School of Information Sciences - University of Pittsburgh

LIS 2685: Advanced Topics for Information Retrieval **[Current as of 8/16/2016]**
Spring 2011

Instructor:

Daqing He, PhD
School of Information Sciences, University of Pittsburgh
Phone: 412-624-2477
E-mail: dah44@pitt.edu
Office: Room 618, Information Science Building
Office Hours: by appointment

Graduate Student Assistant:
Jiepu Jiang – E-mail: jij29@pitt.edu
Office hours: ???
Office: 707 IS building

I. *Course Description:*

This course offers an examination of problems and techniques related to storing and accessing information widely existing in libraries and on the Web. Aiming to provide students more in depth overview of several approaches to information access with a primary focus on search-based information access, this course builds upon the knowledge of related courses on information retrieval and information organization. The content of the course includes more detailed discussions of: the design principles of underneath retrieval engines, database technologies for bibliographic collections, the search strategies and tactics for library and Web searches, the models for describing users' information seeking behaviors, and newly emerged topics related to searching in digital libraries, Web and other interesting areas.

Prerequisite: LIS2005 ORGANIZING AND RETRIEVING INFORMATION

II. *Course Objectives:*

Students will understand:

1. the environment of information retrieval;
2. the principal components of information retrieval systems, Web search engines and online databases;
3. the impacts and implications of digital libraries and Web to the retrieval services in libraries.

Upon successful completion of the course, students will be able to do the following:

1. analyze an information problem and identify appropriate retrieval process;

2. evaluate the emerging information retrieval practices in library services and on the Web.

Please note that this syllabus is subject to change. Any changes will be announced in class and by e-mail and a revised copy of the syllabus will be placed online on the CourseWeb site.

III. CourseWeb Information:

CourseWeb is a Web-based system using BlackBoard software that facilitates course-related communication as well as distribution of course materials and grades. You can access CourseWeb at http://courseweb.pitt.edu . You must log in with your University Computer Account – this is the one that goes with your 'pitt.edu' e-mail address. If you do not have a Pitt account, please contact Computing Services (CSSD) at 412-624-HELP [4357] to find out how to get one. Course-related e-mail will be sent to your Pitt e-mail account. If you do not read e-mail on your Pitt account, you are responsible for forwarding any e-mail received on your Pitt account to the e-mail address that you use. See http://accounts.pitt.edu/ for information on managing your Pitt account and forwarding e-mail. If you have trouble logging in to CourseWeb, you may need to log in to the accounts website above to activate your Pitt e-mail account. Call 412-624-HELP with any problems relating to your account.

IV. Textbook and Readings:

There is no required textbook for this course. However, various parts of the following books will be used in the class:

1. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze, "Introduction to Information Retrieval". Cambridge University Press. 2008. Available at http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html. Referred as "IIR" subsequently.
2. Stefan Büttcher, Charles L. A. Clarke, Gordon V. Cormack, "Information Retrieval: Implementing and Evaluating Search Engines." MIT Press. 2010. Sample chapters are available at http://www.ir.uwaterloo.ca/book/. Referred as "IES" subsequently.
3. Ricardo Baeza-Yates, Berthier Riberiro-Neto, "Modern Information Retrieval", Addison Wesley, 1999. ISBN-10: 020139829X. Referred as "MIR" subsequently.
4. Richard K. Belew, "Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW", Cambridge University Press, 2000. Referred as "FOA" subsequently.

All course content reading will be provided by 3-4 required readings each week. You will be asked to submit a short reading note each week before the class in the blog space you created for this course in blogger.com. The note is informal in style – even bulleted lists can be used when appropriate, however, the response should clearly indicate the context, including the part of the text that triggered your questions. Do not summarize the readings. Instead, discuss your thoughts, ideas, and questions related to them. Please include at least one question each week that you would like to invoke discussion on it. The note for each week's readings should be submitted by 11:59pm the Saturday before the class. As described below, 10 responses are required as part of your final grade, each of which counts for .5 participation point.

Readings will generally be available via CourseWeb (if available in electronic format) and on reserve in the IS Library. Additional readings might be communicated in class and on CourseWeb.

