

INFO-340: Database Management and Information Retrieval

Winter 2007
B.S. Informatics
Information School
University of Washington

Theories and models in system-centered approaches to information retrieval and database management. Information retrieval and database management systems include text and multimedia databases, web search engines and digital libraries. Issues in system design, development and evaluation, and tools for searching, retrieval, user interfaces, and usability. *Prerequisite:* CSE 373

Course website & Listserv

<http://courses.washington.edu/info340/>

info340a_wi07@u.washington.edu

(Archive: https://mailman1.u.washington.edu/mailman/listinfo/info340a_wi07)

Registered students are subscribed automatically using their UW mail account.

Credit Hours

5 (3 lecture hours; 2 lab hours; 10 outside hours)

Meeting times

<i>Lecture</i>	Tuesday/Thursday 130 – 250, MEB 242
<i>Lab</i>	Thursday 330 – 520, MGH 430

Instructor

David Hendry, Assistant Professor
330J Mary Gates Hall
dhendry@u.washington.edu | <http://faculty.washington.edu/dhendry>

Office hours: Tuesday, 330 – 430 or by appointment.

Teaching assistant

Beth Fournier, MLIS Student
bethf2@u.washington.edu

Office hours: Friday, 230 – 330 or by appointment (in TE Lab, MGH 440)

Student services

Dowell Eugenio, Student Services Administrator
470E Mary Gates Hall
deugen3@u.washington.edu

Please note: If you have any concerns about a course or the TA, please see the TA about these issues as soon as possible. If you are not comfortable talking with the TA or not satisfied with the response that you receive, you may contact the instructor of the course. If you are still not satisfied with the response that you receive, you may contact Joseph Janes, the Associate Dean for Academics in 370 Mary Gates Hall, by phone at : (206) 616-0987, or by e-mail at jwj@u.washington.edu. You may also contact the Graduate School at G-1 Communications Building, by phone at (206) 543-5900, or by e-mail at efeetham@u.washington.edu

Overview

Information systems have an enormous impact on our personal and civic lives. We find information systems virtually everywhere we live, work, and play.

Information systems can be examined from many different perspectives. For example, we could study:

- The productivity benefits—or costs—of information systems to a person, a firm or a nation;
- The mathematics and engineering research base that underpins information systems;
- Why information systems often fail and methodologies that mitigate the risk;
- Specific types of information systems in domains such as urban planning, health care, environment science, popular culture, business, etc.

In this class, however, we will leave these important areas of study to the side. This class introduces the theory and practice of information system design. It focuses on two fundamental approaches: Relational Database systems and Information Retrieval (IR) systems. We shall see that these are distinct approaches for solving different problems.

Drawing on your experience in programming, in website development, and knowledge organization this course will enable you to develop working systems and prepare you for advanced courses in database and IR systems.

Textbooks and readings

The textbook for this course is

Connolly, T. M. & Begg, C. E. (2003). *Database Systems: A Practical Approach to Design, Implementation, and Management* (4th Edition) New York: Addison-Wesley Publishing. [ISBN: 0321210255]

This textbook will be used for the first half of the course. It has been selected because of its breadth and depth of coverage of relational databases. It is well written, well designed, and contains many examples. You should find this book to be useful for several years to come.

For the second half of the course, we will draw upon readings from several sources including:

Belew, R. K. (2001). *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW*. New York: Cambridge University Press. [ISBN: 0521630282]

If you are particularly interested in Information Retrieval, consider purchasing the book by Belew. Readings for the second half of the course will be posted on the website.

Learning

Aims

The general aims of this course are to:

1. Develop a conceptual understanding for relational database and information retrieval systems
2. Develop skills in implementing information systems using these two approaches
3. Improve skills in collaboration in technical teams
4. Develop skills for integrating technologies to develop a small but complete application.

