MIS 611D "TOPICS IN DATA AND WEB MINING" - Spring 2019

Hsinchun Chen, Professor, Department of MIS

Instructor: Hsinchun Chen, Ph.D., Professor, Management Information Systems Dept, Eller College of Management, University of Arizona
Time/Classroom: T/TH 12:30PM-1:45PM MCCL 430E
Instructor's Office Hours: T/TH 2:00-3:00PM or by appointment
Office/Phone: MCCL 430X, (520) 621-4153
Email/Web site: hchen@eller.arizona.edu; https://ai.arizona.edu/about/director (email is the best way to reach me!)
Class Web site: http://ai.eller.arizona.edu/hchen/mis611D/ (VERY IMPORTANT!)
Teaching Assistants (TAs):

- Shuo Yu, shuoyu@email.arizona.edu, Ph.D. student (office: MCCL 430 Cubical #34-35)
- Hongyi Zhu, zhuhy@email.arizona.edu, Ph.D. student (office: MCCL 430 Cubical #36-37)

TA Office Hours: TA hours will be announced via email.

CLASS MATERIAL (Optional)

- *Data Mining: Practical Machine Learning Tools and Techniques*, by Witten, Frank, Hall & Pal, 4th Edition, 2017, Morgan Kaufmann (also with a 5-week MOOC course). See more at: http://www.cs.waikato.ac.nz/ml/weka/
- *Artificial Intelligence: A Modern Approach*, by Russel & Norvig, 3rd Edition, 2000, Prentice Hall
- *Deep Learning*, by Goodfellow, Bengio & Courville, 2016, MIT Press
- Additional readings and handouts will be distributed in class and made available through the class web site.

COURSE OBJECTIVES

This Ph.D. level course/seminar aims to achieve the following objectives:

- **Introducing computational design science paradigm, landscape, and research methods for emerging, high-impact, and diverse business intelligence and data analytics applications;**
- **Providing the foundation and knowledge in state-of-the-art data, text, and web mining research. including the emerging AI and deep learning methods**

The course will cover data analytics topics and papers in the intersection of computational MIS and CS and will include readings and lectures for the foundational techniques and computational methods. A midterm will be administered to test this foundational knowledge. Selected case studies and research methodologies in MIS and CS for conducting advanced data and web mining research for emerging business and scientific applications will be introduced in the class. Students will be required to execute an individual data or web mining project to demonstrate their ability to conduct hands-on, original data or web mining research. In addition, students will be required to survey key data and web mining conferences, academic institutions, and major industry labs to identify emerging techniques, topics and applications. Students will perform a comprehensive literature review and develop a research proposal for a data and web mining topic of interest. Proposal writing instruction and training based on major federal funding agencies (e.g., NSF, NIH, DOD) will be provided in the class.

PREREQUISITE FOR THE COURSE

Programming experience in selected modern computing languages (e.g., Java, C, C++, Python) and DBMS (SQL).

COURSE TOPICS

Topic 1: Introduction (the field of MIS & CS)

- From computational design science in MIS to applied data science in CS
- Academic/Ph.D. career & progression, from MIS/CS to NSF/NIH
- Business intelligence and analytics, opportunities & techniques
- Emerging AI applications, from face recognition to autonomous vehicle
- Data, text and web mining overview: AI, ML, deep learning
- Data mining and web computing tools (by TAs): Weka, Tableau, Hadoop, SPARK

Topic 2: Web Mining (the changing world)

- Web 1.0, 1995-: WWW, search engines, surface web, spidering, graph search, genetic algorithms
- Web 2.0, 2005-: deep web, social web, web services & mesh-ups, social media, crowdsourcing systems, network sciences
- Web 3.0, 2010-: mobile web, IoTs, mobile & cloud computing, big data analytics, dark web, mobile analytics, cybersecurity
- Web 4.0, 2015-: AI-empowered society, image/face, translation, drones, autonomous vehicles, health, security

Topic 3: Data Mining (the analytics techniques)

- Symbolic learning: decision trees, random forest
- Statistical analysis: regression, principal component analysis, Naïve Bayes
- Statistical machine learning: Support Vector Machines, Hidden Markov Models, Conditional Random Fields
- Neural networks and soft computing: feedforward networks, self-organizing maps, genetic algorithms
- Network analysis: social network analysis, graph models
- Deep learning: Convolutional NN, Recurrent NN, Long Short-Term Memory
- Representation learning: Transfer Learning, Deep Generative Models

Topic 4: Text Mining (handling unstructured text)

- Digital library and search engines
- Information retrieval & extraction: vector space model, entity & topic extraction
- Authorship analysis: lexical, syntactic, structural, and semantic analysis
- Sentiment and affect analysis: lexicon-based, machine learning based
- Information visualization: scientific, text, and web visualization

Topic 5: Emerging Research in Data and Web Mining (major conferences, groups, opportunities)

- Emerging research in major data and web mining conferences: ACM KDD, IEEE ICDM, WWW, ACM SIGIR, ACM CHI, AAAI, IJCAI, ICML, NIPS, ICLR
- Key journals: MISQ, ISR, IEEE TKDE, JAMIA, JBI, JASIST
- Emerging research in major academic institutions: Stanford, Berkeley, CMU, MIT
- Emerging research in major industry research labs: Google, Facebook, Amazon, Baidu, Microsoft
- Emerging data and web mining applications: health, security, e-commerce, AV, drones, robotics
- Proposal writing instruction and training: NSF, NIH review template & process

GRADING POLICY

• Project proposal	5%
• Midterm exam	30%
• Major conference review	15%
• Research project	40%
• Class attendance and participation	10%
TOTAL	100%

MIDTERM EXAM (30%)

The midterm exam will be closed book, closed notes and in the short-essay format (8-10 questions). The questions will be based mostly on classroom lectures. There will be NO Final Exam for this class. Academic integrity will be strictly enforced. Consequence for cheating will be severe.

