Assignment No. 04

Unit IV

Distributed and Multimedia IR.

Q1. What is distributed IR? Explain it with the source selection.

Ans: A distributed computing system can be viewed as an MMID parallel processor with relatively slow inter-processor communication channel and the freedom to employ a heterogenous collection of processors in the system.

- Source selection is the process of determining which of the distributed document collection most likely to contain relevant documents the current query and therefore should the query for processing.
- Simple processing: Assume that every equally likely to contain the relevant always broadcast the query to all collections. Appropriate used when documents are randomly partitioned or there is significant semantic overlap between the collection.
- The collection can also be ranked to their likelihood of containing relevant documents.
  a. If documents are partitioned into semantically meaningful collection
  b. It is prohibitely expensive to search collection every time.
- Basic Technique:
  a. Treat each collection as if it were assigned large document.
  b. Generate a collection vector for each collection.
  c. Evaluate the query vector against each collection vector to produce a ranked listing of collections.

centralized system

- In a system comprising impedentally heterogenous search servers, the distributed document collection will be built and maintained independentally.
- Distributed document collections in a system with independentally running, heterogenous engines will be made and kept separately There is no centralized authority over the document splitting procedure.

Q.3. Explain in detail the working of MULTOS data model.

Ans:- Multimedia office server (MULTOS) is a multimedia document server with advanced document retrieved capabilities developed in the context of an ESPRINT project in the area of office systems.

- It is based on the client/server architecture
Three different types of document servers

are required
1. Current servers
2. Dynamic servers
3. Archieve servers.

- The MULTOS data model allows:

a. The representation of high level concepts present in the database.

b. The grouping of documents into classes of documents having similar content and structure.

c. The expression of conditions on free text.

- The MULTOS representation of documents and operations on them are based on a format model. At the beginning, a standardized document presentation was assumed i.e. the ODA model.

Q.4. Explain distributed IR Query processing.

Ans. Query processing in a distributed information retrieval system is proceeds as follows:

1. Select collection to search
2. Distribute query to selected selections.
3. Evaluate query at distributed collection in parallel.
4. Combine results from distributed collections into final results.

- As described in the previous section, step 1 may be eliminated if the query is always broadcast to every document collection in the system. Otherwise, one of the previously described selection algorithms is used and the query is distributed to the selected collections.
- Each of the participating search servers then evaluates the query on the selected

algorithm finally the results are merged

- There are number of scenarios as follows:

If the query is boolean and the search server return Boolean result, sets all the sets are simply unioned to create the final resulted.

Q.7. Explain GEMINI approach for multimedia IR

Ans. - GEMINI approach is based on the following two ideas. It is a 'quick-and dirty' test to discard the vast majority of non-qualifying objects.

- Design fast search algorithm that locate objects that match a query object, exactly or approximately. An obvious solution is to apply sequential scanning: for each and every object, we can compute its distance from Q and report the objects with distance $D(Q_1, Q_1) <= E$.

- However, sequential scanning may be slow for two reasons:

1. The distance computation may be expensive.
2. The database size N might be very large.

- GEMINI aims to provide a faster alternative, and is based on two ideas:

1. A quick-and-dirty test, to discard quickly the vast majority of non-qualifying objects.
2. The use of spacial access methods to achieve faster than sequential searching.

Q.9. Explain source selection.

Ans.- Source selection is the process of determining which the distributed collections are most likely to contain

