

INFORMATION RETRIEVAL SYSTEMS

(Professional Elective-V)/ (Common to CSE & IT)

Course Code : 15CT1139

L	T	P	C
3	0	0	3

Pre-requisites:

Database Management Systems

Course Outcomes:

At the end of the Course, the Student will be able to:

- CO 1** Design pre-processing methods for effective information retrieval
- CO 2** Build tolerant information retrieval.
- CO 3** Implement index compression process.
- CO 4** Formulate textual information into vectors
- CO 5** Analyze ranked and unranked search results

UNIT-I

(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**(10 Lectures)****DICTIONARIES AND TOLERANT RETRIEVAL:**

Search structures for dictionaries, Wildcard queries, General wildcard 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**(10 Lectures)****INDEX COMPRESSION:****STATISTICAL PROPERTIES OF TERMS IN INFORMATION RETRIEVAL, HEAPS' LAW:**

Estimating the number of terms, Zipf's law: Modeling the distribution of terms, Dictionary compression, Dictionary as a string, Blocked storage, Postings file compression, Variable byte codes, \tilde{a} -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**(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 tf scaling, 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 BOOK:

Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, “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 REFERENCE:

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