

CSCI-548: Information Integration on the Web

Craig Knoblock

University of Southern California

August 08

University of Southern California

1

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

August 08

University of Southern California

2

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

August 08

University of Southern California

3

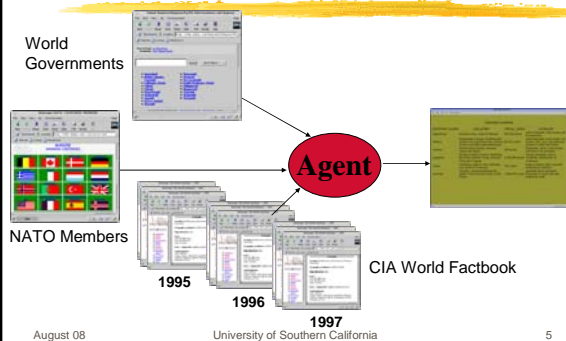
Example of Different Types of Integrated Applications

August 08

University of Southern California

4

Integration to Analyze Data

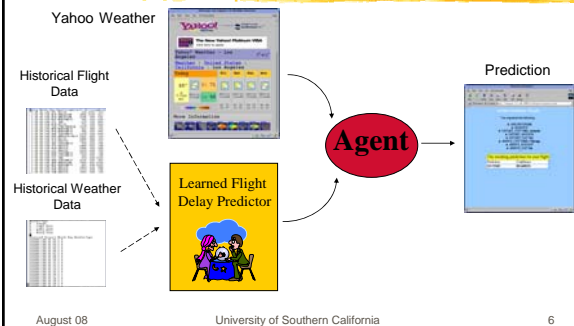


August 08

University of Southern California

5

Mining Integrated Data to Predict Flight Delays



August 08

University of Southern California

6

Monitoring Online Sources to Provide Real-time Notification

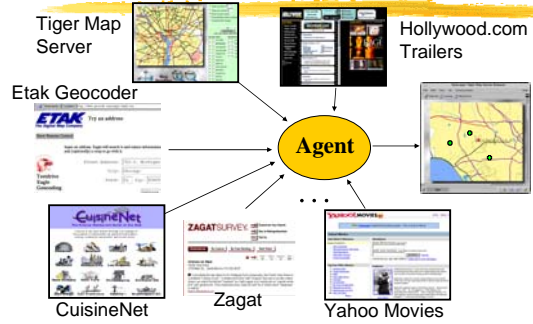


August 08

University of Southern California

7

Integrating Diverse Sources to Provide a Unified View



August 08

University of Southern California

8

Integration to Support Planning

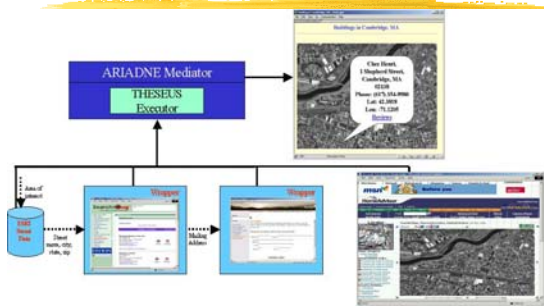


August 08

University of Southern California

9

Integration for Geospatial Visualization

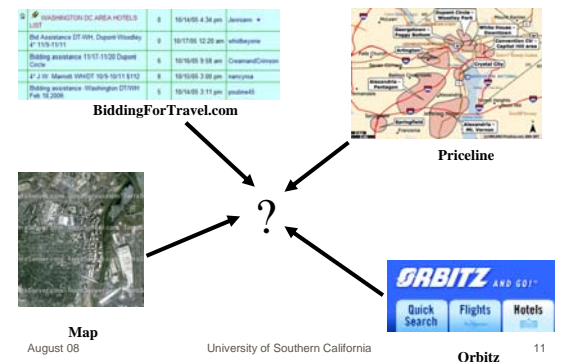


August 08

University of Southern California

10

Integration for Decision Making



August 08

University of Southern California

11

Course Overview

August 08

University of Southern California

12

XML

- ⌘ XML widely used as an internet data interchange language
- ⌘ Xquery – language for manipulating XML documents
- ⌘ In today's class we will cover Xquery

August 08

University of Southern California

13

Wrapper Generation



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

August 08

University of Southern California

14

Wrapper Generation

- ⌘ Turning online sources into structured information
- ⌘ Research Topics
 - ☑ Wrapper Learning
 - ☑ Automatic Wrapper Generation
 - ☑ Wrapper Maintenance
- ⌘ Tools
 - ☑ Dapper
 - ☑ Simile

August 08

University of Southern California

15

Information Extraction (IE)

Example:

"1988 Honda Accord for sale! Only 80k miles, Runs Like New, V6, 2WD... \$2,500 obo. SUPER DEAL."

