

Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur

Information Retrieval (IR)

Day 1

Prof. Niloy Ganguly

Email: niloy@cse.iitkgp.ernet.in

Web: <http://www.facweb.iitkgp.ernet.in/~niloy/>

Course Details

- **Course code:** CS60092
- **Credits (L-T-P):** 3-0-0
- **Class timings:** Wednesday 11:30 AM – 12:25 PM, Thursday 10:30 AM – 11:25 AM, Friday 8:30 AM – 10:25 AM
- **Slot:** E
- **Classroom:** 119, CSE
- **Faculty office:** 313, CSE

Study Source

- **Book:** *"Introduction to Information Retrieval"*, Christopher Manning, Prabhakar Raghavan and Hinrich Schütze, Cambridge University Press, 2008.
- **Web Link:** <http://nlp.stanford.edu/IR-book/information-retrieval-book.html>

Evaluation Details

- **Class Test 1:** 5 marks
- **Mid-semester:** 25 marks
- **Class Test 2:** 5 marks
- **End-semester:** 50 marks
- **Internal Assessment:** 10 marks
- **Attendance:** 5 marks

Teaching Assistants

- **Maunendra Sankar DeSarkar**
 - Email: maunendra@cse.iitkgp.ernet.in
- **Ishani Chakraborty**
 - Email: ishani@cse.iitkgp.ernet.in
- **Rishiraj Saha Roy**
 - Email: rishiraj@cse.iitkgp.ernet.in
- **Abhijnan Chakraborty**
 - Email: chakraborty.abhijnan@gmail.com

Course Overview (1)

■ Boolean retrieval model

- The *Boolean retrieval model* is a model for information retrieval in which we can pose any query which is in the form of a Boolean expression of terms, that is, in which terms are combined with the operators and, or, and not.

■ Vocabulary and posting lists

- The *vocabulary* is a list of terms that the system uses. It stores the occurrence of these terms in a linked data structure called a *posting list*.

■ Indexing

- *Indexing* refers to the storage of data in a memory efficient fashion enabling fast retrieval.

Course Overview (2)

■ Vector space model

- The representation of a set of documents as vectors in a common vector space is known as the *vector space model* and is fundamental to a host of information retrieval operations ranging from scoring documents on a query, document classification and clustering.

■ IR Evaluation

- Information retrieval has developed as a highly empirical discipline, requiring careful and thorough *evaluation* to demonstrate the superior performance of novel techniques on representative document collections.

■ Relevance feedback

- The idea of *relevance feedback* is to involve the user in the retrieval process so as to improve the final result set.

Course Overview (3)

■ Query expansion

- In *query expansion*, users give additional input on query words or phrases, possibly suggesting additional query terms.

■ Language models

- A *language model* is a function that puts a probability measure over strings drawn from the vocabulary of a language, which can be used to explain generation of sentences and corpora in that language.

■ Text classification

- Categorization of documents into a set of predefined classes is called text *classification*.

Course Overview (4)

- Support vector machines (SVMs)
 - An *SVM* is a kind of classifier: it is a machine learning method where the goal is to find a decision boundary between two classes that is maximally far from any point in the training data.
- Clustering
 - *Clustering* partitions documents into groups (which are not predefined, like classification) such that documents within a cluster should be as similar as possible; and documents in one cluster should be as dissimilar as possible from documents in other clusters.

Course Overview (5)

■ Web search

- *Web search* refers to finding documents from the World Wide Web by users by issuing queries to engines.

■ Web crawling

- *Web crawling* is the process by which we gather pages from the Web, in order to index them and support a search engine.

■ Link analysis

- *Link analysis* refers to the study of hyperlinks and the graph structure of the World Wide Web, which has been instrumental in the development of Web search.

Questions?

Thank you!