CS798 (Fall 2007)

Information Retrieval

Instructor

Charles Clarke
Office: DC2597D
Email: claclark@plg.uwaterloo.ca
Phone: (519) 888-4567 (x32184)
Web: plg.uwaterloo.ca/~claclark

Overview

Information retrieval (IR) is concerned with representing, searching and manipulating large collections of electronic text and other human-language data. IR systems and services are now widespread, with millions of people depending on them daily to facilitate business, education and entertainment. Web search engines — Google, Windows Live, Yahoo and others — are by far the most popular and heavily used IR services, providing access to up-to-date technical information, locating people and organizations, and summarizing news and events. Other IR applications include digital libraries, desktop search, enterprise content management, as well as IR components embedded in email clients and other software.

This course provides an overview of IR theory, algorithms, data structures, evaluation, and applications. The material overlaps the areas of databases, natural language processing, machine learning, and HCI, and will be of interest to students in those areas. Course-work will include a number of short written problems, a literature survey, an in-class presentation and a major project.

Topics

- *Basic Techniques:* representation, search, indexing, inverted indices, Boolean retrieval, the vector space model, proximity, test collections, evaluation.
- *Searching and Browsing:* probabilistic retrieval, Okapi BM25, language modeling, divergence from randomness, passage retrieval, classification, clustering, learning to rank, implicit user feedback.
- *Algorithms and Data Structures*: static index construction, dynamic index construction, index compression, query processing, query optimization, simple structural queries.
- *Evaluation:* statistical foundations of evaluation, efficiency, response time, effectiveness, recall, precision, NDCG, other measures.
- *Applications:* architectures, parallel systems, Web retrieval, XML retrieval, spam filtering, filesystem search, digital libraries, multimedia retrieval.

Grading

Your final grade will be based on:

1. short exercises (10%);
2. a literature review of a topic area selected by the student with the agreement of the instructor (30%);
3. a 30 minute in-class presentation covering your literature review (20%);
4. a course project, which will include both individual work and teamwork involving the class as a whole (40%);

Initial Meeting

Monday, September 10, 2007, 2:00-3:00, MC2036
If you are interested in the course, but you cannot attend the meeting, please send me email.