4 UNIT - IV

Distributed and Multimedia IR

Syllabus

Distributed IR : Introduction, Collection Partitioning, Source Selection, Query Processing, web issues.

MULTIMEDIA IR : Introduction, Data Modeling, Query languages, Generic multimedia indexing approach, One dimensional time series, two dimensional color images, Automatic feature extraction, Trends and Research Issue.

4.1 Distributed Information Retrieval - Introduction

University Question

Q. Describe the architecture of distributed IR. **SPPU : May 17, May 19, 8/9 Marks**

- Distributed computing is the application of multiple computers connected by a network to solve a single problem.
- A distributed computing system can be viewed as a MIMD parallel processor with a relatively slow inter-processor communication channel and the freedom to employ a heterogeneous collection of processors in the system.
- In fact, a single processing node in the distributed system could be a parallel computer in its own right. Moreover, if they all support the same public interface and protocol for invoking their services, the computers in the system may be owned and operated by different parties.
- Distributed systems typically consist of a set of server processes, each running on a separate processing node, and a designated broker process responsible for accepting client requests, distributing the requests to the servers, collecting intermediate results from the servers, and combining the intermediate results into a final result for the client.
- This computation model is very similar to the MIMD parallel processing model but the difference here is that :
  1. The subtasks run on different computers and the communication between the subtasks is performed using a network protocol such as TCP/IP rather than the shared memory-based inter-process communication mechanisms.
  2. Ina 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.
- Applications that lend themselves well to a distributed implementation usually involve computation and data that can be split into coarse-grained operations with relatively little communication required between the operations. Parallel information retrieval based on document partitioning fits this profile well.
- Moreover, documents are always grouped into collections, either for administrative purposes or to combine related documents into a single source. Collections, therefore, provide a natural granularity for distributing data across servers and partitioning the computation. Note that since term partitioning imposes greater communication overhead during query processing, it is rarely employed in a distributed system.

- To build a distributed IR system, we need to consider both engineering issues common to many distributed systems and algorithmic issues specific to information retrieval.
- The critical engineering issues involve defining a search protocol for transmitting requests and results; designing a server that can efficiently accept a request, initiate a subprocess or thread to service the request, and exploit any locality inherent in the processing using appropriate caching techniques; and designing a broker that can submit asynchronous search requests to multiple servers in parallel and combine the intermediate results into a final end user response.
- The algorithmic issues include how to distribute documents across the distributed search servers, how to select which servers should receive a particular search request, and how to combine the results from the different servers.
- The search protocol specifies the syntax and semantics of messages transmitted between clients and servers, the sequence of messages required to establish a connection and carry out a search operation, and the underlying transport mechanism for sending messages (e.g., TCP/IP).
- At a minimum, the protocol should allow a client to-
  - Obtain information about a search server, e.g., a list of databases available for searching at the server and possibly statistics associated with the databases.
  - Submita search request for one or more databases using a well defined query language.
  - Receivesearch results in a well defined format.
  - Retrieveitems identified in the search results.
- For closed systems consisting of homogeneous search servers, a custom search protocol may be most appropriate, particularly if special functionality (e.g., encryption of requests and results) is required.
- Alternatively, a standard protocol may be used, allowing the system to interoperate more easily with other search servers.
- The Z39.50 standard for client/server information retrieval defines a widely used protocol with enough functionality to support most search applications.
- Another proposed protocol for distributed, heterogeneous search, called STARTS (Stanford Proposal for Internet MetaSearching), was developed at Stanford University in cooperation with a consortium of search product and service vendors.
- STARTS was designed from scratch to support distributed information retrieval and includes features intended to solve the algorithmic issues related to distributed Information Retrieval, such as merging results from heterogeneous sources.

4.2 Collection Partitioning

University Questions

Q.	Explain collection partitioning with respect to distributed IR.	SPPU : May 12, Dec. 13, 3/4 Marks
Q.	Explain distributed IR with the help collection partitioning.	SPPU : May 13, 3 Marks
Q.	Describe Collection partitioning in distributed IR.	SPPU : May 14, May 15, 5 Marks
Q.	What do you mean by collection partitioning in Distributed IR?	SPPU : Dec. 16, 4 Marks

- The procedure used to assign documents to search servers in a distributed IR system depends on a number of factors. First, we must consider whether or not the system is centrally administered.
- In a system comprising independently administered, heterogeneous search servers, the distributed document collections will be built and maintained independently.
- In this case, there is no central control of the document partitioning procedure and the question of how to partition the documents is essentially debatable. It may be the case, however, that each independent search server is focused on a particular subject area, resulting in a semantic partitioning of the documents into distributed collections focused on particular subject areas. This situation is common in Metasearch systems that provide centralized access to a variety of back-end search service providers.
- When the distributed system is centrally administered, more options are available.
  - **The first option** is simple replication of the collection across all of the search servers. This is appropriate when the collection is small enough to fiton a single search server, but high availability and query processing through put are required. In this scenario, the parallelism in the system is being exploited via multitasking and the broker's job is to route queries to the search servers and balance the loads on the servers.
  - Indexing the documents is handled in one of the following two ways.
    1. In the first method, each search server separately indexes its replica of the documents.
    2. In the second method, each server is assigned a mutually exclusive subset of documents to index and the index subsets are replicated across the search servers. A merge of the subsets is required at each search server to create the final indexes
- In either case, document updates and deletions must be broadcastto all servers in the system. Document additions may be broadcast, or they may be batched and partitioned depending on their frequency and how quickly updates must be reflected by the system.
- **The second option** is random distribution of the documents. This is appropriate when a large document collection must be distributed for performance reasons but the documents will always be viewed and searched as if they are part of a single, logical collection. The broker broadcasts every query to all of the search servers and combines the results for the user.
- **The final option** is explicit semantic partitioning of the documents. Here the documents are either already organized into semantically meaningful collections such as by technical discipline, or an automatic clustering or categorization procedure is used to partition the documents into subject specific collections.

