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:

> Foundations of Databases, by Serge Abiteboul, Richard Hull, and Victor Vianu. Addison Wesley, 1995.

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:

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

Data Integration under relational and ontological constraints, Datalog±	4
Review	1

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".
- P. C. Kanellakis. Elements of relational database theory. In J. Van Leeuwen, Ed., Handbook of Theoretical Computer Science, pages 1074–1156. Elsevier, Amsterdam, 1991.
- E. F. Codd. A relational model of data for large shared data banks. Communications of the ACM, 13(6):377–387, 1970.

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".
- A. K. Chandra and P. M. Merlin. Optimal implementation on conjunctive queries in relational data bases. In Proc. ACM SIGACT Symposium on the Theory of Computing, 77–90, 1977.
- Tutorial on "Logic and Database Queries", M.Y. Vardi, I. Barland, B. McMahan.
- M. Y. Vardi. The complexity of relational query languages. In Proc. ACM SIGACT Symp. on the Theory of Computing, pages 137–146, 1982..

Datalog

- Chapter 12 ("Datalog") of "Foundations of Databases".
- Chapter 13 ("Datalog Evaluation") of "Foundations of Databases".
- Chapter 3, 12 from Jeffrey D. Ullman: Principles of Database and Knowledge-Base Systems, Volume I. Computer Science Press 1988/1989, Volumes 1 and 2

Datalog with negation

- Chapter 14 ("Recursion and Negation") of "Foundations of Databases".
- Chapter 15 ("Negation in Datalog") of "Foundations of Databases".
- P. G. Kolaitis. The expressive power of stratified logic programs. Information and Computation, 90(1):50–66, 1991.

Functional and Inclusion Dependencies

- Chapter 8 ("Functional and Join Dependency") of "Foundations of Databases".
- Chapter 9 ("Inclusion Dependency") of "Foundations of Databases".
- "The Theory of Database Dependencies: A Survey", R. Fagin and M.Y. Vardi.

General Dependencies and the chase algorithm

- Chapter 10 ("A Larger Perspective") of "Foundations of Databases".
- Catriel Beeri and Moshe Y. Vardi. 1981. The Implication Problem for Data Dependencies. In *Proceedings of the 8th Colloquium on Automata, Languages and Programming*, Shimon Even and Oded Kariv (Eds.). Springer-Verlag, London, UK, UK, 73-85.
- C. Beeri and M. Vardi. "A proof procedure for data dependencies", Journal of the ACM (JACM), 1984. R. Meyden.
- R. van der Meyden. "Logical approaches to incomplete information: A survey", logics for databases and information systems, 1998.
- Greco, Sergio, Cristian Molinaro, and Francesca Spezzano. "Incomplete Data and Data Dependencies in Relational Databases." Synthesis Lectures on Data Management, 2012.
- Cali, G. Gottlob, and M. Kifer. "Taming the infinite chase: Query answering under expressive relational constraints." Proc. of KR, 2008.

Description Logics

Readings from Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, Peter F. Patel-Schneider (Eds.): The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press 2003

- 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

- Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, Mariano Rodriguez-Muro, and Riccardo Rosati, "Ontologies and Databases: The DL-Lite Approach". S. Tessaris et al. (Eds.): Reasoning Web 2009, LNCS 5689, pp. 255–356, 2009.
- Calvanese, Diego, et al. "Tractable reasoning and efficient query answering in description logics: The DL-Lite family." Journal of Automated reasoning, 2007.
- Franz Baader, Sebastian Brandt, and Carsten Lutz. Pushing the EL Envelope. In Proc. of the 19th Joint Int. Conf. on Artificial Intelligence (IJCAI 2005), 2005.
- Franz Baader, Sebastian Brandt, and Carsten Lutz. Pushing the EL Envelope Further. In Proc. of the Washington DC workshop on OWL: Experiences and Directions (OWLED08DC), 2008.
- Héctor Pérez-Urbina, Boris Motik, Ian Horrocks: "Tractable query answering and rewriting under description logic constraints". J. Applied Logic 8(2): 186-209 (2010)

- Héctor Pérez-Urbina, Ian Horrocks, Boris Motik: "Efficient Query Answering for OWL 2". International Semantic Web Conference 2009: 489-504
- Roman Kontchakov, Carsten Lutz, David Toman, Frank Wolter, Michael Zakharyaschev: "The Combined Approach to Query Answering in DL-Lite". KR 2010
- Riccardo Rosati, Alessandro Almatelli, "Improving Query Answering over DL-Lite Ontologies". KR 2010.
- Rosati, Riccardo. "Prexto: Query Rewriting under Extensional Constraints in DL–Lite." The Semantic Web: Research and Applications. Springer Berlin Heidelberg, 2012. 360-374.
- Kikot, Stanislav, Roman Kontchakov, and Michael Zakharyaschev. "Conjunctive query answering with OWL 2 QL." Proc. of the 13th Int. Conf. KR. 2012.
- Giorgio Stefanoni, Boris Motik, and Ian Horrocks. "Small Datalog Query Rewritings for EL." Proc. of the 25th Int. Workshop on Description Logics (DL 2012), 2012.

