Research Project Descriptions

REMARKS

Choosing a Topic. By **3 pm Wednesday, January 26**th, please send me a list, ordered by preference, of at least 3 projects that you would like to work on (email `ludaesch@ucdavis.edu` with subject=`ECS289F`). I will aim at accomodating your preferences as much as possible. If you have questions regarding a topic, see me at the office hour on Wednesday (preferred) or send email.

Preparation. The project descriptions given below provide the basic information to help you decide on a topic. However, both theory and practice projects will require follow-up meetings with me (the earlier the better!), e.g., to discuss the focus of a given reading assignment, how to summarize the articles, or how to approach the practice projects (some might need access to restricted resources), etc. You can meet me either during my office hours (Wed+Fri, 11am-noon) or during additional scheduled meetings.

Teamwork. Several projects are closely related and can be done in teams of typically 2 students.

Presentation. All projects require a presentation at class (e.g., using Powerpoint slides, or for LaTeX friends using PROSPER). In addition, theory projects require a (brief) written report, practice projects a system demonstration (or report). Details depend on the particular project.

1 Data Integration

1.1 (T) Foundations of Data Integration & Query Rewriting

The goal of this theory project is to provide an overview of the foundations of query rewriting for data integration, including *Global-as-View* (GAV) and *Local-as-View* (LAV) techniques. Good introductions and the primary sources for this project are [KOCH, 2001, Chapter 3] and [LENZERINI, 2002].

1.2 (T) Special Algorithms in Data Integration

The goal of these theory projects is to study one or more specific algorithms in data integration, for example:

1.2.1 Answering Queries using Views

Here the setting is often LAV, i.e., the content of sources is defined via views over a (virtual) global schema. The user issues a query against the global schema and the system has to find a rewriting that uses only the views [HALEVY, 2001], [DUSCHKA *et al.*, 2000].

1.2.2 Answering Queries with Limited Access Patterns

Sometimes sources can only execute certain queries, e.g., given via a set of *access* (or *binding*) *patterns*. Thus a query plan has to be found that observes those patterns [NASH & LUDÄSCHER, 2004], [DEUTSCH *et al.*, 2005].

1.2.3 Semantic Integration

Here the goal is to study and compare recent ontology-based mediation approaches, e.g., [TZITZIKAS *et al.*, 2002], [PEIM *et al.*, 2002], and [GOBLE *et al.*, 2001]. In these, data integration involves "glue ontologies" in addition to the local and global database schemas. The final list of papers is TBD and might include the above or others from (cf. the survey [WACHE *et al.*, 2001]).

1.2.4 Schema Matching.

In schema matching, correspondences between elements of two or more schemas need to be identified. References include those in the special SIGMOD Record section on semantic integration [SIG, 2004] and others.

1.3 (P) Practice of Data Integration

The goal of this project is to implement a simple GAV mediator prototype for a Global-as-View (GAV) approach, i.e., in which query literals are successively unfolded until an executable query plan is reached. This view unfolding is based on a standard unification algorithm. Advanced constructs, e.g. for handling recursive views can be added. This practice project lends itself to collaboration with Project 1.1).

2 Knowledge Representation

2.1 (T) Biological Ontologies and Pathways

The goal of this theory project is to give an overview of KR techniques used to represent biological information, e.g., the Gene Ontology [CONSORTIUM, 2002] or biological pathway databases such as EcoCyc and BioCyc: [KARP, 1999], [KARP, 2000], [KARP, 2001], [BIO, 2003].

2.2 (T+P) Introduction to Formal Concept Analysis (FCA)

The goal of this project is to provide an introduction to FCA and some of its applications. For the practical part, examples should be prepared using the Toscana system [GANTER & WILLE, 1999], [BURMEISTER, 2003], [TOS, 2003].

2.3 (T+P) Benchmarking Ontology Reasoners

The goal of this project is to use experiment with and compare a number of reasoning engines for ontologies, e.g., Racer, FaCT, Jena, Pellet, or the Jess OWL reasoner. (Note that many of these can be used within Protégé.) [HORROCKS, 1999], [HAARSLEV & MLLER,], [PROGRAMME,], [KOPENA & REGLI,], [MINDSWAP,], [PRO, 2003].

3 Scientific Workflows

Most of these projects are conducted using (and possibly extending) the Kepler scientific workflow system [KEP, 2004]. Since Kepler extends the Ptolemy II system (PTII) [PTO, 2004], it is a good idea to install PTII first, run a number of example models (called *workflows* in Kepler), and then learn how to create your own models/workflows using the graphical user interface (Vergil) [BROOKS *et al.*, 2004]. Then you can start to tackle Kepler. Note that there are Kepler mailing lists and IRC channels on which Kepler developers exchange valuable information.

