Unit 4.
Distributed and Multimedia IR.

- Distributed IR:

- Distributed computing is the application of multiple computer connected by a network to solve a single problem
- Distributed computing system can be viewed as a MIMD parallel processor with a relatively slow inter-processor communication channel and freedom to employ a heterogeneous collection of processor in the system.
- Distributed IR is the practice of distributing the processes and tasks involved in IR across multiple computers or nodes in network.
- This approach often used to improve efficiency and scalability of large-scale IR system by dividing the workload and processing tasks in parallel.
- Distributed IR commonly applied in search engines where indexing, query processing and result ranking can be distributed to optimize performance.

- Computational model is very similar to the MIMD parallel processing model but difference here is that.

1. Sub-tasks runs on different computers and the communication between the subtasks is performed using network protocol such as TCP/IP rather than the shared memory based inter-process communica-

tion mechanism.

2. In a distributed system it is more common to employ a procedure for selecting a subset of the distributed servers for processing a particular request rather than broadcasting every request to every server in the system.

* Architecture of Distributed IR

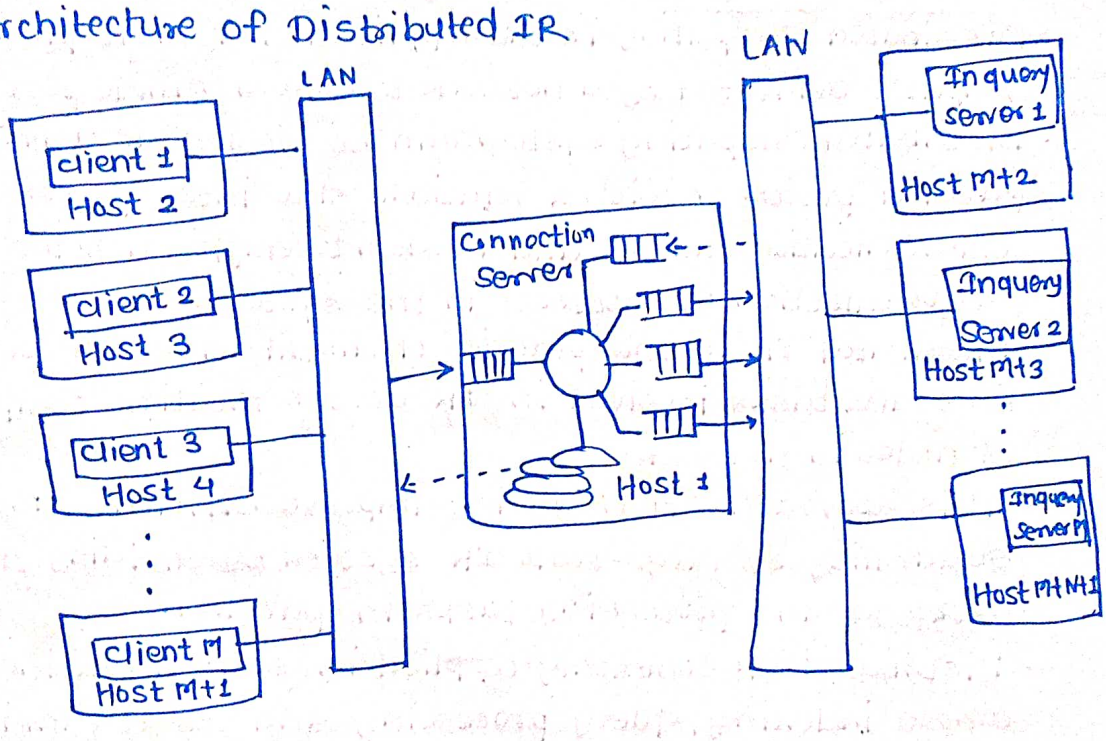


Fig. Architecture of Distributed IR.

- At a minimum, the protocol should allow a client to:
  - Obtain information about search server, e.g. a list of database available for searching at the server and possibly statistics associated with the databases.
  - Submit a search request for one or more databases using a well defined query language.

- Receive search results in a well defined format.
- Retrieve items identified in the search results.

