17CS41E2 - INFORMATION RETRIEVAL

|  |  |  |  |
| --- | --- | --- | --- |
| **Course Category:** | Professional Elective | **Credits:** | 3 |
| **Course Type:** | Theory | **Lecture – Tutorial – Practical:** | 3-0-0 |
| **Prerequisite:** | Require the basics of Database Management System concepts | **Sessional Evaluation:**  **Univ. Exam Evaluation:**  **Total Marks:** | 40  60  100 |
| **Objectives** | * + To cover the foundations of information retrieval, design, analysis and implementation of IR systems.   + To become familiar with difference between Information retrieval and data Base Management Systems | | |

|  |  |  |
| --- | --- | --- |
| **Course Outcomes** | Upon the successful completion of the course, the students will be able to: | |
| CO1 | Understand the fundamentals of Information retrieval and the difficulty of representing and retrieving documents, images etc. |
| CO2 | Specify the design of indexing algorithms which is governed by hardware constraints and benefits of Compression. |
| CO3 | Learn the essentials of a search engine to rank the order of documents. |
| CO4 | Address the formal evaluation methodology that has been developed for evaluating unranked retrieval results. |
| CO5 | Generate the probabilistic approach to information retrieval and language models. |
| CO6 | |  | | --- | | Use the classification for standing queries and deal with web indexes. | |
| **Course Content** | **UNIT – I**  **Introduction**: Overview of Information Retrieval, Motivation, History, Information Retrieval vs. Data Retrieval, Vocabulary and Posting lists, Dictionaries and Tolerant retrieval.  **UNIT – II**  **Index Construction**: Blocked sort-based indexing, single pass in memory indexing, Distributed indexing, Dynamic indexing.  **Index Compression**: Dictionary and Postings file compression.  **UNIT – III**  **Vector Space Model**: Parametric and zone indexes, term frequency and weighting, scoring, ranking, Components of IR systems, Vector Space scoring and query operator interaction.  **UNIT – IV**  **Evaluation of IR:** Standard test collections, Evaluation of unranked and ranked retrieval sets, System quality and user utility, Relevance feedback and query expansion Query Languages, Evaluation of XML retrieval  **UNIT – V**  **Probabilistic information retrieval**: Probability theory, probability ranking, e binary independence model, appraisal and some extensions.  **Language models for information retrieval**: query likelihood model, Language modeling versus other approaches, Extended language modeling  **UNIT – VI**  **Text classification and Naive Bayes**: Naive Bayes text classification, Properties of Naive Bayes, Evaluation of text classification, Vector space classification.  **Web Search**: History of Web, Index size and estimation, Web Crawling, Link Analysis(HITS, Google PageRank), User Interfaces and Visualization | |
| **Text Books and References:** | **TEXT BOOK(S):**   1. "[Introduction to Information Retrieval](http://www-csli.stanford.edu/~schuetze/information-retrieval-book.html)". C.D. Manning, P. Raghavan, and H. Schütze. Cambridge University Press, 2008.   **REFERENCE BOOKS:**   1. "[Modern Information Retrieval](http://people.ischool.berkeley.edu/~hearst/irbook/)". Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Addison Wesley, 1999. 2. "[Mining the Web: Discovering Knowledge from Hypertext Data](http://www.cse.iitb.ac.in/~soumen/mining-the-web/)". Soumen Chakrabarti. Morgan Kaufmann, 2003. 3. "[Information Retrieval: Algorithms and Heuristics](http://www.springer.com/west/home/generic/search/results?SGWID=4-40109-22-35393326-0&referer=www.springeronline.com&SHORTCUT=www.springer.com/sgw/cda/frontpage/0,11855,4-40109-22-35393326-0,00.html)", D. Grossman and O. Frieder. Springer, 2nd Ed., 2004. 4. "[Managing Gigabytes: Compressing and Indexing Documents and Images](http://www.cs.mu.oz.au/mg/)", I. Witten, A. Moffat, and T. Bell. 2nd Ed., Morgan Kaufmann, 1999. | |
| **E-Resources** | * + Lucene Text Search Engine (<http://lucene.apache.org/>)   + Hadoop (https://hadoop.apache.org/)   + Strang's Linear Algebra Course (MIT) (<http://ocw.mit.edu/OcwWeb/Mathematics/18-06Spring-2005/CourseHome/index.htm>)   + Andrew Moore's Statistical Data Mining Tutorials (CMU) (<http://www.autonlab.org/tutorials/>)   + Matei Zaharia's Introduction to MapReduce and Hadoop (Cloud Computing) [(in powerpoint)](http://www.cs.berkeley.edu/~demmel/cs267_Spr09/Lectures/Cloud_MapReduce_Zaharia.ppt) or [(archived video)](mms://media.citris.berkeley.edu/cs267_4-22-09) | |