USC Viterbi School of Engineering

CSCI 548: Information Integration on the Web

Units: 3
Term—Day—Time:
Spring 2015 – MW – 3:30-4:50pm

Location: ZHS 163

Instructor: Jose Luis Ambite
Office: Outside classroom
Office Hours: Immediately after class
Contact Info: ambite@isi.edu, 310-448-8472.

Instructor: Craig Knoblock
Office: AFH B55a
Office Hours: Wednesdays 10-11am
Contact Info: knoblock@usc.edu, 310-448-8786.

Teaching Assistant: Bo Wu
Office: plaza between RTH cafe and EEB
Office Hours: Tuesday 10am-12pm
Contact Info: wubo@usc.edu

Catalogue Course Description

Foundations, techniques, and algorithms for information integration. Topics include Semantic Web, linked data, data integration, entity linkage, source modeling, and information extraction.

Expanded Course Description

This course focuses on foundations, techniques, and algorithms 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 lots of student participation and significant hands-on experience.

Learning Objectives

The learning objectives for this course are:

- Understand the foundations and techniques of the Semantic Web, including RDF, OWL, SPARKL, linked data, and linked services
- Understand the theory and techniques of traditional data integration, including view integration, schema mapping, record linkage
- Understand the algorithms and techniques for data cleaning, source modeling, building mashups, semi-structured extraction, and information extraction
- Understand the theory and application of the state-of-the-art software and tools for information extraction
- For any given integration problem, be able to select and apply the most relevant information integration techniques to solve that problem

Prerequisite(s): CSCI 561
Co-Requisite (s): none
Concurrent Enrollment: none
Recommended Preparation: CSCI 585 and some programming experience

Course Notes

The course will be run as a lecture class with student participation strongly encouraged. There are weekly readings and students are encouraged to do the readings prior to the discussion in class. All of the course materials, including the readings, lecture slides, homeworks will be posted online on Blackboard.

Technological Proficiency and Hardware/Software Required

Students are expected to know how to program in a language such as Java, C++, or Python. Students are also expected to have their own laptop or desktop computer where they can install and run software to do the weekly homework assignments.

Required Readings and Supplementary Materials

Required Textbook: Principles of Data Integration by Doan, Halevy, & Ives, Morgan Kaufmann, 2012
The book is available online from the USC library and is available for purchase.

All of the required readings are listed in the course schedule.

Description and Assessment of Assignments

Homework Assignments

There will be weekly homework assignments for the first 12 weeks of class. The assignments must be done individually. The homework assignments are expected to take 6-8 hours per week. Each assignment is graded on a scale of 0-100 and the specific rubric for each assignment is given in the assignment. The homework topics are listed in the Course Schedule.

Grading Breakdown

Quizzes: There will be weekly quizzes based on the material from the week before. There is no mid-term for this class.
Homework: There will be weekly homework based on the topics of the class each week.
Final Exam: There is a final exam at the end of the semester covering all of the material covered in the class.

Grading Schema:	
Quizzes	25%
Homework	50%
Final:	25%
Total	100%

Grades will range from A through F. The following is the breakdown for grading:

94 - 100 = A	74 - 76 = C
90 – 93 = A-	70 - 73 = C-
87 – 89 = B+	67 - 69 = D+
84 – 86 = B	64 - 66 = D
83 – 83 = B-	60 - 63 = D-
77 – 79 =C+	Below 60 is an F

Assignment Submission Policy

Homework assignments are due at 11:59pm on the due date and should be submitted in Blackboard. You can submit homework up to one week late, but you will loose 20% of the possible points for the assignment. After one week, the assignment cannot be submitted.

