IS392: Web Mining and Information Retrieval

Last updated Dec 29, 2014 (subject to change)

Faculty Instructor: Y.F. Brook Wu, Ph.D.
Office: GITC 5503
Office Hours: Wednesday 3-5:30pm
E-mail: wu AT njit DOT edu

T.A.: Mr. Mingzhu Zhu
Office: GITC 4215 (across from the back elevator)
Office Hours: 3-5:30pm and by appointment.
E-mail: mz59 AT njit DOT edu

Classroom: GITC 1400
Class Meets: Wednesday 6-9pm
Class Site: please go to moodle.njit.edu and login with your UCID. You will find IS 392, if you are enrolled in this class.

Overview

This course introduces the design, implementation and evaluation of search engines and web mining applications. Topics include: automatic indexing, natural language processing, retrieval algorithms, web page classification and clustering, information extraction, summarization, search engine optimization, and web analytics. Students will gain hands-on experience applying theories in case studies.

Prerequisites

- Completion of 100 level course in computing sciences: CS 101 or CS 111 or CS 113 or CS 115 or IS 118.

Learning Goals

1. Acquire a basic understanding of linguistic analysis.
2. Learn various automatic indexing techniques.
3. Obtain knowledge in retrieval models.
4. Learn web crawling.
5. Understand web usage, content and structure mining with emphasis on the first two types.
6. Become familiar with web analytics.
7. Become familiar with applying web mining and analytics to search engine optimization.

Textbook

Search Engines: Information Retrieval in Practice, by Croft, Metzler, and Strohman.
Publisher: Addision-Wesley
ISBN-13: 978—0-13-607224-9

Additional Materials

- Paper 1: What Do People from Information Retrieval?, W. Bruce Croft
- Paper 2: *Search Engine Optimization Starter Guide*, Google, http://static.googleusercontent.com/external_content/untrusted_dlcp/www.google.com/en/us/webmasters/docs/search-engine-optimization-starter-guide.pdf
- Paper 3: *Web Analytics Definitions V4.0*, Web Analytics Association, 2007, http://www.digitalanalyticsassociation.org/Files/PDF_standards/WebAnalyticsDefinitionsVol1.pdf

NJIT University Code on Academic Integrity
http://www.njit.edu/academics/pdf/academic-integrity-code.pdf is strictly enforced.

Grading

- Attendance and class activities 13%
  - Attendance 5%
  - In-class design activity 5%
  - Alternative search engine presentation 3%
- Assignments 42%
  - Assignment 1 Comparing Search Engines 6%
  - Assignment 2 Programming Assignment 1 12%
  - Assignment 3 Programming Assignment 2 12%
  - Assignment 4 Web Mining 12%
- Midterm 20%
- Final 25%

Total: 100%

Weekly Coverage of Material

The following table shows approximately how much time may be devoted to each topic and the corresponding readings from the textbook and papers.

Week	Topics	Materials
1	Course Logistics and Introduction	Paper 1
2	Overview of Search Engine and IR	Ch 1
3	Architecture of search engines	Ch 2
4	Architecture of search engines (cont) Crawls and feeds	Ch 2, 3
5	Crawls and feeds (cont)	Ch 3
6	Processing text	Ch 4

7	**Midterm (March 11, subject to change)**	
8	Processing text Ranking with Indexes	Ch 4, 5
	Spring Break (March 17-23, No Class)	
9	Search Engine Optimization	Paper 2
10	Web Analytics	Paper 2, 3
11	Web Mining	PPT on moodle
12	Web Mining	PPT on moodle
13	Queries and interfaces	Ch 6
14	Evaluating search engines	Ch 8
15	Wrap up	
	Final Exam: Date/Time TBA by Registrar	