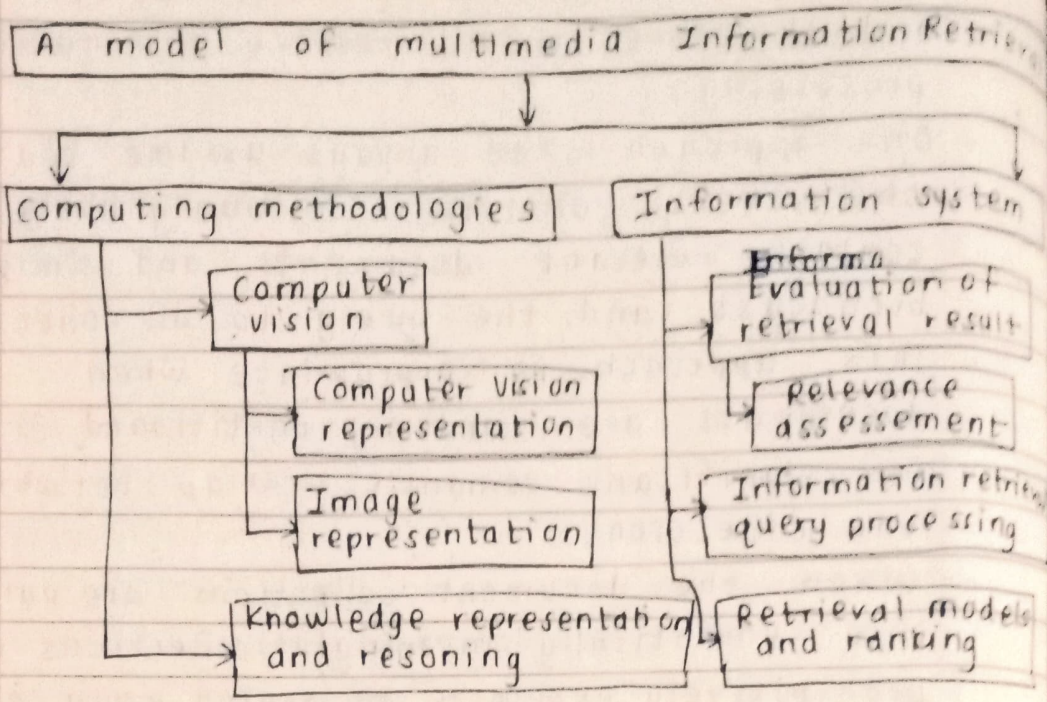
relevant documents for the current query and therefore should receive the query for processing.

- One approach is to always assume that the every collection is equally likely to contain relevant documents and simply broadcast and the query to all collection.
- This approach is appropriate when documents are randomly partitioned or there is significant semantic overlap between the collection.
- When the document collections are partitioned into semantically meaningful collections or it is prohibitively expensive to search every collection every time the collections can be ranked according to their likelihood of continuing the relevant document.
- The basic technique is to treat each collections as if it were a single large documents, index the collections, and query against the collection to produce a ranked listing of collection, we can apply a standard cosine similarity.

Q.14. Explain model of multimedia Information retrieval

Ans: The model provides retrieval functionality across two media and images as well as multiple dimensions, including from content and organization.

- For many years, IR has been an active research area. In the beginning the field concentrated on processing user information needs as expressed in a list of keywords in an efficient and effective manner..



Q 20. What Query languages is with respect to distributed IR. Explain it in detail.

Ans.- The query mechanisms may include free text search. SQL-like querying, icon-based querying based on the entity-relationship (ER) diagram, content-based querying, sound-based querying as well as virtual reality (VR) techniques.

- Before, surveying current development in language for heterogenous multimedia databases and data sources, we will first examine query.
- we will use data fusion examples to exemplify this method as we explain a patent spatial/temporal query language known as summation QL.
- It is projected that many websites on the world wide web (www) will develop into rich sources of spatial/temporal multimedia databases.

- A significant amount of soft real-time, hard, real-time and real-time sources of information need to be processed, checked for consistency, organized and sent to the different agencies.

Q.24. Describe multimedia data support in commercial DBMS.

Ans.- To represent multimedia data, current DBMS support variable length data types. Data type supported by commercial DBMS is standard and DBMS vendor

Federated indexing. Each source maintains its own index, broker merges results.

3. Security and Privacy
   - Security query rating with authentication and encryption.
   - Access control policies enforced at the resource level.

Q. 34. Explain the generic multimedia indexing approach.

Ans. - Indexing approaches for multimedia data retrieval is the process of preparing a database of multimedia objects to allow, for quick access and comparisons based on their derived properties.

