

THE COLLEGE OF NEW JERSEY

Computer Science / Interactive Multi-Media

CSC320/IMM320: Information Retrieval

Fall 2007

Course Goals

The course is aimed at advanced undergraduate students in Computer Science, Interactive Multimedia, Information Science, Business. The course is intended to prepare students to design, use and evaluate information retrieval systems. The course also aims to give students a broad understanding of inner workings of automated information retrieval systems, and how such systems interact with users and affect their productivity.

Course Description

This course will discuss theory and practice of searching and retrieval of text and bibliographic information. Topics covered include automated indexing, statistical and linguistic models, text classification, Boolean and probabilistic approaches to indexing, query formulation and output ranking, information routing and filtering, topic detection and tracking, as well as measures of retrieval effectiveness, including relevance, utility, miss/false-alarm. Techniques for enhancing retrieval effectiveness including relevance feedback, query reformulation, thesauri, concept extraction, and automated summarization. Experimental retrieval approaches from Text Retrieval Conferences (TREC); modern Internet search engines (Google, Yahoo, etc).

Course Prerequisites

The prerequisite for taking this class for CS students is the CSC 230 (CS II) "Computer Science II : Data Structures" course. For IMM students, the prerequisite is the IMM core. For students outside of the CS and IMM majors, a skill equivalent of CS or IMM prerequisites is expected and will be assessed by the instructor before the student is allowed to join the class. Student background is expected to include familiarity with data structures and algorithms, elementary algebra, basic statistics and probability, elements of logic and set theory, having used a catalogue in a library, and having used an Internet search system.

About the Instructor

Dr. Miroslav Martinovic

Brief Biography

Ph.D. 1993 Belgrade University / New York University.
CS Faculty 2000-present, TCNJ
CS/Math Faculty, 1989-2000, Wagner College
CS/Math Faculty, 1983-1988, Belgrade University
Principal Scientist, 1989-present.

Research Interests

Question-Answering Systems
Natural Language Processing
Information Retrieval
Theory of Gaming
Computer Science Education
Logic Programming
Expert Systems

Sponsors

NSF
DARPA
NIST
Microsoft
World University Service
Studenica Foundation

E-mail Address :

mmmartin@tcnj.edu (click to e-mail)

Telephone :

(609) 771-2789.

Office :

Holman Hall 207 / 230.

Class Time:

at Holman Hall

Special Topics : Information Retrieval

Lecture Notes (e-mail instructor for a password)	Monday, Thursday 12:30-1:50	at Holman Hall 253
Instructor supervised assigned work from the Paper / Topic List below	360 minutes at students own schedule.	at Holman Hall 372

Textbooks:

Course Main Text

Modern Information Retrieval R. Baeza-Yates, B. Ribeiro-Neto
 Published by Addison Wesley, ISBN 0-201-39829-X
 2000.

Additional Texts

1. *Readings in Information Retrieval* Karen Sparck-Jones and Peter Willett (editors)
 Morgan-Kaufmann Publishers, 1997.
2. *Natural Language Information Retrieval* Tomek Strzalkowski (editor)
 Kluwer Academic Publishers, 1999.
3. *Information retrieval: data structures & algorithms* William B. Frakes and Ricardo Baeza-Yates
 Englewood Cliffs, N.J.: Prentice Hall, 1992.
4. *Mathematical Foundations of Information Retrieval* by: S. Dominich
 Published by Kluwer Publishing, ISBN 0-7923-6861-4
 1999.

Office Hours :

Monday	Tuesday	Thursday	Friday
9:15-9:45	9:15-12:45	9:15-9:45	3:45-5:45 (by appointment)
11:30-1:30		11:30-1:30	

Grading Policy:

Attendance, Class Participation and Effort	20%
Topic Presentation and Critique	40%
Topic/Paper List (e-mail instructor for a password)	
Paper critique and presentation guidelines (e-mail instructor for a password)	

Final Paper/Project with Presentation and Demo

[Project with guidelines and resources](#) (e-mail instructor for a password)

40%

[Student Accounts Info](#)

(i) The course topics will be examined through readings, discussion, hands-on experience using various information retrieval systems, and through participation in evaluation of different retrieval algorithms on various test collections.

