Automating the Extraction of Data from HTML Tables with Unknown Structure

Stefan Rümmele
PSWIE SS 2005
Overview

• Application range
• Problems
• Extraction ontologies
• Location & extraction solution
Application range

Source:

– Various websites with HTML tables
– Tables have different and unknown structure
– All tables relate to a specific domain of interest

Target:

– Predefined relational schema
### Example: car advertisement

<table>
<thead>
<tr>
<th>Car</th>
<th>Year</th>
<th>Make</th>
<th>Model</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>0001</td>
<td>1999</td>
<td>Pontiac</td>
<td>Firebird</td>
<td>Air Conditioning</td>
</tr>
<tr>
<td>0002</td>
<td>2000</td>
<td>Acura</td>
<td>RL 3.5</td>
<td>Blue</td>
</tr>
<tr>
<td>0003</td>
<td>2002</td>
<td>Honda</td>
<td>Accord EX</td>
<td>White</td>
</tr>
</tbody>
</table>

Viewing and querying web car advertisements through target schema:

- \{Car, Year, Make, Model, Mileage, Price, PhoneNr\}
- \{Car, Feature\}
Pre-Owned Inventory

To see a list of all our cars, trucks, vans and SUV's, [click here].

- Looking for a price quote? Check out our Quick Quote Form.
- Need financing? Try our new Pre-Approval Form.
- Check out our Internet Only Specials.

To search for a specific vehicle or model, use our easy search engine below. Our inventory changes daily, so drop us an email or give us a call if you don't see the car you want. We will make sure you find your dream car! You searched for:

- All vehicles available.

66 matches found. Vehicles 1 to 25 shown.

<table>
<thead>
<tr>
<th>Year</th>
<th>Make and Model</th>
<th>Price</th>
<th>Miles</th>
<th>Exterior</th>
<th>Interior</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>Pontiac Firebird</td>
<td>$10,000</td>
<td>90,000</td>
<td>Blue</td>
<td>Red</td>
</tr>
<tr>
<td>2000</td>
<td>Acura RL 4dr</td>
<td>$15,500</td>
<td>90,450</td>
<td>Silver</td>
<td>Gray</td>
</tr>
<tr>
<td>2002</td>
<td>Honda Accord EX</td>
<td>$19,500</td>
<td>90,450</td>
<td>White</td>
<td>Black</td>
</tr>
<tr>
<td>2002</td>
<td>Honda Passport</td>
<td>$20,998</td>
<td>10,411</td>
<td>Black</td>
<td>Gray</td>
</tr>
<tr>
<td>2009</td>
<td>Acura RDX Type-S</td>
<td>$10,995</td>
<td>14,995</td>
<td>Red</td>
<td>Black</td>
</tr>
<tr>
<td>2006</td>
<td>Chevrolet Silverado 1500</td>
<td>$15,995</td>
<td>90,450</td>
<td>Silver</td>
<td>Black</td>
</tr>
<tr>
<td>2001</td>
<td>Honda Accord Value Package</td>
<td>$13,995</td>
<td>31,711</td>
<td>Silver</td>
<td>Black</td>
</tr>
<tr>
<td>2001</td>
<td>Chevrolet Silverado 1500</td>
<td>$12,995</td>
<td>20,021</td>
<td>Pewter</td>
<td>Gray</td>
</tr>
</tbody>
</table>

Show checked vehicles | New search | Show 25 more

All vehicles subject to prior sale. We reserve the right to make changes without notice, and are not responsible for errors.
Pre-Owned Inventory

At Howard Auto Group we have created a complete and comprehensive Internet sales department to give our customers an alternative buying experience. Once you have found a vehicle you like we will be glad to give you our lowest no haggle price right up front! Then if you like that price you can complete the transaction with your Internet manager. He can also quote you a payment, interest rate and if you have a trade in give you an evaluation of your trade in. Remember the Internet department is designed to provide the fastest and friendliest service to the Internet user and to ensure a totally different buying experience! If you have any questions please feel free to give us a call at 405-926-8955 or toll free 877-944-2842 or drop us an e-mail. Your Internet sales staff at Howard Auto Group is waiting to help you. http://www.honda.com, phone, toll free, email, feliciano, nelson, steve, ramos, jay, jake, and ryan.

Schedule a test drive

Send Me More Information

2002 Honda Accord EX $21,988

Features
- Air Conditioning
- Driver Side Air Bag
- Passenger Side Air Bag
- Anti-Lock Brakes
- AM/FM Cassette
- Security Features
- Alloy Wheels
- Automatic Transmission
- Bucket Seats
- Compact Disc Player
- Cruise Control
- Front Wheel Drive
- Intermittent Wipers
- Moon Roof

Price $21,988
Mileage 13,875 miles
Body Type Car
Body Style Coupe
Exterior White
Transmission Automatic
Engine 3.0L 6 cyl Fuel Injection
Fuel Type Gas
Stock Number 550291A
VIN 1HGE22562A01864
Location problems

- Multiple frames
- Tables for layout
- Table rows not part of the data
- Tables displayed piecemeal
- Tables spanning multiple pages
- No `<table>` tag
Extraction problems 1/2

• Merged attributes/values
• Subsets
• Synonyms
• Extra information
• Linked information
• Externally factored data
• Unexpected multiple values
- List table
- Position of attributes
- Duplicate data
- Missing information
- Attribute as value
Definition:

“An ontology is a specification of a conceptualization.” (Tom Gruber)

Content:

– Object/Relationship-model instance
– Data frame for each object set

Purpose:

 Formal defined system that serves as a wrapper for a narrow domain of interest.
Extraction ontology - example

01. Car [-> object];
02. Car [0:1] has Year [1:*];
03. Car [0:1] has Make [1:*];
04. Car [0:1] has Model [1:*];
05. Car [0:1] has Mileage [1:*];
06. Car [0:1] has Feature [1:*];
07. Car [0:1] has Price [1:*];
08. PhoneNr [1:*] is for Car [0:1];
09. Year matches [4]
   10. constant {extract "\d{2}";
   11.       context "\b[4-9]\d\b";
   12.       substitute "^" -> "19"; },
   13. ...
14. Mileage matches [8]
15. ...
16. keyword "\bmi\s\b", "\bmi\.\", "\bmi\b",
17. "\bmileage\b", "\bodometer\b",
18. ...
1. Locate the table of interest.
   – Table on the main page
   – Tables on linked pages
   – Achieved using a heuristic with several rules based on the ontology

2. Form attribute-value pairs.

   {<Make: ACURA>, <Model: legend>, <Year: 1992>,
    <Colour: grey>, <Price: $9500>, <Auto: Yes>,
    <Air Cond.: No>, <AM/FM: Yes>, <CD: No>}
Solution (2/3)

3. Adjust attribute-value pairs.
   – Attribute-value pairs from linked tables
   – Process boolean indicators
     Make: ACURA; Model: legend; Year: 1992;
     Colour: grey; Price: $9500; Auto; AM/FM;

4. Analyze extraction patterns.
   – Applying the extraction ontology
     {<Car: 0011>, <Year: 1992>, <Make: ACURA>,
      <Model: legend>, <Mileage: >,
      <Price: $9500>, <PhoneNr: >},
     {<Car: 0011>, <Feature: grey>},
     {<Car: 0011>, <Feature: Auto>},
     {<Car: 0011>, <Feature: AM/FM>}

5. Infer Mappings.

- Transformations needed in steps 2-4 are recorded
- Mapping is produced out of this information
- Queries on the target schema can be translated to a query on the source
- Result contains additional values not recognized by the ontology