- The high-dimensional float-valued feature vectors that make up multimedia content by their very nature make the similarity criteria that are employed to compute multimedia, objects frequently quite complex.
- The main objective of multimedia indexing is to efficiently support multimedia similarly search which is the basis of the majority of multimedia application.
- For whole match queries the problem is defined as follows:

i. We have collection of N objects O1, O2, O3.

ii. The distance or dissimilarity between two objects (Oi, Oj) is given by the function D(Oi, Oj). The function is implemented as program.

- The user has to specify the query object Q and the tolerance ϵ.
- Here, we have to find the objects in the collection of N objects O1, O2, ... ON that are within distance ϵ from the query objects, one solution to find such objects is to apply the sequential scanning.

Q.39 Explain in detail automatic feature extraction in Distributed IR.

Ans.- GEMINI approach is useful for setting that

Page No. 1
Date

Information storage and Retrieval

Unit III Distributed and Multimedia IR

Nov-Dec 2022

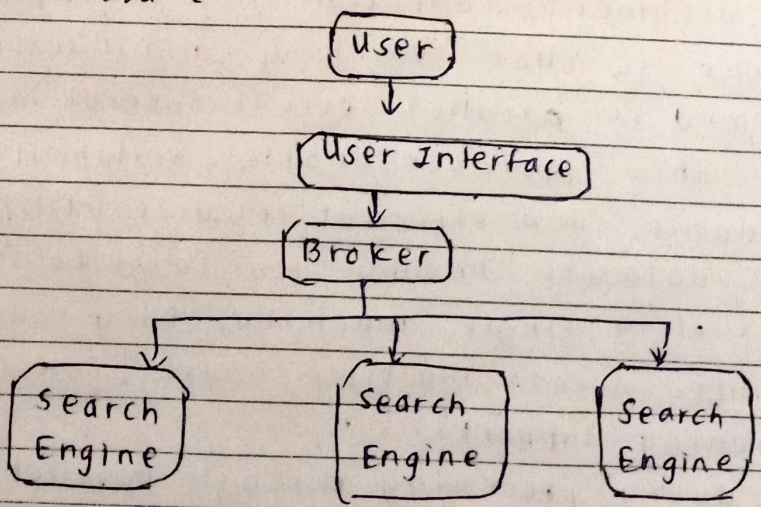
Q.3 a) What is distributed IR? Explain the architecture of distributed IR

Ans. Distributed Information Retrieval (DIR) is an extension of traditional information retrieval where the data to be indexed and searched is distributed across multiple servers, databases, or search engines rather than stored in a single centralized system. It is used when:

- The dataset is too large to store or process in one place.
- The information belongs to different organizations or sources.
- Faster response time and scalability are required

e.g. web search engines, federated search engines, cloud-based search services, and enterprise multi-database search.

- Architecture



- Components

1. User

The end-user who submits a query

2. User Interface (client / Frontend)

Used for inputting queries and displaying results

3 Broker (Query Coordinator / Dispatcher)

- Receives the query from the user interface
- Determines which distributed search engines or databases are relevant
- Forwards the query to selected sources
- Merges results returned from different search engines
- Ranks and sends final results to the user

4 Local Search Engines

Each node

- Stores and indexes its own document collection
- Executes query locally

5. Document Collections

Distributed repositories containing the actual documents or indexed representations

Q 3 b) What is Collection Partitioning with respect to distributed IR. Explain in detail.

Ans Collection Partitioning in Distributed Information Retrieval (IR) refers to the method of dividing a large document collection into multiple smaller partitions so that searching and indexing can be performed in parallel across several machines or servers. This approach enables scalability, improved performance, and efficient resource utilization.