3.1 (T) Dataflow Process Networks

The goal of this project is to provide an overview of the foundations of dataflow process networks; a core reference is [LEE & PARKS, 1995].

3.2 (P) Scientific Data Analysis Workflows

The goal of this project is to study typical scientific "data analysis pipelines" (application areas include, e.g., from genomics or biodiversity informatics) and learn how to interface with external applications such as `R` [R, 2004]. Based on existing examples, a scientific workflow needs to be created that provides additional constructs, e.g., to create an execution log to keep track of the data transformations performed, or to provide an improved interface to `R`.

3.3 (P) Collection Management in Scientific Workflows

The goal of this project is to study the features of the SDSC Storage Resource Broker [SRB, 2004], e.g., file and replica management, metadata-based querying, and authentication. An example workflow is to be created that highlights several of these features.

3.4 (P) Data-Intensive Scientific Workflows

The goal of this project is to experiment with and compare different ways to move data in scientific workflows, e.g., using `scp` (secure copy), `GridFTP` [GRI, 2004], and `SRB` [SRB, 2004]. Kepler already includes components for some of these or can invoke these tools via a command line. Example workflows need to be created that show and compare the different ways of data movement.

3.5 (P) High-Throughput Scientific Workflows

The goal of this project is to study high-throughput Grid workflows, i.e., in which a large number of jobs are executed on a remote cluster computer. The staging of files, job submission, and result fetching is controlled via Kepler. The job scheduling software is NIMROD or Condor [DOUGLAS THAIN & LIVNY, 2004, CON, 2004]. An example workflow exists on which this project can be based.

References

[BIO, 2003] BioCyc Knowledge Library, 2003. `http://biocyc.org/`.

[BROOKS *et al.*, 2004] C. Brooks, E. A. Lee, X. Liu, S. Neuendorffer, Y. Zhao, & H. Zheng. Using Vergil (Chapter 2 from Heterogeneous Concurrent Modeling and Design in Java). Technical Memorandum UCB/ERL M04/27, July 29, 2004, University of California, Berkeley, 2004. `http://embedded.eecs.berkeley.edu/concurrency/ptolemy/usingVergil.pdf`.

[BURMEISTER, 2003] Peter Burmeister. Formal Concept Analysis with ConImp: Introduction to the Basic Features. `http://www.mathematik.tu-darmstadt.de/~burmeister/ConImpIntro.ps`, 2003.

[CON, 2004] Condor Week Presentations. `http://www.cs.wisc.edu/condor/CondorWeek2004/presentations.html`, 2004.

[CONSORTIUM, 2002] Gene Ontology Consortium. GO, 2002. `http://www.geneontology.org/`.

[DEUTSCH *et al.*, 2005] Alin Deutsch, Bertram Ludäscher, & Alan Nash. Rewriting Queries using Views with Access Patterns under Integrity Constraints. In *Intl. Conference on Database Theory (ICDT)*, 2005.

[DOUGLAS THAIN & LIVNY, 2004] Todd Tannenbaum Douglas Thain & Miron Livny. Distributed Computing in Practice: The Condor Experience. *Concurrency and Computation: Practice and Experience*, 2004. http://www.cs.wisc.edu/condor/doc/condor-practice.pdf.

[DUSCHKA *et al.*, 2000] Oliver M. Duschka, Michael R. Genesereth, & Alon Y. Levy. Recursive Query Plans for Data Integration. *Journal of Logic Programming*, 43(1):49–73, 2000.

[GANTER & WILLE, 1999] Bernhard Ganter & Rudolf Wille. *Formal Concept Analysis – Mathematical Foundations*. Springer, 1999. http://www.springer.de/cgi-bin/search_book.pl?isbn=3-540-62771-5.

[GOBLE *et al.*, 2001] C. Goble, R. Stevens, G. Ng, S. Bechhofer, N. Paton, P. Baker, M. Peim, & A. Brass. Transparent Access to Multiple Bioinformatics Information Sources. *IBM Systems Journal*, 40(2):534–551, 2001. http://www.research.ibm.com/journal/sj/402/goble.pdf.

[GRI, 2004] GridFTP, 2004. http://www.globus.org/datagrid/gridftp.html.

[HAARSLEV & MLLER,] Volker Haarslev & Ralf Mller. Racer: Renamed ABox and Concept Expression Reasoner. http://www.sts.tu-harburg.de/~r.f.moeller/racer/.

