



Chapter 1 : Introduction to Information Retrieval		1-1 to 1-12
1.1	Introduction to Information Retrieval.....	1-1
1.2	The Nature of Unstructured and Semi-Structured Text.....	1-3
1.3	Inverted Index.....	1-4
1.4	Boolean Queries in Information Retrieval.....	1-8
1.5	Difference between Information Retrieval (IR) and Information Extraction (IE).....	1-11
1.6	Difference between Information Retrieval (IR) and Data Retrieval (DR).....	1-11
Chapter 2 : Text Indexing, Storage and Compression		2-1 to 2-28
2.1	Text Encoding.....	2-1
2.1.1	Tokenization.....	2-5
2.1.1(A)	Issues in Tokenization.....	2-5
2.1.2	Dropping Common Terms-Stop Words.....	2-6
2.1.3	Stemming and Lemmatization.....	2-7
2.1.4	Phrase Query.....	2-8
2.1.4(A)	Biword index.....	2-9
2.1.4(B)	Phrase (Positional) Index.....	2-9
2.2	Index Compression.....	2-11
2.2.1	Lexicon (Dictionary) compression.....	2-13
2.2.2	Postings lists compression.....	2-16
2.2.2(A)	Gap Encoding.....	2-16
2.2.2(B)	Gamma Codes.....	2-18
2.2.2(C)	Zipf's Law.....	2-19
2.3	Index Construction.....	2-21
2.3.1	Blocked Sort-based indexing.....	2-22
2.3.1(A)	Blocked Sort-based index construction: Sort and Merge.....	2-22
2.3.2	Dynamic indexing.....	2-24
2.3.3	Positional Index.....	2-25
2.3.4	N-gram Index.....	2-26
2.4	Real word Issues of Indexing.....	2-27

**Chapter 3 : Retrieval Models****3-1 to 3-38**

3.1	Overview of Retrieval Models.....	3-1
3.2	Boolean Model.....	3-2
3.2.1	The Extended Boolean Model	3-4
3.2.2	Ranked Retrieval.....	3-4
3.2.3	Scoring of Ranked Retrieval	3-4
3.3	TFIDF	3-4
3.3.1	Bag of words model	3-5
3.3.2	Term Frequency tf.....	3-5
3.3.3	Log-frequency weighting	3-6
3.3.4	Document frequency- idf weight.....	3-6
3.3.5	tf-idf weight	3-7
3.4	Vector Space Model.....	3-8
3.5	Probabilistic Information Retrieval	3-11
3.5.1	The Probabilistic Ranking Principle (PRP)	3-12
3.5.2	Binary independence model	3-12
3.6	OKAPI BM25 : A Non Binary Model	3-14
3.7	Language Modelling.....	3-15
3.7.1	Language Model.....	3-15
3.7.2	Unigram Language Model	3-17
3.7.3	Finite automata and language models.....	3-18
3.7.4	Types of Language Model	3-18
3.8	Latent Semantic Indexing	3-19
3.8.1	Matrix Decomposition	3-19
3.8.2	Low Rank Approximation.....	3-19
3.8.3	Problems in Information Retrieval	3-20
3.8.4	Latent Semantic Indexing (LSI).....	3-20
3.8.5	Singular Value Decompositions	3-21
3.8.5(A)	Computation of SVD.....	3-22
3.9	The Cosine Measure.....	3-25
3.10	Vector Space Scoring.....	3-27



3.10.1	Queries as Vectors	3-28
3.10.2	Computing Vector Scores.....	3-28
3.11	Document Length Normalization	3-29
3.12	Relevance Feedback and Query Expansion	3-31
3.12.1	Relevance Feedback.....	3-31
3.12.2	Query Expansion	3-34
3.12.2(A)	Query Expansion	3-34
3.12.2(B)	Thesaurus-based Query Expansion	3-35
3.12.2(C)	Automatic Thesaurus Generation.....	3-35
3.13	Rocchio	3-36
3.13.1	Basics of Rocchio Algorithm.....	3-36
3.14	Efficiency Considerations	3-37
Chapter 4 : Performance Evaluation		4-1 to 4-12
4.1	Introduction.....	4-1
4.2	Search Engines and Search Scenarios.....	4-1
4.3	User Happiness	4-3
4.4	Performance Evaluation of Search Engines	4-4
4.4.1	Performance evaluation measures.....	4-5
4.4.2	Precision and Recall	4-6
4.4.3	F-measure	4-7
4.5	Issues in the Design of Search Engines.....	4-7
4.6	Creating Test Collections.....	4-10
4.6.1	Kappa Measure for Inter-Judge Agreement.....	4-11
Chapter 5 : Text Categorization and Filtering		5-1 to 5-22
5.1	Introduction to Text Classification	5-1
5.1.1	Text Classification in Information Retrieval.....	5-2
5.1.2	Statistical Text classification	5-3
5.2	Naive Based Model for Text Classification	5-3
5.3	Spam Filtering	5-7
5.4	Vector Space Classification.....	5-10
5.5	K- Nearest Neighbour (KNN) Classification.....	5-13



