# INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)
Dundigal, Hyderabad - 500 043

## COMPUTER SCIENCE AND ENGINEERING

## COURSE DESCRIPTION FORM

| | | | | |
|---|---|---|---|---|
| **Course Title** | **INFORMATION RETREVAL SYSTEM** | | | |
| **Course Code** | **A70533** | | | |
| **Regulation** | **R15 - JNTUH** | | | |
| **Course Structure** | Lectures | Tutorials | Practicals | Credits |
| | 4 | - | - | 4 |
| **Course Coordinator** | Ms. S.J. Sowjanya, Associate Professor, CSE | | | |
| **Team of Instructors** | Mr. N. V. Krishna Rao, Associate Professor, CSE<br>Mr. C. Praveen Kumar, Assistant Professor, CSE | | | |

## I. COURSE OVERVIEW:

The main objective of this course is to present the scientific support in the field of information search and retrieval. 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. It discusses implementation and evaluation issues of new algorithms like clustering, pattern searching, and stemming with advanced data/file structures, indirectly facilitating a platform to implement comprehensive catalogue of information search tools while designing an e-commerce web site.

## II. PREREQUISITE(S):

| Level | Credits | Periods/ Week | Prerequisites |
|---|---|---|---|
| UG | 4 | 4 | Data base Management System |

## III. MARKS DISTRIBUTION:

| Sessional Marks | University End Exam marks | Total marks |
|---|---|---|
| **Midterm Test**<br>There shall be two midterm examinations. Each midterm examination consists of Essay paper, objective paper and assignment.<br>The essay paper is for 10 marks of 60 minutes duration and shall contain 4. The student has to answer 2 questions, each carrying 5 marks.<br>The objective paper is for 10 marks of 20 minutes duration. It consists of 10 multiple choice and 10 fill-in-the blank questions, the student has to answer all the questions and each carries half mark.<br>First midterm examination shall be conducted for the first two and half units of syllabus and second midterm examination shall be conducted for the remaining portion.<br>Five marks are earmarked for assignments. There shall be two assignments in every theory course. Assignments are usually issued at the time of commencement of the semester. These are of problem solving in nature with critical thinking.<br>Marks shall be awarded considering the average of two midterm tests in each course. | 75 | 100 |

## IV. EVALUATION SCHEME:

| S. No | Component | Duration | Marks |
|---|---|---|---|
| 1. | I Mid Examination | 80 minutes | 20 |
| 2. | I Assignment | - | 5 |
| 3. | II Mid Examination | 80 minutes | 20 |
| 4. | II Assignment | - | 5 |
| 5. | External Examination | 3 hours | 75 |

## V. COURSE OBJECTIVES:

**At the end of the course, the students will be able to:**

I. Demonstrate genesis and diversity of information retrieval situations for text and hyper media.
II. Describe hands-on experience store, and retrieve information from www using semantic approaches
III. Demonstrate the usage of different data/file structures in building computational search engines.
IV. Analyze the performance of information retrieval using advanced techniques such as classification, clustering, and filtering over multimedia.
V. Analyze ranked retrieval of a very large number of documents with hyperlinks between them.
VI. Demonstrate Information visualization technologies like Cognition and perception in the Internet or Web search engine.

## VI. COURSE OUTCOMES:

**After completing this course the student must demonstrate the knowledge and ability to:**

1. Describe the objectives of information retrieval systems.
2. Describe models like vector-space, probabilistic and language models to identify the similarity of query and document.
3. Implement clustering algorithms like hierarchical agglomerative clustering and k-means algorithm.
4. Understand relevance feedback in vector space model and probabilistic model.
5. Illustrate how N-grams are used for detection and correction of spelling errors**.**
6. Understand the method of Regression analysis to estimate the probability of relevance.
7. Understand the method to construct thesauri automatically and manually.
8. Understand natural language systems to build semantic networks for text**.**
9. Illustrate algorithms used for natural language processing.
10. Understand the measures to evaluate the performance of cross language information retrieval systems.
11. Understand query, document and phrase translation**.**
12. Design the method to build inverted index.
13. Understand the detection of the duplicate document.
14. Illustrate the information retrieval as relational application.
15. Understand the model of distributed information retrieval.