(ii) There will be periodic assignments and a final paper/project.

(iii) In-class presentations (readings and/or experiments) will require a preparation that includes finding materials outside of base reading and during the assigned work periods.

Final paper will be a technical paper on an IR issue. Topics for a programming project include :

- topic detection through concept extraction / topic tracking
- pivoted normalization weighting in SMART
- query expansion tool
- self-learning concept spotter
- LMI using TTP
- automatic summarizing
- sub-categorization of retrieved set
- question-answering.

CSC320 / IMM320 Tentative Schedule

Week 1

Introduction to Information Retrieval

Reading: MIR Chapters 1 and 2

- A. What is Information Retrieval?
- B. The notion of Relevance.
 - 1. Conceptual
 - 2. Computational
 - 3. Document similarity measures.
- C. The IR tasks:
 - 1. Ad-hoc Querying.
 - 2. Filtering and Routing
 - 3. Topic Detection and Tracking
 - 4. Question Answering.
 - 5. Automated Summarization.
 - 6. Information Fusion.
- D. Conceptual Models of IR systems.
 - 1. Boolean
 - 2. Vector-Space
 - 3. Probabilistic
 - 4. Extended models
 - 5. Natural Language Processing

E. Characteristics of text collections.

1. Structured text and Bibliographic records.
 2. Multimedia documents.
 3. Full Text.
-

Week 2

Readings : KSJ&PW Chapter 5.3, 5.4, 5.5

Conceptual Models Discussion.

Student presentations; IR systems

A. Boolean and Extended Boolean Models; Glimpse

(download, install)

B. The Vector Space Model; SMART, WAIS, Prise systems

(download, install),

C. Probabilistic Models; InQuery, Cheshire, Inktomi

D. Natural Language Processing models of the IR task. (DR.LINK; Evans; Topic) Articles from instructor

Week 3

Readings : MIR chapter 3

Lecture and discussion

Class exercise: evaluating web search engines by pooling method

Evaluation

A. Assumptions in IR performance evaluation.

1. Fully automated vs. interactive systems.
2. Who determines relevance?

B. Evaluation metrics

1. Recall and Precision
2. Miss and False Alarm
3. ROC curves
4. Other measures

C. Reference Collections

1. Classic collections: Cranfield, CACM, ISI, INSPEC, ...
2. Tipster/TREC
3. TDT collections

D. Evaluation methodology

1. What is a good IR experiment?
2. Experimental design.
3. Selection of test collections.
4. Running experiments and collecting results.
5. Standard analyses and analysis tools.
6. Graphical and tabular display of results.

E. Standard Evaluations and results

1. TREC
 2. MUC
 3. SUMMAC
 4. TDT
-

Week 4 & 5

Readings : MIR chapters 7, 8, 9, 10

Automated Indexing

Readings : MIR chapters 7, 8 + other sources

2 lectures (A-D) and student presentations (E-G)

- A. Properties of language collections.
 - 1. Statistical distributions; Zipf's law,
 - 2. Stochastic language models.
 - 3. Metadata and Markup Languages (SGML, HTML, XML)
 - 4. Multimedia
 - 5. Full-text Documents vs. Bibliographic records.
- B. Data and File Structures for Information Retrieval.
 - 1. Inverted files.
 - 2. Signature file.
 - 3. Other file structures (PAT trees, Grid Files, Hashing).
 - 4. DBMS-based Information Retrieval.
- C. Indexing goals.
 - 1. Passage vs. document retrieval.
 - 2. Segmentation approaches
 - 3. Salton's "Blueprint for automatic indexing"
- D. Indexing and storage issues.
 - 1. Index compression.
 - 2. Automating Hypertext linkages.
- 3. Positional information in indexes.
- E. Text Analysis (student presentations & discussion)
 - 1. Stoplists and Stemming algorithms.
 - 2. Phrases and collocations
 - 3. Disambiguation
- F. Linguistically Motivated Indexing
(student presentation: KSJ article; Readings)
 - 1. Stemming and Morphological analysis.
 - 2. Part-of-speech tagging and Parsing
 - 3. Phrase recognition and variants (term paper)
 - 4. Concept extraction
 - 5. Case studies: MUC, SCISOR, FERRET, NLIR
- G. Thesaurus Construction
(student presentations)
 - 1. Collection-sensitive thesauri.
 - 2. Manually derived thesauri
(WordNet, Snomed, MESH, LCSH).
 - 3. Automatically derived thesauri.
 - a. Term associations
(Spark Jones, Grefenstette, Strzalkowski).
 - b. Latent Semantic Indexing.

