MIS 510 "Web Computing and Mining" - Spring 2014
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: Section 1 M/W 11:00-12:15PM MCCL 123; Section 2 M/W 3:30-4:45PM MCCL 123
Instructor's Office Hours: M/W 2:00-3:00PM or by appointment
Office/Phone: MCCL 430X, (520) 621-4153
Email/Web site: hchen@eller.arizona.edu; http://ai.arizona.edu/hchen (email is the best way to reach me!)
Class Web site: http://ai.arizona.edu/mis510 (VERY IMPORTANT!)
Teaching Assistants (TAs):

- Jonathan Shan Jiang, jiangs@email.arizona.edu, MIS Ph.D. student (office: MCCL 430)
- Julian Chenhui Guo, chguo@email.arizona.edu, MIS Ph.D. student (office: MCCL 430)

Desert Angels (DA) coordinator: Justin Williams, jwilliams@CMI.arizona.edu, Founder, Startup Tucson
TA/DA Office Hours: TAs will be available 2/5-2/26 and 4/9-4/30. The Desert Angels group will also provide office hours for discussion of project ideas in mid-February. Actual time/place will be announced.

CLASS MATERIAL (Optional)

- "Data Mining: Concepts and Techniques," Jiawei Han and Micheline Kamber, Morgan Kaufmann, Third edition, 2011.
- "Data Mining: Practical Machine Learning Tools and Techniques," Ian H. Witten, Eibe Frank, and Mark Hall, Morgan Kaufmann, Third edition, 2011.
- The WEKA Data Mining Software, Ian H. Witten and Eibe Frank. http://www.cs.waikato.ac.nz/ml/weka/index.html
- Please consult class web site for many excellent past web mining class project ideas and implementations. http://ai.eller.arizona.edu/mis510/syllabus/4_ProjectResources/index.asp
- Additional readings and handouts will be distributed in class and made available through the class web site. http://ai.eller.arizona.edu/hchen/class510.htm

OBJECTIVES OF THE COURSE

This course introduces web computing and mining techniques, systems, and applications that are suitable for developing web-based information systems in e-commerce, knowledge management systems, web/data/text mining, business intelligence, security informatics, and health informatics. The course contains lectures, readings, lab sessions, and two hands-on group system development projects. The course will cover **web mining, data mining, and text mining**. In web mining, we will introduce web architecture, search engines, search algorithms, web services/APIs, Web 2.0/3.0, cloud computing, and mobile web. State-of-the-art data and text mining algorithms are discussed in the context of modern and emerging information systems in business, security, and health informatics. Selected data mining algorithms such as neural networks, decision trees, statistical learning, and social network analysis will be presented for clustering, classification, and predictive analytics problems. Information retrieval, natural language processing, sentiment analysis, authorship analysis, and information visualization will be discussed in text mining, especially for emerging business intelligence and big data applications.

Selected algorithms will be introduced in the classroom using English-like pseudo-code. This course requires hands-on web-based system development and business analytics. The focus of the course is on web computing and mining concepts and applications. Two TAs with good technical skills will be provided to assist in guest lectures and lab sessions for system development and analytics. Members of the highly-regarded Tucson Desert Angels will also help develop and evaluate the final group web mining project ideas and business models. **The class will prepare students to gain cutting-edge web computing and mining knowledge and hands-on experiences that are critical for future careers at leading Internet companies (e.g., Google, Facebook, Twitter, Amazon, Expedia, Microsoft, IBM) and/or for future web entrepreneurial activities.**

PREREQUISITE FOR THE COURSE

Programming experience in Java and DBMS and good Web knowledge; capability and aspiration for learning.