Course Schedule: A Weekly Breakdown

	Topics/Daily Activities	Readings	Quizzes & Homeworks	Instructor
Week 1 Jan 12 Jan 14	**Course Introduction** **RDF, Graph data model**	Frank Manola and Eric Miller. Rdf primer. Technical report, W3C, February 2004. http://www.w3.org/TR/2004/REC-rdf-primer-20040210/. Tim Berners-Lee. Why rdf model is different from the xml model. Technical report, W3C, 1998. http://www.w3.org/DesignIssues/RDF-XML.html.	Homework 0: Academic Integrity (due on Friday)	Professor Ambite Professor Ambite
Week 2 Jan 21	**RDF Schema**	Rdf vocabulary description language 1.0: Rdf schema. Technical report, W3C, February 2004. http://www.w3.org/TR/2004/REC-rdf-schema-20040210/. Ben Adida, Ivan Herman, Manu Sporny, and Mark Birbeck. Rdfa 1.1 primer rich structured data markup for web documents. Technical report, W3C, June 2012. http://www.w3.org/TR/2012/NOTE-rdfa-primer-20120607/.	Quiz 1 (on Wednesday) Homework 1: Creating a Wrapper (due on Friday)	Professor Knoblock
Week 3 Jan 26 Jan 28	**SPARQL query language** **OWL 2**	Steve Harris and Andy Seaborne. Sparql 1.1 query language. Technical report, W3C, January 2012. http://www.w3.org/TR/2012/PR-sparql11-query-20121108/. Krtzsch Markus, Simancik Frantisek, and Horrocks Ian. A description logic primer. 2012. http://arxiv.org/pdf/1201.4089.pdf.	Quiz 2 (on Monday) Homework 2: RDF (due on Friday)	Professor Ambite Professor Ambite
Week 4 Feb 2	**Linked Data**	Aduna B.V. Http communication protocol for sesame 2. In System documentation for Sesame 2.x, chapter 8. October 2013. http://www.csee.umbc.edu/courses/graduate/691/spring14/01/examples/sesame/openrdf-sesame-2.6.10/docs/system/ch08.html. Chimezie Ogbuji. Sparql 1.1 graph store http protocol. Technical report, W3C, May 2012. http://www.w3.org/TR/sparql11-http-rdf-update/.	Quiz 3 (on Monday) Homework 3 SPARQL (due on Friday)	Professor Knoblock

Feb 4	**Data Cleaning**	Wrangler: Interactive visual specification of data transformation scripts. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2011. http://vis.stanford.edu/papers/wrangler. Bo Wu, Pedro Szekely, and Craig A. Knoblock. Minimizing user effort in transforming data by example. In Proceedings of the International Conference on Intelligent User Interface, 2014. http://www.isi.edu/integration/papers/wu14-iui.pdf. Open Refine, Explore data. http://youtu.be/B70J_H_zAWM. Open Refine, Clean and transform data. http://youtu.be/cO8NVCs_Ba0. Open Refine, Reconcile and match data. http://youtu.be/5tsyz3ibYzk.		Mr. Wu
Week 5 Feb 9 Feb 11	**Database theory basics** **Logical Data Integration**	AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Principles of Data Integration, chapter 2.1, 2.2, 2.3 and 2.4. Morgan Kaufmann, 2012. http://proquest.safaribooksonline.com.libproxy.usc.edu/book/databases/9780124160446.	Quiz 4 (on Monday) Homework 4: OWL (due on Friday)	Professor Ambite Professor Ambite
Week 6 Feb 18	**Scalable data integration**	Alon Halevy and Rachel Pottinger. A scalable algorithm for answering queries using views. The VLDB Journal The International Journal on Very Large Data Bases, 2001. http://www.vldb.org/conf/2000/P484.pdf. Scalable query rewriting: a graph-based approach, 2001. http://www.isi.edu/ ambite/konstantinidis2011-sigmod.pdf.	Quiz 5 (on Monday) Homework 5: Data Cleaning (due on Friday)	Mr. Konstantinidis
Week 7 Feb 23 Feb 25	**Schema Mapping** **Linked Services**	Jayant Madhavan, Shawn R. Jeffery, Shirley Cohen, Xin (Luna) Dong, David Ko, Cong Yu, and Alon Halevy. Web-scale data integration: You can only afford to pay as you go, 2007. http://www.docin.com/p-47000224.html. Barry Norton and Reto Krummenacher. Consuming dynamic linked data. In Proceedings of the 1st International Workshop on Consuming Linked Data, 2010. http://ceur-ws.org/Vol-665/NortonEtAl COLD2010.pdf. Mohsen Taheriyan, Craig A. Knoblock, Pedro Szekely, and Jose Luis Ambite. Rapidly integrating services into the linked data cloud. In Proceedings of the 11th International Semantic Web Conference (ISWC 2012), 2012. http://www.isi.edu/integration/papers/taheriyan12-iswc.pdf. AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Principles of Data Integration, chapter 5. Morgan Kaufmann, 2012. http://proquest.safaribooksonline.com.libproxy.usc.edu/book/databases/9780124160446.	Quiz 6 (on Monday) Homework 6: Logical Data Integration (due on Friday)	Professor Ambite Mr. Taheriyan

