

**OBJECTIVES:**

- To understand the basics of information retrieval with pertinence to modeling, query operations and indexing.
- To understand the various applications of information retrieval giving emphasis to multimedia IR, web search.
- To learn measuring effectiveness and efficiency of information retrieval techniques.
- To get used to performing Parallel Information Retrieval.
- To understand the concepts of digital libraries.

**UNIT I INTRODUCTION 9**

Basic Concepts – Practical Issues – Retrieval Process – Architecture – Boolean Retrieval – Retrieval Evaluation – Open Source IR Systems–History of Web Search – Web Characteristics–The impact of the web on IR –IR Versus Web Search–Components of a Search engine

**UNIT II RETRIEVAL MODELING 9**

Taxonomy and Characterization of IR Models – Boolean Model – Vector Model – Term Weighting – Scoring and Ranking –Language Models – Set Theoretic Models – Probabilistic Models – Algebraic Models – Structured Text Retrieval Models – Models for Browsing

**UNIT III INDEXING 9**

Static and Dynamic Inverted Indices – Index Construction and Index Compression. Searching – Sequential Searching and Pattern Matching. Query Operations –Query Languages – Query Processing – Relevance Feedback and Query Expansion – Automatic Local and Global Analysis – Measuring Effectiveness and Efficiency

**UNIT IV EVALUATION AND PARALLEL INFORMATION RETRIEVAL 9**

Traditional Effectiveness Measures – Statistics in Evaluation – Minimizing Adjudication Effect – Nontraditional Effectiveness Measures – Measuring Efficiency – Efficiency Criteria –Queueing Theory – Query Scheduling – Parallel Information Retrieval – Parallel Query Processing – MapReduce

**UNIT V SEARCHING THE WEB 9**

Searching the Web –Structure of the Web –IR and web search – Static and Dynamic Ranking – Web Crawling and Indexing – Link Analysis – XML Retrieval Multimedia IR: Models and Languages – Indexing and Searching Parallel and Distributed IR – Digital Libraries

**TOTAL : 45 PERIODS****OUTCOMES:**

**Upon completion of the course, the student will be able to**

- Build an Information Retrieval system using the available tools.
- Identify and design the various components of an Information Retrieval system.
- Measure effectiveness and efficiency of information retrieval techniques.
- Use parallel Information Retrieval approaches in real world problems.
- Design an efficient search engine and analyze the Web content structure.

**REFERENCES:**

1. Ricardo Baeza – Yates, Berthier Ribeiro – Neto, “Modern Information Retrieval: The Concepts and Technology behind Search”, (ACM Press Books), Second Edition, 2011.
2. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze, “Introduction to Information Retrieval”, Cambridge University Press, First South Asian Edition, 2008.
3. Stefan Buttcher, Charles L. A. Clarke, Gordon V. Cormack, “Information Retrieval Implementing and Evaluating Search Engines”, The MIT Press, Cambridge, Massachusetts London, England, 2010.

CO	PO						PSO		
	1	2	3	4	5	6	1	2	3
1.	√		√	√			√	√	
2.	√		√				√	√	
3.	√		√				√	√	
4.	√		√	√		√	√	√	√
5.	√		√	√			√	√	√

**BD5151**

**BIG DATA MINING AND ANALYTICS**

**L T P C**  
**3 0 0 3**

**OBJECTIVES:**

- To understand the computational approaches to Modeling, Feature Extraction.
- To understand the need and application of Map Reduce.
- To understand the various search algorithms applicable to Big Data.
- To analyze and interpret streaming data.
- To learn how to handle large data sets in main memory.
- To learn the various clustering techniques applicable to Big Data.

**UNIT I DATA MINING AND LARGE SCALE FILES**

**9**

Introduction to Statistical modeling – Machine Learning – Computational approaches to modeling – Summarization – Feature Extraction – Statistical Limits on Data Mining – Distributed File Systems– Map-reduce – Algorithms using Map Reduce – Efficiency of Cluster Computing Techniques.

**UNIT II SIMILAR ITEMS**

**9**

Nearest Neighbor Search – Shingling of Documents – Similarity preserving summaries – Locality sensitive hashing for documents – Distance Measures – Theory of Locality Sensitive Functions – LSH Families – Methods for High Degree of Similarities.

**UNIT III MINING DATA STREAMS**

**9**

Stream Data Model – Sampling Data in the Stream – Filtering Streams – Counting Distance Elements in a Stream – Estimating Moments – Counting Ones in Window – Decaying Windows

**UNIT IV LINK ANALYSIS AND FREQUENT ITEMSETS**

**9**

Page Rank –Efficient Computation – Topic Sensitive Page Rank – Link Spam – Market Basket Model – Apriori algorithm – Handling Larger Datasets in Main Memory – Limited Pass Algorithm – Counting Frequent Item sets.

**UNIT V CLUSTERING**

**9**

Introduction to Clustering Techniques – Hierarchical Clustering –Algorithms – K-Means – CURE – Clustering in Non – Euclidean Spaces – Streams and Parallelism – **Case Study:** Advertising on the Web – Recommendation Systems

**TOTAL : 45 PERIODS**