- Example of DIR: (Google search Engine)
- Federated Search

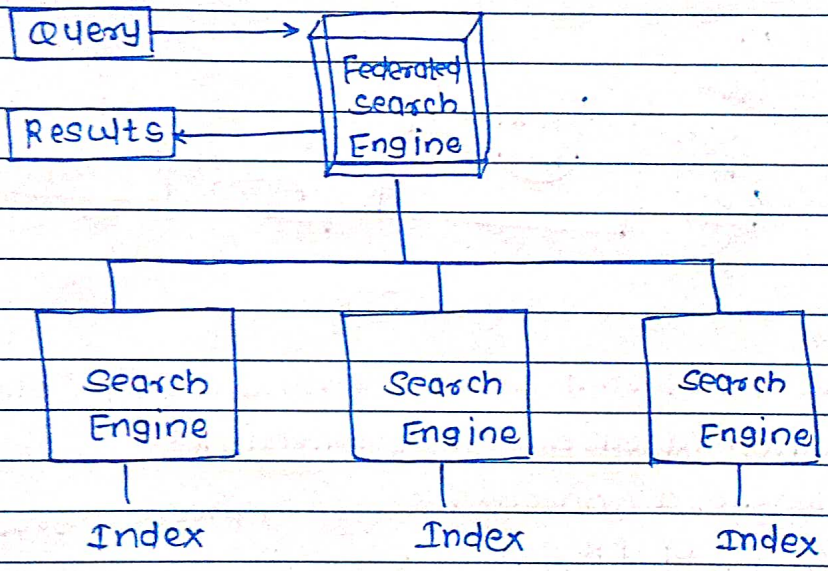


Fig. Example of DIR
(Google search Engine)

- Federated search is another name of DIR.
- It has its own collection of data.
- Instead of Crawling a site, Federated search systems send a user query to a resource's search tools.
- User can search multiple data sources simultaneously using Federated search by sending a single query.
- The Federator gathers result from one or more search

engines and then present all the results in the single user terminal.

