

CS495 (Future CS429) – Introduction to Information Retrieval Systems

Last Updated - 02/22/02

Course Manager – Dr. Nazli Goharian, Clinical Assistant Professor

3 credit hours; elective for CS & CPE; 150 min. lecture each week

Current Catalog Description - Overview of fundamental issues of information retrieval with theoretical foundations. The Information-retrieval techniques and theory, covering both effectiveness and run-time performance of information-retrieval systems are covered. The focus is on algorithms and heuristics used to find documents relevant to the user request and to find them fast. The course covers the architecture and components of the search engine such as parser, stemmer, index builder, and query processor. The students learn the material by building a prototype of such a search engine. Prerequisite: CS331 or CS401 and Strong Programming knowledge.

Textbook

- D. Grossman and O. Frieder, *Information Retrieval: Algorithms and Heuristics*, Kluwer Academic Press, 1998.
- D. Grossman and O. Frieder, *Anatomy of a Search Engine: A Java-Based Introduction to Scalable Information Retrieval*, Prentice Hall (in process)

References - other textbooks or materials

- none

Course Goals - Students should be able to:

- Explain the information retrieval storage methods (Inverted Index and Signature Files)
- Explain retrieval models, such as Boolean model, Vector Space model, Probabilistic model, Inference Networks, and Neural Networks.
- Explain retrieval utilities such as Stemming, Relevance Feedback, N-gram, Clustering, and Thesauri, and Parsing and Token recognition.
- Design and implement a search engine prototype using the storage methods, retrieval models and utilities.
- Apply the research ideas into their experiments in building a search engine prototype.

Prerequisites by Topic

- Data Structures, Algorithm and Strong Object Oriented Programming.

Major Topics Covered in Course

1. Introduction, basic terminology (relevance ranking, recall, precision, average precision).	3 hours
2. Architecture of a search engine (parser, index builder, query processor).	3 hours
3. Retrieval utilities: parsing (token recognition), stemming, N-grams.	6 hours
4. Storage methods: building inverted index, signature files.	3 hours
5. Retrieval models and relevance ranking: Boolean, vector space model, probabilistic model, inference networks, and neural networks and different similarity measures.	7 hours
6. Retrieval utilities: relevance feedback, clustering, and thesauri.	3 hours
7. Efficiency issues: inverted index compression, posting list thresholding.	3 hours
8. Discussions on the search engine prototype implementation issues.	3 hours
9. Text retrieval evaluation standards and benchmarks.	2 hours
10. Advanced topics	3 hours
11. Paper presentation	6 hours
Midterm Exam	3 hours
Final Exam	-

Laboratory projects (specify number of weeks on each)

- none

Estimate CSAB Category Content in Credit Hours

	CORE	ADVANCED		CORE	ADVANCED
Data Structures		1	Computer Organization and Architecture		
Algorithms		1	Concepts of Programming Languages		
Software Design		1			

Oral and Written Communications - Every student is required to submit at least 5 written reports (not including exams, tests, quizzes, or commented programs) of typically 3 pages and to make 1 oral presentations of typically 20 minutes duration. Include only material that is graded for grammar, spelling, style, and so forth, as well as for technical content, completeness, and accuracy.

- Design and Summarization of the prototype implementation.
- Research paper presentation

Social and Ethical Issues - Please list the topics that address the social and ethical implications of computing covered in all course sections. Estimate the class time spent on each topic. In what ways are the students in this course graded on their understanding of these topics (e.g., test questions, essays, oral presentations, and so forth)?

- Ethical issues in information search, information hiding and information security. 1 hour.

Theoretical Foundations - Please list the types of theoretical material covered, and estimate the time devoted to such coverage in contact (lecture and lab) hours.

- Retrieval models and algorithms, 7 hrs.
- Retrieval utilities, 9 hrs.
- Efficiency methods, 4 hrs.
- Storage Methods, 3 hrs.
- Advanced and research related material, 6

Problem Analysis - Please describe the problem analysis experiences common to all course sections.

- Analysis of the experimental results obtained by the implementation of different algorithms and techniques in the search engine prototype projects.

Solution Design - Please describe the design experiences common to all course sections.

- Design modules for implementation of various information retrieval strategies, utilities, and efficiency techniques for a search engine.

Other Course Information

- none