CSCI-548: Information Integration on the Web

Craig Knoblock
and
Kristina Lerman

University of Southern California
Introduction

Information Integration

Information Integration

- Integrating data from heterogeneous sources

- Challenges:
  - Accessing the data
  - Resolves differences at the schema level
  - Resolving differences at the data level
  - Efficiently performing the integration
Introduction

...on the Web

- Web provides an incredible sources 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 of Different Types of Integrated Applications
Integration to Analyze Data

World Governments

NATO Members

Agent

CIA World Factbook

1995

1996

1997
Mining Integrated Data to Predict Flight Delays

Yahoo Weather

Historical Flight Data

Historical Weather Data

Agent

Learned Flight Delay Predictor

Prediction
Monitoring Online Sources to Provide Real-time Notification

Send Email Notification

New Listing: 3br 2bath 200K
Integrating Diverse Sources to Provide a Unified View

Tiger Map Server

Etak Geocoder

CuisineNet

Zagat

Yahoo Movies

Hollywood.com Trailers

Agent
Integration to Support Planning
Integration for Geospatial Visualization

ARIADNE Mediator

THESEUS Executor

Area of Interest

ESRI Street Data

Wrapper

Sheet name, address

Wrapper

Mailing Address

Buildings in Cambridge, MA

Chet Henri,
1 Shepard St,
Cambridge, MA
02139
Phone: (617) 354-8990
Lat: 42.3818
Lon: -71.1205
Reviews
### Integration for Decision Making

<table>
<thead>
<tr>
<th>Washington DC Area Hotels List</th>
<th>0</th>
<th>10/14/05 4:34 pm</th>
<th>Jennann</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bid Assistance 4th WH, Dupont-Woodley 4th 11/9-11/11</td>
<td>0</td>
<td>10/17/05 12:20 am</td>
<td>whidbeyone</td>
</tr>
<tr>
<td>Bidding assistance 11/17-11/20 Dupont Circle</td>
<td>6</td>
<td>10/15/05 5:58 pm</td>
<td>CreamandCrimson</td>
</tr>
<tr>
<td>J.W. Marriott: WH/DT 10/9-10/11 $112</td>
<td>8</td>
<td>10/15/05 3:08 pm</td>
<td>nancyrea</td>
</tr>
<tr>
<td>Bidding assistance -Washington D/T/WH Feb 18,2006</td>
<td>5</td>
<td>10/14/05 3:11 pm</td>
<td>pruten45</td>
</tr>
</tbody>
</table>

**Map**

June 09 University of Southern California

**BiddingForTravel.com**

**Priceline**

**Orbitz**

**Quick Search**

**Flights**

**Hotels**
Course Overview
XML

- XML widely used as an internet data interchange language
- Xquery – language for manipulating XML documents
- This course assumes that you already know XML and Xquery
Wrapper Generation

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

- Turning online sources into structured information
- Research Topics
  - Wrapper Learning
  - Automatic Wrapper Generation
  - Wrapper Maintenance
- Tools
  - Dapper
  - OpenKapow
Information Extraction (IE)

Example:

“1988 Honda Accord for sale! Only 80k miles, Runs Like New, V6, 2WD... $2,500 obo. SUPER DEAL.”
Information Extraction

- How to find the structure in unstructured text

- Research Topics
  - Extraction using NLP techniques
  - Extraction with Conditional Random Fields
  - Exploiting reference sets for extraction
Data Integration

Mediator

Outlook Server

CDW

Yahoo Laptops

Mediator

Timeline Server

Local sources & services

Remote sources & services
Data Integration

- 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
  - IBM Data Integrator
Semantic Web
Semantic Web

- How do we create a semantic layer on the web

Research Topics
- Organizing knowledge
- Adding a semantic layer to the web
- Reasoning and querying over the semantic web data
Dataflow Execution

Wrapper Vote-Smart → Select senators, house reps → Wrapper Yahoo News → Wrappers
OpenSecrets (names page) → Wrapper OpenSecrets (member page) → Wrapper OpenSecrets (funding page) → Join name

- Barbara Boxer
- Dianne Feinstein
- Jane Harman
- Boxer
- Feinstein
- Harman

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

address → senators & house reps → combined results

recent news

graph URL

member URL

funding URL

4676 Admiralty Way Marina del Rey CA

June 09 University of Southern California 22
Dataflow Execution

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

- Tools
  - Theseus agent execution system
### Record Linkage

**Zagat’s Restaurant Guide Source**

- Art’s Deli
- California Pizza Kitchen
- Campanile
- Citrus
- Grill, The
- Philippe The Original
- Spago

**Department of Health Restaurant Source**

- Art’s Delicatessen
- Ca’ Brea
- CPK
- The Grill
- Patina
- Philippe’s The Original
- The Tillerman

---

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

- Align information across sources

Research Topics:
- Blocking
- Matching individual attributes
- Matching entire records
### Aligning Schemas and Modeling Sources

**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 “office” occurs in name => office-phone

If “fantastic” & “great” occur frequently in data instances => description
Aligning Schemas and Modeling Sources

- Given two different sources with different schemas, how do we automatically align the information?
- Given a new source, how do we construct a model of the source for integration?