Week 6

Query Languages and Query operations

Readings : MIR chapters 4 and 5 + other readings

Lecture

- A. Keyword queries
- B. Bag-of-words queries
- C. Boolean queries
- D. Enhanced BOW queries
- E. Relevance feedback
- F. Query expansion & user interaction (term paper)
- G. Cross-language retrieval (term paper)

Week 7

NLP Tools : Parsers (A Parser for English)

Paper presentation and critique with a demonstration session

Papers : [Papers/APParser/manual.ps](#),
[Papers/APParser/APParser.htm](#)

Resource directory (springfield) :
[/projects/mmmartin/Information Retrieval/NYU Parser/](#)

Week 8

NLP Tools : Electronic Lexicons (WordNet)

Paper presentation and critique with a demonstration session

Documentation :
<http://www.cogsci.princeton.edu/~wn/doc.shtml>
Resource directory :
[~mmmartin/www/CMSC485/Papers/WordNet/](#)

Week 9

Automatic Classification

Readings: TBD

Lecture and presentations
Project: topic detection by concept extraction

- A. Manual classification.
 - 1. Classification schemes.
 - 2. The manual classification task.
- B. Automatic Classification - Routing and Filtering
 - 1. Standing queries and profiles
 - 2. Routing task
 - 3. Filtering task
- C. Automatic Classification -- Clustering.
 - 1. The cluster hypothesis.
 - 2. Hierarchical Classification
 - a. Single Link.
 - b. Complete link.
 - 3. Heuristic classification
 - a. Rocchio's method.
 - b. Datola method.
 - c. Scatter-gather.
- D. Using classification for searching and retrieval. (student presentation)
 - 1. Library classification as an organizing principle
 - 2. Yahoo and WWW classification.
 - 3. Cheshire 2-stage retrieval.
- F. Topic Detection and Tracking

Project Presentations and Demos

Week 14

Additional Texts:

1. Gerard Salton. Automatic text processing: the transformation, analysis, and retrieval of information by computer. Reading, Mass. : Addison-Wesley, 1988.
2. C. J. van Rijsbergen. Information retrieval. London : Butterworths, 1975.
3. Text Retrieval Conference (TREC) proceedings (copies from instructor)
4. ACM SIGIR Conference Proceedings (copies from instructor)
5. Technical journals:
 - a. Information Processing & Management, Pergamon Press
 - b. Information Retrieval, Kluwer Academic Publishers
 - c. Computational Linguistics, MIT Press
 - d. Journal of the ASIS

Attachments :

List of Papers/Topics

Topic Paper and Demonstration Materials	Presenter	Presentation Date
1. NLP Tools - SMART IR System : Paper Presentation and a Demonstration Session Paper : Papers/SMART/SmartCourse.html	Lester Wolfgang, Corey Shaffer	10/18
2. Web Information Retrieval : Google's Success Paper : Papers/Google/Google.pdf	Nathan Iyer, Hartigan, Christine, Ialeggio, David	10/22
3. Text Annotation Techniques Paper : Papers/TAT/	Kirsten Gerbehy	11/1

