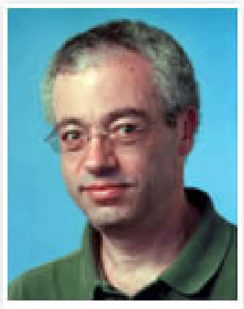
Information Integration on the Web (CS 548)

Fall 2014

Craig Knoblock (http://www.isi.edu/~knoblock)



Pedro Szekely (http://www.isi.edu/~szekely)

Administrative Information

Prerequisites (can be waived):	▪ CS561 ▪ CS585
Location:	▪ 5:30-650@THH212 ▪ 3:30-4:50@VKC100
Teaching assistants:	▪ Bo Wu ▪ Hao Wu
Graders:	▪ Himanshu Kulkarni ▪ Purva Tadphale
Blackboard:	TBD

This course will focus on foundations and techniques for information extraction, modeling and integration. Topics covered include semantic web (RDF, OWL, SPARQL), linked data and services, mash-ups, theory of data integration, schema mappings, record/entity linkage, data cleaning, source modeling, and information extraction. The class will be run as a lecture course with significant hands-on experience.

Lectures

Introduction and course overview

2014-08-25

Overview of issues that make information integration challenging and the techniques that we will cover in the course to address these challenges. Disucssion of the grading policy, including the quizzes, homeworks and the final project.

Craig Knoblock
Pedro Szekely

RDF, graph data model

2014-08-27

Introduction to the Resource Description Framework (RDF), the langauage of the Semantic Web and the Linked Data Cloud. We will review basic concepts of XML and contrast XML with RDF. Will introduce the concept of URIs, the global graph of data that RDF supports, and the syntax to write RDF documents.

Pedro Szekely

RDF Schema and basic inference

2014-09-03

RDF Schema defines the semantics of RDF graphs. We will introduce the elements of the RDF Schema language and show how these elements enable an RDF database (triple store) to make inferences to derive new information.

Pedro Szekely

SPARQL query language

2014-09-08

SPARQL is the query language of RDF databases. We will introduce the types of queries that can be expressed in SPARQL and the main elements of the language. We will show how one can query the Linked Data cloud using SPARQL.

Pedro Szekely

Linked Data, common vocabularies/ontologies

2014-09-10

Tim Berners Lee introduced the concept of Linked Data more than 10 years ago. Since then, the Linked Data cloud has grown to more than 30 billion facts covering pretty much any topic you can think of. We will review the principles of Linked Data and provide an overview of the main datasets and vocabularies available in the Linked Data cloud.

Pedro Szekely

Data cleaning

2014-09-15

Data cleaning is a major headache in information integration because most datasets contain errors that need to be corrected before the information is usable. We will cover the main types of errors and common techniques for detecting and correcting data errors. We will provide an overview of Google Refine, a popular tool for data cleaning.

Bo Wu

Social Data Analysis

2014-09-17

Invited lecture illustrating applications of information integration to social network analysis.

Hao Wu

Database theory basics: queries, query containment, Datalog

2014-09-22

We review the basics of the relational data model, query languages (conjunctive queries and recursive queries - Datalog), reasoning about queries (query containment), and data modeling constructs (keys, functional dependencies, referential integrity constraints, and their generalizations: tuple-generating dependencies and equality-generating dependencies). We use queries to describe the contents of sources and to map data structured according to the schemas of the sources into a common harmonized target schema.

Craig Knoblock

Logical data integration: answering queries using views (GAV, LAV, st-tgds)

2014-09-24

We describe the formal approach to data integration: (1) defining mappings between source and target schemas, using logical rules/queries/st-tgds, and (2) reasoning with these rules to rewrite queries posed over the target schema to executable queries over the source schemas. We describe the virtual approach to data integration where the data remains at remote sources and the data integration system (mediator) retrieves the data in real time in response to user queries.

