CSCI-548: Information Integration on the Web

Craig Knoblock

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 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

1995

1996

1997
Predicting Flight Delays

Yahoo Weather

Historical Flight Data

Historical Weather Data

Agent

Learned Flight Delay Predictor

Prediction
Real Estate Notifications

New Listing:
3br 2bath
200K

Send Email Notification
TheaterLoc Entertainment Agent

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

Diagram showing the flow of information and services.
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
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
  - EasyBuilder from Fetch Technologies
  - Simile/PiggyBank
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

CDW

Mediator

Timeline Server

Yahoo Laptops

Outlook 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
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

Join

Wrapper OpenSecrets (names page)
Wrapper OpenSecrets (member page)
Wrapper OpenSecrets (funding page)

Barbara Boxer
Dianne Feinstein
Jane Harman

Boxer
Boxer
Feinstein
Harman

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

address

senators & house reps

recent news

combined results

all officials

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

4676 Admiralty Way Marina del Rey CA

funding URL

member URL

graph URL
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

listed-price | contact-name | contact-phone | office | comments

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 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 Integration

Street Vector Data
Corrected Tiger Line Files

Constraint Satisfaction

Initial Hypothesis
Result After Constraint Satisfaction

Geocoded Houses

Los Angeles County Assessor’s Site
Property Tax Records

Data Extracted from Online Site

Census Master Address File

Satellite Image
Terraserver

Street Address City, State Zipcode
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
Geospatial Data Integration

- How do we integrate data across geospatial sources

- Research topics
  - Fusion different geospatial layers
  - Finding and identifying geospatial data
  - Reasoning with the geospatial sources
And other topics

- Intellectual Property
- Data mining from the Web
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: 26%

* Quizzes: 20% (2 pts per quiz)
  - Last 10 minutes of every class
  - There are no makeups if you miss the quiz

* Final Exam: 30%
  - Final: Thursday, Dec 7, 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 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
- 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 – SAL 200c

- Microsoft Instructional Lab – SAL 200c
- Lab fee: $175
- You must pay the fee even if you don’t use the lab – this makes it possible to do the hands on homeworks
- 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
- TA 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
💖 You will work in groups on the course project
   ✗ All 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 expulsion 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 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: last 10 minutes of each class
  - Final: 2 hours
Course Project

- A single unified class project to build mashups for finding apartments
  - Divide the class up into teams
  - Each team will work on one of the aspects that we cover in class
  - Next week I’ll ask you to select the topics you want to work on
Mashup (from the Wikipedia)

- A **mashup** is a website or web application that seamlessly combines content from more than one source into an integrated experience.
- Content used in mashups is typically sourced from a third party via a public interface or API.
Example Mashup for Bidding on Priceline

<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 DT-WH, Dupont-Woodley 4&quot; 11/5-11/11</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 am</td>
<td>CreamandCrimson</td>
</tr>
<tr>
<td>4&quot; J.W. Marriott WH/DT 10/9-10/11 $112</td>
<td>8</td>
<td>10/15/05 3:00 pm</td>
<td>nancyrea</td>
</tr>
<tr>
<td>Bidding assistance -Washington DT/WH Feb 18,2006</td>
<td>5</td>
<td>10/14/05 3:11 pm</td>
<td>pruitne45</td>
</tr>
</tbody>
</table>