GRADING POLICY

• Group cyber analytics project (2/26)	20%
• Midterm exam (3/24)	30%
• Group web mining demo V1.0 (4/7)	10%
• Group web mining demo V2.0 (4/30-5/7)	30%
• Group web mining paper (5/14)	5%
• Class attendance and participation	5%
TOTAL	100%

GROUP CYBER ANALYTICS PROJECT (20%)

Cyberspace has transformed the daily lives of people for the better. The rush to adopt cyberspace, however, has exposed its fragility and vulnerabilities: corporations, agencies, national infrastructure and individuals have been victims of cyber-attacks. Cyber Analytics offers a unique opportunity for understanding the vulnerabilities of information infrastructures and the ecosystem and dynamics of the international hacker community.

In this team project (each team will consist of 4 members, the same as the final Group Web Mining Project), you are required to perform analytics of relevance to emerging cybersecurity concerns using two large-scale security testbeds: the Shodan and the Hacker Web. **Shodan** is a search engine that lets you find specific types of computers (routers, servers, etc.) or internet-enabled devices (Internet of Things) using a variety of search filters. Large numbers of devices and computer systems are increasingly connected to the Internet. Shodan users are able to find systems including traffic lights, security cameras, home heating systems as well as control systems for water parks, gas stations, water plants, power grids, nuclear power plants and particle-accelerating cyclotrons; most have little security. Several Shodan APIs are available for accessing the database. The **Hacker Web**, developed by the AI Lab of the University of Arizona, contains multilingual forum postings from members of the international hacker community. The tetsbed will allow researchers and practitioners to: (1) detect, classify, measure and track the formation, development and spread of topics, ideas, and concepts in cyber attacker social media communication; (2) identify important and influential cyber criminals and their interests, intent, sentiment, and opinions in online discourses; and (3) induce and recognize hacker identities, online profiles/styles, communication genres, and interaction patterns. Both testbeds contain multi-million records of great value to cybersecurity research.

Each project team will be required to develop critical and relevant cybersecurity related research questions or hypotheses. They will access either or both testbeds via APIs or database languages (SQL) to extract contents of relevance to their research questions. Statistical analysis or data mining techniques need to be applied to derive insights relating to the research questions and draw scientific and systematic conclusions. A set of research slides (20-30 slides) and a 10-page research paper will be submitted on February 26, 2014. Selected interesting projects will be requested to summarize their findings in class. TAs will be available for issues relating to accessing testbeds, APIs, and analytics.

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.

GROUP WEB MINING PROJECT ("Business Web Mining and Analytics: Cloud and Mobile Applications"): (45%)

Most e-commerce firms such as Amazon, eBay and Google have opened up their billion-dollar data troves through web Application Programming Interfaces (APIs). Thanks to the emergence of new software standards known collectively as Web Services, web sites can swap functions, features, and data with one another like never before, all in a highly automated fashion. Increasingly, e-commerce firms are behaving like competing software platforms (i.e., cloud computing) that can be programmed as easily as the operating system on your PC. The rise of Web 2.0 (social), 3.0 (mobile, cloud), open data, and Web Services goes even further, holding out the promise of automating the links between online businesses.

Your final group project will be about "Business Web Mining and Analytics." Each project team will consist of 4 members of diverse skills, who will participate in the design, coding, implementation, analysis, and maintenance of a prototype web-based business. You will be using various cloud computing platforms, e.g., Amazon, Google, Saleforce.com for your projects. You are required to create a web business, with a complete web site and business functionalities for specific customers using at least one of the three main e-commerce APIs (Amazon, eBay or Google) and others. You need to present a strong business case and design attractive system features. Your project will be judged based on its system functionalities, novelty, and business feasibility. Students are required to use 10+ diverse and multimedia Web APIs (i.e., mash-ups). Web 2.0 and 3.0 applications are strongly encouraged, e.g., YouTube, Facebook, twitter, Flickr, Instagram, etc., especially for creative and societally beneficial (e.g., health, lifestyle, environmental) problems. Your project also needs to have a "mining" or analytics component, based on contents or data provided for your web business. Open source data mining tools (e.g., Weka, Mahout) are encouraged, in addition to existing commercial database and analytics software (e.g., MS SQL, SPSS, MySQL).