- Architecture of Distributed IR

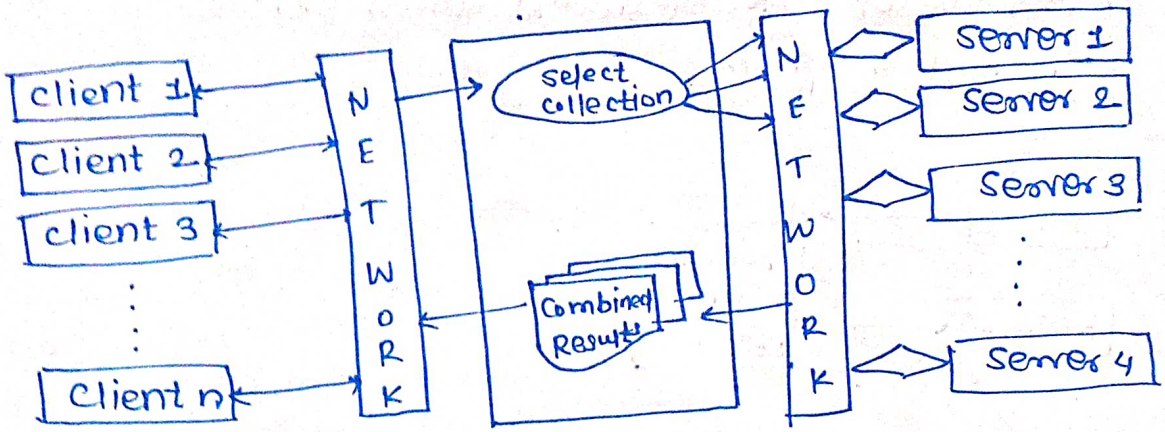


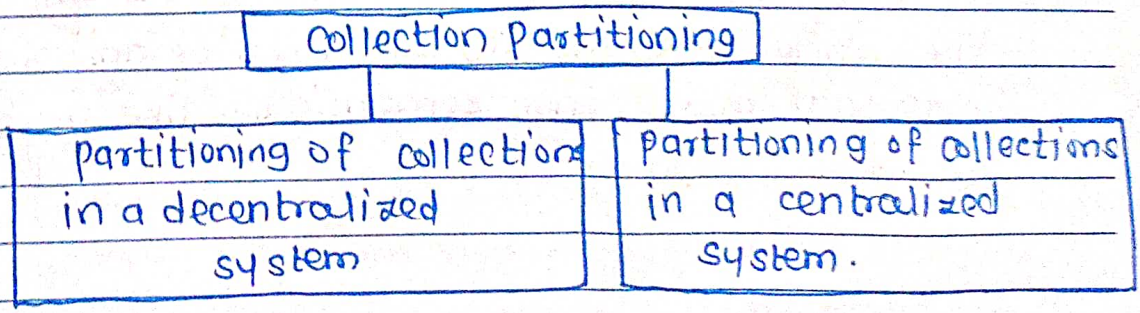
Fig. Architecture of DIR.

- DIR conceptual model that enables several users (clients) to continously search various document collections.
- In DIR architecture, a connection server connects a group of client (users) to a group of IR system.
- Distributed system typically consist of a set of server processes, each running on a seperate processor node, and a designated broker process for accepting client requests, distributing the requests to the servers, collecting intermediate results from the servers, and combining the intermediate into a find result for the client.
- In DIR system, prototype that enables several users to simultaneously search various documents search.

- All communications betn clients and (servers) IR systems are handled by connection server.
- Additionally, if is simply for us to modify the architecture and assess its success. Our findings demonstrate that our architecture can support high number of clients.

• Collection Partitioning:

* What exactly a collection?
  - DIR, when there are multiple collections, searching.
  - Local contexts, for instance, partitioning a sizable collection.
  - Wide area settings, such as company network or internet.
  - Divide huge collections among processors to speed up processing - due to governmental or regulatory requirements.
  - Heterogenous environments, many IR systems, networks with hundreds or thousands of collections and collections that are indexed on the web.
  - Economic expenses of conducting a full site search.
  - Economic costs of conducting a full network search.



1. Partitioning of collections of in a Decentralized System:
   - Distributed document collections in a system with independently running, heterogeneous search Engines will be made and kept seperately.
   - There is no centralized authority over the document splitting procedure.
   - It's feasible that each search engine has a specific area of expertise.

2. Partitioning Collection in a centralized system:
   - When the collection is small enough to fit on a single search server but high availability and query processing throughput are required, collection partitioning in a centralized system is appropriate
   - The goal of the broker is to direct inqueries to the search servers and balance the loads on the servers, which is how the parallelism in the system is being utilized through multitasking.
   - Every query is broadcast to every search server by the broker, who then complies the results for the user.
   - Explicit semantic partitioning of the documents, which may be done either manually or automatically.
   - The articles may already have been organized into collections that make sense semantically, like groupings by technical discipline.

- Source Selection:

- The process of choosing which of the distributed document collections is most likely to contains document relevant to the present query and as a result should receive the query for processing is known as Source selection.
- One technique is to simply broadcast the query to all collections, assuming that every collection has an equal likelihood of having relevant items.
- This approach is suitable when are randomly sorted into groups or when there is a significant amount of semantic overlap.
- Source selection is the process of deciding which scattered document collection is most likely to contain the current query's pertinent documents (and therefore should receive the query for processing).
- It is suitable to employ the straightforward method of assuming that every collection has an equal possibility of containing relevant documents and broadcasting the query to all of them when documents are randomly partitioned or when there is sufficient semantic overlap between the collections.
- The collections can also be ranked based on how likely they are to succeed.
- This is feasible if texts are grouped into groupings that have semantic importance, otherwise, searching through each collections indivually would be excessively expensive.

- The basic strategy is to generate a collection vector for each collection and compare it to the query vector treating each collection as if it were a single, enormous document.

- Query Processing:

A DIR system's query processing

1. Choose the collections to search.
2. Send the query to the chosen collections.
3. Parallel evaluation of queries at dispersed collections.
4. Combine results from the distributed collections into final product.