V. Course Schedule At a Glance

Date	**Unit**	**Notes**
Jan 11	1. Information Retrieval Environment and Course Overview	
Jan 18	2. IR Theories and Systems: Collection and Query Processing	Introduction of Term Projects
Jan 25	3. IR Theories and Systems: Matching Models	Assignment 1 Out
Feb 1	4. Search Strategies, Tactics and Relevance Feedback	
Feb 8	No class, on travel	
Feb 15	5. Virtual Reference Services	Guest Lecturer by Jongdo Park
Feb 22	6. Evaluation of Search Results	Assignment 2 Out
Mar 1	7. Database technologies I – Basic Relational model	
Mar 8	Spring Break	
Mar 15	8. Database technologies II – Simple SQL	Assignment 3 Out
Mar 22	9. Users and Information Seeking Models	Take home Exam Out
Mar 29	10. Information Retrieval in Digital Libraries	
Apr 5	11. Information Retrieval on the Web	
Apr 12	12. Intelligent Information Retrieval	
Apr 19	13. Information Retrieval Services and Future Trends	

VI. Course Schedule in Detail

Unit 1: Information Retrieval Environment and Course Overview

Objectives: After this class, you should be able to

- explain the basic concepts of information retrieval
- restate the relationship between IR process and its surrounding environment
- explain and identify the major factors that affect IR process
- restate the expectations and requirements of the course
- make decision on whether attending the course

Required Readings

1. Gary Marchionini. “Information Seeking in Electronic Environments”. Chapter 3, page 32-49. Cambridge University Press, 1995. (available in CourseWeb)

2. MIR sections 1.1-1.4 (available at http://people.ischool.berkeley.edu/~hearst/irbook/)
3. FOA Section 1.1 in Chapter 1. PDF version of Ch. 1. on author's site http://www.cs.ucsd.edu/~rik/foa/.
4. N.J. Belkin, R.N. Oddy, H.M. Brooks, "ASK for information retrieval: Park 1 background and theory", Journal of document 38(2), 1982. (available in CourseWeb)

Unit 2: IR Theories and Systems: Collection and Query Processing

Objectives: After this class, you should be able to

- explain the basic process involved in collection and query processing
- explain how documents are processed into inverted files, and
- write inverted files based on some simple document collections

Required Readings:

1. IIR sections 1.2, 1.3, chapters 2 and 3
2. FOA Sections 1.2-1.5 in Chapter 1. PDF version of Ch. 1. on author's site http://www.cs.ucsd.edu/~rik/foa/.

Unit 3 : IR Theories and Systems: Matching Models

Objectives: After this class, you should be able to

- explain the basic ideas of major matching models
- select right match models for a given query

Required Readings:

1. IIR sections 1.4, chapters 6, 11 and 12

Unit 4 : Search Strategies and Tactics

Objective: After this class, you should be able to

- develop an ability to analyze searcher's information needs, formulate effective search strategies and to provide techniques and practice so that students are able to perform an efficient search on commonly used information retrieval systems.
- Develop ability to utilize devices provided by the system for improving retrieval effectiveness Boolean logic, proximity searching, truncation and other tools

Required Readings:

1. Borgman C.L., Moghdem D. and Corbett, P.K. (1984). Effective Online Searching: A Basic Text. New York: M. Dekker. Chapter 2: Characteristics of a Good Searcher, pp. 13-18. (Available in CourseWeb)
2. Hawkins D.T. and Wagers, R. (1982). Online bibliographic search strategy development. Online 6(3): 12-19. (Available in CourseWeb)
3. Basch, R. (1989). The Seven deadly sins of full-text searching. Database, 12(4): 15-23. (Available online through InfoTrac- at Pitt E-Journal http://ug4fn7ck2h.search.serialssolutions.com/)
4. Spink, A., & Saracevic, T. (1997). Interaction in information retrieval: Selection and effectiveness of search terms. Journal of the American Society for Information Science, 48(8), 603-609. http://www.scils.rutgers.edu/~tefko/JASIS1997.pdf

---

Unit 5: Virtual Reference Services

Objective: After this class, you should be able to

- to explain the difference between traditional reference service and digital reference service focusing technologies.
- to describe key aspects of digital reference.
- to discuss the relationship between digital reference and information retrieval.

Required Readings:

1. Penka, J. T. (2003). The Technological Challenges of Digital Reference: An Overview. D-Lib Magazine, 9(2). Available from http://www.dlib.org/dlib/february03/penka/02penka.html
2. Lankes, R. D. (2003). The Digital Reference Research Agenda. *Journal of the American Society for Information Science and Technology, 55(4),* 301-311.
3. Nicholson, S. (2003). Exploring the Future of Digital Reference through Scenario Planning. In Lankes, R. D., Nicholson, S., and Goodrum, A. (Ed.), *The digital reference research agenda*. (pp. 177-182). Association of College & Research Libraries. Available from http://www.bibliomining.com/nicholson/nicholsonpdfs/refscen.pdf
4. Pomerantz, Jeffrey (2004). The current state of digital reference: validation of a general digital reference model through a survey of digital reference services. Information processing & management, 40 (2).

