CSCI-548: Information Integration on the Web

Craig Knoblock

University of Southern California
Introduction

Information Integration

Integrating data from heterogeneous sources

Challenges:
- Accessing the data
- Resolving differences at the schema level
- Resolving differences at the data level
- Efficiently performing the integration
Introduction
...on the Web

- Web provides an incredible source of data

- However, new challenges arise:
  - Need to turn web pages into structured data
  - Don’t have control over the data
  - Sources have input/output constraints
  - Distributed nature of the web can make integration slow
Example Applications
Integrating Country Information

World Governments

NATO Members

Agent


CIA World Factbook
Predicting Flight Delays

Yahoo Weather

Historical Flight Data

Historical Weather Data

Learned Flight Delay Predictor

Agent

Prediction
Real Estate Notifications

New Listing:
3br 2bath
200K

Send Email Notification
TheaterLoc Entertainment Agent

- Tiger Map Server
- Etak Geocoder
- CuisineNet
- Zagat
- Yahoo Movies
- Hollywood.com Trailers

...
Travel Planning Assistant
Geospatial Data Integration
WorldInfo Assistant
Course Overview
XML

- XML widely used as an internet data interchange language
- Xquery - language for manipulating XML documents
- In this class I will cover the Xquery language
Wrappers

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

Appears in the Category: Restaurants

Jump to Top
Wrappers

Turning online sources into structured information

Research Topics

- Wrapper Learning
- Automatic Wrapper Generation
- Wrapper Maintenance

Tools

- AgentBuilder
- AgentRunner
Plan Execution

4676 Admiralty Way Marina del Rey CA

address

senators & house reps

recent news

combined results

Wrapper Vote-Smart

Select senators, house reps

Wrapper Yahoo News

Wrapper OpenSecrets (names page)

Wrapper OpenSecrets (member page)

Wrapper OpenSecrets (funding page)

Join name

graph URL

Boxer Anthrax investigation continues...
Boxer Bay area politicians meet...
Feinstein Bay area politicians meet...
Harman Life in LA is just too sunny...

Barbara Boxer
Dianne Feinstein
Jane Harman
George Bush
Dick Cheney
Barbara Boxer
Dianne Feinstein
Jane Harman
James Hahn

4676 Admiralty Way Marina del Rey CA

January 05 University of Southern California 16
Plan Execution

Research Topics

- Streaming dataflow execution systems
- Optimizing execution systems
  - Adaptive execution strategies
  - Speculative Execution

Tools

- Theseus agent execution system
Data Integration

- Mediator
- Outlook Server
- Timeline Server
- CDW
- Yahoo Laptops

Local sources & services
Remote sources & services
Data Integration Systems

- Information mediators
  - Used to automatically select and compose information across sources

- Research Topics
  - Global-as-view vs. Local-as-view integration
  - Optimizing query plans

- Tools
  - Prometheus information mediator
Record Linkage

How can the same objects be identified when they are stored in inconsistent text formats?
Record Linkage

- Align information across sources
- Research Topics:
  - Matching individual attributes
  - Matching entire records
- Tools
  - Apollo Record Linkage System
Aligning Schemas and Ontologies

**Mediated schema**

<table>
<thead>
<tr>
<th>price</th>
<th>agent-name</th>
<th>agent-phone</th>
<th>office-phone</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>listed-price</td>
<td>contact-name</td>
<td>contact-phone</td>
<td>office</td>
<td>comments</td>
</tr>
</tbody>
</table>

**Schema of realestate.com**

<table>
<thead>
<tr>
<th>listed-price</th>
<th>contact-name</th>
<th>contact-phone</th>
<th>office</th>
<th>comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>$250K</td>
<td>James Smith</td>
<td>(305) 729 0831</td>
<td>(305) 616 1822</td>
<td>Fantastic house</td>
</tr>
<tr>
<td>$320K</td>
<td>Mike Doan</td>
<td>(617) 253 1429</td>
<td>(617) 112 2315</td>
<td>Great location</td>
</tr>
</tbody>
</table>

**homes.com**

<table>
<thead>
<tr>
<th>sold-at</th>
<th>contact-agent</th>
<th>extra-info</th>
</tr>
</thead>
<tbody>
<tr>
<td>$350K</td>
<td>(206) 634 9435</td>
<td>Beautiful yard</td>
</tr>
<tr>
<td>$230K</td>
<td>(617) 335 4243</td>
<td>Close to Seattle</td>
</tr>
</tbody>
</table>

If “fantastic” & “great” occur frequently in data instances => description

If “office” occurs in name => office-phone
Aligning Schemas and Ontologies

- Given two different sources with different schemas, how do we automatically align the information?