Objectives

On the successful completion of this course, you should be able to:

1. Describe the components of an information system and list risks for why information systems fail
2. Describe and practice techniques in conflict management and describe a developmental model for teams
3. Describe the functions and organization of database management systems
4. Describe the relational model, including the data structure and algebra
5. From problem statements, derive SQL statements for querying, updating and creating databases
6. Create Entity-Relationship and Enhanced-Entity Relationship models for small systems
7. Read an ER diagram as a specification and implement a database system for it
8. Describe the problems of data redundancy and update anomalies and be able to normalize a database to 3NF to avoid these problems
9. Describe a three-tier information system
10. Outline a methodology for designing database applications
11. Implement a small database-driven website using ODBC as the middle tier using the following open source tools: NetBeans 5.5, JSP, JDBC, and PostgreSQL
12. Describe the function and organization of an information retrieval (IR) system, including documents, document collections, terms, queries, matching, ranking, and results
13. Explain how an inverted file works and describe its basic space and time complexity
14. Describe the difference between Boolean and ranked retrieval
15. Know the formula for the Zipf distribution and recognize its curve in a data set
16. Explain inverse document frequency and other methods for weighting terms in documents
17. Describe PageRank
18. Given several weighting functions, explain their difference
19. Using existing components from the Lucene framework, implement a search interface for a website
20. Describe concepts for evaluating information systems, including system performance metrics (e.g., coverage, precision, recall, etc.) and usability metrics (task completion time, number of errors, etc.)
21. Describe methods for evaluating information systems (e.g., usability evaluations, log file analysis, etc.)

Academic accommodations

To request academic accommodations due to a disability, please contact Disabled Student Services: 448 Schmitz, 206-543-8924 (V/TTY). If you have a letter from DSS indicating that you have a disability which requires academic accommodations, please present the letter to me so we can discuss the accommodations you might need in the class.

Academic accommodations due to disability will not be made unless the student has a letter from DSS specifying the type and nature of accommodations needed.

For additional information, see *Statements to Ensure Equal Opportunity and Reasonable Accommodation*, downloaded March 5, 2003, <http://www.washington.edu/admin/eoo/eoost.html>

Academic honesty

The essence of academic life revolves around respect not only for the ideas of others, but also their rights to those ideas and their promulgation. It is therefore essential that all of us engaged in the life of the mind take the utmost care that the ideas and expressions of ideas of other people always be appropriately handled, and, where necessary, cited. For writing assignments, when ideas or materials of others are used, they must be cited. The format is not that important—as long as the source material can be located and the citation verified, it's OK. What is important is that the material be cited. In any situation, if you have a question, please feel free to ask. Such attention to ideas and acknowledgment of their sources is central not only to academic life, but life in general.

Please acquaint yourself with the University of Washington's resources on academic honesty: <http://depts.washington.edu/grading/issue1/honesty.htm>

Students are encouraged to take drafts of their writing assignments to the Writing Center for assistance with using citations ethically and effectively. Information on scheduling an appointment can be found at:

<http://www.uwtc.washington.edu/resources/eiwc/>

Copyright

All of the expressions of ideas in this class that are fixed in any tangible medium such as digital and physical documents are protected by copyright law as embodied in title 17 of the United States Code. These expressions include the work product of both: (1) your student colleagues (e.g., any assignments published here in the course environment or statements committed to text in a discussion forum); and, (2) your instructors (e.g., the syllabus, assignments, reading lists, and lectures). Within the constraints of "fair use", you may copy these copyrighted expressions for your personal intellectual use in support of your education here in the iSchool. Such fair use by you does not include further distribution by any means of copying, performance or presentation beyond the circle of your close acquaintances, student colleagues in this class and your family. If you have any questions regarding whether a use to which you wish to put one of these expressions violates the creator's copyright interests, please feel free to ask the instructor for guidance.

Privacy

To support an academic environment of rigorous discussion and open expression of personal thoughts and feelings, we, as members of the academic community, must be committed to the inviolate right of privacy of our student and instructor colleagues. As a result, we must forego sharing personally identifiable information about any member of our community including information about the ideas they express, their families, life styles and their political and social affiliations. If you have any questions regarding whether a disclosure you wish to make regarding anyone in this course or in the iSchool community violates that person's privacy interests, please feel free to ask the instructor for guidance.

Knowing violations of these principles of academic conduct, privacy or copyright may result in University disciplinary action under the Student Code of Conduct.

Student Code of Conduct

Good student conduct is important for maintaining a healthy course environment. Please familiarize yourself with the University of Washington's Student Code of Conduct at:

<http://www.washington.edu/students/handbook/conduct.html>

Assessment

The assignments strike a balance between theory and practice and between individual and group work.

