Programme: BTech /PG Year: 2015 Semester: Final year/First Year

**Course code: Information Retrieval and Web Search**

Course: Program Elective Credits: 04 Hours: 40

**Course Context and Overview (100 words):**

**The growth of data is exponential in nature and the Internet is undisputedly the world’s largest collection of information. In such a huge pool of web data, search engines are becoming the key technologies to extract useful material among the billions of available resources. This requires automated way to find, analyze, store, and extract relevant and comprehensive data. The course focuses on techniques used to retrieve useful information from repositories such as the Web. It introduces standard concepts in information retrieval such as documents, queries, collections, and relevance. At later stage it covers efficient indexing, retrieval techniques and their applications.**

**Prerequisites Courses: Any Programming Language, Data Structure, basics of Math-III**

**Course outcomes (COs):**

|  |
| --- |
| **On completion of this course, the students will have the ability to:** |
| **CO1: Understand the common algorithms and techniques for information retrieval (document indexing and retrieval, query processing, etc).** |
| **CO2: Understand quantitative evaluation methods for the IR systems and data mining techniques.** |
| **CO3: Understand and apply popular probabilistic retrieval methods and ranking principles for analysis of retrieved data.** |
| **CO4: Understand the techniques and algorithms existing in practical retrieval and data mining systems such as those in web search engines and recommender systems.** |

**Course Topics:**

|  |  |  |
| --- | --- | --- |
| **Topics** | **Lecture Hours** | |
| **UNIT - I**   1. **Topic** : **Basic IR Models*:*** |  |  |
| **1.1 Boolean and vector-space retrieval models; Ranked retrieval; Text similarity metrics.** | 03 | 10 |
| **1.2 TF-IDF (term frequency/inverse document frequency) weighting; Cosine similarity. Basic Tokenizing Indexing, and Implementation of Vector-Space Retrieval.** | 03 |
| **1.3 Simple tokenizing, stop-word removal, and stemming; inverted indices; efficient processing with sparse vectors.** | 04 |  |
| **UNIT - II**   1. **Topic: Experimental Evaluation of IR** |  | 10 |
| * 1. **Performance metrics: recall, precision, and F- measure;** | 02 |
| * 1. **Evaluations on benchmark text collections.** | 03 |
| * 1. **Query Operations and Languages: Relevance feedback; Query expansion; Query languages**. | 05 |
| **UNIT - III**   1. **Topic** |  | 10 |
| * 1. **Text Representation: Word statistics; Zipf's law; Porter stemmer; Morphology; Index term selection; Using thesauri. Metadata and markup languages (SGML, HTML, XML).** | 03 |
| * 1. **Text Categorization and Clustering: Categorization algorithms: Naive Bayes, Decision trees; Clustering algorithms: Agglomerative clustering; k-means;** | 03 |
| * 1. **Expectation maximization (EM); Applications to information filtering, organization, and relevance feedback; SVM, Matrix decomposition, Latent Semantic Indexing.** | 04 |
| **UNIT - IV**   1. **Topic: Web Search*:*** |  | 10 |
| * 1. **Search engines; Spidering; Metacrawlers; Directed spidering;** | 03 |
| * 1. **Link analysis; Shopping agents; Focused crawling.** | 03 |
| * 1. **4.3 Advanced Topics: Covers recent concepts and algorithms** | 04 |

**Textbook references (IEEE format):**

**Text Books:**

1. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze: ***Introduction to Information Retrieval****,* Cambridge University Press (Available online)
2. Soumen Chakrabarti: ***Mining the web***, Morgan Kaufmann.

**Reference Books:**

1. D. Grossman and O. Frieder: ***Information Retrieval: Algorithms and Heuristics***, Springer
2. R. Baeza-Yates and B. Ribeiro-Neto*:* ***Modern Information Retrieval***, Addison Wesley
3. ***Research papers*** which will be communicated during the course

**Evaluation Methods:**

|  |  |
| --- | --- |
| Item | Weightage |
| Assignments | 40% |
| Implementations of Algorithms |
| Seminar Presentations |
| Midterm | 20% |
| Final Examination | 40% |

**Prepared By:**

**Update: ­21/04/2015**