- Research Topics
  - Automatic schema alignment based on structure and naming
  - Automatic alignment based on the source contents
Constraint Integration
Constraint Integration Frameworks

- Approach to tightly integrating closely related sources

- Research:
  - Constraint propagation and constraint satisfaction techniques

- Tools
  - Heracles constraint integration system
Geospatial Data Integration

- Satellite Image Terraserver
- Street Vector Data Corrected Tiger Line Files
- Constraint Satisfaction
- Initial Hypothesis Result After Constraint Satisfaction
- Geocoded Houses

Census Master Address File

Los Angeles County Assessor’s Site Property Tax Records

Data Extracted from Online Site

<table>
<thead>
<tr>
<th>Address</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>642 Penn St</td>
<td>33.923413</td>
<td>-118.409809</td>
</tr>
<tr>
<td>640 Penn St</td>
<td>33.923412</td>
<td>-118.409809</td>
</tr>
<tr>
<td>636 Penn St</td>
<td>33.923412</td>
<td>-118.409809</td>
</tr>
<tr>
<td>604 Palm Ave</td>
<td>33.923414</td>
<td>-118.409810</td>
</tr>
<tr>
<td>610 Palm Ave</td>
<td>33.923414</td>
<td>-118.409810</td>
</tr>
<tr>
<td>604 Sierra St</td>
<td>33.923415</td>
<td>-118.409810</td>
</tr>
<tr>
<td>639 Sierra St</td>
<td>33.923412</td>
<td>-118.409810</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Address</th>
<th># units/</th>
<th>Area (sq ft)</th>
<th>Lot size</th>
</tr>
</thead>
<tbody>
<tr>
<td>642 Penn St</td>
<td>3</td>
<td>1793</td>
<td>135.72 * 53.33</td>
</tr>
<tr>
<td>610 Palm Ave</td>
<td>1</td>
<td>884</td>
<td>69 * 42</td>
</tr>
<tr>
<td>624 Penn</td>
<td>1</td>
<td>629</td>
<td>62, Sierra</td>
</tr>
<tr>
<td>639 Sierra St</td>
<td>1</td>
<td>1408</td>
<td>121 * 53.5</td>
</tr>
</tbody>
</table>

610, Palm or 645, Sierra
645, Sierra or 639, Sierra
633, Sierra or 629, Sierra
628, Sierra or 623, Sierra
610 or 610
642 or 610
642, Penn or 636, Penn
610, Palm or 645, Sierra
645, Sierra or 639, Sierra
633, Sierra or 633, Sierra
639, Sierra or 633, Sierra
642, Penn
628, Penn
624, Penn
623, Sierra
624, Penn
623, Sierra
642, Penn St El Segundo, CA 90245
640 Penn St El Segundo, CA 90245
636 Penn St El Segundo, CA 90245
604 Palm Ave El Segundo, CA 90245
610 Palm Ave El Segundo, CA 90245
645 Sierra St El Segundo, CA 90245
639 Sierra St El Segundo, CA 90245

Address Latitude Longitude
642 Penn St 33.923413 -118.409809
640 Penn St 33.923412 -118.409809
636 Penn St 33.923412 -118.409809
604 Palm Ave 33.923414 -118.409810
610 Palm Ave 33.923414 -118.409810
645 Sierra St 33.923413 -118.409810
639 Sierra St 33.923412 -118.409810

- Los Angeles County Assessor’s Site Property Tax Records
- Data Extracted from Online Site

- Palm Ave
- Sierra St
- Mariposa Ave
- Penn St
Application Areas

- Geospatial data integration
  - Includes satellite imagery, maps, vector data and many related online sources

- Biological data integration
  - Huge number of sources on gene-related information
  - Many sources available as web services

- In this course we will focus on the first application area
And other topics

- Semantic Web
- Data mining from the Web
- Information extraction
Course Details
Where to find me...

Research Associate Professor
Computer Science Department
PHE 416 (Only for office hour after class)

Senior Project Leader
Information Sciences Institute
Marina del Rey
ISI 922 (Office the rest of the time)
TA, Grader & Office Hours

❖ Professor: Craig Knoblock (Knoblock@isi.edu)
  ▲Office Hours:
    ✗Tuesday 5-6pm (PHE 416)
    ✗Thursday 3–4pm (ISI 922 or 310-448-8786)
❖ TA: Martin Michalowski (martinm@isi.edu)
  ▲Office Hours: Monday 1-2:30pm (SAL 200c)
❖ TA: Anshuman Chakravartty (achakrav@usc.edu)
  ▲Office Hours: (all in SAL 200c)
    ✗Tue: 11-12:30pm, Wed: 1-2:30pm, Th: 10-11:30pm, Fri: 2-3:30pm
❖ Grader: Junaid Chaudhry (chaudhry@isi.edu)
Course Web Pages