BiddingForTravel.com

Map

December 06

University of Southern California
alcmenetest

address

2799 Jefferson Davis Highway

city

Crystal City

state

Virginia

zipcode

22202

checkin

11/04/05

checkout

11/06/05

BiddingForTravel
<table>
<thead>
<tr>
<th>area</th>
<th>star</th>
<th>hostname</th>
<th>biddingfortravel_price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crystal City</td>
<td>3</td>
<td>Marriott Crystal City</td>
<td>$70</td>
</tr>
<tr>
<td>Crystal City</td>
<td>3</td>
<td>Sheraton Crystal City</td>
<td>$72</td>
</tr>
<tr>
<td>Crystal City</td>
<td>3</td>
<td>Hilton Crystal City</td>
<td>$65</td>
</tr>
<tr>
<td>Crystal City</td>
<td>2.5</td>
<td>Courtyard Marriott Crystal City</td>
<td>$49</td>
</tr>
<tr>
<td>Crystal City</td>
<td>3</td>
<td>Crystal Gateway Marriott</td>
<td>$57</td>
</tr>
<tr>
<td>Crystal City</td>
<td>3</td>
<td>Crown Plaza</td>
<td>$75</td>
</tr>
<tr>
<td>Crystal City</td>
<td>3</td>
<td>Hyatt Regency Crystal City</td>
<td>$85</td>
</tr>
<tr>
<td>Crystal City</td>
<td>3</td>
<td>Doubletree Crystal City</td>
<td>$60</td>
</tr>
<tr>
<td>Hotel Name</td>
<td>Whole Address</td>
<td>Orbitz</td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>--------------------------------------------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>Marriott Crystal City</td>
<td>1999 Jefferson Davis Highway, City</td>
<td>$119</td>
<td></td>
</tr>
<tr>
<td>Sheraton Crystal City</td>
<td>1800 Jefferson Davis Highway, City</td>
<td>$149</td>
<td></td>
</tr>
<tr>
<td>Hilton Crystal City</td>
<td>2399 Jefferson Davis Highway, City</td>
<td>$139</td>
<td></td>
</tr>
<tr>
<td>Courtyard Marriott City</td>
<td>2899 Jefferson Davis Highway, City</td>
<td>$109</td>
<td></td>
</tr>
<tr>
<td>Crystal Gateway Man</td>
<td>1700 Jefferson Davis Highway, City</td>
<td>$149</td>
<td></td>
</tr>
<tr>
<td>Crown Plaza</td>
<td>1480 Crystal Drive, Crystal City, City</td>
<td>$89</td>
<td></td>
</tr>
<tr>
<td>Hyatt Regency Crystal</td>
<td>2799 Jefferson Davis Highway, City</td>
<td>$119</td>
<td></td>
</tr>
<tr>
<td>Doubletree Crystal</td>
<td>300 Army Navy Drive, Crystal City, City</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>lat</td>
<td>lon</td>
<td>hotelname</td>
<td>wholeaddress</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>38.8556</td>
<td>-77.051</td>
<td>Marriott Cry</td>
<td>1999 Jefferson Dr. 3</td>
</tr>
<tr>
<td>38.8575</td>
<td>-77.051</td>
<td>Sheraton Cr</td>
<td>1800 Jefferson Dr. 3</td>
</tr>
<tr>
<td>38.8523</td>
<td>-77.052</td>
<td>Hilton Crystl</td>
<td>2399 Jefferson Dr. 3</td>
</tr>
<tr>
<td>38.8461</td>
<td>-77.052</td>
<td>Courtyard M</td>
<td>2899 Jefferson Dr. 2.5</td>
</tr>
<tr>
<td>38.8583</td>
<td>-77.051</td>
<td>Crystal Gate</td>
<td>1700 Jefferson Dr. 3</td>
</tr>
<tr>
<td>38.8607</td>
<td>-77.049</td>
<td>Crown Plaza</td>
<td>1480 Crystal Dr. 3</td>
</tr>
<tr>
<td>38.8484</td>
<td>-77.052</td>
<td>Hyatt Regency</td>
<td>2799 Jefferson Dr. 3</td>
</tr>
<tr>
<td>38.8635</td>
<td>-77.052</td>
<td>Doubletree</td>
<td>300 Army Navy Dr. 3</td>
</tr>
</tbody>
</table>
Goal of Course Project

- Create a application for creating mashups for finding apartments
  - Create the tools for rapidly creating a apartment mashup for anywhere in the US
  - Each individual mashup will:
    - Pull apartment data from structured sites (e.g., Craig’s List) as well as unstructured sources such as newspapers
    - Extract the details of the listings
    - Geocode the addresses
    - Place the listings on the maps
    - Provide related information such as restaurants, schools, crime stats
Project Topics

- Wrapping structured web sites
- Extraction of the apartment data
- Modeling the data from the extracted sources
- Linking of data across sources
- Integration across sources
- Geospatial integration
- Reasoning and querying using the Semantic Web
- Integrating the components
Grading of Projects

Projects are worth 26% of your grade

Grade based on:

- Overall project
- Team’s contribution
  - Proposal
  - Software component
  - Class presentation (in the last two classes)
  - Project report (on the web)
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