Unit 6: Evaluating Search Results and System Performance

Objective: After this class, you should be able to

- explain the evaluation criteria and general issues relating to the evaluation of search results
- adopt an evaluation framework for your own need

Required Readings:

1. IIR chapter 8
2. Barry, Carol L. (1994). User-defined relevance criteria: An exploratory study. Journal of the American Society for Information Science, 45(3):149-159. (available in CourseWeb)
3. Saracevic, T. (1975). Relevance: A review and framework for the thinking on the notion in information science. Journal of the American Society for Information Science, 26: 321-343. (available in CourseWeb)

Interesting Readings:

4. Ellen M. Voorhees , Report on TREC-9, ACM SIGIR Forum . http://portal.acm.org/citation.cfm?id=381260&coll=ACM&dl=ACM&CFID=50124169&CFTOKEN=39429722.

Unit 7: Database Technologies I – Relational Database

Objective: After this class, you should be able to

- understand the logical view of data in relational model
- understand the characteristics of relational operators
- able to calculate the outcomes of relational operators
- tell how to define and refine relationships in database design
- understand how to develop a ER diagram for database design
- understanding what normalization is and its role in database design
- tell 1NF, 2NF, 3NF, BCNF, and 4NF

Required Readings:

1. Sharon Allen and Evan Terry.. (2005). Beginning Relational Data Modeling. Chapter 2 and 3. http://books.google.com/books?id=62CFtFea0NsC&printsec=frontcover#v=onepage&q=&f=false
2. Mike Hillyer. An Introduction to Database Normalization. http://dev.mysql.com/tech-resources/articles/intro-to-normalization.html
3. Paul Litwin. Fundamentals of Relational Database Design. http://www.deeptraining.com/litwin/dbdesign/FundamentalsOfRelationalDatabaseDesign.aspx

---

Unit 8: Database Technologies II – Simple SQL

Objective: After this class, you should be able to

- understanding the basic commands and functions of SQL
- able to use SQL for data administration (e.g. create tables, indexes)
- able to use SQL for data manipulation (e.g., add, modify, delete data)

Required Readings:

1. Philip Greenspun. SQL for Web Nerds, chapter 3 “simple queries” http://philip.greenspun.com/sql/
2. W3school. SQL tutorial. http://www.w3schools.com/SQl/default.asp

Unit 10: Users and Information Seeking Models

Objective: After this class, you should be able to

- Restate theories and research on people's information seeking behavior
- Explain the characteristics of different user groups in seeking information

Required Readings:

1. Kuhlthau, C. (1991). Inside the search process: Information seeking from the user's perspective. Journal of the American Society for Information Science, 42(5), 361-371. (Available online through Pitt E-Journal http://ug4fn7ck2h.search.serialssolutions.com/)
2. Tenopir, C. (1997). Common End User Errors. Library Journal 122 (8): 31-32. (Available online through Pitt E-Journal http://ug4fn7ck2h.search.serialssolutions.com/)
3. Borgman, C.L. (1989). All Users of Information Retrieval Systems are not Created Equal: An Exploration into Individual Differences. Information Processing & Management 25 (3): 237-251. (Available online through Pitt E-Journal http://ug4fn7ck2h.search.serialssolutions.com/)

Interesting Readings:

4. Marchionini, G. (1995). Information Seeking in Electronic Environments. Cambridge, UK: Cambridge University Press. Chapter 1, Information and Information Seeking, pp. 1-10.
5. Bates, M. (1989). The design of browsing and berrypicking techniques for the online search interface. Online Review, 13(5), 407-424. http://www.gseis.ucla.edu/faculty/bates/berrypicking.html

Unit 11: Retrieving information in Digital Libraries

Objectives: After this class, you should be able to

- identify the similarity and difference between search in digital libraries and in traditional libraries
- explain the basics of retrieval techniques for digital libraries
- search effectively and efficiently using some DL search tools

Required Readings:

1. Papadakis, I. et al (2009) Subject-based Information Retrieval within Digital Libraries Employing LCSHs. D-Lib magazine, 15(9/10). http://www.dlib.org/dlib/september09/papadakis/09papadakis.html
2.