[HALEVY, 2001] Alon Halevy. Answering Queries Using Views: A Survey. *VLDB Journal*, 10(4):270–294, 2001. http://www.cs.washington.edu/homes/alon/site/files/view-survey.ps.

[HORROCKS, 1999] Ian Horrocks. The FaCT System, 1999. http://www.cs.man.ac.uk/~horrocks/FaCT/.

[KARP, 1999] Peter D. Karp. EcoCyc: The Resource and the Lessons Learned. In *Bioinformatics Databases and Systems*. Kluwer, 1999. http://www.ai.sri.com/pkarp/pubs/ecocyc-lessons.ps.

[KARP, 2000] Peter D. Karp. An Ontology for Biological Function Based on Molecular Interactions. *Bioinformatics*, 16(3):269–285, 2000. http://www.ai.sri.com/pubs/files/887.ps.

[KARP, 2001] Peter D. Karp. Pathway Databases: A Case Study in Computational Symbolic Theories. *Science*, 293:2040–2044, 2001. http://www.ai.sri.com/pubs/full.php?id=880.

[KEP, 2004] KEPLER: A System for Scientific Workflows, 2004. http://kepler-project.org.

[KOCH, 2001] Christoph Koch. *Data Integration against Multiple Evolving Autonomous Schemata*. PhD thesis, Technische Universität Wien, Austria, 2001. http://www.dbai.tuwien.ac.at/staff/koch/download/thesis_20010516_1500_final.pdf.

[KOPENA & REGLI,] Joseph Kopena & William Regli. OWLJessKB: A Semantic Web Reasoning Tool. http://edge.cs.drexel.edu/assemblies/software/owljesskb/.

[LEE & PARKS, 1995] Edward A. Lee & Thomas Parks. Dataflow Process Networks. *Proceedings of the IEEE*, 83(5):773–799, May 1995. http://citeseer.nj.nec.com/455847.html.

[LENZERINI, 2002] M. Lenzerini. Data Integration; A Theoretical Perspective. Tutorial at the ACM Symposium on Principles of Database Systems (PODS), 2002. http://www.sigmod.org/pods/tut/le.pdf.

[MINDSWAP,] Mindswap. Pellet OWL Reasoner. http://www.mindswap.org/2003/pellet/index.shtml.

[NASH & LUDÄSCHER, 2004] Alan Nash & Bertram Ludäscher. Processing Unions of Conjunctive Queries with Negation under Limited Access Patterns. In *Intl. Conference on Extending Database Technology (EDBT)*, Heraklion, Crete, Greece, 2004.

[PEIM *et al.*, 2002] Martin Peim, Enrico Franconi, Norman W. Paton, & Carole A. Goble. Query Processing with Description Logic Ontologies Over Object-Wrapped Databases. In *14th Intl. Conference on Scientific and Statistical Database Management (SSDBM)*, Edinburgh, Scotland, 2002. `http://citeseer.nj.nec.com/peim01query.html`.

[PRO, 2003] Protégé 2000 Project Page, 2003. `http://protege.stanford.edu/`.

[PROGRAMME,] HP Labs Semantic Web Programme. Jena A Semantic Web Framework for Java. `http://jena.sourceforge.net/`.

[PTO, 2004] PTOLEMY II project and system. Department of EECS, UC Berkeley, 2004. `http://ptolemy.eecs.berkeley.edu/ptolemyII/`.

[R, 2004] R – Statistical Data Analysis, 2004. `http://www.r-project.org`.

[SIG, 2004] SIGMOD Record, Special Section on Semantic Integration, December 2004. `http://www.sigmod.org/sigmod/record/issues/0412/`.

[SRB, 2004] SDSC Storage Resource Broker, 2004. `http://www.npaci.edu/DICE/SRB/`.

[TOS, 2003] ToscanaJ SourceForge Project, 2003. `http://toscanaj.sourceforge.net/`.

[TZITZIKAS *et al.*, 2002] Yannis Tzitzikas, Nicolas Spyratos, & Panos Constantopoulos. Translation for Mediators over Ontology-based Information Sources. In *Second Hellenic Conf. on Artificial Intelligence (SETN)*, 2002. `citeseer.nj.nec.com/tzitzikas02query.html`.

[WACHE *et al.*, 2001] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, & S. Hübner. Ontology-Based Integration of Information – A Survey of Existing Approaches. In *Proc. of the IJCAI-01 Workshop: Ontologies and Information Sharing*, 2001. `http://www.cs.vu.nl/~heiner/public/ois-2001.pdf`.