Assessment	% Grade
Four individual assignments	25%
Group project	20%
Midterm exam	15%
Final exam	30%
Spirit and participation in classes/labs	10%

Individual assignments

You will complete four individual assignments:

Assignment	Due	Week
A1. Search Interface with the Google Base API	Jan 23	#4
A2. SQL Data Definition and Manipulation	Jan 30	#5
A3. Database Design	Feb 06	#6
A4. Information Retrieval: Matching & Ranking	Mar 01	#9

A1 is worth 10%. A2, A3, and A4 are each worth 5%.

Please note: All assignments are due at the beginning of class.

Social Bookmarking Project

Working in groups of 2-3, you will develop an information system. The deliverables:

Deliverable	Due	Week
P1. System Design, Rough Draft	Jan 25	#4
P2. Design & Code: Functions 1–2	Feb 22	#8
P3. Final System	Mar 08	#10
P4. Demonstration (to be arranged with Beth)	Mar 08-09	#10

Notes

1. P1 and P2 are worth 5% each and final system is worth 15%.
2. Deliverables are due at the beginning of class.
3. Groups will be selected by the instructor and teaching assistant on the basis of a skills profile.

Midterm and final exam

The mid term and final exam will assess your knowledge the conceptual foundations of relational database systems and information retrieval systems.

Midterm exam	15%	Thr, Feb 8, in class
Final exam	30%	Friday, March 16, 2007, 230 – 430, MEB 242*

*** Please note: We will seek to change this date to Tue March 13, 1:30 – 3:30 but this will require the unanimous consent of all students.**

Spirit and participation in classes/labs

It is important to the instructor and teaching assistant that you help make INFO-340 fun, interesting, and challenging. With spirit and a professional manner, we can create a supportive and rewarding learning environment.

Among the things you can do are:

1. Treat all with respect – be constructive in all discussions
2. Come to class prepared – read carefully and be ready for discussion
3. Be an active listener – be attentive, be engaged, use in-class technology with discretion
4. Ask challenging questions, participate in discussion
5. Comment, build on, or clarify what others have done or said
6. Help your classmates use development tools and technologies
7. Post useful or interesting information to the class discussion list
8. Visit the instructor during office hours to chat, to ask questions, or to give feedback.

Deliverable	Due	Week
Personal Statement	Mar 8	#10

Please write a 2 or 3 paragraph personal statement on how you contributed to the class. Attendance at labs will be taken and your lab website will be checked periodically. Participation is graded because responsiveness and involvement are crucial elements of your development. Your participation is worth 10% of your final grade.

Grading criteria

Work in this course will be graded to criteria. In other words, you won't be graded on a curve. Each deliverable is designed to test your achievement against one or more of the learning objectives. Different assignments emphasize different learning objectives. The meanings of grades are described below.

General grading information for the University of Washington is available at:

- http://www.washington.edu/students/genocat/front/Grading_Sys.html

The iSchool has adopted its own criteria for grading graduate courses. The grading criteria used by the iSchool is available at:

- <http://depts.washington.edu/grading/practices/guidelin.htm>

Grade	Performance Quality*
3.9 - 4.0	Superior performance in all aspects of the course with work exemplifying the highest quality. Unquestionably prepared for subsequent courses in field.
3.5 - 3.8	Superior performance in most aspects of the course; high quality work in the remainder. Unquestionably prepared for subsequent courses in field.
3.2 - 3.4	High quality performance in all or most aspects of the course. Very good chance of success in subsequent courses in field.
2.9 - 3.1	High quality performance in some of the course; satisfactory performance in the remainder. Good chance of success in subsequent courses in field.
2.5 - 2.8	Satisfactory performance in the course. Evidence of sufficient learning to succeed in subsequent courses in field.
2.2 - 2.4	Satisfactory performance in most of the course, with the remainder being somewhat substandard. Evidence of sufficient learning to succeed in subsequent courses in field with effort.
1.9 - 2.1	Evidence of some learning but generally marginal performance. Marginal chance of success in subsequent courses in field.

*Taken from Faculty Resource on Grading, downloaded March 5, 2003, <http://depts.washington.edu/grading/practices/guidelin.htm>

Standard cover sheet

To protect your privacy when exercises are returned and to facilitate communication, submitted work must have a cover sheet. The cover sheet must include the following information and be formatted nicely:

- Course name
- Quarter, program, department, and university
- Assignment name
- Your name and e-mail address
- A date
- A web site address (if relevant).