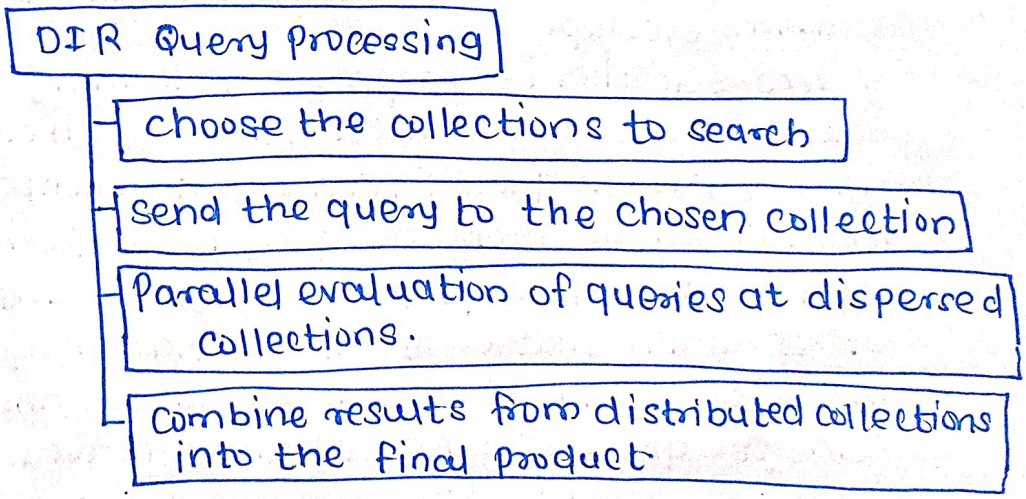


Fig. DIR Query Processing.

- Merging the Results:

There are various situations for merging the results:

- The final result set is the union of the result sets if the query is boolean and the search servers return Boolean result sets.

- Several strategies are available, ranging from straightforward to sophisticated and accurate, if the inquiry incorporates free-text ranking.
- The earliest method is to use round robin interleaving to merge the ranking result lists.
  1. 1st document from the first list
  2. 1st document from the second list.

  N: 1st document from Nth list.
  N+1: 2nd document from the first list.

- Because hits from irrelevant collections are assigned the same status as hits from highly relevant collections, this method is likely to produce low-quality results.
- It is better to combine the result lists depending on relevance score.
- Re ranking is necessary if the document collections are semantically divided or are managed by seperate parties.
- We risk getting the wrong findings if adequate global term statistics are not employed to calculate the document ratings.

You can calculate a collection's weightusing formula:

$$W = 1 + |C| \cdot (S - S^{\wedge}) / S^{\wedge}$$

where, $|C|$ - no. of collections you searched.
S - Collection score , $S^{\wedge}$ - mean of those scores.

- Issues in DIR

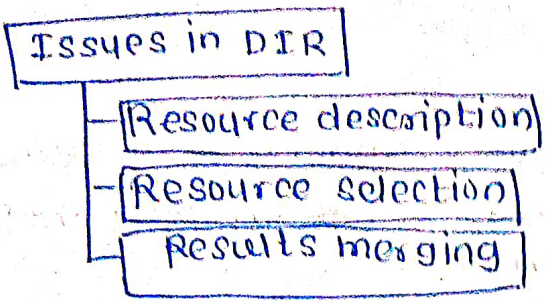


Fig. Issues in DIR

- The scattered location and control of information in a wide area computer network are reflected in that multi-database model of IR.
- However, it is also more complicated than the IR model using single database necessitating the solution to a number of additional issues:

1. Resource description:
   - Each text databases contents must be explained.

2. Resource selection:
   - choosing which database (s) to search require consideration of an information demand and list of resource descriptions.

3. Merging Results:
   - combining ranked lists that each database returned to create a single cohesive ranked list.

- During a 5 year period, addresses many of problems that aries in DIR system.
  1. Difficulty of briefly explaining the contents of each database or resource that is available.
  2. Proposes a scoring system for databases based on how well they will likely meet information needs.
  3. Issue of combining the results produced by various search systems is covered. Examines the process by which a DIR system in a multi-party setting. obtains resources descriptions.

Multimedia IR

Introduction

Multimedia Information Retrieval (MIR) involves the retrieval of various forms of multimedia content such as images, audio, video, and text from large datasets. It combines techniques from information retrival, computer vision, audio processing and natural language processing to enable efficient and effective searching, organizing, and retrieving of multimedia data.

* Data Modelling

Data Modelling in multimedia information retrieval (MIR) refers to the process of creating structured representations of multimedia data in a way that allows efficient storage, organization, and retrieval of the content.