Craig Knoblock

Logical data integration: (advanced)

2014-09-29

We describe the formal approach to data integration: (1) defining mappings between source and target schemas, using logical rules/queries/st-tgds, and (2) reasoning with these rules to rewrite queries posed over the target schema to executable queries over the source schemas. We describe the virtual approach to data integration where the data remains at remote sources and the data integration system (mediator) retrieves the data in real time in response to user queries.

George Konstantinidis

Schema Mapping

2014-10-01

Different datasets often use different schemas even when they contain information about the same topic. In order to integrate these datasets it is useful to identify a mapping between the attributes in these datasets so that one can combine them into one integrated dataset. In this class we cover techniques for automatically identifying mappings between schemas.

Craig Knoblock

RDF mapping tools

2014-10-06

A popular and useful technique for integrating datasets (databases, delimited text files, XML or JSON) is to map them to RDF. Once the datasets are mapped to RDF it becomes possible to query them in an integrated way using SPARQL. In this class we cover techniques for mapping different types of datasets to RDF.

Pedro Szekely

Karma, semi-automatic source modeling

2014-10-08

Karma is our own tool for semi-automatically mapping a variety of sources to a domain model defined using an ontology or an RDF Schema. In this class we will cover the algorithms that Karma uses to learn from previous mappings and the algorithms to automatically suggest models for new sources.

Pedro Szekely

Automatic source modeling

2014-10-13

Mapping datasets to RDF can be difficult and time-consuming. In this class we cover different techniques for automatically mapping a dataset to a domain ontology (RDF Schema).

Craig Knoblock

OWL2: Description Logics, Inference

2014-10-15

OWL, the Web Ontology Language is the standard language for defining ontologies in the Semantic Web. OWL ontologies define the semantics of classes and properties and allow inferences to be made to derive new facts. In this class we cover the main OWL constructs and how they allow a system to infer new facts.

Pedro Szekely

Linked Services

2014-10-20

A significant amount of data on the Web is available via Web APIs. These APIs typically return their data in XML or JSON. In this class we review techniques for modeling Web APIs to make them easy to discover and to wrap them so that they can consume and produce RDF. This makes it possible to integrate them with other Linked Data sources.

Mohsen Taheriyan

Big Data

2014-10-22

Lately, there has been significant interest in NoSQL databases that support parallel distributed analysis of large datasets on commodity hardware. In this lecture we explore the latest ideas for using NoSQL databases for storing Linked Data at large scale.

Jason Slepicka

Ontology-based data integration

2014-10-27

We consider the problem of data integration when the target schema is expressed as an ontology, in particular, the OWL2QL and OWL2EL profiles. We discuss approaches to answer queries over such ontologies and the computational effects of different ontology languages.

José Luis Ambite

OWL2 Profiles: QL, EL, RL

2014-10-29

Pedro Szekely

Full OWL reasoning is computationally expensive (exponential) and thus impractical for most applications. The OWL profiles define subsets of OWL where the reasoning algorithms perform efficiently. In this class we introduce the most common OWL profiles, review the reasoning capabilities that they support and illustrate them in practical settings.

Record linkage – string matching

2014-11-03

Craig Knoblock

Record Linkage is the problem of identifying when two different records refer to the same real world entity (e.g., whether two records about restaurants refer to the same restaurant). Record Linkage is one of the most important problems in information integration. In this class we begin our study of Record Linkage by covering several techniques for fuzzy matching text strings.

Record linkage – record matching

2014-11-05

Craig Knoblock

We continue our study of Record Linkage by reviewing techniques for matching records consisting of multiple attributes. We will review the main algorithms for record linkage and cover optimization technques to avoid the N-squared comparison problem.

Mashups: principles

2014-11-10

Pedro Szekely

Mashups are applications that combine information from multiple Web sites and Web APIs to build an interesting new application. In this calls we discuss the main techniques that have been used to create mashups and describe interesting research systemts that incorporate these techniques.