4.3 Source Selection

University Questions

Q. Explain source selection with respect to distributed IR. **SPPU : May 12, Dec. 13, 3/4 Marks**

Q. Explain distributed IR with the help source selection. **SPPU : May 13, 3 Marks**

Q. Describe Source Selection in distributed IR. **SPPU : May 14, May 15, 5 Marks**

Q. What do you mean bysource selection in Distributed IR? **SPPU : Dec. 16, 4 Marks**

- Source selection is the process of determining which of the distributed document 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 every collection is equally likely to contain relevant documents and simply broadcast the query to all collections.
- This approach is appropriate when documents are randomly partitioned or there is significant semantic overlap between the collections.
- When 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 containing relevant documents.
- The basic technique is to treat each collection as if it were a single large document, index the collections, and evaluate the query against the collections to produce a ranked listing of collections. We can apply a standard cosine similarity measure using a query vector and collection vectors.
- To calculate a term weight in the collection vector using tf-idf style weighting term frequency tfi,j is the total number of occurrences of term i in collection j, and the inverse document frequency idfi for term i is log(N/ni), where N is the total number of collections and ni is the number of collections in which term i appears.
- Problem of this approach is that although a particular collection may receive a high query relevance score, there may not be individual documents within the collection that receive a high query relevance score, essentially resulting in a false drop and unnecessary work to score the collection.
- To avoid this problem Moffat and Zobel proposed a solution by indexing each collection as a series of blocks, where each block contains B documents.
- When B equals 1, this is equivalent to indexing all of the documents as a single, monolithic collection.
- When B equals the number of documents in each collection, this is equivalent to the original solution. By varying B, a trade off is made between collection index size and likelihood of false drops.
- An alternative to searching a collection index was proposed by Voorheesusing training queries to build a content model for the distributed collections. When a new query is submitted to the system, its similarity to the training queries is computed and the content model is used to determine which collections should be searched and how many hits from each collection should be returned.

4.4 Query Processing

University Questions

Q. Explain query processing with respect to distributed IR. **SPPU : May 12, 4 Marks**

Q. How are queries processed in distributed IR? **SPPU : Dec. 12, May 16, 9 Marks**

Q. Explain distributed IR with the help query processing. **SPPU : May 13, 4 Marks**

Q. Explain Query processing in distributed IR system. **SPPU : Dec. 13, Dec. 16, 8 Marks**

Q. Write short note on query processing in a distributed IR system. **SPPU : May 14, Dec. 14, 8 Marks**

Q. Explain with example how Query Processing is done in distributed IR system. **SPPU : May 15, 10 Marks**

- Query processing in a distributed information retrieval system is proceeds as follows :
  1. Select collections to search.
  2. Distribute query to selected collections.
  3. Evaluate query at distributed collections in parallel.
  4. Combine results from distributed collections into final result.

- 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 collections using its own local search algorithm. Finally, the results are merged.
- There are a number of scenarios as follows :

1. If the query is Boolean and the search servers return Boolean result sets, all of the sets are simply unioned to create the final result set.
2. If the query involves free-text ranking, a number of techniques are available ranging from simple/naive to complex/accurate.

- 1) The simplest approach is to combine the ranked hit-lists using round robin interleaving. This is likely to produce poor quality results since hits from irrelevant collections are given status equal to that of hits from highly relevant collections. An improvement on this process is to merge the hit-lists based on relevance score.
- As with the parallel process described for Document Partitioning, unless proper global term statistics are used to compute the document scores, we may get incorrect results. If documents are randomly distributed such that global term statistics are consistent across all of the distributed collections, the merging based on relevance score is sufficient to maintain retrieval effectiveness.
- If the distributed document collections are semantically partitioned or maintained by independent parties, then re ranking must be performed.
- Callan proposes re ranking documents by weighting document scores based on their collection similarity computed during the source selection step. The weight for a collection is computed as :

$$\omega = 1 + |C| \cdot (s - \bar{s}) / \bar{s}$$

Where, $|C|$ - The number of collections searched,

s - The collection's score,

$\bar{s}$ - The mean of the collection scores.

