# INFORMATION RETRIEVAL SYSTEMS

***(Professional Elective-VI)/(Common for CSE,IT)***

COURSE CODE: 15CT1139[1]

| L | T | P | C |
|---|---|---|---|
| 3 | 0 | 0 | 3 |

Pre-requisites: *Database Management Systems*

**COURSE OUTCOMES:**
At the end of the course the student shall be able to

**CO1:** Design pre-processing methods for effective information retrieval
**CO2:** Build tolerant information retrieval.
**CO3:** Implement index compression process.
**CO4:** Formulate textual information into vectors
**CO5:** Analyze ranked and unranked search results

**UNIT-I** **(8-10 Lectures)**

**BOOLEAN RETRIEVAL:** An example information retrieval problem, A first take at building an inverted index, Processing Boolean queries, The extended Boolean model versus ranked retrieval.
**THE TERM VOCABULARY AND POSTINGS LISTS:** Document delineation and character sequence decoding, obtaining the character sequence in a document, choosing a document unit, Determining the vocabulary of terms ,Tokenization, Dropping common terms: stop words, Normalization (equivalence classing of terms) stemming and lemmatization, Faster postings list intersection via skip pointers, Positional postings and phrase queries , Bi-word indexes , Positional indexes, Combination schemes.

**UNIT-II** **(8-10 Lectures)**

**DICTIONARIES AND TOLERANT RETRIEVAL:** Search structures for dictionaries , Wildcard queries, General wild card queries, k-gram indexes for wildcard queries, Spelling correction ,Implementing spelling correction, Forms of spelling correction , Edit distance , k-gram indexes for spelling correction, Context sensitive spelling correction , Phonetic correction.
**INDEX CONSTRUCTION:** Hardware basics, Blocked sort-based indexing , Single-pass in-memory indexing , Distributed indexing , Dynamic indexing , Other types of indexes

**UNIT-III** **(8-10 Lectures)**

**INDEX COMPRESSION:** Statistical properties of terms in information retrieval, Heaps' law: Estimating the number of terms, Zipf's law: Modeling the distributionof terms, Dictionary compression, Dictionary as a string, Blockedstorage, Postings file compression, Variable byte codes, γ-codes.SCORING, TERM WEIGHTING:Parametric and zone indexes, Weighted zone scoring, Learning weights, The optimal weight g, Term frequency and weighting ,Inverse document frequency, Tf-idf weighting.

**UNIT-IV** **(8-10 Lectures)**

**THE VECTOR SPACE MODEL:** The vector space model for scoring, Dot products, Queries as vectors, Computing vector scores, Variant tf-idf functions, Sub linear tfscaling , maximum tf normalization, Document and query weighting schemes, Pivoted normalized document length

**UNIT-V**

**EVALUATION IN INFORMATION RETRIEVAL:** Information retrieval system evaluation, Standard test collections, Evaluation of unranked retrieval sets, Evaluation of ranked retrieval results, Assessing relevance, Critiques and justifications of the concept Relevance, A broader perspective: System quality and user utility, System issues , User utility , Refining a deployed system, Results snippets.

**TEXT BOOKS:**

1. Christopher D. Manning, PrabhakarRaghavan, HinrichSchütze, "An Introduction to Information Retrieval" , 1st Edition, Cambridge University Press, 2008.

**REFERENCES:**

1. G G Chowdhury "Introduction to Modern Information Retrieval", 3rd Edition, neal-schuman publishers, 2010.
2. Gerald J.Kowalski, Mark T.Maybury, "Information storage and Retrieval systems: theory and implementation", 2nd Edition, kluwer academic publishers, 2009.

**WEB REFERENCES:**

1. http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

***