636 - Information Retrieval

Course Objectives: Basic and advanced techniques for text-based information systems: efficient text indexing; Boolean and vector based retrieval models; Web search including crawling.

Course Contents:

Overview of Information Retrieval: Function of an IR system, Kinds of IR systems, Components of an IR system, Problems in designing an IR system.The nature of unstructured and semi-structured text.

Text Analysis and Indexing: Preliminary stages of text analysis and document processing, tokenization, stemming, lemmatization, stop words, phrases, Indexing: Boolean IR models, inverted files, indexing, signature files, PAT trees, Positional indices. Vector-based IR models: TF/IDF term weighing, similarity measures, test collections and issues.

Index construction and Compression: Postings size estimation, merge sort, dynamic indexing, positional indexes, n-gram indexes. Index compression: lexicon compression and postings lists compression. Gap encoding, gamma codes, Zipf's Law. Blocking. Extreme compression.

Query Processing: Query expansion: spelling correction and synonyms. Wild-card queries, permuterm indices, n-gram indices. Edit distance, soundex, language detection.

Matching techniques: Similarity between documents and queries, Parametric or fielded search. Document zones. The vector space retrieval model, tf.idf weighting. Scoring documents, vector space scoring, the cosine measure, efficiency considerations, reduced dimensionality approximations, Latent Semantic Indexing (LSI), random projection, Page Ranking and HITS.

Information Extraction: Information extraction, Named entity extraction, Question Answering. Summarization - Qualities of good summary, summary types, extract summary.

Evaluation of IR systems: Assessment of the performance of IR systems - Precision, Recall, F-Measure. Criteria for evaluation, measuring 'goodness', tests of IR systems. Presentation of search results, display of search results, manipulation of search results.

Relevance feedback: User modeling and information need: user profiling, Relevance judgments. Additional term selections to the system, Dynamic respond ally to judgments and selections, Personalization of search.

Taxonomy and Ontology: Creating domain specific ontology, Ontology life cycle.

Distributed and Parallel IR: Relationships between documents, Identify appropriate networked collections, Multiple distributed collections simultaneously.

Web Search Engines: Web crawlers, robot exclusion, Web data mining, Metacrawler, Collaborative filtering, Web agents (web shopping, bargain finder,..), Economic, ethical, legal and political issues.

Multimedia IR: Techniques to represent audio and visual document, Query databases of multimedia documents, Display the results of multimedia searches.

Main Reading:

1. Managing Gigabytes, by I. Witten, A. Moffat, and T. Bell.
2. Modern Information Retrieval, by R. Baeza-Yates and B. Ribeiro-Neto.
3. Information Retrieval: Algorithms and Heuristics by D. Grossman and O. Frieder