Query answering under TGDs and EGDs

- J.F. Baget, M. Leclere, M.L. Mugnier, et al. "Walking the decidability line for rules with existential variables". In Proc. 12th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR10), 2010.
- J.F. Baget, M. Leclere, M.L. Mugnier, E. Salvat, et al. "Extending decidable cases for rules with existential variables". In Proc. of IJCAI, 2009.
- M.L. Mugnier. "Ontological query answering with existential rules". Web Reasoning and Rule Systems, 2011.

Introduction to Data Integration

- Alon Y. Halevy, "Answering queries using views: A survey". VLDB Journal. 2001.
- Maurizio Lenzerini. "Data integration: a theoretical perspective". In PODS '02
- Jeffrey D. Ullman, "Information integration using logical views", Theoretical Computer Science, May 2000.
- Friedman, Marc, Alon Levy, and Todd Millstein. "Navigational plans for data integration." Proceedings of the National Conference on Artificial Intelligence, 1999.
- (Optional) Chapters 2 and 3 of: AnHai Doan, Alon Y. Halevy, Zachary G. Ives: "Principles of Data Integration". Morgan Kaufmann 2012.

Data Integration Algorithms

- Oliver M. Duschka, Michael R. Genesereth, Alon Y. Levy, "Recursive query plans for data integration", The Journal of Logic Programming, April 2000.
- Abiteboul, Serge, and Oliver M. Duschka. "Complexity of answering queries using materialized views." Proceedings of PODS. ACM, 1998.
- F.N. Afrati, C. Li, and J.D. Ullman. "Generating efficient plans for queries using views". In ACM SIGMOD, 2001.
- Rachael Pottinger & Alon Halevy, "A Scalable Algorithm for Answering Queries Using Views", VLDB, 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

- R. Fagin, P.G. Kolaitis, R.J. Miller, and L. Popa. "Data exchange: semantics and query answering". Theoretical Computer Science, 2005.
- R. Fagin, P.G. Kolaitis, and L. Popa. "Data exchange: getting to the core". ACM Transactions on Database Systems (TODS), 2005.
- Ronald Fagin, Laura M. Haas, Mauricio Hernández, Renée J. Miller, Lucian Popa, Yannis Velegrakis. Clio: Schema Mapping Creation and Data Exchange. Conceptual Modeling: Foundations and Applications. Lecture Notes in Computer Science Volume 5600, 2009.
- Balder ten Cate, Laura Chiticariu, Phokion G. Kolaitis, Wang Chiew Tan: "Laconic Schema Mappings: Computing the Core with SQL Queries". PVLDB 2009.
- Giansalvatore Mecca, Paolo Papotti, and Salvatore Raunich. 2012. Core schema mappings: Scalable core computations in data exchange. Information Systems: 37 (7). November 2012.

Data integration under relational and ontological constraints

- Duschka, Michael R. Genesereth, Alon Y. Levy, "Recursive query plans for data integration", The Journal of Logic Programming, April 2000.
- Jarek Gryz. "Query Rewriting Using Views in the Presence of Functional and Inclusion Dependencies". Information Systems 24(7): 597-612 (1999).
- Alin Deutsch, Lucian Popa, Val Tannen "Query reformulation with constraints". SIGMOD Record (SIGMOD) 35(1):65-73 (2006). Oliver M.
- Alin Deutsch, Lucian Popa, Val Tannen. Chase & Backchase: A Method for Query Optimization with Materialized Views and Integrity Constraints. Technical Report MS-CIS-01-16, 2001.
- Lucian Popa, Alin Deutsch, Arnaud Sahuguet, and Val Tannen. "A chase too far?" In ACM SIGMOD, 2000.
- C. Koch. "Query rewriting with symmetric constraints". Foundations of Information and Knowledge Systems, 2002.
- A. Cali. "Query answering by rewriting in GLAV data integration systems under constraints". Semantic Web and Databases, 2005.
- Antonella Poggi, Domenico Lembo, Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, Riccardo Rosati: "Linking Data to Ontologies". Journal of Data Semantics, 2008.
- F.N. Afrati and N. Kiourtis. "Computing certain answers in the presence of dependencies". Information Systems, 2010.
- A. Cali, G. Gottlob, T. Lukasiewicz, B. Marnette, and A. Pieris. "Datalog+/-: A family of logical knowledge representation and query languages for new applications". In Logic in Computer Science (LICS), 2010.
- A. Cali, G. Gottlob, T. Lukasiewicz, and A. Pieris. "A logical toolbox for ontological reasoning". SIGMOD Record, 2011.
- Calì, Andrea, Georg Gottlob, and Thomas Lukasiewicz. "Datalog+/-: a unified approach to ontologies and integrity constraints." Proc. of the 12th International Conference on

Database Theory. ACM, 2009.

- Gottlob, Georg, Giorgio Orsi, and Andreas Pieris. "Ontological queries: Rewriting and optimization". ICDE, 2011.

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/. 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/.