Staple the exercise pages to the cover sheet.

Late policy

1. If you will miss the deadline, you should inform the instructor as soon as you can, indicating when you will submit the work. The instructor will try to accommodate your needs. You should use this clause only for extraordinary personal reasons.
2. It is at the instructor's discretion to accept late work or assign late penalties (see 1 above). For any late assignment, 10% will be taken off your work per day. After five days, your work will not be accepted.
3. Late work must be handed to the instructor or teaching assistant in person. You may also be able to hand work in at the front desk of the Information School and at student services but this cannot be guaranteed.

Work that is handed in late is penalized for two reasons. First, to be fair, all students should be given the same time limits. Second, if you spend too much time on one assignment, it is quite likely that you will have insufficient time to spend on subsequent assignments.

Right to revise

The instructor reserves the right to revise this syllabus.

Re-grading policy

To have work re-graded, you must submit a Re-grade Request within five days of when your work was returned. The request must be a single page, printed on paper or sent by e-mail. It should contain the following information:

- Re-grade Request
- The information contained on the standard cover sheet
- An explanation for why you believe you deserve a higher grade.

The instructor, possibly in collaboration with the teaching assistant, will consider your request. If the instructor is convinced by your argument, your work will be re-graded. If not, the instructor will send you e-mail explaining why. No re-grades will be considered for late work.

Guidelines on using e-mail

When communicating with the instructor or teaching assistant, please follow these guidelines:

- You are welcome to give feedback to the instructor and teaching assistant about the course, to ask a question about an assignment, to share an interesting article or resource, to report that you will be absent from a class/lab, to request additional time for an assignment (because of significant health, personal, or educational matter), or similar communication ;
- Whenever appropriate, please copy the class listserv with your question or comment;
- E-mail concerning assignments might not be replied to if it is sent within 36hr of the assignment due date;
- If your e-mail concerns your grade, please follow the re-grading policy (see above);
- E-mail that is sent on Friday afternoon or over the weekend is not replied to until Monday or Tuesday of the following week;
- If you don't receive a reply within 2 days or so, please resend your e-mail or ask about it during class or lab.

Class Schedule

Week 1: Overview

- Read C & B, Chap. 1-2
- L1 Greetings;
- L2 Introduction to Relational Database Systems
- Lab Development Tools, I

Week 2: Relational model

- Read C & B, Chap. 3-4 (4.2 optional)
- L1 Relational Data Structure & Relational Integrity
- L2 Relational Algebra
- Lab Development Tools, II

Week 3: SQL

- Read C & B, Chap. 5-6, Appendix C
- L1 Introduction to Storage and Indexing
- L2 SQL Query Language
- Lab SQL

Week 4: ER-Modeling

- Read C & B, Chap. 9, 11-12,
- L1 Entity Relationship Modeling, I
- L2 Entity Relationship Modeling, II
- Lab Introduction to JDBC, I

Week 5: Normalization

- Read C & B, Chap. 13
- L1 Database Normalization
- L2 Review of Database Systems
- Lab Introduction to JDBC, II

Week 6: Introduction to IR Systems

- L1 Introduction to Information Retrieval (IR) systems
- L2 Midterm**
- Lab Teamwork and Project Work

Week 7: Documents and indexing

- Read • Belew, Chap. 1-2
• Barroso, L. A., Dean, J., & Hölzle, U. (2003). Web Search for a Planet: The Google Cluster Architecture *IEEE Micro* 23 (2), 22-28.
- L1 Documents, Metadata & Document Surrogates
- L2 Indexing
- Lab Introduction to Lucene, I

Week 8: Queries and matching

- Read Belew, Chap. 3
- L1 Inverted File Structure
- L2 Weighting and Matching
- Lab Introduction to Lucene, II

Week 9: Evaluation

- Read Baeza-Yates & Ribeiro-Neto, Chap. 3
- L1 Precision/Recall experimentation
- L2 Usability Evaluation
- Lab Project Work

Week 10: Review

- L1 DB Review
- L2 Review & Future Directions
- Lab DEMO lab (optional)