- The most accurate technique for merging ranked hit-lists is to use accurate global term statistics. This can be accomplished in one of a variety of ways one of this is discussed in following paragraph.
- If the collections have been indexed for source selection, that index will contain global term statistics across all of the distributed collections. The broker can include these statistics in the query when it distributes the query to the remote search servers. The servers can then account for these statistics in their processing and produce relevance scores that can be merged directly.
- If a collection index is unavailable, query distribution can proceed in two rounds of communication. In the first round, the broker distributes the query and gathers collection statistics from each of the search servers. These statistics are combined by the broker and distributed back to the search servers in the second round. Finally, the search protocol can require that search servers return global query term statistics and per-document query term statistics. The broker is then free to rerank every document using the query term statistics and a ranking algorithm of its choice.

- The end result is a hit-list that contains documents from the distributed collections ranked in the same order as if all of the documents had been indexed in a single collection.

4.5 Web Issues

Challenges

1. **Dynamically Evolving and Expanding Data**
   - Today the data is in order of peta-bytes size.
   - Everyday its size is increasing in multi folds.
   - This scenario is a great challenge for Web based IR.
2. **Highly Relevant Results**
   - Users need highly effective and relevant result for the queries.
   - Queries vary in their keywords and languages.
3. **Fast Response Time**
   - Queries are expected to be processed in very short time.
   - Results need to be compiled after retrieval in short time.
4. **Scalability**

   The retrieval system should be ready to scale as per requirement
5. **Compiling Results**
   - Results from various sources and on various systems needs to be handled, to generate
   - A single list of results for the user. Huge number of duplicates in different lists is a challenge.

4.6 Multimedia Information Retrieval - Introduction

University Question

Q. What is multimedia information retrieval ? **SPPU : May 13, Dec. 14, 4 Marks**

- Multimedia information system is widely recognized as one of the most promising and growing field in the area of information management as it is rapidly used in many application environment like offices, CAD/CAM.
- The most important characteristic of multimedia information systems is the variety of data it supports. Multimedia system should have the capability to store, retrieve, transport and present data with very heterogeneous characteristics such as text, images, graphs and sound.
- And because of these reasons the development of multimedia system is more complex than traditional information system.

4.6.1 Multimedia Information System vs. Traditional System

- The data model, the query language, the access and storage mechanisms of a Multimedia information system supports object with very complex structure. While traditional system or conventional system deals with simple data type such as strings or integers.
- Multimedia system handles multimedia data while traditional system handles textual unstructured data. Traditional systems are unable to support the mix of unstructured and structured data and different kinds of media.

- Traditional system does not support metadata information such as that provided by data base schema which is fundamental component in a database management system. A multimedia information retrieval system requires some form of database schema because multimedia applications need to structure their data at least partially.
- Multimedia information retrieval system requires handling metadata which is crucial for data retrieval. Whereas traditional information retrieval system don't have such requirement.
- Traditional system handles attribute bases queries i.e. set of attributes. But multimedia information retrieval system answer attribute based as well as content based queries i.e. set of features.
- In traditional system object retrieved by query processing are exact and precise. While in multimedia information retrieval system exact match cannot be applied means there is no guarantee that the object retrieved by this type of predicate are 100% correct or precise.
- Queries in traditional type system are 'Retrieve all names having Ids between EMP10 to EMP100'. While in multimedia information retrieval system queries are of the type 'Retrieve all the cars manufactured by same company and with different colour'.

4.6.2 Data Modelling

- To ensure fast retrieval a multimedia information retrieval system should be able to represent and store multimedia objects.
- The system should be therefore able to deal with different kinds of media and with semi-structured data i.e. data whose structure may not match or only partially match, the structure prescribed by the data schema.
- The system must extract some feature from multimedia objects to represent semi-structured data.

4.6.3 Data Retrieval

University Questions

Q. Explain the steps multimedia IR. **SPPU : May 13, 8 Marks**

Q. Explain query specification query processing with respect to multimedia information retrieval. **SPPU : May 13, 4 Marks**

Q. Discuss steps on which of data retrieval relies. **SPPU : Dec. 14, 4 Marks**

Q. Explain the steps involved for retrieving data in multimedia IR systems. **SPPU : Dec. 14, 8 Marks**

Q. What is feature extraction in multimedia information retrieval? How is it helpful for data retrieval ? **SPPU : May 15, 9 Marks**

- The main goal of a multimedia information retrieval system is to efficiently perform retrieval based on user requests by exploiting not only data attributes but also content of multimedia objects
- Data retrieval depends on the following steps as shown in Fig. 4.6.1.