Research Topics

- Automatic schema alignment based on structure and naming
- Automatic alignment based on the source contents
- Automatic modeling of the inputs/outputs and function of a source or service
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 Fusion

Imagery of Interest

Available raster maps

Imagery with aligned layers
Geospatial Data Fusion

- How do we integrate data across geospatial sources
- Research topics
  - Map search and extraction
  - Geospatial source registration and alignment
  - Integrating text documents with imagery
  - Constraint satisfaction for geospatial reasoning
And other topics

- Intellectual Property
- Mashup Construction
- Social Networking
- Folksonomies
Course Details
Where to find me...

- Research Professor
  Computer Science Department
  Outside THH 212 (Immediately before and after class)

- Senior Project Leader
  Information Sciences Institute
  Marina del Rey
  ISI 922 (by appointment)
  310-448-8786

- Email: knoblock@isi.edu
Where to find Professor Lerman...

- Research Assistant Professor
  Computer Science Department
  Outside THH 212 (Immediately BEFORE class)

- Project Leader
  Information Sciences Institute
  Marina del Rey
  ISI (by appointment)
  310-448-8714

- Email: lerman@isi.edu
TA & Grader Office Hours

- TA: Yao-Yi Chiang (yaoyichi@isi.edu)
  - Office Hours: Wednesdays and Fridays
    11am-12pm
  - Location: TBD
- Grader: TBD
  - Office Hours: TBD
Course Web Pages

- Blackboard – blackboard.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 try to provide helpful suggestions
  - 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: 20%
  - 8 homework assignments – 2.5pts each
  - Must be turned in the week they are due
  - Partial credit for one week extension only (1 pt off)

- Course project: 30%

- Quizes: 20% (2 pts per quiz)
  - Beginning or end of each Monday class (sometimes Wednesday instead)
  - There are no make ups if you miss the quiz, but there will be 11 quizzes and you can drop the lowest grade

- Final Exam: 30%
  - Final: Friday, May 8, 2-4pm (Check for conflicts!)
More on Grading

- This is a hard class, but you will learn a lot!
  - Lots of technical reading – there is no good textbook
  - Lots of homework
  - Quizzes every week
  - Final exam and course projects
- I do give B’s, C’s, and even D’s and F’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
- Please read all required readings before the class they are covered
- Quizzes may cover lectures, readings, and/or homeworks
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

- No lab this year!
- This means you don’t have to pay a lab fee.
- But it does mean that you need to find access to a computer where you can install software
  - Not possible in the USC controlled labs
- If you don’t have access to a computer you can use for the homeworks, please see me right away
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
- You can work in pairs on the course project
  - Both students must participate and present
Cheating

- Not tolerated!
- No second chances – all infractions will be reported
  - First offense is automatic failure in the class
  - Second offense is suspension from the University
- 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 or to vibrate mode in class
Quizes & Exams

- The quizzes 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 have learned in class
- Be creative!
- An ideal project is one that you could publish a paper about
  - I have had several students turn projects into published papers
  - One even won a best paper award!
- Four components to a project:
  - Proposal
  - Demonstration (you will make a video, which you can use in your presentation)
  - Presentation (either an oral or poster presentation for the class)
  - Paper (written in the form of a conference paper)
Grading of Projects

- Overall: 30%
  - Proposal – no grade, but determines oral vs. poster presentation
  - Presentation – 10%
  - Demonstration – 10%
  - Paper – 10%
    - Projects graded on innovation, creativity, applied ideas learned in class, etc…

- Written proposals: March 11 @3pm (submit online)
- Project presentations: April 20, 22, 27
- Demonstrations submitted online on day of presentation
- All papers due on May 1 @ midnight
Project Presentation

- Presentations are either posters or oral presentations to the class.
- We will determine whether it will be an oral or poster presentation based on your proposal.
- You can still get full credit for a poster, but you should aim to get an oral presentation.
- This is the same thing that happens at conferences.
Example Projects

- A new approach to wrapping structured web sites
- An extension to one of the tools for extracting data from text documents
- An evaluation and detailed analysis of the various tools for automatically modeling sources
- A novel approach to linking data across sources
- A semantic web implementation that integrates data across various apartment sources
- An extension to one of the mashup building tools
- Anything that builds on or extends the ideas and tools that we cover in class…be creative!
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 (for PhD students)
- Recommendation letters (anyone that gets at least an A-)
- Positions at related companies
  - Fetch Technologies has hired a number of students that took the course in the past
  - Other companies are often looking for students