- Blackboard – totale.usc.edu
  - Your USC login works on this account
  - If you are registered for 548, you will have access
- All readings, slides, homeworks, etc will be posted on the site page
- Please check for announcements and read the discussion board on a regular basis
- All questions should be posted (not emailed!)
  - If you know the answer to a posted question, please answer it!
  - But please don’t post answers to homeworks!
Prerequisites & Recommendations

Prerequisites
- CS561 or CS573 -- Introduction to AI
- CS585 – Database Systems

Recommended Courses
- CS571 – Issues of Programming Language Design
- CS573 – Advanced AI
Grading

- **Homework:** 24%
  - 8 homework assignments – 3pts each
  - Must be turned in the week they are due
  - Partial credit for one week extension only
- **Course project:** 35%
- **Quizes:** 11% (1pt per quiz)
  - First or last 10 minutes of every class (don’t be late)
  - There are no makeups if you miss the quiz
- **Final Exam:** 30%
  - Final: May 3, 2-4pm (Check for conflicts!)
More on Grading

- This is a hard class!
  - Lots of very technical reading – there is no good textbook
  - Lots of homework
  - Quizzes every week
  - Final exam and course projects

- I do give B’s and C’s

- Grade distribution will be roughly half A’s and B’s (I consider a C a failing grade)

- If you get 90pts or more you will definitely get an A
Readings

Posted on the site each week

You can read it online or print them

Try to read all required readings before the class they are covered

Quizzes may cover material that is only presented in the readings!
Slides

* Available online by midnight of the day before the lecture
* These are not intended as a replacement for the lecture
* You can print these out and make notes on them
  * I suggest you print 6 slides per page to save paper
Course Lab – SAL 200c

- Microsoft Instructional Lab – SAL 200c
- Lab fee: $175
- You must pay the fee even if you don’t use the lab – if you don’t think this is fair, don’t take the course
- All registered students should have an account
- Shared with other courses, so plan ahead
- You are encouraged to use your own computers, but you will need Windows 2000 or XP for wrapper tools
- TAs will hold office hours in the lab
Working Together

- Each person must do their own homework
  - We will check for overlap in homeworks
  - If we find any plagiarism, all parties lose credit so
    - Don’t share your answers
    - Don’t leave printouts in the trash with your answers
    - Don’t give out your password
    - Don’t copy others (they may have the wrong answer anyway!)

- You can ask the TAs for help

- Encouraged to work in pairs on the course project
  - Both students must participate and present
  - Expectations are the same for individual and joint projects
Cheating

- Not tolerated!
- No second chances – all infractions will be reported
- Examples:
  - Turning in someone else’s homework
  - Copying from someone else during a quiz or exam
  - Doing a project that uses someone else’s work without giving them credit
Cell Phone Use

- If it makes noise, turn it off in class
Quizes & Exams

- The quizes and exam will cover the material in the lectures and the readings
- Format: problems and short answers
- If you keep up with the readings and participate in class, the exams won’t be too hard
- Timing:
  - Quizes: first or last 10 minutes of each class
  - Final: 2 hours
Course Projects

- Information integration project based on what you learned in class
- Be creative!
- An ideal project would be one you could publish a paper about
- Four components to this project:
  - Proposal
  - Demonstration (presented in SAL 200c)
  - Presentation (short presentation to the entire class)
  - Paper (written in the form of a conference paper)
    - 6-8 pages
Grading of Projects

[Overall: 35%
  Proposal 5%
  Paper 10%
  Demo 5%
  Presentation 5%
  Applied techniques learned from class 5%
  Innovation and creativity 5%]

[Written proposals: March 1 at 1:50pm (submit online)
Proposal presentations: April 19 and 26
Demos due on date of presentation
Papers dues on April 26]
Project Presentation

- Presentations are either posters or Powerpoint presentations to the class
- I will determine posters or presentations based on your project proposal
- You can still get full credit for a poster, but everyone should try to get a presentation
- This is the same thing that happens at conferences
Example Projects

▷ A system that took an arbitrary web page and built a map showing all of the locations
▷ A meta comparison shopping engine
▷ A real estate notification agent that emailed a satellite image and map of the property
▷ A new approach for linking records across sources
▷ An improvement on an algorithm that we learn about in class
▷ An empirical evaluation of different approach to some task
When the Course is Over

- Directed research (1-2 MS or PhD Students)
- M.S. Thesis
- Summer interns (MS or PhD)
- Research Assistantships (1-2 PhD Students)
  - I can also recommend you for positions in other groups
- Teaching Assistantships (both PhD & MS)
- Recommendation letters (anyone that gets at least an A-)
- Positions at related companies
  - Last year Fetch Technologies hired two students that took the course in the past and made an offer to a third