Wrapper Learning Tools

Martin Michalowski
CSCI 548, Spring 2008
February 5th

Today's Topics
- What are wrapper tools?
- Why are they important?
  - The "Big Picture"
  - How are they made?
- Creating your own wrappers
  - Open Kapow

What are Wrapper Tools?
- "Free-form" information
- Structured "Queriable"

Why are they important?
- Allow computers access online information
- Easier manipulation of data
- Selection/Filtering
- Integration of multiple sources
- Can turn the web into a database
- Reduce the work required of a designer
  - Depending on the tool of course!

What are Wrapper Tools?
- "Free-form" information
- Customized Content

What are Wrapper Tools?
- Google
- Apple
- Skiing

Why are they important?

Book A Ski Trip With Red Lion Today. Incredible Packages For $189 Only!
Ski Packages. Complete Ski Vacations including Air, Hotel, Lift Tickets & Rentals.
Custom Ski Vacation Packages for Val Thorens Ski Resort 866-521-4179
The Big Picture

Data Integration Application

Web pages -> Integrate and Process data -> Web service

Wrapper Tools

Web service GUI application Etc.

How are they made? (In general)

- Attempt to find structure in a webpage
  - Regular expressions, DOM structure, etc.
- Exploit this structure to extract data
  - Look for a pattern, pull out information
- Allow you to define the format of extracted data
  - XML, HTML, etc.

Open Kapow

- [http://www.openkapow.com](http://www.openkapow.com)
- Wrapper generation tool
  - Robots/Agents
  - Relies on the DOM structure of a page
  - Mimics a person browsing the web
  - RoboMaker tool for visual generation
  - Ability to
    - Publish robots for everyone to use
    - Use other published robots
  - Three types of robots: RSS, REST, WebClips

Open Kapow: RSS Agent

- Grab top story from Digg
  - Modified to get all front page stories from Digg
- Extract title, url, description
- Show how to iterate, correct iterations
- Contend with errors (missing data)
- Publish online and call

Digg.com RSS Agent

Select the Agent type

Digg.com RSS Agent

Enter beginning URL
Digg.com RSS Agent
Run debugger to test your robot

Digg.com RSS Agent
Getting all page stories, not just the top one

Digg.com RSS Agent
Fix loop to only get data from relevant sections

Digg.com RSS Agent
Errors getting descriptions for all articles
Digg.com RSS Agent
Allow empty descriptions

Now we have an agent that gets all digg stories

Publish the agent to openkapow.com (account req')

Call agent from its URL

Open Kapow: REST Agent
- Snowforecast.com
- Take URL as input
- Go through all resorts and get resort info
- Use URL to load details page
  - Get weather
- Return as XML (or HTML, JSON, CSV)
- Add group name to divide into elements
- Publish and use

REST Agent
Create a REST web service robot
Define output variables

Loop through all resorts

Extract name, base, top, acres, run, snowfall

Load page from a URL on this page
REST Agent

Loop over both rows of weather

REST Agent

Loop over three columns in each row

REST Agent

Extract weather for each day

REST Agent
REST Agent
Test the agent

Alter agent to take a URL as input

Provide an initial URL

Publish agent and run it

Results, not divided into separate elements

Assign output group name to divide into elements
REST Agent
New results grouped into different elements

OpenKapow.com Resources
- Demos (most don’t work)
- Tutorials
- Other Robots/Agents
- You can download and edit agents
- Forums

Thank you!

Questions?