When datasets become too large to store or process on a single machine, they must be distributed across multiple nodes. Collection partitioning supports:

- Faster query processing through parallel searching
- Scalability as data grows
- Fault tolerance (if one node fails, others continue)
- Efficient storage and load balancing

* Types

1 Document Partitioning (Horizontal Partitioning)
- Documents are split across multiple servers.
- Each server indexes only a subset of documents
- A query is sent to all partitions, and results are merged.

e.g Node A → Docs 1 - 1M
Node B → Docs 1M - 2M
Node C → Docs 2M - 3M

* Advantages
- Simple to implement
- Easy to scale by adding more nodes

* Disadvantages
- Requires merging results from all nodes
- Query latency increases if many partitions must be searched

2. Team Partitioning (Vertical Partitioning)
- Index is split by terms, not documents
- Each node stores postings for certain keywords

Node A → A-H
Node B → I-P
Node C → Q-Z

* Advantages
- Only some nodes need to be queried based on query items.
- Reduces communication and processing

* Disadvantages
- High overhead in combining partial results
- Query needs access to all term partitions

3. Hybrid Partitioning
- Combines term and document partitioning
- Each node manages both a subset of documents and a subset of terms
- Goal: Achieve better load balancing and lower latency

* Working

Step 1 User submits query

Step 2 Distributed broker determines which partitions to query

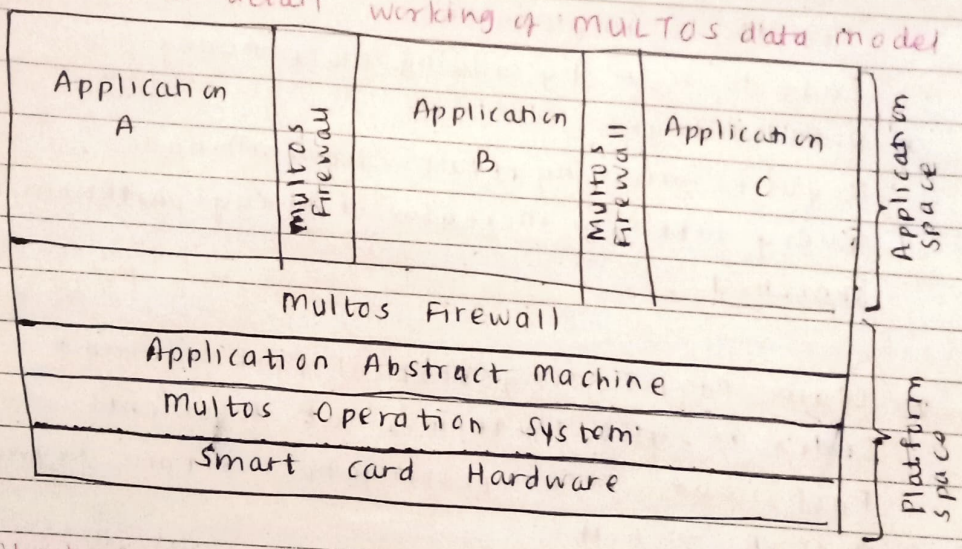
Step 3 Queries are executed in parallel across partitions

Step 4 Results are ranked and merged

Step 5 Final ranked list returned to user

Q.4a) Explain in detail working of MULTOS data model

Ans



Q.4a) What is multimedia IR? Explain the architecture of multimedia IR in detail

Ans. Multimedia Information Retrieval (MIR) refers to the process of searching, indexing, and retrieving multimedia content, such as images, audio, text, video, graphics and animations

Unlike traditional text-based IR, multimedia IR must extract and interpret raw content features, because multimedia objects often lack sufficient textual metadata.