August 08

University of Southern California

16

Information Extraction

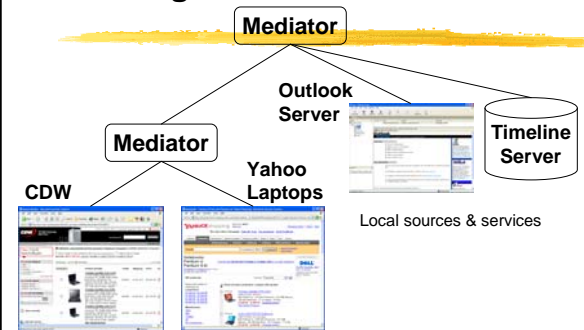
- ⌘ How to find the structure in unstructured text
- ⌘ Research Topics
 - ☑ Extraction using NLP techniques
 - ☑ Extraction with Conditional Random Fields
 - ☑ Exploiting reference sets for extraction

August 08

University of Southern California

17

Data Integration



August 08

University of Southern California

18

Data Integration

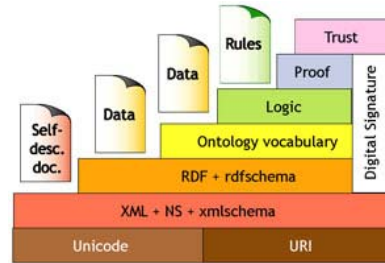
- ⌘ Information mediators
 - ☑ Used to automatically select and compose information across sources
 - ☑ Research Topics
 - ☑ Global-as-view vs. Local-as-view integration
 - ☑ Optimizing query plans
- ⌘ Tools
 - ☑ Prometheus information mediator
 - ☑ IBM Data Integrator

August 08

University of Southern California

19

Semantic Web



August 08

University of Southern California

20

Semantic Web

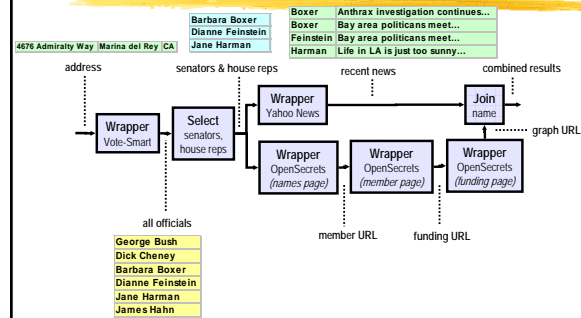
- ⌘ How do we create a semantic layer on the web
- ⌘ Research Topics
 - ☑ Organizing knowledge
 - ☑ Adding a semantic layer to the web
 - ☑ Reasoning and querying over the semantic web data

August 08

University of Southern California

21

Dataflow Execution



August 08

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

August 08

University of Southern California

23

Record Linkage



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

Record Linkage

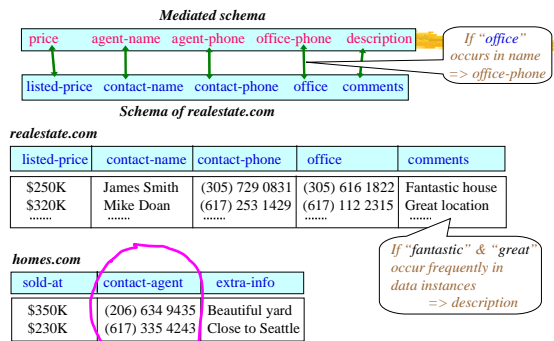
- ⌘ Align information across sources
- ⌘ Research Topics:
 - ☑ Blocking
 - ☑ Matching individual attributes
 - ☑ Matching entire records

