Generating Wrappers with Fetch Agent Platform 3.2

Matthew Michelson and Craig A. Knoblock
CSCI 548: Lecture 2
Starting our example

- Extract list of cars from Craig’s List:
  - Post text
  - Link to details page
    - Get the timestamp
    - Get the contact email address
  - Navigate through next links
First Piece: Building Agents

- Start up Agent Builder 3.2....
Add an entry connector

- How should the agent start to get information?
  - Entry is just a link to a page
    - Google News
  - Entry is a form:
    - Craig’s list search box
Adding our connector

- Click "Connectors tab"
  - Click green globe with plus sign (Add New Connector...)
- Select the "Form"
  - Then "Use example of connector on page" since we want it to find the form elements for us.
Adding connector

- Click `<FORM>` for page
  - highlights what we want.
- Click Next.
  - List of form elements we can define
  - Name, Type, Value for each of the fields.
- Select the query field, it will highlight.
- Then click the value box, and select a new parameter.
  - name it “searchValue.”
  - Click OK → next.
  - Name the connector and then make sure to check the box that says "when I finish start the connectivity wizard."
Name this variable
Connecting the search...

- Next screen, click "Create new wrapper."
  - Name it PostsPage,
  - check box that says "when I finish start the add new samples wizard."
  - Start with 5 samples, and you're off...

- New screen with connector's name, and a value.
  - Click each row and add a value
  - Wizard grabs the pages returned when you submit this value in the form.
  - Now you have 5 example pages to learn to extract from…
What you put here is what is submitted in the form.
What’s the flow like?

- Start with entry connector, query on a search term, get pages back…
- View → Agent Layout.
  - We will come back to this later when we define a "next" link.
  - To get back into our wrapper, just double click it.
Front page of CList cars

Search term flows from form to result pages

Set of pages returned for search term
Defining Extraction Schema

- We want the details link, the text, …
- Define our elements that we want to extract from the source.
  - 2 main elements from results pages
    - next link, that lets us navigate from this page to the next one.
    - List of items for sale that matched our search query.
      - link to the details page
      - text for the link.
Defining Extraction Schema

- Click “Add new item” (in “Define” tab page)
  - This brings up a dialog box that lets you define items.
- Do our List first
  - Put "Posts" as the name, and check "List"
- Add items to a list,
  - Right click list to add a new item
    - "detailsLink" → "Data -- URL"
    - "linkText" → "Data -- text" item.
- Add the details link
  - right click on the top level item "Data Schema"
  - "nextLink" → "data url"
Click to add new schema
• Right click to add items to list
• Right click on Data Schema to add top level elements
Schema Definitions

- We have defined all our schema elements and are ready to begin extraction!
Training AB for Extraction

- Top right corner
  - select training page
  - right click and select "Use for training"
- Note we are in “Train” tab
Training AB for Extraction

- Drag and drop elements from HTML to Agent
  - Can have null values (check in Validation too!)
  - Drag and Drop troubles: go directly into the html
    - Click the Source tab
    - Highlight the item yourself and drag it over
  - Personally: I define more than 1, 2, last elements for Lists
    - Right click on item, add element after, repeat…
- Add new pages for training:
  - Click “Add pages…” and supply URLs.
Add more pages to train

Adding more rows

Drag and Drop
Learning Extraction Rules

- Train some pages
  - Different cases = more robust rules
- Learn extraction rules.
  - Click on the Owl icon "Learn Rules"
  - Be patient: hard rules take time to learn!
  - (Remember we’re in the “train” tab)
Yikes…Errors!

- Relax, take a deep breath…
- Mark-Up errors
  - User mistakes: You accidentally mark the wrong stuff
    - It can help you find these
  - Source errors: Sometimes weirdness in pages
    - Trick: Disable all training files (right click → Disable)
    - Enable them one at a time and retrain
Congratulations!

- You can now extract:
  - Lists of detail links and their text
  - Next links
  - You are officially an information extractor!