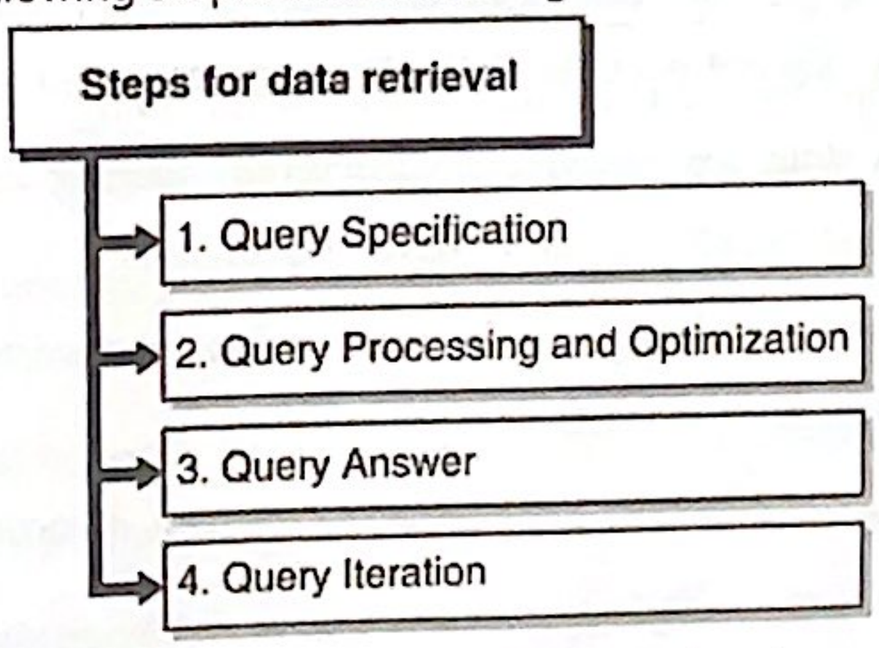

Fig. 4.6.1 : Steps for data retrieval

1. Query Specification

- In this step, the user specifies the request. The query interface should allow the user to express.
- Fuzzy predicates for proximity searches for example find all images similar to a car.
- Content based predicates for example find multimedia objects containing an apple.
- Conventional predicates on the object attributes for example conditions on the attribute 'color' of an image such as 'find all red images'.
- Structural predicates for example find all multimedia objects containing a video clip.

2. Query Processing and Optimization

- Like traditional system in multimedia information retrieval system query is parsed, compiled and optimized using best evaluation plan to generate internal representation. Due to the fuzzy terms, content-based predicates and structural predicates, query processing is a very complex activity. As compare with traditional and spatial databases a very little work is done on query processing strategies for multimedia information system.
- The main problem is the heterogeneity of data: different query processing strategies one for each data type should be combined together in some way.

3. Query Answer

- The retrieved objects are returned to the user in decreasing order of relevance.
- Relevance is the distance function from the query object to the stored object.

4. Query Iteration

- In traditional system when the system returns answer to the user the query process ends.
- While in multimedia information retrieval system due to the inevitable lack of precision in the user request the query execution is iterated until the user is satisfied.
- At each iteration, the user supplies the system with additional information by which the request is refined reducing or increasing the number of returned answers.

4.7 Data Modelling in Multimedia Information Retrieval

The integration of multimedia data in a traditional DBMS is not an easy task due to :

- Multimedia data is inherently different from conventional data. The main difference is that information about the content of multimedia data are usually not encoded into attributes provided by the data schema.
- As text, image, video and audio data are typically unstructured; specific methods are needed to identify and represent content features and semantic structures of multimedia data.
- Multimedia system requires large storage; one single image requires several Kbytes of storage where as single second of video requires several Mbytes of storage.
- The content of multimedia data is difficult to analyze and compare in order to be actively used during query processing.

- In order to resolve data modelling issues the framework of multimedia information retrieval systems entails two main tasks :

1. A data model should be defined for the user to specify the data to be stored into the system. Such data model should be able to give integrated support for both conventional and multimedia data types. It also able to provide methods to analyze, retrieve and query such data.
2. The system should provide model for the internal representation of multimedia data.

- As far as the first task is concerned, a promising technology with respect to the modelling requirements of multimedia data is the object oriented.
- The richness of data model provided by OODBMSs makes them more suitable than relational DBMSs for modelling both multimedia data types and their semantic relationships.
- The concept of class can be naturally used to define adhoc data types for multimedia data in that a class is characterized by both a set of attributes and a set of operations that can be performed on these attributes.
- As the classes are related into inheritance hierarchies it helps to allow the definition of a multimedia class as specialization of one or more super classes.
- In terms of storage techniques, query processing and transaction management the performance of OODBMS is not comparable to that of relational DBMS.
- One more drawback of OODBMS is that it is non standard.
- Even though a standard has been defined by Object Database Management Group a very few system support it.
- Because of the above mentioned reasons a lot of work has been devoted to the extension of the relational model with capabilities for modelling complex objects.
- The goal of object relational technology is to extend the relational model with ability to represent complex data types and to maintain the performance and simplicity of relational DBMS and related query languages.
- The possibility of defining abstract data types inside the relational model allows one to define ad hoc data types for multimedia data. And such data types provide support for content-dependant queries.
- The second problem which is related with data modelling is how to represent multimedia data inside the system.
- It is not sufficient to describe multimedia data through a set of attributes but some information should be extracted from the objects and used during query processing due to particular nature of multimedia data.
- The extracted information is typically represented as a set of features; each multimedia object is therefore internally represented as a list of features, each of which represents a point in a multidimensional space.
- Multi-attribute access methods can then be used to index and search for them. Features can be assigned to multimedia objects either manually by the user or automatically by the system.
- In general a hybrid approach is used by which the system determines some of the values and the user corrects or augments them.
- In both the cases, values assigned to some specific features such as the shape of an image or the style of an audio object are assigned to the object by comparing the object with some previously classified objects.
- To establish whether an image represents a car or a house the shape of the image is compared with the shapes of already classified cars and houses before taking a decision.

