COLLEGE OF COMPUTING AND INFORMATION
DEPARTMENT OF INFORMATION STUDIES

IIST 433-1
Information Storage and Retrieval

Fall 2008
Room: SS 134

Instructor:
Alon Friedman, PhD
E-mail: af534922@albany.edu
Office Hours: Tuesdays 1:15pm to 2:00pm and on Thursdays between 2:10 to 3:30 in Room: LI-72, otherwise via email only
Class Hours: Thursdays 2:00pm to 5:30pm in Room # 134

COURSE DESCRIPTION:

This course will introduce the student to the fundamentals of information storage and retrieval (ISR) systems, including components, models, structures, information representation, vocabulary control, search strategies (query language/operation, indexing and search), User interface and visualization and User dimension and evaluation. We will also cover some practical information retrieval (IR) systems on the Web and for the libraries. A good familiarity with computers and some programming experience is highly desirable but not necessary.

COURSE OBJECTIVES:

- To identify the various components of an information storage and retrieval system.
- To become familiar with different models and structures an IR system may take.
- To understand the theoretical foundations of various IR methods.
- To be able to design and implement IR systems using off-the-shelf software packages.
- To examine the factors that influences the performance of an IR system.
- To be aware of the current research in the field and on the Web.

TEXTBOOKS:

1. Heting Chu. *Information Representation and retrieval in the digital age*. Medford, NJ: Information Today Press 2003. (ASIS&T monograph Series).
2. Charles T. Meadow, Bert R. Boyce, Donald H. Kraft, Carol L Barry. *Text Information Retrieval Systems*. Third Edition (Library and Information Science) (Library and Information Science). Academic Press 2007-01-23. ISBN: 0123694124/ 0-12-369412-4

CLASS MEETING:

The course will meet once a week for 13 weeks. As a result, the students are requested to attend every single class.

CLASS ATTENDANCE:

Attendance will be taken each class. Students need to attend class. In the unavoidable event of an absence, students should make arrangements with other students to pick up class notes and assignments. The instructor will allow time the first meeting of class to find study partners. Students who miss more than two classes will have their final grade dropped by 15 points.

READING:

Students are expected to read the assignments before upcoming classes.

ASSIGNMENTS:

All work is due at the time assigned on each homework assignment and will have the grade reduced by 5 points if no previous permission for lateness was obtained from the instructor. Groups of individuals may work on the problem sets, but each individual must hand in a completed assignment.

PROJECT(S):

Each student in the class will submit three projects: (1) Use a conventional database software package (e.g., Microsoft Access, EndNote, ProCite, etc.) to design and implement an IR system consisting of about 50 records, on a topic of your choice (e,g., Music, Books, Staples etc.). (2) Create a website that will support your database file and provide the ability to search. (3). Evaluation of an ISR system. Dates for submission are marked on the class schedule.

FINAL EXAM:

The final exams must be taken when scheduled! If an exam is missed, a grade of zero (0) will be recorded. If a student has a valid excuse or has notified the professor in advance of the absence the exam may be made up.

INCOMPLETES:

Students who do not turn their final projects in on time should expect their grades will be reduced by 25%. Late homework assignments lose 5 points at the discretion of the professor.

PLAGIARISM AND CHEATING:

Due to the intensive nature of this course, students are encouraged to form study groups and to work together on assignments. Learn by interacting with one another—support and help one another. However, unannounced quizzes will clearly be expected to reflect individual effort—you are expected to neither give nor receive assistance from anyone. As a policy for this course, plagiarism, self-plagiarism or cheating will result in a failing grade for the course. High standards of academic honesty will be upheld in this class at all times. Plagiarism (in writing or code) will result in a zero for the assignment in which the plagiarism occurred, a zero for the course and a referral to the Dean of Undergraduate Studies. After two referrals to the Dean's office for plagiarism, students are automatically referred to the Office of Judicial Affairs.

CLASS COMFORT:
Please turn off your cell phone. If absolutely necessary leave it on, but exit the room as quietly as possible.

GRADING:

Tasks	Percentage
1. Assignments	20%
2. Three Projects	30%
3. Unannounced Quizes	20%
3. Final Exam	30%

A	96-100
A-	90-95
B+	88-89
B	84-87
B-	80-83
C+	78-79
C	74-77
C-	70-73
D+	68-69
D	60-67
E	Below 60

CLASS SCHEDULE:

Date	Topics	Reference	Homework Assignments
Aug 28	Review Syllabus. Introduction: * What is information * What is information explosion and overload * Pioneers and milestones in ISR	Read: Meadow et. al: Ch. 1	Homework assignment will be given after each class
Sept. 4	Basics of Information Storage and Retrieval: * Key concepts of ISR * Components of an ISR system * Types of ISR systems	Read: Meadow et. al: Ch. 2 Chu: Ch. 1-2	Homework assignment will be given after each class
Sept. 9	Structure of ISR:	Read: Meadow et. al:	Homework