- What to do with next links?
  - Want recursive navigation
    - Get pages and pages of Toyotas…
Adding in anchor connectors

- Switch to agent layout (View → Agent Layout).
- Right click on Wrapper2
  - Add a navigation connector
  - Select anchor, select a currently defined data item: the “nextLink”.
  - Select next, name it "NextLink"
  - Uncheck the connectivity wizard, since you already have pages from the site. (Don't worry, we will use this wizard again in a second...)
Method of Connector Definition
This step defines how the connector will be built.

Please select the method for building the new connector.

- Use a currently defined data item
- Use an example of the connector on a page
- Define each connector element individually
Our new connector

- Mouse over it and you can see its name, etc.
- Right click on it
  - "Add New Path" which gives you a little connector object.
  - Just connect it back into Wrapper2.
What about the detailsLink?

- What can we do?
  - Iterate through a full list, following next links
- What more do we want?
  - For each item in our list, get the details page
- What do we have so far?
  - detailsLink tells us where to go
- What’s needed?
  - Follow detailsLink to a page and extract stuff from it
Create new wrappers and hook them up

- New connector object, “detailsLink” out of Wrapper2.
- Select “Connectivity Wizard,” check “Create new Wrapper”
- ENSURE “Add new samples wizard” is checked
  - How many pages do you want to add?
  - Check “Only ___ Samples” since we only need a few
  - Pick them at random, do not allow duplicates → ensure as many cases as possible
Define the Next Navigation Target

This step defines the type of target wrapper for the current connector.

You have selected the connector called "detailsLink" for current wrapper "Wrapper2". Please specify the downstream wrapper for this connector using one of the following options:

- Create a new wrapper
- Select an existing wrapper
Add New Samples Wizard

Number of Samples To Fetch

This step determines the number of new sample pages to retrieve.

Please specify the number of sample pages that the connector will fetch:

- All remaining (no duplicates)
- Only [10] samples

- [ ] Pick them at random
- [ ] Allow duplicates

Note. Potential samples using this connector: 262
Samples already fetched using this connector: none

<< Back  Next >>  Cancel
We can navigate and get details pages now!

- What do we want?
  - Reply-to email address
  - Time stamp of post

- How do we define the schema?

- How do we train the extractor?
Finishing up…

- We can do all our extraction, but we’re not quite done
- Go into Agent Layout (View → Agent Layout)
- Click on the wand (“Generate the plan…”)
  - Save the agent (if you haven’t)
    - name the agent: Lastname_firstname_hwX
    - name the plan: Production
- My mistake: always forget to keep the name Production!
Things to remember

- Name your agents
  Lastname_Firstname_HWX
- If you don’t see a needed button, check which tab you are in
- Remember to name your plan Production!
Second Piece: Running Agents

- Agent Runner
  - Hosts and runs agents
Once you copy your agent…

- Start Agent Runner (if it’s not running)
  - Start → Programs → Fetch → Agent Runner → start Agent Runner (wait for it to start…)
- Agent Runner → Web Interface
  - Start → Programs → Fetch → Agent Runner → Web Interface
### Repository Administration

<table>
<thead>
<tr>
<th>Item</th>
<th>Refresh Repository</th>
<th>Reload Selected</th>
<th>Enable Selected</th>
<th>Disable Selected</th>
<th>Show Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>CraigslistCarsSF</td>
<td><img src="arrow-up-green" alt="Refresh" /></td>
<td><img src="clock" alt="Reload" /></td>
<td><img src="on" alt="Enable" /></td>
<td><img src="off" alt="Disable" /></td>
<td><img src="eye" alt="Show Filter" /></td>
</tr>
</tbody>
</table>

Copyright © 2003 [Fetch Technologies, Inc.](http://www.fetchtech.com) All rights reserved.
Enabling your agent

- In Admin tab
  - Click on Grey arrow, it will turn Green
  - Agent is now enabled
- Click Execution tab
  - If your agent is enabled, you will see it with a lightening bolt next to it
  - Click the lightening bolt to get to your entry connector starting point
Agent Runner - Microsoft Internet Explorer

Address: http://localhost:8080/agent/list_items_for_execution

Item Execution

Show Filter

Item

 craigslistcars.sf

Show Filter

Copyright © 2005 Fetch Technologies, Inc. All rights reserved.
Run it!

- Now you can add your inputs and hit run
- Agent returns XML:
  - Remember your XQuery!