MIS 611D "TOPICS IN DATA AND WEB MINING" - Spring 2016

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: M/W 11:00AM-12:15PM MCCL 430E
Instructor's Office Hours: M/W 10:00-11:00AM 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):

- Weifeng Li, weifengli@email.arizona.edu, MIS Ph.D. student (office: MCCL 430)
- Sagar Samtani, sagars@email.arizona.edu, MIS Ph.D. student (office: MCCL 430)

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

CLASS MATERIAL (Optional)

- *Data Mining with Weka*, Ian H. Witten and Eibe Frank (also with a 5-week MOOC course). http://www.cs.waikato.ac.nz/ml/weka/
- *Visual Insights: A Practical Guide to Making Sense of Data*, MIT Press, Katy Börner & David E. Polley, 2014 (also with a 7-week MOOC course). http://info.ils.indiana.edu/~katy/S637/
- *An Introduction to Statistical Learning, with Applications in R,* Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, Springer, 2013. http://www.StatLearning.com/
- Additional readings and handouts will be distributed in class and made available through the class web site.

COURSE OBJECTIVES

This Ph.D. level course aims to provide the foundation and knowledge in state-of-the-art data, text, and web mining research for various structured, unstructured and web-based, data-centric applications. Students will become familiar with key data, text and web mining computational methods and techniques. They will also learn to apply such analytical techniques and related methodologies in advanced business, scientific, or web research.

The course will cover data mining, text mining, and web mining methods and case studies. For most enterprises, database management systems (DBMS) are fully deployed and the mostly structured contents (e.g., transaction records, sales, credit ratings) have been used to perform data mining for various business analytics purposes. Increasingly with the overwhelming amount of unstructured contents collected in many e-commerce and web applications (e.g., email, forum postings, clinical discharge summaries, customer comments), text mining techniques have become critical for analyzing such valuable business assets as well. More recently, many critical business applications have been deployed on the modern web computing platforms (e.g., web services, mesh-ups, cloud computing, mobile platform, big data, Internet of Things), data collection, mining, and analytics in a streaming and dynamic web environment has become critically important. MIS Ph.D. students and selected MS students will need to gain

knowledge in these important and high-impact areas of data mining, text mining, and web mining.

In data mining, the topics will include various computational paradigms for association, clustering, classification, anomaly detection, and predictive modeling, including: decision trees, statistical regression, statistical machine learning, graph models, neural networks, and soft computing. The exact methods will be selected to avoid duplication with data mining methods already covered in MIS 611A ("Design Science, Analytical and Computational Research Methodologies") in the Fall semester. In text mining, the topics will include: information retrieval, text segmentation, information extraction, sentiment analysis, authorship analysis, and information visualization. In web mining, the topics will include: search engines, social media systems, web services and APIs, deep web spidering, cloud computing, mobile web, and big data analytics.

The course 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 Computational Sciences for conducting advanced data and web mining research for emerging e-commerce 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

- From design science to data science
- Business intelligence and analytics
- Data, text and web mining overview

Topic 2: Data Mining (selected methods will be introduced to avoid duplication with MIS 611A)

- Symbolic learning: decision trees, random forest
- Statistical analysis: regression, discriminant analysis, principal component analysis
- Neural networks and soft computing: perceptron, self-organizing maps, genetic algorithms, deep learning
- Statistical machine learning and graph models: Support Vector Machines, Hidden Markov Models, Conditional Random Fields
- Network Analysis: social network analysis, graph models, exponential random graph models
- Evaluation and validation: hold-out sampling, cross-validation
- Meta learning: ensemble methods

Topic 3: Text Mining

- Digital library and search engines
- Information retrieval: vector space model, bag of words, text segmentation
- Information extraction: entity extraction, relation extraction, topic extraction
- Question-answering systems: Watson, Siri
- Authorship analysis: lexical, syntactic, structural analysis
- Sentiment and affect analysis: lexicon-based, machine learning based, emoticon analysis
- Multilingual processing: machine translation, cross-lingual IR
- Information visualization: scientific, text and web visualization

Topic 4: Web Mining

- Web 1.0, 2.0, 3.0
- Search engines: ranking, search logs, search algorithms
- Deep web spidering: forums, IRCs, dark web
- Social media and crowdsourcing systems: wisdom of the crowd, sentiment analysis
- Web systems and computing: web services, APIs, and mesh-ups
- Cloud computing and big data analytics: Hadoop, MapReduce, Mahout, Spark
- Mobile computing: Android, iOS, app development
- Internet of Things: mobile sensors, mobile security

Topic 5: Emerging Research in Data and Web Mining

- Emerging research in major data and web mining conferences: ACM KDD, IEEE ICDM, WWW, ACM SIGIR, ACM CHI, AAAI, IJCAI, AMIA, WH
- Emerging research in major academic institutions: Stanford, UIUC, Berkeley, CMU, MIT
- Emerging research in major industry research labs: Google, Facebook, Amazon, EBay, Baidu, Microsoft
- Emerging data and web mining applications: health informatics, security informatics
- Proposal writing instruction and training: NSF, NIH, DOD review template

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 integrality 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 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 13	Syllabus & registration	class roster, student form (TA)
Jan 18 (M)	Martin Luther King Jr. Holiday	NO CLASS
Jan 20	Class overview and expectation	syllabus
Jan 25 (M)	MIS & design science	readings, discussions
Jan 27	BI, analytics, big data	readings, discussions
Feb 1 (M)	Major conferences, journals, NSF	handouts, resources
Feb 3	Web computing & mining	overview, web evolution
Feb 8 (M)	Web 1.0, Surface Web	overview
	PROJECT PROPOSAL DUE (CONFERENCE & RESEARCH, 5%)	
Feb 10	Search engine, deep web	graph search, genetic algorithms
Feb 15 (M)	Web 2.0, Social Web	readings, overview
Feb 17	Social networking, dark web	readings, lecture
Feb 22 (M)	Web 3.0, Mobile Web	readings, SoLoMo web
Feb 24	IoTs, apps, cybersecurity	readings
Mar 1 (M)	Cloud, Hadoop, Spark	TA lecture
Mar 3	Data mining	overview
Mar 7 (M)	Open source tool, Weka	TA lecture
Mar 8	LAST DAY TO DROP (with "W")	
Mar 9	Classification algorithms	Regression, discriminant analysis
*Mar 14-18	SPRING RECESS	NO CLASS
Mar 21 (M)	**MIDTERM EXAM (30%)**	
Mar 23	Classification algorithms	ID3, Random Forest, SVM
Mar 28 (M)	Statistical machine learning	HMM, NN, deep learning
Mar 30	Clustering algorithms	K-means, SOM, LDA
Apr 4 (M)	**CONFERENCE REVIEW PRESENTATION (15%)**	
Apr 6	**CONFERENCE REVIEW PRESENTATION**	
Apr 11 (M)	Text mining	overview
Apr 13	Topic extraction and authorship	techniques
Apr 18 (M)	Sentiment analysis	opinion mining, social media
Apr 20	Q/A systems	IBM, Watson
Apr 25 (M)	Data visualization	HCI
Apr 27	**RESEARCH PROJECT PRESENTATION (30%)**	
May 2 (M)	**RESEARCH PROJECT PRESENTATION**	
May 4	**RRESEARCH PROJECT PRESENTATION**	
May 6-12	FINAL EXAM WEEK	NO EXAM FOR MIS 611D
May 11	**FINAL PROJECT PAPER DUE (10%)**	