Each team member will be required to participate in all aspects of the project (contributing at least 25% of total project effort). At the end of the semester, each group will present their work in actual demos and presentations. Demo V1.0 will be presented (and graded, 10%) on 4/7/2014. This initial prototype should include 5+ web APIs, initial database, and web interface to the system and data. Demo V2.0 will be presented (and graded, 30%) during the last week of the semester (4/30-5/7, 2014). This version should demonstrate complete and robust system functionality, strong business case, and novelty. Each group will be rated by the instructor, Desert Angel members, and other groups during their presentation, and each team member will be rated by other members in the same team. Students can consult past project for ideas; however novelty (something new) is highly valued. For 2014, students are encouraged to develop their applications on cloud platforms and with social media and/or mobile apps focus. Your final report (5%, 10-15 pages) should summarize the complete project and detail the role and actual contributions of each team member. The final report is due 5/14/2014 (the last day of the Final Exam week).

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, web architecture, web APIs, mash-ups, cloud computing platforms, MapReduce, Hadoop, Weka, text mining, etc. Selected guest speakers will present in the class.

COURSE OUTLINE (tentative)

DATE	TOPIC	CONTENT/NOTES
Jan 15	Course introduction	syllabus, roster, overview
Jan 20 (M)	Martin Luther King Jr. Holiday	NO CLASS
Jan 22	Web mining overview	class projects, cloud computing
Jan 24 (F)	**Team building sessions**	project pitch, Desert Angels
Jan 27 (M)	Web applications, APIs, Amazon	TA guest lecture
Jan 29	Web 1.0, search engines	architecture and components
Jan 31 (F)	**Lab sessions**	system, APIs, cloud
Feb 3 (M)	Cybersecurity	overview, Shodan
	WEB MINING PROPOSAL DUE	
Feb 5	Cyber analytics	Hacker Web
Feb 10 (M)	Search algorithms	graph search
Feb 12	Cloud, MapReduce, Hadoop	TA guest lecture
Feb 17 (M)	Google & Facebook Story	readings, business models
Feb 19	Web 2.0	readings, overview
Feb 24 (M)	Social networking, crowdsoucing	readings, systems
Feb 26	Web 3.0	readings, SoLoMo web
	CYBER ANALYTICS PROJECT DUE (20%)	
Mar 3 (M)	Mobile web	Android, virtual web
Mar 5	BI & analytics	readings, Big Data
Mar 10 (M)	Data mining overview	overview, applications
Mar 11	LAST DAY TO DROP (with "W")	
Mar 12	Classification algorithms	Regression, neural networks
*Mar 15-23	SPRING RECESS	NO CLASS
Mar 24 (M)	**MIDTERM EXAM (30%)**	
Mar 26	Classification algorithms	ID3, SVM, HMM
Mar 31 (M)	Clustering algorithms	K-means, hierarchical, SOM
Apr 2	Open source analytics	Weka, Mahout
Apr 7 (M)	**WEB MINING PROJECT PRESENTATION/DEMO 1.0 (10%)**	
Apr 9	Text mining overview	NLP, Watson
Apr 14 (M)	Digital library/IR	DL initiatives
Apr 16	Natural language processing	overview, tools
Apr 21 (M)	Topic extraction and authorship	techniques
Apr 23	Sentiment analysis	opinion mining, social media
Apr 28 (M)	Information visualization	HCI
Apr 30	**WEB MINING PROJECT PRESENTATION/DEMO 2.0 (30%)**	
May 5 (M)	**WEB MINING PROJECT PRESENTATION/DEMO 2.0**	
May 7	**WEB MINING PROJECT PRESENTATION/DEMO 2.0**	
May 9-15	FINAL EXAM WEEK	NO EXAM FOR MIS 510
May 14	**FINAL PROJECT PAPER DUE (5%)**	