CSCI 599: Foundations of Databases, Knowledge Representation, Data Integration and Data Exchange

Basic Information

Place and Time: Fall 2013, Tuesdays and Thursdays, 5:00-6:20pm
Instructors: José Luis Ambite, ambite@isi.edu, 310-448-8472, www.isi.edu/~ambite/
George Konstantinidis; konstant@usc.edu, www-scf.usc.edu/~konstant/

Course Description

This course will be an introduction to the fundamental theoretical principles of database and knowledge representation systems, and the recent theory underlying practical applications such as data integration and exchange. The course will cover the relational model, dependency theory, and description logics. We will discuss recent semantic web languages, such as the OWL2 Profiles (QL, EL, and RL), that provide tractable reasoning and query, and their connections to relational models under constraints. We will discuss different query languages and their expressive power and complexity of query answering. We will cover fundamental results on conjunctive queries, including containment and equivalence algorithms (homomorphism theorem). We will describe recursive languages, such as Datalog. We will discuss dependency theory (tuple- and equality- generating dependencies) and query evaluation and containment under constraints. We will introduce the classical chase algorithm, and its variants, as a tool for reasoning under constraints. This will set up the stage to describe the foundations of logical data integration and exchange, an important practical application of this theory. In the world of BigData semantic data integration is critically needed. While the discussion of BigData often focuses on the volume (massive quantity) of the data, or its velocity (rapid generation or collection), addressing its variability (different data types, formats, schemas) is required to perform meaningful analysis. Data integration and exchange provide the tools to make sense and effectively query large amounts of data. We will cover the fundamentals of schema mappings (source-to-target dependencies: LAV, GAV, GLAV), and algorithms for query answering using views (Bucket, Inverse Rules, Minicon, MCDSAT, and our own GQR) with and without constraints, including description logic constraints, highlighting the connections between the classical relational model, description logics, and Semantic Web knowledge representation W3C standards (RDF/S, OWL).

We expect that this course will be of interest to students in databases and theory. Students in databases, even if focused on system aspects, will benefit from understanding practically important results such as query containment, logical query optimization, and data integration and data exchange. Theory students may gain a different perspective on the complexity of query languages and their relationship to other formalisms to describe computation.

Course structure, requirements and grades

The course will consist on readings on the topic, instructor and student-led presentations, complemented with exercises, and a class project. The project may range from a literature review on a specific topic to original research. The optimal outcome of the class would be a
strong paper submission to one of the leading database, knowledge representation, artificial intelligence, or semantic web conferences.

The required textbook will be:


The book is freely available at: http://webdam.inria.fr/Alice/

**Prerequisites:** The course is self-contained and has no pre-requisites. However, some knowledge of database systems, logic and computational complexity is needed in order to keep up with the pace of the course. Recommended courses include:

CSCI-561 Introduction to AI
CSCI-585 Database Systems
CSCI-548 Information Integration on the Web
CSCI-581 Logic and its Applications
CSCI-586 Database Systems Interoperability

**Evaluation:** Students will be evaluated on the basis of class presentations, weekly quizzes on the readings, homework assignments, a class project or term paper, and class participation. Grading breakdown:

- Class project: 30%
- Class presentations: 20%
- Homework assignments: 20%
- Weekly quizzes: 20%
- Class participation: 10%

**Course Plan**

The overview of topics and expected days devoted to each topics follows:

<table>
<thead>
<tr>
<th>Topic</th>
<th>Number of lectures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction to logic</td>
<td>1</td>
</tr>
<tr>
<td>Introduction to the relational model</td>
<td>2</td>
</tr>
<tr>
<td>Conjunctive and First-order Queries: evaluation and containment</td>
<td>2</td>
</tr>
<tr>
<td>Recursive queries: Datalog. Recursion and negation</td>
<td>2</td>
</tr>
<tr>
<td>Dependency Theory: Functional and Inclusion Dependencies</td>
<td>1</td>
</tr>
<tr>
<td>General Dependencies (TGDs, EGDs) and the Chase algorithm</td>
<td>2</td>
</tr>
<tr>
<td>Description Logics: representation and reasoning</td>
<td>2</td>
</tr>
<tr>
<td>Ontology Based Data Access</td>
<td>4</td>
</tr>
<tr>
<td>Query answering under TGDs and EGDs</td>
<td>1</td>
</tr>
<tr>
<td>Introduction to Logical Data Integration, Exchange, Schema Mappings</td>
<td>1</td>
</tr>
<tr>
<td>Data Integration: query rewriting algorithms</td>
<td>4</td>
</tr>
<tr>
<td>Data Exchange</td>
<td>2</td>
</tr>
</tbody>
</table>
Data Integration under relational and ontological constraints, Datalog±