Special Topics : Information Retrieval

4. Image Retrieval : Paper Presentation Paper Resources : Papers/ImageIR/	Rupert, Jeffrey; Ramos, Daniel	11/5
5. Thesauri in Information Retrieval Papers : Papers/IRThesauri/AutoDerofThes.ppt	McConnell, Christian, Krutchkoff, Alexander	11/8
6. YAHOO Radio Service Paper Resource : http://www.yahoo.com/	Brian Glaz, Tom Winnicki	11/12
7. Question / Answer Taxonomies and Categorizations in QA Systems Paper Resource : Papers/QuestionAnswerCategorization/	Vercruyssen, Mathew	11/15
8. MURAX and ASKJEEVES Paper Resource : Papers/MurAskJ/	Lee, Brandon; Ozol, Matthew	11/19
9. Topic Detection and Tracking Paper Resources: Papers/TDT/	Steven Cook	11/22
10. CYC – A Large Common Sense Knowledge Base in Information Retrieval Paper Resources : ~mmartin/ResearchCyc	Craig Hidders	11/29
11. Information Technology at MOMA: a Case Study Resource: http://www.moma.org	Britney Pringle	12/3

[Paper critique and presentation guidelines](#)

Paper Critique Guidelines

Each critique should be no more than one page long. Less than a page is OK. The purpose of a critique is not to summarize the paper; rather you should choose one or two points about the work that you found interesting.

Examples of questions that you might address are:

- What problem does this paper solve, and what are the strengths and limitations of its approach?
- Is the evaluation fair? Does it achieve it support the stated goals of the paper?
- Does the method described seem mature enough to use in real applications? Why or why not? What applications seem particularly amenable to this approach?
- What good ideas does the problem formulation, the solution, the approach or the research method contain that could be applied elsewhere?
- What would be good follow-on projects and why?
- Are the paper's underlying assumptions valid?
- Which important issues in the field does this paper illuminate and how?
- Did the paper provide a clear enough and detailed enough description of the proposed methods for you to be able to implement them? If not, how is additional clarification or detail needed?

to be able to implement them? If not, where is additional clarification or detail needed?

Your critique should be typed (single space) and should list the title of the paper and its authors at the top, along with your name.

Avoid **unsupported** value judgments, like "I liked..." or "I disagreed with..." If you make judgments of this sort, explain why you liked or disagreed with the point you describe.

Be sure to distinguish comments about the writing of the paper from comment about the technical content of the work.

Paper Presentation Guidelines

Length : class period (60-80 minutes)

Medium : PowerPoint, HTML slides, PDF slides or alike.

Paper Critique Presentation Guidelines

Length : class time (talk of up to 40 minutes to be followed by an up to 40 minutes discussion mediated by the presenter)

Medium : PowerPoint, HTML slides, PDF slides or alike.

Note about how the preparedness for other students presentations affects the grade

- (i) All listed papers must be read by every student in class.
 - (ii) The discussion following the paper presentation and paper critique presentation demonstrates that the student has read the paper.
 - (iii) Student's involvement and competence in the discussion from (ii) will directly affect the "Attendance, Class Participation and Effort"'s 20% of student's total grade for the entire course.
-

CSC 320 / IMM 320

Term Project/Paper Suggestions

I. Design and Implementation of an IR System Project

1. Selection of a document collection for IR system.
2. Marking up or unifying mark ups of collection documents.
3. Environment preparation for document collection.
4. Creating SMART system specification files.
5. Processing document collection using SMART system and specs from 4.
6. Executing different search queries using SMART system for retrieval.
7. Reporting on phases 1.-6. and evaluating the IR system built

II. Research Projects

These projects will investigate an open issue in IR and prepare a research paper outlining existing approaches, their strengths and weaknesses, and offer a new approach to be investigated.

1. Natural language processing approaches in IR
2. Question answering
3. Evaluation of IR performance
4. Automatic summarization methods
5. Automated classification
6. Machine learning in IR
7. Cross-lingual IR
8. Cross-lingual summarization
9. Multi-media retrieval (speech, video, web pages)
10. Information fusion.