Week 8 Mar 2 Mar 4	**RDF Mapping Tools** **Semi-Automatic Source Modeling**	[2] R2rml: Rdb to rdf mapping language. http://www.w3.org/TR/r2rml/. Craig A. Knoblock, Pedro Szekely, Jose Luis Ambite, , Aman Goel, Shubham Gupta, Kristina Lerman, Maria Muslea, Mohsen Taheriyan, and Parag Mallick. Semi-automatically mapping structured sources into the semantic web. In Proceedings of the Extended Semantic Web Conference, Crete, Greece, 2012. http://www.isi.edu/integration/papers/knoblock12-eswc.pdf.	Quiz 7 (on Monday) Homework 7: Triple Stores (due on Friday)	Professor Ambite Professor Ambite
Week 9 Mar 9 Mar 11	**Source Modeling** **Data Warehousing**	Mark James Carman and Craig A. Knoblock. Learning semantic descriptions of web information sources. In Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI), January 2007. http://www.isi.edu/integration/papers/carman07-ijcai.pdf. Jos´e Luis Ambite, Sirish Darbha, Aman Goel, Craig A. Knoblock, Kristina Lerman, Rahul Parundekar, and Thomas Russ. Automatically constructing semantic web services from online sources. In Proceedings of the 8th International Semantic Web Conference (ISWC 2009), 2009. http://www.isi.edu/integration/papers/ambite09-iswc.pdf. [TBD]	Quiz 8 (on Monday) Homework 8: Karma (due on Friday)	Professor Knoblock Professor Ambite
Week 10 Mar 23 Mar 25	**String Matching** **Record Linkage**	[1] AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Principles of Data Integration, chapters 4 & 7. Morgan Kaufmann, 2012. http://proquest.safaribooksonline.com.libproxy.usc.edu/book/databases/9780124160446.	Quiz 9 (on Monday) Homework 9: String Similarity (due on Friday)	Professor Knoblock Professor Knoblock
Week 11 Mar 30 Apr 1	**Mashup principles** **Mashup tools**	Shubham Gupta and Craig A. Knoblock. Building geospatial mashups to visualize information for crisis management. In Proceedings of the 7th International Conference on Information Systems for Crisis Response and Management, 2010. http://www.isi.edu/integration/papers/gupta10-iscram.pdf. Jeffrey Wong and Jason I. Hong. Making mashups with marmite: towards end-user programming for the web. In ACM SIGMOD Record, 2007. http://repository.cmu.edu/cgi/viewcontent.cgi?article=1063&context=hcii.	Quiz 10 (on Monday) Homework 10: Record Linkage (due on Friday)	Professor Knoblock Professor Knoblock

		Rob Ennals, Eric Brewer, Minos Garofalakis, Michael Shadle, and Prashant Gandhi. Intel mash maker: join the web. 2007. http://23.30.224.201/publications/intel-mash-maker-join-web. Huynh David, Mazzocchi Stefano, and Karger David. Piggy bank: Experience the semantic web inside your web browser. 2007. http://simile.mit.edu/papers/iswc05.pdf. Leo Sauermann and Richard Cyganiak. Cool uris for the semantic web. Technical report, 2008. http://www.w3.org/TR/cooluris/.		
Week 12 Apr 6 Apr 8	**Information Extraction**	Matthew Michelson and Craig A. Knoblock. Semantic Annotation of Unstructured and Ungrammatical Text. In Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-2005), Edinburgh, Scotland, 2005. http://www.isi.edu/integration/papers/michelson05-ijcai.pdf Andrew McCallum. Information Extraction: Distilling Structured Data from Unstructured Text . ACM Queue, volume 3, Number 9, November 2005. http://people.cs.umass.edu/~mccallum/papers/acm-queue-ie.pdf Charles Elkan, Tutorial on Log-linear Models and Conditional Random Fields. http://videolectures.net/cikm08_elkan_llmacrf/	Quiz 11 (on Monday) Homework 11: Mashups (due on Friday)	Professor Knoblock Professor Knoblock
Week 13 Apr 13 Apr 15	**OWL Profiles** **Ontology-based Data Integration**	Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. Dl-lite: tractable description logics for ontologies. In Proc. of the 20th National Conference on Artificial Intelligence, 2005. http://www.aaai.org/Papers/AAAI/2005/AAAI05-094.pdf. Hector Prez-Urbina, Ian Horrocks, and Boris Motik. Efficient query answering for owl 2. In International Semantic Web Conference, 2009. Efficient Query Answering for OWL 2. https://www.cs.ox.ac.uk/boris.motik/pubs/puhm09query-OWL2.pdf.	Quiz 12 (on Monday) Homework 12: Information Extraction (due on Friday)	Professor Ambite Professor Ambite