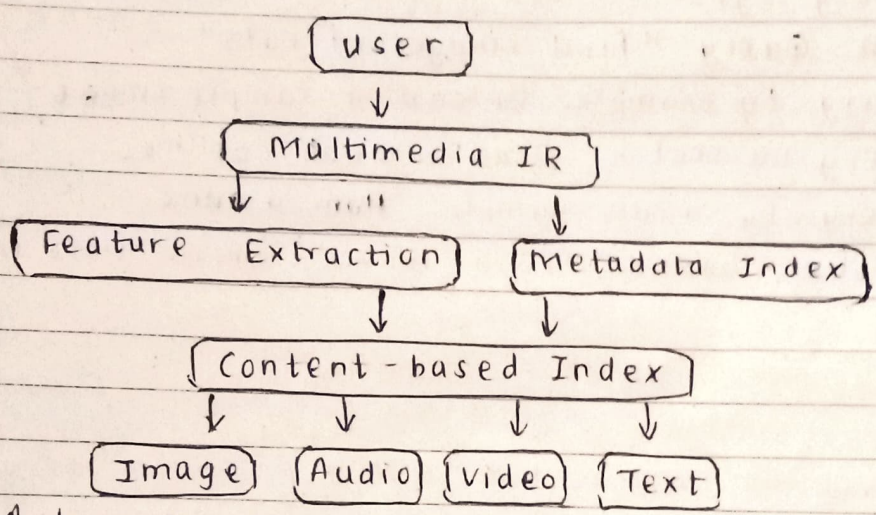
* Goals

- Retrieve relevant multimedia objects based on user queries
- Enable searching by keywords, examples, or similarity

- Support multimodal queries combining text + content based features

e-g. Searching images by providing an example image (CBIR), ~~search~~ searching videos by keywords or visual scenes, searching audio by humming or sample audio clips

★ Architecture



A typical Multimedia IR system includes following components:

1. User Interface
   Allows user to submit queries (text, query, sample media, semantic query) and view results
2. Query Processor
   Interprets the query and decides how to process and retrieve results
3. Feature Extraction
   Extract content-based features
   - Image: Color, texture, shape
   - Audio: Pitch, frequency, MFCC
   - Video: Motion patterns, scenes
   - Text: Keywords, semantics
4. Indexing Module
   Creates Metadata Index (tags, annotations) and Content-based index (visual/audio features)
5. Database / Multimedia Repository
   Stores multimedia objects and their feature vectors

6. Matching and Ranking Engine
Compares extracted features with stored features and ranks results by similarity
7. Result Presentation
Displays results visually or interactively

* Query Types
- Text Query: "Find images of cats"
- Query by example: Upload a sample image
- Query by sketch: Draw a rough outline
- Query by audio sample: Hum a tune
- Concept-based: "Find movies about space travel"

May-Jun 2024

Q.3a) What is Distributed IR? Explain in detail automatic feature extraction in Distributed IR

Ans In Distributed Information Retrieval (DIR), data is stored across multiple servers or locations. When dealing with multimedia data (images, audio, video, etc.), the system must extract meaningful information automatically to support searching and indexing. This process is called Automatic Feature Extraction.

Automatic feature extraction is the process of computing important characteristics or patterns directly from raw multimedia objects without requiring human annotation. These extracted features are then used to build indexes that enable efficient retrieval and similarity matching.

* Importance in DIR
- Manual tagging of multimedia is slow, expensive, and subjective
- Distributed systems handle extremely large datasets where automation is essential.

- Enables content-based retrieval (e.g searching an image database using an example image)

* Types of features extracted

- Images: Color histograms, texture patterns, shapes, edges, deep CNN features
- Audio: Pitch, spectral energy, MFCC coefficients, tempo, silence segments
- Video: Motion vectors, scene transitions, keyframes, object tracking
- Text: keywords, TF-IDF weights, semantic embeddings

* Working

1. Data acquisition from distributed sources
2. Preprocessing (noise removal, normalization)
3. Feature extraction on each node using algorithms or machine learning
4. Local indexing based on extracted features
5. Global index merging at central broker
6. Parallel similarity matching during queries

* Advantages

- Faster and scalable multimedia search
- Less dependency on manual metadata
- Enables similarity-based search rather than exact match
- Supports machine-learning based retrieval

Q 3 b) Describe Multimedia data support in Commercial DBMS

Ans Modern commercial Database Management systems (DBMS) provide built-in support for storing, indexing, and retrieving multimedia data such as images, audio, video, and documents. This support is essential for enterprise-level multimedia applications

* MM support Features in DBMS

- BLOB Data Types: Multimedia stored as Binary Large Objects

Traditional text indexing

- Specialized Data Types E.g. Oracle ORDImage, SQL Server VARBINARY(MAX)
- Content based indexing: Extracts features and indexes multimedia attributes
- Full-text search: Indexing text inside documents (PDF, Word, etc)
- Streaming support: Audio/video playback support using buffer streaming
- Integration with external libraries ML-based recognition and imaging libraries

e.g. Oracle Multimedia (Oracle 12c): Image, audio, video storage + content-based retrieval API