- A weight is usually assigned to each feature value representing the uncertainty of assigning such a value to that feature. For example if we are 80% sure that a shape is a square we can store this value together with the recognized shape.
- It follows that data modelling in multimedia information retrieval system is must take into account the complex structure of data and the need of representing features extracted from multimedia objects.
- Next we will give an overview of the support for multimedia data provided by commercial DBS then an example of the data model developed in the context of the MULTOS project.

4.7.1 Multimedia Data Support in Commercial DBMS

University Question

Q. Describe Multimedia Data Support in Commercial DBMS. **SPPU : Dec. 13, May 15, May 16, May 19, 9/10 Marks**

- Most of the current relational DBMSs support variable length data types which are useful to represent multimedia data. But the way by which data is supported is not standard.
- Each DBMSs vendor uses different names for such data types and provides support for different operations.

For example -

- The Oracle DBMS provide the VARCHAR2 data type to represent variable length character strings with maximum length of data is 4000 bytes.
- The RAW and LONG RAW data types are used for data that is not interpreted by Oracle. These data types can be used to store sounds, graphics or unstructured objects.
- LOB (Large Object) data types can be used to store large unstructured data objects up to 4 GB in size.
- BLOB (Binary Object) are used to store unstructured binary large objects.
- CLOB (Character Large Object) are used to store character large object data.
- The Sybase SQL server supports IMAGE and TEXT data types to store images and unstructured text and also provides a limited set of functions for searching and manipulation.
- But the support provided by the above mentioned data type is very limited and operations can be performed on such data by means of the built in functions provided by the DBMS are very simple.
- The major improvement have been provided by SQL3 with respect to its predecessor SQL-92 like :
  - Support for an extensible type system. Extensibility is achieved by providing constructs to define user dependent abstract data types like object oriented manner.
  - Each type specification consists of both attribute and function specifications.
  - A strong encapsulation is provided in that attribute values can only be accessed by using some system functions.
  - User defined functions can be either visible from any object or only visible in the object they refer to.
  - Both single and multiple inheritances can be defined among user-defined types and dynamic late binding is provided.
  - Provides three types of collections data types as: sets, multisets and lists.
  - Several system defined operations are provided to deal with collections.

- Provides a restricted form of object identifier that supports sharing and avoid data duplication.

Examples of SQL3

Most of the commercial products have already implemented their proprietary versions of SQL3. For example:

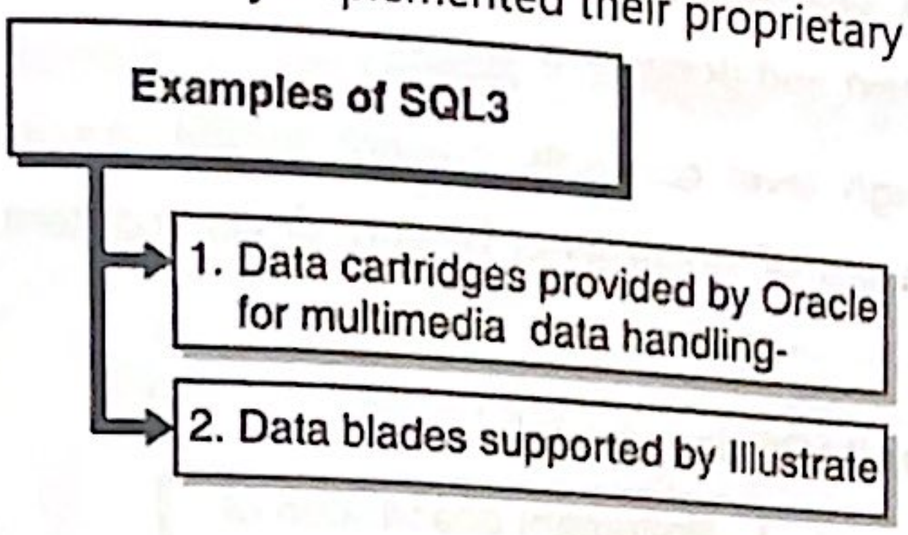

Fig. 4.7.1 : Examples of SQL3

1. **Data cartridges provided by Oracle for multimedia data handling**

- Oracle provide data cartridge for text, spatial data, image, audio and video data.
- Oracle8 provides a ConText cartridge which is text management solution combining data management capabilities of traditional system with advanced retrieval and natural-language process technology.
- ConText supports the document formats like ASCII, MSWord and HTML.
- ConText has ability to find document of specific type.

2. **Data blades supported by Illustrate**

- Provides support for 3D and 2D spatial data blades for modelling spatial data.
- Provides boxes, vectors, quadrangles data types
- Provides INTERSECT, CONTAINS, OVERLAPS, CENTER operations.
- Implements R-tree for performing efficient spatial queries.
- Supports a data blade which can be used to query images by content.
- The object relational technology and its extensive type system are now widely used in industrial and research projects for example LA Scala. The goal of this project is the development of the multimedia archive of Teatro alla Scala.

4.7.2 MULTOS Data Model

University Questions

Q. How image analysis and image access accomplished in MULTOS data model. **SPPU : May 12, 8 Marks**

Q. Write a note on MULTOS data model. **SPPU : May 14, May 16, May 17, 8/9 Marks**

Q. Explain MULTOS data model with example. **SPPU : Dec. 14, 8 Marks**

Q. Write short note on MULTOS query language. **SPPU : May 15, 8 Marks**

- MULTOS i.e. 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.
- Based on client-server architecture.