	* Record structure * File structure * Sequential file * Database structure ** Hierarchical ** Network ** Relational	Ch. 3-4, 6 Date. J.: An introduction to database systems. Ch. 1-2.	assignment will be given after each class
Sept. 18	ISR representation: * What is information representation * What are information representation attributes and values	Read: Meadow et. al: Ch. 3-4 and Chu: Ch. 2-3	I. Homework assignment will be given after each class II. First Project is Due
Sept. 25	Language in ISR * Natural language vs. controlled vocabulary * Types of controlled vocabulary ** Thesauri ** Subject heading ** Classification schemes	Read: Meadow et. al.: Ch. 5-7, 12 and Chu: Ch. 4	Homework assignment will be given after each class
Oct. 2	Retrieval Techniques and Query representation * Retrieval techniques ** Query representation ** Search techniques	Read: Meadow et. al.: Ch. 7-9 and Chu: Ch. 5	I. Homework assignment will be given after each class. II. Second Project is Due
Oct. 9	Retrieval Approaches * Retrieval by searching * Retrieval by browsing * The hybrid approach * The integrated approach	Read: Meadow et. al.: Ch. 10-11 and Chu: Ch. 6	Homework assignment will be given after each class
Oct. 16	Information Retrieval Models * Boolean logic * Vector space * Probabilistic * Extensions of major ISR models * Web Search engine models	Read: Meadow et. al.: Ch. 12-13 and Chu: Ch. 7	Homework assignment will be given after each class
Oct. 23	Retrieval of Information Unique in Content and Format * Multilingual information * Multimedia information * Hypertext and hypermedia	Read: Meadow et. al.: Ch. 14-15 and Chu: Ch. 8-9	Homework assignment will be given after each class
Nov. 30	IR Systems: * Library systems * Computer systems * CD-ROM systems * Web and Internet systems	Read: Chu: Ch. 9 plus Begum and Ahmed (2001).	I. Homework assignment will be given after each class II. Third Project is Due
Nov. 6	Trends in ISR * Artificial Intelligence * Natural language processing * Expert Systems	Read Chu: Ch. 10 plus Meadow et. al.: Ch. 15	Homework assignment will be given after each class.

	* Semantic Web		
Nov. 13	The future of information retrieval		
Nov. 20	First review	Read: Chu Ch. 1-5 and Meadow et. al. 1-4	
Nov.27	No class		
Dec. 4	Final Review	Read: Chu Ch. 6-10 and Meadow et. al. 5-15	
Dec. 9	Reading day		
Dec. 13	FINAL EXAM		

PARTIAL REFERENCE LIST:

Baeza-Yates, Ricardo, and Ribeiro-Neto, Berthier. (1999). *Modern information retrieval*. New York: ACM Book Press.

Begum, Suraiya and Ahmed, Zabed (2001). Development of Web-based IR Systems: A Review.
Information Society Today. URL:http://www.infosciencetoday.org/ir.htm

Chowdhury, Gobinda G. (2004). *Introduction to modern information retrieval*. London: Facet.

Date, J. C. 1986. An introduction to database systems: vol. I (4th ed.), Addison-Wesley Longman Publishing Co., Inc., Boston, MA

Harter, Stephen P. (1985). *Online information retrieval: Concepts, principles, and techniques*. New York: Academic Press.

Ingwersen, Peter, and Järvelin, Kalervo. (2005). *The turn: Integration of information seeking and retrieval in context*. Dordrecht: Springer.

Lancaster, F.W., and Warmer, Amy J. (1993). *Information retrieval today*. Arlington, VA: Information Resources Press

Karen Sparck Jones. (1999). The role of artificial intelligence in information retrieval**.** Journal of the American Society for Information Science. Volume 42 Issue 8, Pages 558 - 565

Marchionini, Gary. (1995). *Information seeking in electronic environments*. New York: Cambridge University Press. (Cambridge Series on Human-Computer Interaction 9).

Meadow, Charles T., Boyce, Bert R., and Kraft, Donald H. (1999). *Text information retrieval systems*. Orlando, FL: Academic Press.

Pao, Miranda Lee (1989). *Concepts of information retrieval*. Englewood, CO: Libraries Limited.

Berkeley Digital Library SunSITE. (1999). Guidelines for Web document style and design. http://sunsite.berkeley.edu/Web/guidelines.html.

Berners-Lee, Tim, Hendler, James, and Lassila, Ora. (May 2001). The Semantic Web. *Scientific American*, 40-43. Also available at http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&sc=I100322.

Boston, Tony. (2005). Exposing the deep Web to increase access to library collections. Refereed paper at Ausweb 2005 Conference. http://ausweb.scu.edu.au/aw05/papers/refereed/boston/paper.html.

Brooks, Terrence A. (2003). Web search: How the Web has changed information retrieval. *Information Research*, 8(3), paper no. 154. http://InformationR.net/ir/8-3/paper154.html.

Bruce, Harry. (2005). Personal, anticipated information need. *Information Research*, 10(3), paper 232. http://InformationR.net/ir/10-3/paper232.html.

Downie, J. Stephen. (2003). Music information retrieval. Annual Review of Information Science and Technology, 37, 295-340. Also available from http://music-ir.org/downie_mir_arist37.pdf .

Hendler, James. (March 11, 1999). Is there an intelligent agent in your future? *Nature*, http://www.nature.com/nature/webmatters/agents/agents.html.

Nielsen, Jakob. (1996, Updated 2004). Top ten mistakes in Web design. http://www.useit.com/alertbox/9605.html .

Nielsen, Jakob. (1999a). "Top ten mistakes" revisited three years later. http://www.useit.com/alertbox/990502.html .

Nielsen, Jakob. (1999b). Top ten new mistakes of Web design. http://www.useit.com/alertbox/990530.html .

Nielsen, Jakob. (1999c). Ten good deeds in Web design. http://www.useit.com/alertbox/991003.html.

Nielsen, Jakob. (2002). Top ten Web-design mistakes in 2002. http://www.useit.com/alertbox/20021223.html.

Nielsen, Jakob. (2003). Top ten Web-design mistakes in 2003. http://www.useit.com/alertbox/20031222.html.

Text Retrieval Conference (TREC). (On-going). http://trec.nist.gov/ .

Svenonius, Elaine. (1990) *Design of Controlled Vocabularies*. in *Encyclopedia of Library and Information Science*, v. 45, sup. 10.

Future of information Retrieval discussion:
http://technorati.com/videos/youtube.com%2Fwatch%3Fv%3DXpXtRu0XfeA