MAJOR CONFERENCE REVIEW AND PROPOSAL (20%)

Each student will be required to select a major data/text/web mining or computing related conference of interest to him/her. He/she will study significant recent (past 4 years) papers published in the selected conference and provide a systematic review, illustration, and analysis of these papers. Based on some of these papers, each student will also propose an individual research project for the class. The instructor will suggest selected major conferences for consideration. Each student will be welcome to suggest other major conferences of potential interest. The instructor will also provide tangible instruction for proposal preparation and writing based on the National Science Foundation (NSF) guideline. A conference review and project proposal (5 pages) will be needed by the third week of the semester and the conference review presentation (10-12 minutes) will be held in the second half of the semester.

RESEARCH PROJECT PRESENTATION AND PAPER (40%)

Each student will be required to propose and execute an individual, original, and data-driven research project in data/text/web mining for emerging applications of interest to the student. Projects will be judged based on the novelty and originality of the chosen or proposed algorithms and the novelty and impact of the chosen applications. Each student will present at the end of the semester (15-20 minutes) and a final research paper (8 pages, IEEE format) will be submitted after all presentation sessions. The instructor will provide details about the final paper format and structure, mostly based on significant IEEE or ACM conferences. The instructor will also discuss with students about the suitability of selected algorithms and applications.

LECTURES, ATTENDANCE, AND ACADEMIC INTEGRITY

Students are required to attend all lectures on time and honor academic integrity. Missing classes will result in loss of points or administrative drop by the instructor. Students are required to send excuse notes (via email) to the instructor before missing classes. Students are permitted to bring laptop to classroom for note taking purposes, but not for checking email or web surfing. Professional attitude and strong work ethics are needed for this class. Students are encouraged to consult the instructor for advice and help.

LAB SESSIONS and GUEST SPEAKERS

Selected lab sessions will be provided during the semester on the following topics: Web services, cloud computing platforms, Hadoop, Weka, etc. Selected guest speakers will present in the class.

COURSE OUTLINE (tentative)

DATE	TOPIC	CONTENT/NOTES
Jan 10	Syllabus & registration	Class roster, syllabus
Jan 15 (T)	MIS, CS, design science	Readings, discussions
Jan 17	Conferences, journals, NSF	Readings, discussions
Jan 22 (T)	Big Data, BI, data analytics	Readings, discussions
Jan 24	AI, deep learning, applications	Readings, discussions
	PROPOSAL DUE (CONFERENCE & RESEARCH, 5%)	
Jan 29 (T)	**Web Computing & Mining**	Overview, discussions
Jan 31	Cloud, Hadoop, SPARK	TA session
Feb 5 (T)	Web 1.0, Surface Web	Overview, discussions
Feb 7	Search engine, graph search, GA	Readings, lecture
Feb 12 (T)	Web 2.0, Social Web	Overview, discussions
Feb 14	Deep web, social media, SNA	Readings, lecture
Feb 19 (T)	Web 3.0, Mobile Web, IoT, dark web	Overview, discussions
Feb 21	Web 4.0, AI Web	Overview, discussions
Feb 26 (T)	**Data Mining**	Overview, discussions
Feb 28	Symbolic learning, AI, decision trees	ID3, RF
Mar 4-8	SPRING RECESS	NO CLASS
Mar 12 (T)	**MIDTERM EXAM (30%)**	
Mar 14	Statistical analysis, PCA, Bayes	Overview, discussions
Mar 19 (T)	DM tools, Weka, Tableau	TA session
Mar 21	Statistical ML, SVM, CRF, HMM	Readings, lecture
Mar 26 (T)	Neural networks, Backprop, SOM	Readings, lecture
Mar 28	Deep learning, CNN, RNN, LSTM	Readings, lecture
Apr 2 (T)	**CONFERENCE REVIEW PRESENTATION (15%)**	
Apr 4	**CONFERENCE REVIEW PRESENTATION**	
Apr 9 (T)	Transfer & representation learning	Readings, lecture
Apr 11	**Text Mining**	Overview, discussions
Apr 16 (T)	IR/IE, Sentiment/authorship analysis	Readings, lecture
Apr 18	Topic Modeling, Info. Visualization	Readings, lecture
Apr 23 (T)	**RESEARCH PROJECT PRESENTATION (30%)**	
Apr 25	**RESEARCH PROJECT PRESENTATION**	
Apr 30 (T)	**RRESEARCH PROJECT PRESENTATION**	
May 3-9	FINAL EXAM WEEK	NO EXAM FOR MIS 611D
May 9	**FINAL PROJECT PAPER DUE (10%)**	