- Supports different type of servers i.e. current servers, dynamic servers and archive servers. Each supported server differs in storage capacity and document retrieval speed.
- Server's supports filling and retrieval of multimedia objects based on document collections, document types, document attributes, document text and document images.
- Allows the representation of high level concepts present in the documents contained in the database, the grouping of documents into classes of documents having similar content and structure and the expression of conditions on free text.
- Each document is described by MULTOS data model.

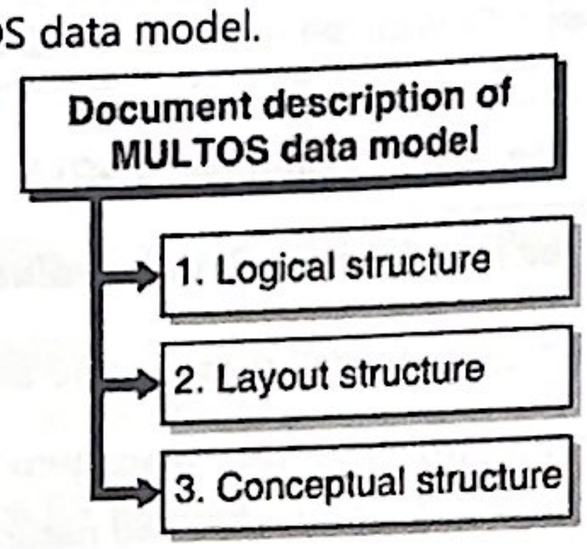

Fig. 4.7.2 : Document description of MULTOS Data model

1. Logical structure

- Determines arrangement of logical components for example title, introduction, chapter, section.
- Allows syntax oriented description of the document content.

2. Layout structure

- Deals with the layout of the document content.
- Contains components such as pages, frames etc.
- Allows syntax oriented description of the document content.

3. Conceptual structure

- Allows semantic oriented description of the document content as opposed to the syntax oriented description provided by the logical and layout structure.
- Added to provide support for document retrieval by content.
- For document retrieval it plays the role of the database schema which enables the use of efficient access structure.
- Are basis for formulating queries at an abstract level.
- MULTOS provides a formal model based on a data structuring tool available in semantic data models, to define the document conceptual structure.
- The logical and layout structures are defined according to the ODA document representations.
- Documents having similar conceptual structures are grouped into conceptual types.
- Conceptual types are maintained in a hierarchy of generalization to handle types. Subtype inherits from its super types the conceptual structure and then refines it.

- Types can be strong and weak. A strong type completely specifies the structure of its instances.
- While weak type partially specifies the structure of its instances. Components of unspecified type called as spring component type can also appear in document definition.
- The conceptual structure of the type Generic Letter is as shown in Fig. 4.7.3.