In MIR, data modelling involves :

1) Feature Extraction : Extracting meaningful features from multimedia data. For images, this could be color, texture, or shape features

2) Feature Representation
Representing these extracted features as numerical values or vectors, which can be used for indexing and retrival.

3) Indexing : Storing these feature representation in a way that allows for efficient retrieval. Indexing structures may include databases, trees, or hash tables.

4) Semantic Information:
Incorporating semantic information if needed. This could involve associating metadata, tags, or annotations with the multimedia data.

5) Query processing : Developing algorithms & methods for comparing query features with the stored representation to retrive relevant multimedia content.

* Query Language

- In Multimedia Information Retrieval (MIR), a query language is a formal way for users to express their information needs when searching for multimedia content within a database

- A query language in MIR typically provides ways to express:

1) Content-based queries : users can describe the visual or auditory characteristics they are seeking in the multimedia content. For instance, they might specify

colors, shapes, or audio patterns.

2) Semantic Queries : Users can express their information needs using terms or concepts related to the content they're looking for.

3) Spatial and Temporal Constraints : In the case of videos or images, users might specify spatial regions of interest or time intervals for their search.

4) Combination of Criteria : Users can often combine different criteria to refine their search, such as specifying both visual & textual requirements.

* Background - Spatial Access Method

- The Background - Spatial Access Method (BSAM) is an indexing technique used in Multimedia Information Retrieval (MIR) to effectively retrieve multimedia content, particularly images, based on their spatial characteristics within the image.

- BSAM is designed to work with content-based image retrieval systems, where users can specify spatial relationships between objects within an image.

- In BSAM, images are divided into multiple regions, and each region is described using relevant visual features. These features might include color histograms, texture descriptors, or other attributes that capture the characteristics of the image content. The spatial relationships between regions are also recorded

- The key idea behind BSAM is to create an index that allows for efficient retrieval of images based on their spatial arrangement. This is particularly useful when users want to find images where certain objects or features are located in specific position relative to each other

* Generic Multimedia Indexing Approach

- The Generic Multimedia Indexing Approach (GMIA) is a framework or methodology used in Multimedia Information Retrieval (MIR) to provide a unified way of Indexing & retrieving various types of multimedia content such as images, videos, and text.

- The goal of GMIA is to create a common indexing structure & techniques that can be applied across different types of ~~multi~~ media, making it easier to manage and search multimedia data

GMIA typically involves the following steps:
1) Feature Extraction
2) Feature Representation
3) Indexing
4) Query Processing
5) Semantic Information

- The aim of Genric Multimedia Indexing Approach is to provide a standardized approach that can handle different types of media while sharing common indexing and retrieval techniques. This helps reduce the complexity of managing multimedia data and enables more efficient search and retrieval in multimedia databases.

* One Dimensional Time Series

- A one dimensional time series refers to a sequence of data points collected over time, where each data point corresponds to a single numerical value of feature at a specific time instance.
- One - dimensional time series are commonly used to represent temporal patterns & variations in multimedia data, such as audio signals, video frames, or other time-dependent information.

For example :

- Heart Rate : A person's heart rate measured at regular intervals can be represented as a one-dimensional time series, showing variations in heart rate over time.

One-dimensional time series in MIR are used for various purposes :

1) Pattern Recognition : Detecting recurring patterns or trends within the time series data.

2) Similarity Search : Finding other time series that are similar in behaviour to a given time series.

3) Event Detection : Identifying specific events or transitions in the time series.

4) classification : Assigning a category or label to a time series based on its behaviour.

5) Anomaly Detection : Detecting unusual or unexpected behaviour in the time series.

* Two-Dimensional color Images

- Two-dimensional color Images are a type of multimedia data that represent visual content using two dimensions (width & height) & include information about color.

In a two dimensional color image :

1) Two dimensional : The image is structured with a width and a height, forming a grid of pixels. Each pixel is located at a specific position within the grid.

2) Color Information : Each pixel in the Image is associated with color information. This color information can be represented using various color models such as RGB (Red, Green, Blue), CMYK (Cyan, Magenta, Yellow, Black), or HSV (Hue, Saturation, Values).