August 08

University of Southern California

25

Aligning Schemas and Modeling Sources



August 08

University of Southern California

26

Aligning Schemas and Modeling Sources

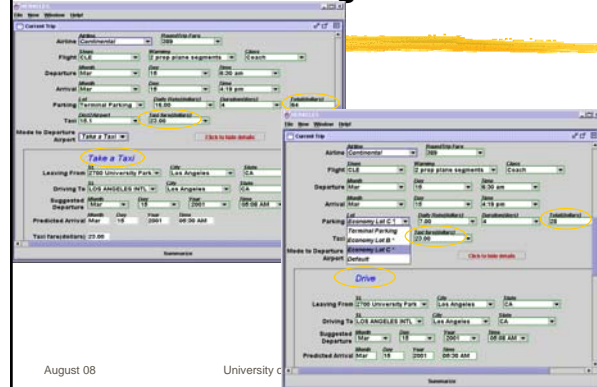
- ⌘ Given two different sources with different schemas, how do we automatically align the information
- ⌘ Given a new source how do we construct a model of the source for integration
- ⌘ Research Topics
 - ☑ Automatic schema alignment based on structure and naming
 - ☑ Automatic alignment based on the source contents
 - ☑ Automatic modeling of the inputs/outputs and function of a source or service

August 08

University of Southern California

27

Constraint Integration



August 08

University of Southern California

Constraint Integration Frameworks

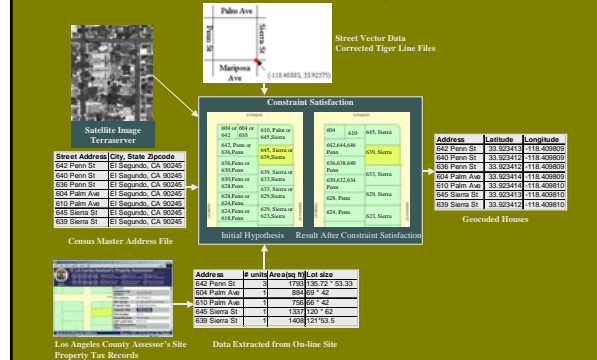
- ⌘ Approach to tightly integrating closely related sources
- ⌘ Research:
 - ☑ Constraint propagation and constraint satisfaction techniques
- ⌘ Tools
 - ☑ Heracles constraint integration system

August 08

University of Southern California

29

Geospatial Data Integration



Geospatial Data Integration

- ⌘ 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

August 08

University of Southern California

31

And other topics

- ⌘ Intellectual Property
- ⌘ Mashup Construction
- ⌘ Social Networking

August 08

University of Southern California

32

Course Details

August 08

University of Southern California

33

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

August 08

University of Southern California

34

TA, Grader & Office Hours

- ⌘ Professor: Craig Knoblock (Knoblock@isi.edu)
 - ☑ Office Hours:
 - ☑ Tuesdays before and after class (Outside THH 212)
 - ☑ By Appointment (ISI 922 or 310-448-8786)
- ⌘ TA: Anon Plangprasopchok (plangpra@isi.edu)
 - ☑ Office Hours: Mondays and Wednesdays
1-2pm in SAL 100
- ⌘ TA: Rattapoom Tuchinda (pipet@isi.edu)
 - ☑ Office Hours: TBD

August 08

University of Southern California

35

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!

August 08

University of Southern California

36

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

⌘ Course project: 30%

⌘ Quizzes: 20% (2 pts per quiz)

- ☑ Last 10 minutes of every class
- ☑ There are no make ups if you miss the quiz

⌘ Final Exam: 30%

- ☑ Final: Tuesday, May 13, 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

⌘ 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

- ☑ No 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 loose 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

August 08

University of Southern California

43

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

August 08

University of Southern California

44

Cell Phone Use

- ⌘ If it makes noise, turn it off or to vibrate mode in class

August 08

University of Southern California

45

Quizzes & 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:
 - ☒ Quizzes: last 10 minutes of each class
 - ☒ Final: 2 hours

August 08

University of Southern California

46

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 a oral or poster presentation for the class)
 - ☒ Paper (written in the form of a conference paper)

August 08

University of Southern California

47

Grading of Projects

- ⌘ Overall: 30%
 - ☒ Proposal – no grade, but determines oral vs. poster presentation
 - ☒ Presentation – 5%
 - ☒ Demonstration – 5%
 - ☒ Paper – 10%
 - ☒ Project overall – 10%
 - ☒ Innovation, creativity, applied ideas learned in class, etc...
- ⌘ Written proposals: March 4 @3pm (submit online)
- ⌘ Project presentations: April 22 & 29
- ⌘ Demonstrations submitted online on day of presentation
- ⌘ All papers due on May 2 @ midnight

August 08

University of Southern California

48

Project Presentation

- ⌘ Presentations are either posters or oral presentations to the class
- ⌘ I 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

August 08

University of Southern California

49

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!

August 08

University of Southern California

50

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

August 08

University of Southern California

51