DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY PATNA

Ashok Raj Path, PATNA 800 005 (Bihar), India

Phone No.: 0612 – 2372715, 2370419, 2370843, 2371929, 2371930, 2371715 Fax – 0612- 2670631 Website: www.nitp.ac.in

MC46XX12 **Information Retrieval** **L-T-P-Cr: 2-0-2-3**

Pre-requisites: Students must have a minimal concept of Data Base Management Systems, fundamental knowledge of Data structure and programming concepts of Python and they must also have a minimal knowledge of Natural language such as the thesaurus, synonyms, etc.

Course Objectives.

1. To present the scientific support in the field of information search and retrieval.
2. This course explores the fundamental relationship between information retrieval, hypermedia architectures, and semantic models, thus deploying and testing several important retrieval models such as vector space, Boolean, and query expansion.
3. It also discusses different methods for clustering and classifying documents to enhance the efficiency of the retrieval system.
4. To understand the XML and web applications of the information retrieval systems.

Course Outcomes – After completing this course, students should be able to:

CO-1. *Define and explain* the basics of information retrieval, and the heart of search engines, inverted indexes, and shows how simple Boolean queries can be processed using such indexes.

CO-2. *Describe* several algorithms for constructing the inverted index from a text collection with particular attention to highly scalable and distributed algorithms that can be applied to very large collections and techniques for compressing dictionaries and inverted indexes.

CO-3. Determine an information retrieval system based on the relevance of the documents it retrieves, and compare the relative performances of different systems on benchmark document collections and queries.

CO-4. *Evaluate* the methods by which retrieval can be enhanced through the use of techniques like relevance feedback and query expansion, which aim at increasing the likelihood of retrieving relevant documents.

CO-5. *Identify* and *examine* the structured retrieval by reducing it to the vector space scoring methods which invokes probability theory to compute scores for documents on queries.

CO-6. *Design and develop* specific Information Retrieval problems using open-source information libraries and develop software using a high-performance, and full-featured text search engine.

Course Outcomes–Cognitive Levels–Program Outcomes Matrix –

[H: High relation (3); M: Moderate relation (2); L: Low relation (1)]

Course Outcomes	Program Outcomes											
	PO-1 (Engineering knowledge)	PO-2 (Problem analysis)	PO-3 Design/development of solutions)	PO-4 (Conduct investigations of complex problems)	PO-5 (Modern tool usage)	PO-6 (The engineer and society)	PO-7 (Environment and sustainability)	PO-8 (Ethics)	PO-9 Individual and team work)	PO-10 Communication)	PO-11 (Project management and finance)	PO-12 (Life-long learning)
CO-1	3	3	3	3	3	3		1	3	3	1	3

CO-2	3	3	3	3	3	3	2	1	3	3	1	3
CO-3	3	3	3	3	3	3	1	1	3	3	1	3
CO-4	3	3	3	3	3	2			3	3	1	3
CO-5	3	3	3	3	3	3	2	1	3	3	1	3
CO-6	3	3	3	3	3	1	1	1	3	3	2	2

UNIT 1 Introduction to Information Retrieval: **Lectures: 6**

Early Developments; Inverted index and Boolean queries; Term vocabulary and posting lists; document delineation and character sequence decoding, determining the vocabulary of terms, faster postings list intersection via skip pointers, positional postings, and phrase queries.

UNIT 2 Text Indexing, Storage and Compression: **Lectures: 9**

Tokenization, stemming, stop words, phrases, index optimization; Index construction; blocked sort-based indexing, single-pass in-memory indexing, distributed indexing, dynamic indexing, and Index compression: Gap encoding, gamma codes, Statistical properties; Zipf's Law, Heap's Law, Dictionary compression, posting file compression.

UNIT 3 Information Retrieval Models: **Lectures: 10**

Boolean; Vector space; Parametric and zone indexes, Term frequency; TF-IDF; variant TF-IDF functions, probabilistic information retrieval; Probability Ranking Principle, Binary Independence model, Okapi BM25; language modeling; Text classification and Naive Bayes, Vector space scoring; The cosine measure; Relevance feedback and query expansion; Rocchio;.

UNIT 4 Performance Evaluation of Information Retrieval: **Lectures: 7**

Information retrieval system evaluation, standard test collections, Evaluating search engines; User happiness, accuracy, precision, recall, F-measure, kappa measure, Evaluation of ranked retrieval results, Assessing relevance, System quality and user quality, and Results snippets.

UNIT 5 Text Clustering: **Lectures: 5**

Clustering versus classification; Partitioning methods; k-means clustering; Mixture of Gaussians model; Flat clustering; Hierarchical clustering**.**

UNIT 6 XML retrieval and Web Information Retrieval: **Lectures: 5**

XML concepts, vector space model, evaluation of XML retrieval, text-centric vs. data-centric in XML, industrial applications of XML retrieval, Hypertext; web crawling; search engines; ranking; link analysis; PageRank.

References

1. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. Introduction to Information Retrieval, Cambridge University Press, 2008.
2. Stefan Büttcher, Charles L. A. Clarke, Gordon V. Cormack. Information Retrieval: Implementing and Evaluating Search Engines, MIT Press.
3. Baeza-Yates and Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley.
4. William B. Frakes, Information Retrieval: Data Structures and Algorithms