Course Readings

The following is a tentative collections of readings organized by topic. The list will be adjusted to address recent developments.

Introduction to Logic

- Chapter 2 (“Theoretical Background”) of “Foundations of Databases”.
- Introduction to Logic, by Mike Genesereth (notes from Stanford CS 157)

Introduction to the Relational Model

- Chapter 3 (“The relational model”) of “Foundations of Databases”.
- Chapter 4 (“Conjunctive Queries”) of “Foundations of Databases”.

Conjunctive and First-Order Queries: Evaluation, Containment, and Equivalence

- Chapter 4 (“Conjunctive Queries”) of “Foundations of Databases”.
- Chapter 5 (“Adding Negation: Algebra and Calculus”) of “Foundations of Databases”.
- Tutorial on "Logic and Database Queries", M.Y. Vardi, I. Barland, B. McMahan.

Datalog

- Chapter 12 (“Datalog”) of "Foundations of Databases".
- Chapter 13 (“Datalog Evaluation”) of "Foundations of Databases".

Datalog with negation

- Chapter 14 (“Recursion and Negation”) of "Foundations of Databases".
- Chapter 15 (“Negation in Datalog”) of "Foundations of Databases".
Functional and Inclusion Dependencies

- Chapter 8 ("Functional and Join Dependency") of "Foundations of Databases".
- Chapter 9 ("Inclusion Dependency") of "Foundations of Databases".

General Dependencies and the chase algorithm

- Chapter 10 ("A Larger Perspective") of "Foundations of Databases".

Description Logics


- Chapter 1 ("An introduction to Description Logics") of "The Description Logic Handbook: Theory, Implementation, and Applications"
- Chapter 2 ("Basic Description Logics") of "The Description Logic Handbook: Theory, Implementation, and Applications"
- Chapter 4 ("Relationships with other Formalisms") of "The Description Logic Handbook: Theory, Implementation, and Applications"
- Chapter 16 ("Description Logics for Databases") of "The Description Logic Handbook: Theory, Implementation, and Applications"

Ontology Based Data Access


Query answering under TGDs and EGDS

Introduction to Data Integration
• Maurizio Lenzerini. "Data integration: a theoretical perspective". In PODS '02

Data Integration Algorithms
• F.N. Afrati, C. Li, and J.D. Ullman. “Generating efficient plans for queries using views”. In ACM SIGMOD, 2001.
• Yolifé Arvelo, Blai Bonet, Maria-Esther Vidal, “Compilation of Query-Rewriting Problems into Tractable Fragments of Propositional Logic”. In AAAI 2006.
• George Konstantinidis and José Luis Ambite. “Scalable Query Rewriting: A Graph-Based
Approach”. In SIGMOD 2011. Athens, Greece.

Data exchange


Data integration under relational and ontological constraints

- Cali, Andrea, Georg Gottlob, and Thomas Lukasiewicz. “Datalog+/:- a unified approach to ontologies and integrity constraints.” Proc. of the 12th International Conference on


**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. The phone number for DSP is (213) 740-0776.

**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, contains the Student Conduct Code in Section 11.00, while the recommended sanctions are located in Appendix A: [http://www.usc.edu/dept/publications/SCAMPUS/gov/](http://www.usc.edu/dept/publications/SCAMPUS/gov/). Students will be referred to the Office of Student Judicial Affairs and Community Standards for further review, should there be any suspicion of academic dishonesty. The Review process can be found at: [http://www.usc.edu/student-affairs/SJACS/](http://www.usc.edu/student-affairs/SJACS/).