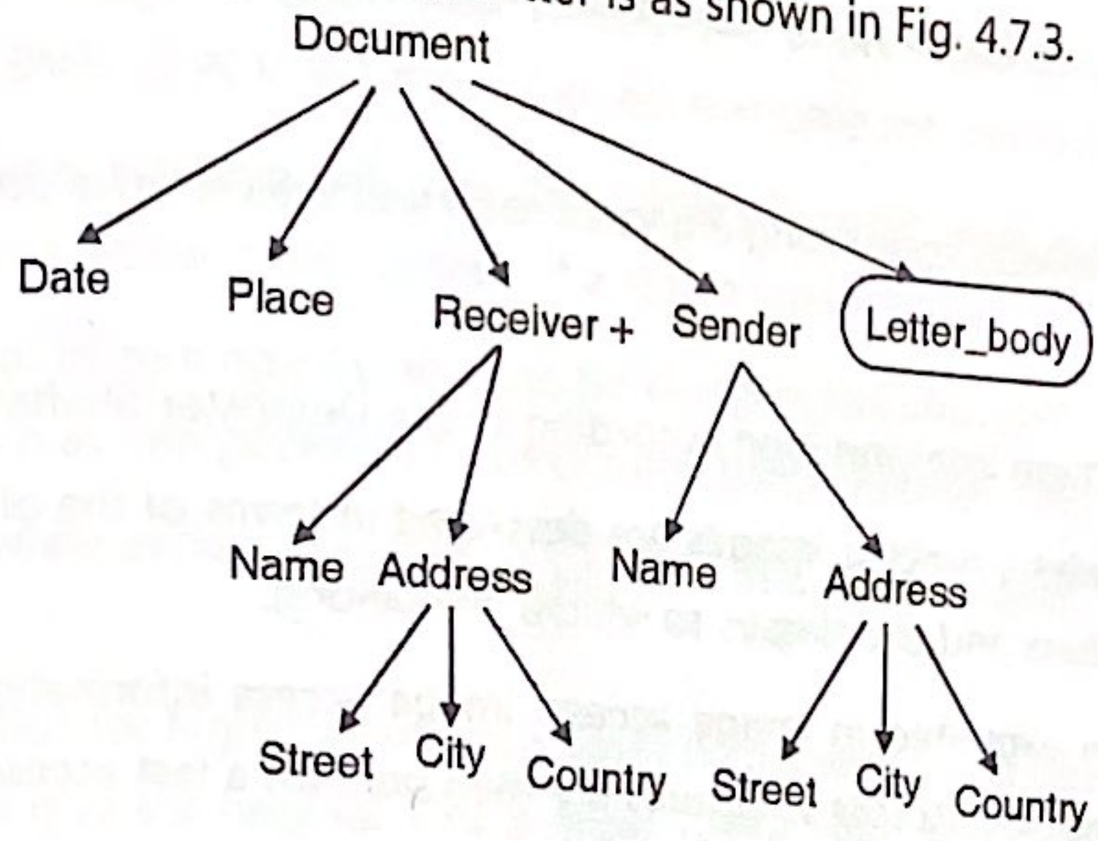

Fig. 4.7.3 : Conceptual Type Structure of Generic_Letter

- The node Letter_Body is a spring conceptual component.
- The complete conceptual structure corresponds to the type Business_Product_Letter is as shown in Fig. 4.7.4.

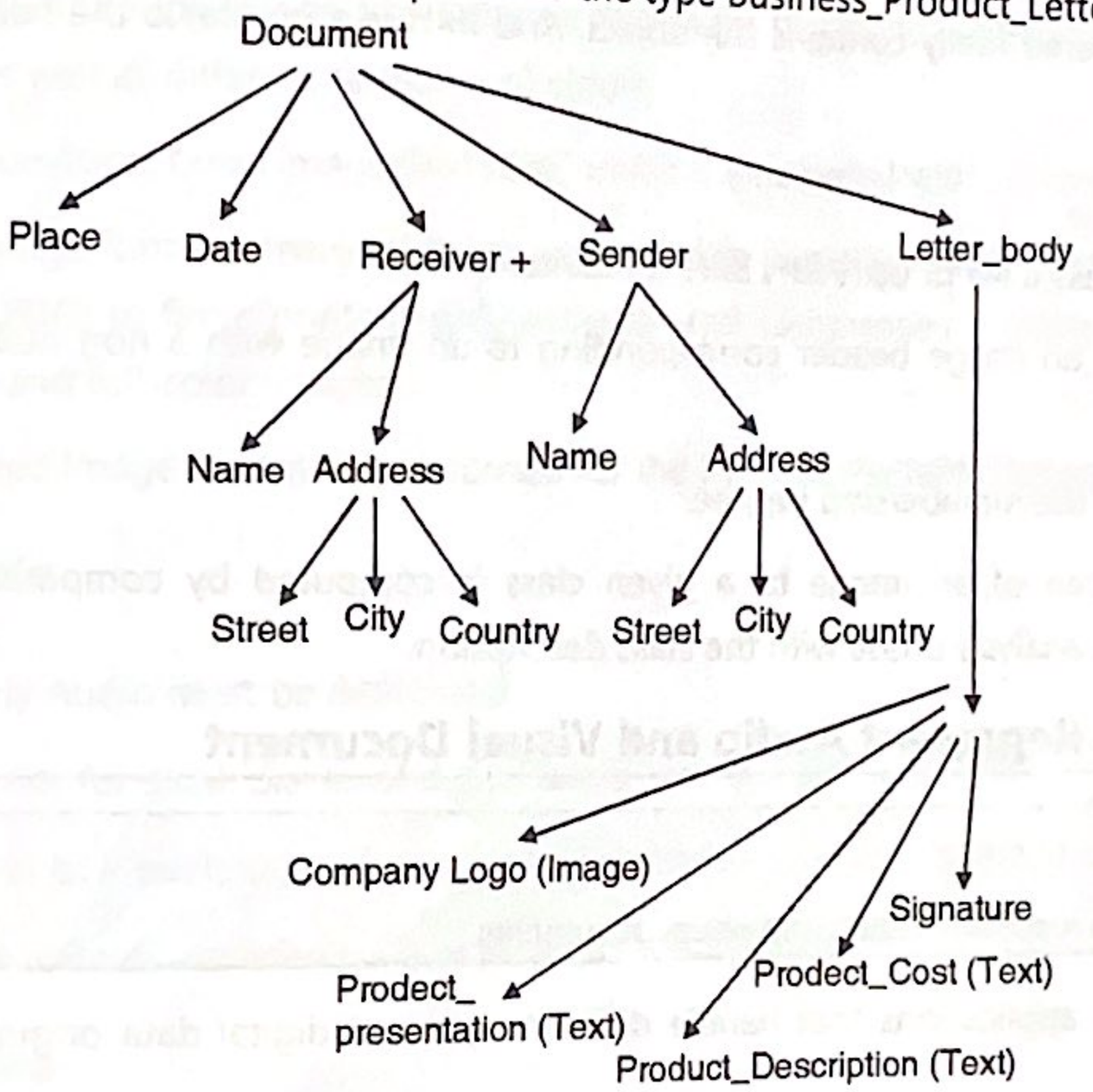

Fig. 4.7.4 : Complete Conceptual Structure of the Type Business_Poduct_Letter

- This type has been obtained from Generic_Letter by specialization of Letter_Body into a complex conceptual component, defined as an aggregation of five conceptual components.
- According to the conceptual model the document type Business_Product_Letter is linked to the document type Generic_Letter by an is-a relationship.
- In this example the '+' symbol attached to the Receiver component means that it is mutivalued. The Name and the Address appear in two subtrees having as roots the conceptual components Receiver and Sender.

