev</th>
<th>Next</th>
<th>Last</th>
</tr>
</thead>
</table>
Whitepages Wrapper

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 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

...<br>

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>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>

USC Information Sciences Institute

/ISI/
### 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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>First</th>
<th>Prev</th>
<th>Next</th>
<th>Last</th>
</tr>
</thead>
</table>

### Wrapper Applied to Wrapper Applied to Changed Source

...  
**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>

**USC Information Sciences Institute**
After Reinduction

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

ADDRESS item

Begin_Rule

ST

End_Rule

NAME         ADDRESS   CITY
Andrew Philpot 600 S Curson Ave   Los Angeles
Lindbergh
by A. Scott Berg

List Price: $30.00
Our Price: $21.00
You Save: $9.00 (30%)

Availability: This title usually ships within 2-3 days
Need this by December 24? No problem. Select shipping method (U.S. addresses).

Click for larger picture

Hardcover - 620 pages (September 1998)
Putnam Pub Group (T); ISBN: 0399144498; Dimensions (in inches): 1.97 x 9.38 x 6.47
Other Editions: Paperback, Audio Cassette (Abridged)

Amazon.com Sales Rank: 3,539
Popular in: U.S. Senate (#5), Laguna Beach, CA (#12). See more
Avg. Customer Review:
Number of Reviews: 80

<table>
<thead>
<tr>
<th>AUTHOR</th>
<th>TITLE</th>
<th>PRICE</th>
<th>AVAILABILITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Scott Berg</td>
<td>Lindbergh</td>
<td>21.00</td>
<td>This title usually ships...</td>
</tr>
</tbody>
</table>
Lindbergh
by A. Scott Berg

List Price: $30.00
Our Price: $21.00
You Save: $9.00 (30%)

Availability: This title usually ships within 2-3 da

Need this by December 24? Select Next D
shipping method (U.S. addresses).

See larger photo

Hardcover - 623 pages (September 1998)
Putnam Pub Group (T); ISBN: 0399144408 ; Dimensions (in inches): 1.97 x 0.36 x 6.47
Other Editions: Paperback, Audio Cassette (Abridged)

Amazon.com Sales Rank: 3,711
Popular in: U.S. Senate (#5), Laguna Beach, CA (#12). See mo
Avg. Customer Review: 
Number of Reviews: 81

<table>
<thead>
<tr>
<th>AUTHOR</th>
<th>TITLE</th>
<th>PRICE</th>
<th>AVAILABILITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIL</td>
<td>NIL</td>
<td>21.00</td>
<td>This title usually ships...</td>
</tr>
</tbody>
</table>

USC Information Sciences Institute
**After Reinduction**

**AUTHOR**  | **TITLE**  | **PRICE**  | **AVAILABILITY**
--- | --- | --- | ---
A.Scott Berg  | Lindbergh  | 21.00  | This title usually ships...
Wrapper Reinduction

Results

- Monitored 10 distinct sources
- There were 8 changes over ~ 1 year
- Extracting examples:
  - 277/338 correct (82%)
  - 31 false positives/30 false negatives
- Reinduction:
  - Average recall = 90%
  - Average precision = 80%
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