- Two dimensional color images are commonly used in various applications within MIR

- Image Retrieval : Two-dimensional color images are often used as queries or target content for retrieval systems, where users search for images similar to a given query image

- Content Analysis : Color information is valuable for content analysis tasks, such as identifying dominant colors, detecting objects, or recognizing patterns in images.

- Visual Search : Image-based search engines user color features to help users find images that match their visual preferences.

- Object Recognition : Color features can contribute to identifying and categorizing objects within images.

TRENDS AND RESEARCH ISSUE

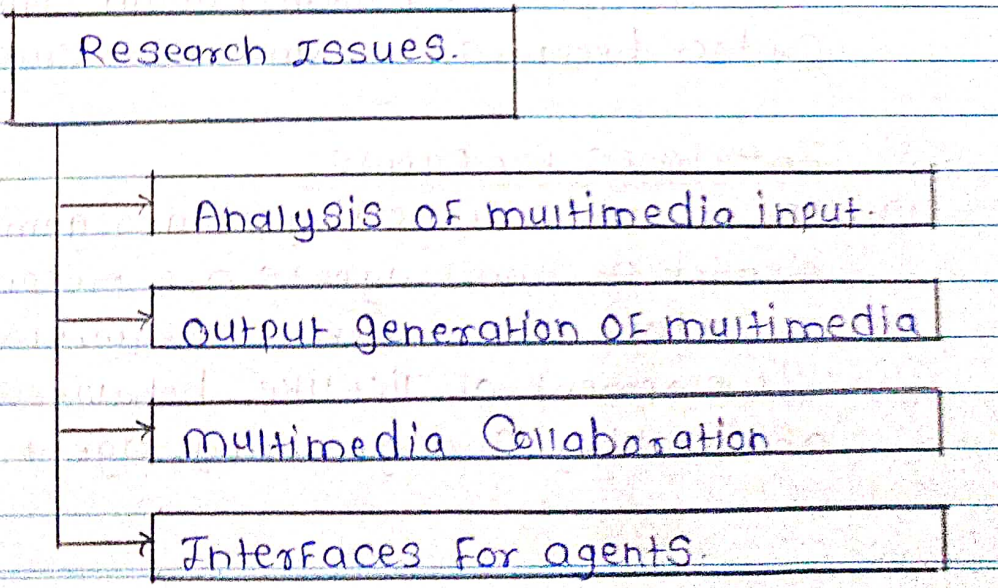
* Trends.

① One of the most difficult and rapidly expanding study fields is content, extraction, indexing, and retrieval of multimedia data.

② Specifically, we require scalable browsing algorithms that enable access to very large multimedia data bases, reliable methods for indexing / retrieving and compressing multimedia data, and semantic visual interfaces that integrate the aforementioned elements into unified multimedia browsing and retrieval systems.

③ The performance of content-based retrieval is improved for retrieval of multimedia by combining several integrated media types.

* Research Issues.



Page No.
Date

(A) Analysis of multimedia Input-

(1) In domains like inter-media segmentation, partial input parsing and interpretation, and partial multimedia, reference resolution, there are still many. Research issues to be solved.

(2) It is necessary to create and test new interactive tool (such force, and facial expression detectors) to open up fresh possibilities, like the detection and tracking of human emotional states.

(B) Output Generation for multimedia.-

(1) There are still significant unanswered problems surrounding the best ways to choose effective media, content and modality allocation (such as language, non-speech audio, or gesture to guide attention? eg. realizing language as visual text or aural speech.)

(C) multimedia Collaboration-

(1) As our world grows more Networked, finding more efficient ways of human-human Computer-mediated Contact becomes more and more crucial.

(D) Interfaces for Agents.

(1) In educational settings, video games, and customer service software, agents are present

(2) Important user interface requirements include the development of lifelike behaviors and the provision of speaking and gesturing agent displays.

Page No.
Date

Multimedia Data Support in Commercial DBMS -

① A multimedia database is a collection of related multimedia data that may include text, graphics (sketches, drawings), photos, animations, video, audio, and other elements.

② It may also contain large amounts of multisource multimedia data. the term "multimedia database management system" refers to the framework that controls various multimedia data types that can be provided, stored, and used in various ways.