Unit 12: Retrieving information on the World Wide Web

Objectives: After this class, you should be able to

- identify the similarity and difference between search on the Web and other medias
- explain the basics of Web retrieval techniques
- search effectively and efficiently using some Web search tools

Required Readings:

1. Blachman, Nancy (n.d.) Google Guide. Review 2 sections listed under "Printable Versions" - I: Query Input and II: Understanding Results. < http://www.googleguide.com/toc.html > (Date Accessed: 10/20/2004).Note: This will help with the assignment!
2. Sullivan, Danny. (2003). Search Engine Watch. Review 2 short pieces: "How Do Search Engines Work?" at http://searchenginewatch.com/webmasters/article.php/2168031 and "How Search Engines Rank Web Pages" at http://searchenginewatch.com/webmasters/article.php/2167961>. (Date accessed: 8/19/2005)
3. Drabenstott, K.M. (2001). Web Search Strategy Development. Online, (25) 4: 18-24. (Available online through Pitt E-Journal http://ug4fn7ck2h.search.serialssolutions.com/)

Interesting Readings:

4. Bellardo-Hahn, T. (1996). Pioneers of the online age. Information Processing and Management, 32(1), 33-48.
5. Eysenbach, G. & Kohler, C. (2002, March 9). How do consumers search for and appraise health information on the World Wide Web? British Medical Journal, 324, 7337, 573+
6. McJunkin, M. C. (1995). Precision and recall in title keyword searches. Information Technology and Libraries, 14(3), 161-171. (Available online through InfoTrac- at Pitt E-Journal http://ug4fn7ck2h.search.serialssolutions.com/)

Unit 13: Intelligent Information Retrieval

Objectives: After this class, you should be able to

- identify the major ideas for integrating intelligence into information retrieval
- explain the basics of intelligent information retrieval techniques

Required Readings:

1. Susan Gauch, Mirco Speretta, Aravind Chandramouli and Alessandro Micarelli. User Profiles for Personalized Information Access. Chapter 2 in Brusilovsky, P., Kobsa, A., Neidl, W. (eds.) (2007) The Adaptive Web: Methods and Strategies of Web Personalization. Lecture Notes in Computer Science, Vol. 4321. Springer-Verlag, Berlin Heidelberg New York.
2. Michael J. Pazzani and Daniel Billsus. Content-Based Recommendation Systems. Chapter 10 of Brusilovsky, P., Kobsa, A., Neidl, W. (eds.) (2007) The Adaptive Web: Methods and Strategies of Web Personalization. Lecture Notes in Computer Science, Vol. 4321. Springer-Verlag, Berlin Heidelberg New York.
3. Ahn, J., Brusilovsky, P., He, D., Grady, J., and Li, Q. (2008). Personalized web exploration with task models. In Proceeding of the 17th international Conference on World Wide

Web (Beijing, China, April 21 - 25, 2008). WWW '08. ACM, New York, NY, 1-10. DOI= http://doi.acm.org/10.1145/1367497.1367499.

4. He, D., Brusiloviksy, P., Grady, J., Li, Q., and Ahn, J. (2007). How Up-to-date should it be? the Value of Instant Profiling and Adaptation in Information Filtering. In Proceedings of the IEEE/WIC/ACM international Conference on Web intelligence(November 02 - 05, 2007). Web Intelligence. IEEE Computer Society, Washington, DC, 699-705. DOI= http://dx.doi.org/10.1109/WI.2007.135

Unit 14: New Trends in Information Retrieval Services

Objectives: After this class, you should be able to

- examine challenges and issues related to manage information retrieval services, including allocating resources, establish policies, teaching retrieval to others, etc.
- explain major issues and challenges as well as future trends in information retrieval and information technology and their implications for professional library and information services.

Required Readings

1. O'Leary, Mick. (1993). New roles for information searchers. Online, 17(1), 10 May. (Available online through Pitt E-Journal http://ug4fn7ck2h.search.serialssolutions.com/)
2. Paepcke, Andreas. (1996). Digital libraries: Searching is not enough. D-Lib Magazine, May. http://www.dlib.org/dlib/may96/stanford/05paepcke.html
3. Block, Marylaine (2002). "My Rules of Information." Searcher (10) 1: 61-67. (Available online at http://www.infotoday.com/searcher/jan02/block.htm)

VII. Term Projects

Introduction:

The term project is designed for students to integrate and extend knowledge acquired throughout the course and to apply that knowledge to solve a problem of substantial scope. Students are required to work individually.

Your task is to propose, plan and carry out a study of a practical retrieval problem, the focus of the study can be either or all of the following:

1. concentrate on existing problems or issues related to information retrieval in library services, traditional or digital,
2. concentrate on an innovative application of an information retrieval technology to a well known problems in library services,
3. concentrate on applying an open source IR system on resolving a practical problems.