Week 14 Apr 20 Apr 22	**Wrapper Learning** **Wrapper Generation**	Ion Muslea, Steve Minton, and Craig A. Knoblock. A hierarchical approach to wrapper induction. In Proceedings of the 3rd International Conference on Autonomous Agents, Seattle, WA, 1999. http://www.isi.edu/integration/papers/muslea99-agents.pdf. AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Principles of Data Integration, chapter 9. Morgan Kaufmann, 2012. http://proquest.safaribooksonline.com.libproxy.usc.edu/book/databases/9780124160446. W. Crescenzi, G. Mecca, and P. Merialdo. RoadRunner. Towards automatic data extraction from large web sites. 2001. http://www.vldb.org/conf/2001/P109.pdf. B. Cenk Gazen and Steven Minton. Overview of autofeed: An unsupervised learning system for generating webfeeds. In Proceedings of AAAI, 2006. http://www.isi.edu/integration/courses/csci548/Papers/gazen06-aaai.pdf.	Quiz 13 (on Monday)	Professor Knoblock Professor Knoblock
Week 15 Apr 27 Apr 29	**Intellectual Property** **Course Review**	Thomas P. Vartanian and Robert H. Ledig. Scrape it, scrub it and show it: The battle over data aggregation. http://web.archive.org/web/20070818130311/http:/www.ffhsj.com/bancmail/bmarts/aba art.html. Kembrew McLeod. Intellectual property law, freedom of expression, and the web, 2003. http://www.electronicbookreview.com/thread/technocapitalism/proprietary. Electronic frontier foundation. http://www.eff.org/issues/intellectual-property.	Quiz 14	Professor Knoblock Professors Knoblock & Ambite
FINAL May 8 2-4pm	**Final Exam**		During assigned time in the *Schedule of Classes* at www.usc.edu/soc.	

Statement on Academic Conduct and Support Systems

Academic Conduct

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in *SCampus* in Section 11, *Behavior Violating University Standards* https://scampus.usc.edu/1100-behavior-violating-university-standards-and-appropriate-sanctions. Other forms of academic dishonesty are equally unacceptable. See additional information in *SCampus* and university policies on scientific misconduct, http://policy.usc.edu/scientific-misconduct.

Discrimination, sexual assault, and harassment are not tolerated by the university. You are encouraged to report any incidents to the *Office of Equity and Diversity* http://equity.usc.edu or to the *Department of Public Safety* http://capsnet.usc.edu/department/department-public-safety/online-forms/contact-us. This is important for the safety of the whole USC community. Another member of the university community – such as a friend, classmate, advisor, or faculty member – can help initiate the report, or can initiate the report on behalf of another person. *The Center for Women and Men* http://www.usc.edu/student-affairs/cwm/ provides 24/7 confidential support, and the sexual assault resource center webpage http://sarc.usc.edu describes reporting options and other resources.

Support Systems

A number of USC's schools provide support for students who need help with scholarly writing. Check with your advisor or program staff to find out more. Students whose primary language is not English should check with the *American Language Institute* http://dornsife.usc.edu/ali, which sponsors courses and workshops specifically for international graduate students. *The Office of Disability Services and Programs* http://sait.usc.edu/academicsupport/centerprograms/dsp/home_index.html provides certification for students with disabilities and helps arrange the relevant accommodations. If an officially declared emergency makes travel to campus infeasible, *USC Emergency Information* *http://emergency.usc.edu* will provide safety and other updates, including ways in which instruction will be continued by means of blackboard, teleconferencing, and other technology.