## VII. HOW PROGRAM OUTCOMES ARE ASSESSED:

| Program Outcomes | | Level | Proficiency assessed by |
|---|---|---|---|
| PO1 | **Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems. | H | Assignments, Tutorials |
| PO2 | **Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences. | H | Assignments |
| PO3 | **Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations. | S | Mini Projects |
| PO4 | **Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions. | S | Projects |
| PO5 | **Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations**.** | S | Mini Projects |
| PO6 | **The engineer and society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice. | S | Assignments |
| PO7 | **Environment and sustainability:** Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development. | N | -- |
| PO8 | **Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice. | N | -- |
| PO9 | **Individual and team work:** Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings. | N | -- |
| PO10 | **Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions. | S | Assignments |
| PO11 | **Project management and finance:** Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments. | N | -- |
| PO12 | **Life-long learning:** Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change. | S | Projects |

**N - None** **S - Supportive** **H - Highly Related**

## VIII. HOW PROGRAM SPECIFIC OUTCOMES ARE ASSESSED:

| Program Specific Outcomes | | Level | Proficiency assessed by |
|---|---|---|---|
| PSO1 | **Professional Skills:** The ability to research, understand and implement computer programs in the areas related to algorithms, system software, multimedia, web design, big data analytics, and networking for efficient analysis and design of computer-based systems of varying complexity. | H | Lectures, Assignments |
| PSO2 | **Problem-Solving Skills:** The ability to apply standard practices and strategies in software project development using open-ended programming environments to deliver a quality product for business success. | H | Mini Projects |
| PSO3 | **Successful Career and Entrepreneurship:** The ability to employ modern computer languages, environments, and platforms in creating innovative career paths, to be an entrepreneur, and a zest for higher studies. | S | Guest Lectures |

**N - None** **S - Supportive** **H - Highly Related**

## IX. SYLLABUS:

**UNIT - I**
Introduction: Retrieval Strategies: Vector Space Model, Probabilistic retrieval strategies: Simple term weights, Non binary independence model, Language models.

**UNIT – II**
Retrieval Utilities: Relevance feedback, Clustering, N-grams, Regression analysis, Thesauri.

**UNIT – III**
Retrieval Utilities: Semantic networks, parsing.
Cross-Language Information Retrieval: Introduction, Crossing the Language barrier.

**UNIT – IV**
Efficiency: Inverted Index, Query Processing, Signature files, Duplicate document detection.

**UNIT – V**
Integrated Structured Data and Text: A Historical progression, Information retrieval as a relational application, Semi-structured search using a relational schema.
Distributed Information retrieval: A Theoretical model of distributed retrieval, Web search.

**Text Books:**

1. David A. Grossman, Ophir Frieder, Information Retrieval- Algorithms and Heuristic, Springer , 2nd edition(Distributed by Universities Press),2004.

**Reference Books:**

1. Gerald J. Kowalski, Mark T. Maybury. Information Storage and Retrieval Systems Springer 2000
2. Soumen Chakrabarti, Mining the Web: Discovering Knowledge from Hypertext Data, Morgan- Kaufmann Publishers, 2002.
3. Christopher D. Manning, Prabhakar Raghavan, Hinrich. Schutze, Introduction to Information Retrieval, Cambridge University Press, Cambridge, England, 2009.

**Reference:**


1. Douglass R. Cutting, David R. Kargerd, Jan O. Pedersen, John W. Tukey: Scatter/Gather: a cluster-based approach to browsing large document collections – 1992 Cited 423 times.
2. Jay M. Ponte, W. Bruce Croft: A language modeling approach to information retrieval - 1998 Cited 721 times.
3. Thomas Hofmann: Probabilistic latent semantic indexing - 1999 Cited 768 times.
4. Jinxi Xu, W. Bruce Croft: Query expansion using local and global document analysis - 1996 Cited 412 times.
5. Yiming Yang, Xin Liu: A re-examination of text categorization methods - 1999 Cited 643 Times.
6. Kalervo jarvelin, Jaana Kekalanin: IR evaluation methods for retrieving highly relevant documents - 2000 Cited 379 times.
7. Jaime Carbonell, Jade Goldstein: The use of MMR, diversity-based reranking for reordering documents and producing Summaries - 1998 Cited 478 times.
8. Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl: An algorithmic framework for performing collaborative filtering - 1999 Cited 578 times.
9. Chengxiang Zhai, John Lafferty: A study of smoothing methods for language models applied to Ad Hoc information Retrieval - 2001 Cited 451 time.