Content of multimedia Database management System-

① media data- the real data used to portray an object

② media Format data- Details Concerning the Format of media data after it has been acquired processed, and encoded. Such as Sampling rate, Resolution, encoding technique, etc.

③ media keyword data- keywords that describe how data is generated. It also goes by the name of "Content descriptive data" Example : the Recording's date, time and location.

④ media Feature data- Information that is depending on the Content, such as the distribution of Color's types of texture, and Shapes.

⑤ There are still many Challenges to multimedia databases. Some of which are:-

Page No.
Date

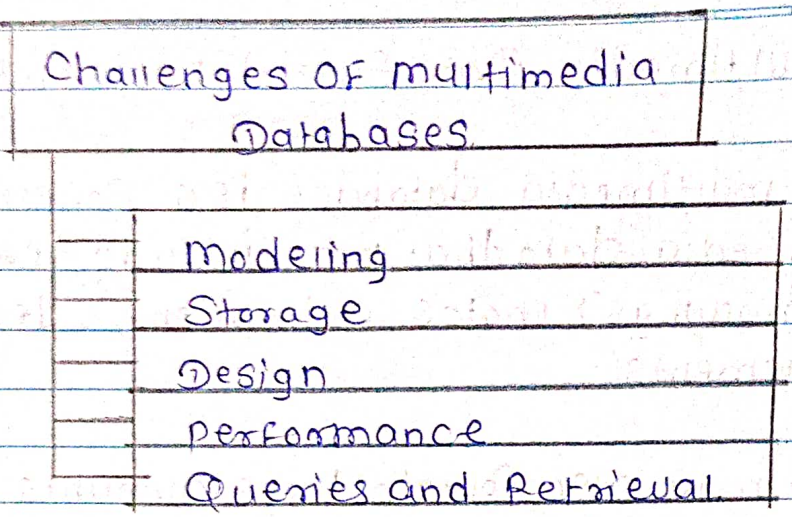


Fig- Challenges of multimedia Databases

① Modeling- By enhancing database Retrieval methods rather than information Retrieval techniques, this Field of study is specialized and warrants special attention.

② Design- Because multimedia databases use a range of Formats, Including JPEG, GIF, PNG, and MPEG, which are difficult to Convert between one another the Conceptual, Logical and physical design.

③ Storage- The Representation, Compression, mapping to device hierarchies, archiving, and buffering during input-output operations Create Challenges when Storing multimedia databases on any Conventional disc

④ performance- In a situation where Video playback or audio-video Synchronization is involved, physical Constraints take precedence.

(5) Queries and Retrieval -
For multimedia material, such as photos, video and audio, through the use of queries, data can be accessed, but this raises a number of challenges that need to be addressed, including effective query formulation, query execution and query optimization.

Imp MULTOS - (The multimedia Filing System)

(1) MULTOS (MULTimedia OFFice Server) is a multimedia document server with advanced document retrieval capabilities, developed in the context of an ESPRIT project in the area of office systems. MULTOS is based on a client/server architecture.

(2) The objective of MULTOS was to provide a pratical and affordable system for storing and retrieving multimedia content in an office setting.

(3) A conceptual model providing a semantic-oriented description of documents forms the the basis of much of the processing within the MULTOS system.

(4) The content-oriented processing, categorization, and retrieval algorithms used by MULTOS are all built on top of this semantic document model.

(5) Document modeling, document type handling, query language design and query processing optimization, image data modeling and retrieval, knowledge based document classification, document management,

The main goals were to,

① Create a System For the Filing and Retrieval OF multimedia documents that is effective and affordable integrate Optical Storage media for the filing OF numerous documents.

② Implement methods For Retrieving text and attribute data based on Content and look into methods for Retrieving picture data based on Content.

③ Implement document magnagement Features like access Control, Security, integrity, and Version Support.

④ There are two different kinds OF document Servers that Correspond to two Separate Sets OF Documents that have various Retrieval Requirements.

⑤ while the document on the archive Server are Reliable and Read far less Frequently than those on the dynamic Server, the dynamic Server allows For document updates.

⑥ magnetic Storage is used to Store documents produced or Collected in the Client environment

⑦ The Classification enables portions OF the document to be linked to Conceptual elements for use in Later document Retrieval.