GUJARAT TECHNOLOGICAL UNIVERSITY

INFORMATION TECHNOLOGY
DATA COMPRESSION AND DATA RETRIVAL
SUBJECT CODE: 2161603
B.E. 6thSEMESTER

Type of course: Core

Prerequisite: None

Rationale: Data compression refers to the process of encoding information such that memory/transmission capacity requirements are minimized. Though there is an exponential growth in memory and transmission capacity, many high-bandwidth applications, such as digital storage and transmission of video, would not work without compression.

Teaching and Examination Scheme:

Teaching Scheme			Credits	Examination Marks						Total Marks
L	T	P	C	Theory Marks			Practical Marks			
				ESE (E)	PA (M)		ESE (V)		PA (I)	
					PA	ALA	ESE	OEP		
3	0	2	5	70	20	10	20	10	20	150

Content:

Sr. No.	Content	Total Hrs	% Weightage
1	Compression Techniques :Lossless Compression , Lossy Compression ,Measures of Performance	**2**	5
2	Mathematical Preliminaries for Lossless Compression Models : Physical Models Probability Models Markov Models Composite Source Model Coding Uniquely Decodable Codes Prefix Codes Algorithmic Information Theory Minimum Description Length Principle	**4**	10
3	Huffman Coding The Huffman Coding Algorithm 41 Minimum Variance Huffman Codes Adaptive Huffman Coding Update Procedure Encoding Procedure Decoding Procedure Golomb Codes Rice Codes	**6**	15

	Tunstall Codes Applications of Huffman Coding Lossless Image Compression Text Compression Audio Compression		
4	Arithmetic Coding Introduction Coding a Sequence Generating a Tag Deciphering the Tag Generating a Binary Code Uniqueness and Efficiency of the Arithmetic Code Algorithm Implementation Integer Implementation Comparison of Huffman and Arithmetic Coding Adaptive Arithmetic Coding	**5**	10
5	Dictionary Techniques Static Dictionary Digram Coding Adaptive Dictionary The LZ77 Approach The LZ78 Approach Applications File Compression—UNIX compress Image Compression—The Graphics Interchange Format (GIF) Image Compression—Portable Network Graphics (PNG) Compression over Modems—V.42 bis	**6**	15
6	Predictive Coding: Prediction with Partial match (ppm): The basic algorithm, The ESCAPE SYMBOL, Length of context, The Exclusion Principle, The Burrows-Wheeler Transform: Move-to-front coding Lossless Image Compression CALIC, JPEG-LS, Multi-resolution Approaches Facsimile Encoding Dynamic Markoy Compression.	**6**	10
7	Mathematical Preliminaries for Lossy Coding Distortion criteria, Models, The Quantization Problem Uniform Quantizer Adaptive Quantization Forward Adaptive Quantization Backward Adaptive Quantization Nonuniform Quantization pdf-Optimized Quantization Companded Quantization	**06**	10
8	Vector Quantization Advantages of Vector Quantization over Scalar Quantization The Linde-Buzo-Gray Algorithm	**07**	10

	Initializing the LBG Algorithm The Empty Cell Problem Use of LBG for Image Compression Tree-Structured Vector Quantizers Design of Tree-Structured Vector Quantizers Pruned Tree-Structured Vector Quantizers Structured Vector Quantizers Pyramid Vector Quantization Polar and Spherical Vector Quantizers Lattice Vector Quantizers		
9	Boolean retrieval An example information retrieval problem A first take at building an inverted index Processing Boolean queries The extended Boolean model versus ranked retrieval The term vocabulary and postings lists Document delineation and character sequence decoding Obtaining the character sequence in a document Choosing a document unit Determining the vocabulary of terms Tokenization Dropping common terms: stop words Normalization (equivalence classing of terms) Stemming and lemmatization Faster postings list intersection via skip pointers Positional postings and phrase queries Biword indexes Positional indexes	**04**	10
10	XML retrieval Basic XML concepts Challenges in XML retrieval A vector space model for XML retrieval Evaluation of XML retrieval Text-centric vs. data-centric XML retrieval	**02**	5

Suggested Specification table with Marks (Theory):

Distribution of Theory Marks					
R Level	U Level	A Level	N Level	E Level	C Level
15	**35**	**15**	**5**	**00**	**00**

Legends: R: Remembrance; U: Understanding; A: Application, N: Analyze and E: Evaluate C: Create and above Levels (Revised Bloom's Taxonomy)

Note: This specification table shall be treated as a general guideline for students and teachers. The actual distribution of marks in the question paper may vary slightly from above table.

Reference Books:

1. Introduction to Data Compression, Khalid Sayood, Morgan Kaufmann
2. Publishers
3. The Data Compression book, Mark Nelson,Jean Loup Gaily
4. Data Compression : The Complete Reference", David Saloman, Springer
5. An Introduction to Information Retrieval, Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, Cambridge University Press,Cambridge, England
6. Information storage and retrieval , Robert Korfhage, WILEY

Course Outcome:

After learning the course the students should be able to:

1. Understand and apply various coding techniques for compression.
2. Differentiate between Lossy and Lossless compression.
3. Understand basic concept of information retrieval

List of Experiments:

1. Write a program that compresses and displays uncompressed windows BMP image file.
2. Write a program to generate binary code in case of arithmetic coding.
3. Implement Huffman Code(HC) to generate binary code when symbol and probabilities are given.
4. Implement Huffman code which can compress given file and decompress compressed file.
5. Implement adaptive Huffman program to compress decompressed file.
6. Write a program to Implement LZ77 algorithm.
7. Write a program to Implement LZ55 algorithm.
8. Write a program to Implement LZ78 algorithm
9. Write a program which performs JPEG compression, process step by step for given 8x8 block and decompression also.
10. Write a program to find tokens from the files and eliminate stop words.
11. Write a program to implement vector space model for XML retrieval.

Design based Problems (DP)/Open Ended Problem:

1. Design an architecture and algorithm for data compression in cache and main memory.
2. Design an algorithm for compressing photo or video that is shared across social media.
3. Design an algorithm for compressing data at sensor which is reporting temperature data.

Major Equipment:

Computer ,Laptop

List of Open Source Software/learning website:

1) http://ocw.usu.edu/Electrical_and_Computer_Engineering/Information_Theory/

ACTIVE LEARNING ASSIGNMENTS: Preparation of power-point slides, which include videos, animations, pictures, graphics for better understanding theory and practical work – The faculty will allocate chapters/ parts of chapters to groups of students so that the entire syllabus to be covered. The power-point slides should be put up on the web-site of the College/ Institute, along with the names of the students of the group, the name of the faculty, Department and College on the first slide. The best three works should submit to GTU.