5.6	Support Vector Machine Classifiers	5-15
5.7	Kernel Functions	5-18
5.7.1	Different types of Kernel in Support Vector Machine.....	5-19
5.8	Boosting.....	5-19
5.8.1	How Boosting Algorithm Works?.....	5-20
5.8.2	Types of Boosting Algorithms.....	5-20

Chapter 6 : Text Clustering**6-1 to 6-20**

6.1	Clustering Versus Classification.....	6-1
6.1.1	Classification	6-1
6.1.2	Clustering.....	6-2
6.1.3	Difference between Classification and Clustering.....	6-3
6.1.4	Clustering in Information Retrieval.....	6-4
6.2	Partitioning Methods	6-4
6.3	k-Means Clustering.....	6-5
6.4	Gaussian Mixture Model for Clustering.....	6-7
6.5	Hierarchical Agglomerative Clustering.....	6-10
6.5.1	Hierarchical clustering.....	6-10
6.5.2	Types of Hierarchical Clustering	6-12
6.6	Clustering Terms Using Documents.....	6-19

Chapter 7 : Advanced Topics**7-1 to 7-22**

7.1	Summarization	7-1
7.1.1	Approaches for Automatic Summarization.....	7-2
7.1.2	Extractive Summarization	7-2
7.1.3	Abstractive Summarization	7-3
7.1.4	Why Automatic Text Summarization is useful?	7-4
7.2	Topic Detection and Tracking.....	7-4
7.2.1	Tasks of Topic Detection and Tracking.....	7-5
7.3	Personalization	7-6
7.3.1	Personalised Search.....	7-7
7.3.2	Google Personalized Search.....	7-8
7.4	Question and Answering.....	7-9



7.4.1	Question Answering System in Information Retrieval.....	7-10
7.4.2	Architecture of Question Answering System.....	7-10
7.4.2(A)	Question Processing	7-10
7.4.2(B)	Passage Retrieval	7-12
7.4.2(C)	Answer Processing.....	7-12
7.5	Cross Language Information Retrieval.....	7-13
7.5.1	Query Translation Approach.....	7-14
7.5.2	Document Translation Approach.....	7-17
7.5.3	Dual Translation (Both Query and Document Translation Approach)	7-18
7.6	Challenges in Common Language Information Retrieval.....	7-18
7.6.1	Tools for Common Language Information Retrieval.....	7-19
7.6.2	Applications of Common Language Information Retrieval.....	7-20
Chapter 8 : Web Information Retrieval		8-1 to 8-16
8.1	Architecture of Web Application.....	8-1
8.2	Hypertext and Hyperlink.....	8-4
8.3	Web Crawling.....	8-5
8.3.1	Architecture of the Web Crawler	8-6
8.3.2	Functions of a web crawler	8-7
8.3.3	Applications of a web crawler	8-8
8.3.4	Features of a web crawler	8-9
8.3.5	Challenges in Web crawling.....	8-9
8.4	Search Engines	8-9
8.5	Ranking.....	8-11
8.5.1	Static Ranking for Web Retrieval.....	8-12
8.5.2	Page Rank Algorithm.....	8-12
8.5.3	Dynamic Ranking.....	8-13
8.5.4	Factors affecting the ranking	8-14
8.6	Link Analysis	8-14
8.6.1	The Web Graph	8-15
8.7	Page Rank	8-16
8.8	HITS	8-16



Chapter 9 : Retrieving Structured Documents		9-1 to 9-16
9.1	XML Retrieval	9-1
9.2	Introduction to XML	9-2
9.2.1	Features of XML.....	9-2
9.2.2	XML Tree Structure.....	9-4
9.2.3	XML DTD	9-5
9.2.4	XML Schema.....	9-6
9.2.5	XML CSS	9-6
9.3	Challenges in XML Retrieval.....	9-8
9.4	Vector Space Model for XML Retrieval.....	9-11
9.5	Evaluation of XML Retrieval.....	9-13
9.6	Semantic Web	9-13
9.6.1	Architecture of Semantic Web.....	9-15