Requirements to the final report

The outcome of the term project:

1. in the case of 1 and 2 is a 8-12 page essay, which includes:

a. a statement of the problem
b. the discussion of the related retrieval technology
c. the statement of how the technology can be used to resolve the problems
d. the implementation and evaluation details if there is any
e. any other issues that are worth mentioning in the report

2. in the case of 3, besides the above essay, it would be good to have a demo system.

<u>Milestones for the project</u>:

Introduction of term project:	January 18
Final project report due:	April 19

VIII. Course Assessment

Participation 10%

Class attendance is required for success in this course, as material will be covered in class that is not included in the readings. Participation is based on active participation in each week's "my reading notes" before the class and "my muddiest points" after the class (5 participation points total).

Assignment 30%

There are total three assignments, each of which will count 10% in the final course score. You are required to make a clear presentation about your ideas.

Exam 30%

Mid-term will last 90 minutes, and covers all the topics taught in the weeks before it. Common exam questions include multiple choices, short definitions, and discussion questions.

Term Project 30%

Please see section VII for detail description of term project.

Course Grading Scale:

The final grade depends on the percentage of points you have earned, and the definition of letter grades is:

- 90 <= A- < 94, 94 < A <= 98, 98 < A+ <= 100
- 80 <= B- < 84, 84 < B <= 88, 88 < B+ < 90
- 70 <= C- < 74, 74 < C <= 78, 78 < C+ < 80
- 60 <= D < 70,
- F < 60

IX. Course Polices:

Ground rules for class discussion:

On-class interaction and discussion will be an important means of learning in this course, therefore, it is important that we work together to create a constructive environment by observing these rules:

- You should participate in the discussion of ideas.
- You should respect diverse points of view.
- You should aware the diverse backgrounds of peers.
- You may not belittle or personally criticize another individual for holding a point of view different than your own
- Your use of language should be respectful of other individuals or groups

Academic Integrity:

It is expected that the work you submit in this course will be your own. While collaboration is allowed for the course project, it should be approved in advance and the nature of each contribution should be specified in the project proposal and the final submission.

The following statement is taken from *The Teaching Assistant Experience: A Handbook for Teaching Assistants and Teaching Fellows at the University of Pittsburgh* (A.P. Haley and J.M. Nicoll, eds.)]

> Plagiarism means submitting work as your own that is someone else's. For example, copying material from a book or other source without acknowledging that the works or ideas are someone else's and not your own is plagiarism. If you copy an author's words exactly, treat the passage as a direct quotation and supply the appropriate citation. If you use someone else's ideas, even if you paraphrase the wording, appropriate credit should be given. You have committed plagiarism if you purchase a term paper or submit a paper as your own that you did not write[1].
>
> Plagiarism is a violation of the University of Pittsburgh's standards on academic honesty, and violations of this policy are taken seriously. **From the *Guidelines on Academic Integrity: Student and Faculty Obligations and Hearing Procedures* (effective September, 1995):**

A student has an obligation to exhibit honesty, and to respect the ethical standards of the historical profession in carrying out his or her academic assignments. Without limiting the application of this principle, a student may be found to have violated this obligation if he or she:

- Presents as one's own, for academic evaluation, the ideas, representations, or words of another person or persons without customary and proper acknowledgment of sources.
- Submits the work of another person in a manner which represents the work to be one's own. [Quotation ellipsed.] [2]

[1] B. G. Davis, *Tools for Teaching* (San Francisco: Jossey-Bass, 1993), 300.

[2] University of Pittsburgh, *Guidelines on Academic Integrity: Student and Faculty Obligations and Hearing Procedures* (Pittsburgh: University of Pittsburgh, 1995), 7-8.

Special Needs:

Students with disabilities who require special accommodations or other classroom modifications should notify the instructor and the University's Office of Disability Resources & Services (DRS) no later than the 2nd week of the term. Students may be asked to provide documentation of their disability to determine the appropriateness of the request. DRS is located in 216 William Pitt Union and can be contacted at 648-7890 (Voice), 624-3346(Fax), and 383-7355(TTY). Students who must miss an exam or class due to religious observances must notify the instructor ahead of time and make alternative arrangements.

Students in this course will be expected to comply with the University of Pittsburgh's Policy on Academic Integrity. Any student suspected of violating this obligation for any reason during the semester will be required to participate in the procedural process, initiated at the instructor level, as outlined in the University Guidelines on Academic Integrity. This may include, but is not limited to, the confiscation of the examination of any individual suspected of violating University Policy. Furthermore, no student may bring any unauthorized materials to an exam, including dictionaries and programmable calculators.