## X. COURSE PLAN:

**At the end of the course, the students are able to achieve the following course learning outcomes:**

| Lecture No. | Topics to be covered | Course Learning Outcomes | Reference |
|---|---|---|---|
| 1-2 | Introduction to Information Retrieval Systems. Precision and Recall. | Understand the Functional overview of IRS | T1: 1 |
| 3-4 | Retrieval Strategies: Vector space model. | Illustrate the model for a query with different documents. | T1: 2.1 |
| 5-7 | Probabilistic retrieval strategies: simple term weights. | Applying weights to terms. | T1: 2.2 – 2.2.1 |
| 8 | Non binary independence model | Illustrate normalization of document length. | T1: 2.2.2 |
| 9-11 | Language models: Smoothing | Understand probability for each term with smoothing. | T1: 2.3 |
| 12-14 | Retrieval Utilities: Relevance feedback in the vector space model and probabilistic model | Understand relevance feedback . | T1: 3.1 |
| 15-18 | Clustering | Understand different clustering algorithms. | T1: 3.2 |
| 19-20 | N grams | Understand the N gram data structure. | T1: 3.4 |
| 21 | Regression Analysis | Understand the Probability of relevance. | T1: 3.5 |
| 22-25 | Thesauri | Understand the Construction of and generating thesauri. | T1: 3.6 |
| 26-27 | Retrieval Utilities: Semantic networks | Illustrate about different distance measures | T1: 3.7 |
| 28-29 | Parsing | Understand different parses | T1: 3.8 |

| | | | |
|---|---|---|---|
| 30-32 | Cross-Language Information retrieval: Introduction | Understand the evaluation of Cross-Language Information retrieval | T1: 4.1 |
| 33-36 | Crossing the language barrier | Understand query and document translation | T1: 4.2 |
| 37-39 | Efficiency: Inverted index | Illustrate the construction and compression of inverted index | T1: 5.1 |
| 40-43 | Query processing | Illustrate index modifications and simplifications | T1: 5.2 |
| 44-46 | Signature of files | Describe the scanning to remove false positives | T1: 5.3 |
| 47-48 | Duplicate document detection | Describe exact and similar duplicates | T1: 5.4 |
| 49-50 | Integrated structured data and text: A Historical progression | Understand user defined operators | T1: 6.2 |
| 51-52 | Information retrieval as a relational application | Illustrate the proximity searches | T1: 6.3 |
| 53-55 | Semi structured search using a relational schema | Explain the storage and tracking XML documents | T1: 6.4 |
| 56-57 | Distributed Information Retrieval: A Theoretical model of distributed retrieval | Understand the distributed retrieval models | T1: 8.1 |
| 58-60 | Web search | Understand different web searches | T1: 8.2 |

## XI. MAPPING COURSE OBJECTIVES LEADING TO THE ACHIEVEMENT OF PROGRAM OUTCOMES AND PROGRAM SPECIFIC OUTCOMES:

| Course Objectives | Program Outcomes | | | | | | | | | | | | Program Specific Outcomes | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PO1 | PO2 | PO3 | PO4 | PO5 | PO6 | PO7 | PO8 | PO9 | PO10 | PO11 | PO12 | PSO1 | PSO2 | PSO3 |
| I | H | | | | | S | | | | | | S | H | | |
| II | S | | S | | | | | | | | | | H | | |
| III | H | | | S | | | | | | H | | | | H | |
| IV | | S | | | | S | | | | | | | S | | |
| V | H | S | | | S | | | | | S | | | H | | S |
| VI | | | H | | | H | | | | | | | | H | |

**S – Supportive** **H - Highly Related**

## XII. MAPPING COURSE OUTCOMES LEADING TO THE ACHIEVEMENT OF PROGRAM OUTCOMES AND PROGRAM SPECIFIC OUTCOMES:

| Course Outcomes | Program Outcomes | | | | | | | | | | | | Program Specific Outcomes | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PO1 | PO2 | PO3 | PO4 | PO5 | PO6 | PO7 | PO8 | PO9 | PO10 | PO11 | PO12 | PSO1 | PSO2 | PSO3 |
| 1 | H | | | H | | | | | | S | | | S | | |
| 2 | | H | | | | | | | | | | | | H | |
| 3 | | H | H | | | | | | | | | | | S | |
| 4 | S | | | | | | | | | H | | | H | | S |
| 5 | | | | H | | | | | | | | | | H | |
| 6 | H | | | | | S | | | | | | H | | S | |
| 7 | | | H | | | | | | | | | | | H | |
| 8 | | | | H | | | | | | H | | | | S | |
| 9 | H | | | | | | | | | S | | H | S | | |
| 10 | | H | S | | | | | | | | | | | H | |
| 11 | H | | | | | | | | | H | | | H | | |
| 12 | | | H | | | | | | | | | | | H | |
| 13 | | H | | | | | | | | | | | | H | |
| 14 | | | H | | | | | | | | | | H | | |
| 15 | H | | | | | | | | | | | | | H | |

**S – Supportive** **H - Highly Related**

**Prepared by:**

Ms. S. J. Sowjanya, Associate Professor, CSE.
Mr. N. V. Krishna Rao, Associate Professor, CSE.
Mr. C. Praveen Kumar, Assistant Professor, CSE.

**HOD, COMPUTER SCIENCE AND ENGINEERING**