- IBM DB2 Net Search Extender / Image Extender: Image Recognition, text scratch
- Microsoft SQL Server: FileStream storage, full-text search, spatial and multimedia data
- PostgreSQL with extensions: pgsphere, PostGIS, image analysis plugins
- MySQL: BLOB storage with plugin support

* Working

1. Multimedia objects stored as BLOB/CLOB
2. Metadata and features stored in separate relational tables
3. Queries may reference:
   - keywords (metadata search)
   - similarity search (feature-vector matching)
4. Results ranked and returned to the user

* Applications

✓ Digital Libraries
✓ Video streaming platforms
✓ Medical imaging databases
✓ GIS and satellite imaging
✓ Face recognition / security systems

- Background - Spacial Access Methods

Traditional text indexing uses inverted files, but multimedia objects require storing feature vectors in multi-dimensional space (e.g., an image might have a 20-dimensional color-texture vector) spacial access methods enable efficient search in high-dimensional feature spaces

- Common spacial Access Methods
- R-Tree: Groups multidimensional bounding rectangles, used for image regions and spacial queries
- R-Tree*: Improved R-Tree with better split/merge heuristics for performance
- KD-Tree: Binary space partition for nearest-neighbor search
- Quad-Tree / Oct-Tree: Recursively partitions 2D / 3D space
- VP-Tree / M-Tree: For similarity search using metric distances

Spacial access structures help in:

- Similarity search (find images / videos closest to query)
- Range queries (retrieve all images with texture value range)
- Reducing search time in high-dimensional spaces

A Generic Multimedia Indexing Approach

A general multimedia indexing system follows these steps:

Multimedia Object → Feature Extraction → Feature Vector → Indexing → Similarity Matching

- Feature Extraction: Extract numerical descriptions (color histogram, shape vectors, MFCC, motion vectors)
- Feature vector Representation: Converts features into multidimensional vectors
- Indexing: Uses spatial / metric index structures (R-Tree, KD-Tree, hashing)
- Distance Metrics Euclidean, cosine similarity, Manhattan distance

- Similarity Retrieval: Rank objects based on closeness in feature space

This approach enables Content-Based Multimedia Retrieval (CBMR/CBIR) e.g., search by example image.

- One-Dimensional Time series

- One-dimensional time series represent data varying with time (audio signal, ECG, stock prices)

Features used

- Pitch / Frequency
- MFCC (Mel-Frequency Cepstral Coefficients)
- Amplitude and energy distribution
- Zero-crossing rate

* Indexing Methods

- Dynamic Time Warping (DTW): Aligns time sequences with different speeds
- Fourier / Wavelet Transform: Reduces dimensionality for indexing
- VP-Trees / M-Trees: Similarity search using metric distances

* Applications

- Music recognition (Shazam)
- Stock market pattern search
- Medical signals (ECG, EEG)

- Two-Dimensional Color Images

* Extracted Features

- Color: Color histogram, color moments
- Texture: Co-occurrence matrix, Gabor filters, wavelets
- Shape: Contours, edge detection, boundary descriptors
- Deep Learning Features: CNN embeddings (ResNet, VGG, etc.)

* Similarity Search
  - Euclidean distance between color vectors
  - Histogram intersection
    Structural similarity metric (SSIM)
  - CNN-based feature distance
* Applications
  - Google image search by example
  - Face recognition
  - Image-based authentication

- Trends and Issues in Multimedia IR
  - Semantic Gap: Difference between low-level features and human interpretation (cat recognition vs pixel values)
  - Deep Learning: CNNs, transformers and attention models for multimodal feature learning
  - Cross-modal Retrieval: Search video by text query; search image by audio input
  - Large-scale indexing: Handling huge datasets using hashing, ANN (Approximate Nearest-Neighbor) search
  - Multimodal Fusion: Combining text + image + motion + audio
  - Relevance Feedback: System learns from user input to refine results
  - Privacy and Copyright: Rights management of multimedia content
    Real-time retrieval: Streaming-video search and indexing