

CSI 550: Information Retrieval

I. Introduction to Information Retrieval

Dates: 9/04 (introductory class) and 9/11

Reading: SEIRP Chapters 1 and 2

- A. What is Information Retrieval?
 - B. The notion of Relevance.
 - 1. Conceptual
 - 2. Computational
 - 3. Document similarity measures.
 - C. IR tasks:
 - 1. Ad-hoc Querying.
 - 2. Filtering and Routing
 - 3. Topic Detection and Tracking
 - 4. Question Answering.
 - 5. Automated Summarization.
 - 6. Information Fusion.
-

II. Conceptual Models of IR systems.

Dates: 9/25 & 10/02

**** LUCENE tutorial 9/25 ****

**** Pre-project: Use Lucene to index data and run queries ****

**** Reading: SEIRP Chapter 7 – sections 1-3*

- 1. Boolean
 - 2. Vector-Space
 - 3. Probabilistic
 - 4. Extended models
-

III. Evaluation

Dates: 10/09 & 10/16 – lectures and student presentations

**** Project in-class pitches 5 mins max on 10/09 ****

Reading: SEIRP chapter 8

- A. Assumptions in IR performance evaluation.
 - 1. Fully automated vs. interactive systems.
 - 2. Who determines relevance?

- B. Evaluation metrics
 - 1. Recall and Precision, Miss and False Alarm, ROC curves
 - C. Reference Collections
 - 1. Classic collections: Cranfield, CACM, ISI, INSPEC, ...
 - 2. Tipster/TREC, TDT collections
 - D. Evaluation methodology
 - 1. Experimental design, running experiments and collecting results.
 - 2. Standard analyses and analysis tools.
 - E. Standard Evaluations and results
 - 1. TREC & TREC tracks, TDT
 - 2. SUMMAC, DUC, TAC, etc.
-

IV. Automated Content Indexing

Dates: 10/23, 10/30 & 11/06 lectures and student presentations

SEIRP chapters 4 & 5; + additional sources

**** Quiz #1 on 10/30 (1 hour) ****

**** Project progress reports due 11/06 ****

- A. Properties of language and media collections.
 - 1. Statistical distributions; Zipf's law; Stochastic language models.
 - 2. Metadata and Markup Languages
 - 3. Multimedia (text, video, audio, graphics)
 - 4. Full-text vs. Bibliographic records.
 - 5. Passage vs. document retrieval
- B. Indexing and storage issues.
 - 1. Index compression.
 - 2. Automating Hypertext linkages.
 - 3. Positional information in indexes.
- C. Data and File Structures for Information Retrieval.
 - 1. Inverted files and other indexing structures
 - 2. File structures (PAT trees, Grid Files, Hashing).
- E. *Text Analysis & Linguistically Motivated Indexing (LMI)*
 - 1. Basic text processing: Stoplists and Stemming, Phrases and collocations
 - 2. Morphological and lexical analysis, Part-of-speech tagging and Parsing
 - 3. Phrase recognition, syntactic structures, concept extraction
 - 4. Case studies: MUC, SCISOR, FERRET, NLIR
- G. Thesaurus Construction (possible student presentations)
 - 1. Collection-sensitive thesauri.
 - 2. Manually derived thesauri (WordNet, Snomed, MESH, LCSH).
 - 3. Automatically derived thesauri.

- a. Term associations.
- b. Latent Semantic Indexing.

H. Genetic Information Retrieval (student presentation)

- 1. searching genetic sequences
 - 2. issues with genetic retrieval
-

V. Query Languages and Query operations

Date: 11/13

SEIRP Chapter 6 + additional readings

**** Project progress in-class report outs (5 mins max) 11/20 ****

- Keyword queries
 - Bag-of-words queries
 - Boolean queries
 - Enhanced BOW queries
 - Relevance feedback
 - Query expansion & user interaction (term paper)
 - Cross-language retrieval (term paper)
-

VI. Automatic Classification and Clustering

Date: 11/20

Reading: SEIRP Chapter 9

- A. Automatic Classification
 - 1. Standing queries and profiles
 - 2. Routing task
 - 3. Filtering task
 - 4. Clustering.
 - B. Topic Detection and Tracking
 - C. Automated Summarization
 - D. Information Fusion
-

VII. Web Search and Browsing

Dates: 12/04

SEIRP Chapter 3 and 4.5 + additional materials

**** Quiz #2 on 12/04 (1 hour) ****

- A. About the Web and hypertext
- B. Types of web search engines
- C. Web crawling and indexing

- D. Page ranking approaches
 - E. E-commerce applications, shopping bots, etc.
-

VIII. Question Answering/Collaborative Systems

Dates: 12/04 (optional)

Reading: SEIRP Chapters 10 & 11

- Classical Q&A problem (student presentation)
 - Modern factoid QA systems (student presentation)
 - Interactive Q&A
 - Social Search and Collaborative QA
 - Case studies: LCC, IBM, Albany
-

IX. CSI 550 Workshop

Date: 12/11

**** Project presentations and demonstrations ****

**** Final project reports due ****

A technical mini-conference with 10 minutes presentations/demonstrations of course project work.

Textbooks:

- (SEIRP) *Search Engines: Information Retrieval in Practice*. W. Bruce Croft, Donald Metzler and Trevor Strohman. Addison-Wesley, 2009. ISBN-13: 978-0136072249
- (MIR) *Modern Information Retrieval*, Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Addison-Wesley/ACM Press, 1999, ISBN 0-201-39829-X – currently out of print

Some Additional Texts and Sources:

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. T. Strzalkowski. *Natural Language Information Retrieval*. Kluwer/Springer, 1999.
4. Text Retrieval Conference (TREC) proceedings (on the web trec.nist.gov)
5. ACM SIGIR Conference Proceedings (copies from instructor)
6. Technical journals:
 - a. *Information Processing & Management*, Pergamon Press (Dewey Library)
 - b. *Information Retrieval*, Kluwer Academic Publishers (Library)
 - c. *Computational Linguistics*, MIT Press (copies from the instructor)
 - d. *Journal of the ASIST* (Dewey Library)
 - e. *Natural Language Engineering* (Cambridge U Press)