Mashups: Yahoo Pipes & YQL, REST

2014-11-12

Pedro Szekely

In this class we focus on techniques for building mashups using Web APIs. We begin by introducing the principles of RESTful APIs and illustrating these principles using examples. Then we cover two interesting systems that allow users to quickly and easily construct mashups using Web APIs.

Extracting entities and relations from text

2014-11-17

Craig Knoblock

Many structured sources contain unstructured fields (e.g., the biography of a person). We will provide a brief overview of the main techniques to extract entities (people, places and organizations) from text and review several off-the-shelf tools for entity extraction.

Semi-structured Data: Wrapper Generation

2014-11-19

Most of the information on the Web is available on Web pages, and often is not available in a structured format or a Web API. Web wrappers are programs that extract data from Web pages and return it in a structured format, typically XML. Several libraries exist for programmers to develop such wrappers. There has been significant research on techniques to generate wrappers by example so that users who don't know how to program can generate wrappers. In this lecture we cover several systems that use this approch and discuss the algorithms that they use.

Craig Knoblock

Semi-structured Data: Wrapper Learning

2014-11-24

An alternative approach to generating wrappers from examples is to automatically learn the wrappers given a set of Web pages. In this class we discuss techinques that enable a system to discover differences in pages and use this information to automatically generate a wrapper to extract information from Web pages.

Craig Knoblock

Intellectual Property

2014-12-01

TBD

Craig Knoblock

Review

2014-12-03

In this class we review all the topics covered during the semester highlighting the main ideas and techniques.

Craig Knoblock
Pedro Szekely

Homeworks

Homeworks must be done individually and they are due at 11:59pm on the due date. You can submit one homework up to one week late without penalty. Once you use your "free late homework" all late homeworks will receive a zero grade.

Due Date	Topic
2014-09-02	Creating a Wrapper for a Web Site
2014-09-07	RDF Graphs
2014-09-14	SPARQL the RDF Query Language
2014-09-21	Using Data Cleaning Tools to Clean a Dataset
2014-09-28	Undertanding GLAV rules
2014-10-05	Working with Triple Stores
2014-10-12	Using Karma to Convert Datasets Into RDF
2014-10-19	Writing an OWL ontology for a Domain
2014-10-26	Writing a Map/Reduce job to Analyze a Dataset

2014-11-09	Using Record Linkage Tools to Link Datasets
2014-11-16	Mashup Tools: Yahoo Pipes and YQL and Google Refine
2014-11-23	Using Information Extraction Tools to Extract Entities from Text

Evaluation

Quizzes	25%
Homeworks	50%
Final Exam	25%

Statement on Academic Integrity

USC seeks to maintain an optimal learning environment. General principles of academic honesty include the concept of respect for the intellectual property of others, the expectation that individual work will be submitted unless otherwise allowed by an instructor, and the obligations both to protect one's own academic work from misuse by others as well as to avoid using another's work as one's own. All students are expected to understand and abide by these principles. SCampus, the Student Guidebook, (www.usc.edu/scampus or http://scampus.usc.edu) contains the University Student Conduct Code (see University Governance, Section 11.00), while the recommended sanctions are located in Appendix A.

Statement for Students with Disabilities

Any student requesting academic accommodations based on a disability is required to register with Disability Services and Programs (DSP) each semester. A letter of verification for approved accommodations can be obtained from DSP. Please be sure the letter is delivered to me (or to TA) as early in the semester as possible. DSP is located in STU 301 and is open 8:30 a.m.–5:00 p.m., Monday through Friday. Website and contact information for DSP: http://sait.usc.edu/academicsupport/centerprograms/dsp/home_index.html, (213) 740- 0776 (Phone), (213) 740-6948 (TDD only), (213) 740-8216 (FAX) ability@usc.edu.

Emergency Preparedness/Course Continuity in a Crisis

In case of a declared emergency if travel to campus is not feasible, USC executive leadership will announce an electronic way for instructors to teach students in their residence halls or homes using a combination of Blackboard, teleconferencing, and other technologies.