- To deal with image data MULTOS provides sophisticated approach-
  - An image analysis process is performed, consisting of two phaseslow level image analysis and high level image analysis.
  - Two types of index are constructed- object index and cluster index.

a. Low level image analysis

During this phase the basic objects composing a given image and their relative positions are identified.

b. High level image analysis

- This phase deals with image interpretation according to the Dempster-Shafter theory evidence.
- At the end of image analysis process images are described in terms of the objects recognized with associated belief and possibility values and the classes to which they belong.
- The information is then exploited in image access. Image access information is stored in an image header associated with the image file. Access structures are then built for a fast access to image headers.

c. Object index

- For each object a list is maintained.
- Each element of the list is a pair BI, IMH. Where, BI is the associated belief interval representing the probability that image considered really contains the object. And IMH is a pointer o the header of the image containing the object.

d. Cluster index

- For each image class a list of pairs MF, IMH is maintained.
- IMH is pointer to an image header corresponding to an image with a non null degree membership to the class.
- MF is the value of the membership degree.
- Membership degree of an image to a given class is computed by comparing the image interpretation resulting from the analysis phase with the class description.

4.8 Techniques to Represent Audio and Visual Document

University Question

Q. Discuss techniques to represent audio and visual documents. **SPPU : Dec. 12, 8 Marks**

- Multimedia stands for applications that handle different types of digital data originating from distinct types of media.
- The most common types of media in multimedia applications are text, sound, images, and video (which is an animated sequence of images) as well as binary data.
- The digital data originating from each of these four types of media is quite distinct in volume, format, and processing requirements (for instance, video and audio impose real time constraints on their processing).
- Different types of formats are necessary for storing each type of media.
- In this section we cover formats for multimedia applications. In contrast with text formats, most formats for multimedia are partially binary and hence can only be processed by a computer.

- Also, the presentation style is almost completely defined, perhaps with the exception ofsome spatial or temporal attributes.

4.8.1 Format for Images

- There are several formats for images. The simplest formats are direct representations of a bit-mapped (or pixel-based) display such as XBM, BMP, or PCX. But those formats consume too much space.
- For example, a typical computer screen which uses 256 colours for each pixel might require more than 1 Mb (one megabyte) in storage just for describing the content of a single screen frame.
- In practice, images have a lot of redundancy and can be compressed efficiently. So, most popular image formats incorporate compression such as Compuserve's Graphic Interchange Format (GIF).
- GIF is good for black and white pictures, as well as pictures that have a small number of colours or gray levels (say 256).
- To improve compression ratios for higher resolutions, lossy compression was developed. That is, uncompressing a compressed image does not give the original. This is done by the Joint Photographic Experts Group (JPEG) format, which tries to eliminate parts of the image that have less impact on the human eye. This format is parametric, in the sense that the loss can be tuned.
- Another common image format is the Tagged Image File Format (TIFF). This format is used to exchange documents between different applications and different computer platforms. TIFF has fields for metadata and also supports compression as well as different numbers of colors.
- Yet another format is Truevision Targa image file (TGA), which is associated with video game boards.
- There are many more image formats, many of them associated to particular applications ranging from fax (bi-level image formats such as JBIG) to fingerprints (highly accurate and compressed formats such as WSQ) and satellite images(large resolution and full-color images).
- In 1996 a new bit-mapped image format was proposed for the Internet: Portable Network Graphics (PNG).

4.8.2 Format for Audio

- In order to store properly Audio must be digitalized.
- The most common formats for small pieces of digital audio are AU, MIDI, and WAVE.
- MIDI is a standard format to interchange music between electronic instrumentsand computers.
- For audio libraries other formats are used such as RealAudio or CD formats.

4.8.3 Format for Video

- There are several formats for animations or moving images (similar to videoor TV), but here we mention only the most popular ones.
- The main one is MPEG (Moving Pictures Expert Group) which is related to JPEG. MPEG works by coding the changes with respect to a base image which is given at fixed intervals. In this way, MPEG profits from the temporal image redundancy that any video has. Higher quality is achieved by using more frames and higher resolution. MPEG specifies different compression levels, but usually not all the applications support all of them. This format also includes the audio signal associated with the video.

- Other video formats are AVI, FLI, and QuickTime. AVI may include compression (CinePac), as well as QuickTime, which was developed by Apple. As for MPEG, audio is also included.

4.9 Query Languages

- In relational or object database systems queries are based on the exact match mechanism.
- With this mechanism system will be able to return only those tuples or objects that satisfy some well specified criteria given in the query expressions i.e. query predicates specify which values the object attributes must contain.
- User should be able to query the content of multimedia objects by specifying values of semantic attributes and also able to specify additional conditions about the content o multimedia data.
- Thus exact match is only one of the possible ways of querying multimedia objects.
- A similarity based approach is applied that considers both the structure and the content of the objects.
- A query of the later type is called as content based queries. These queries retrieve multimedia object depending upon their global content.
- Information on the global content of an object is not represented as attribute values in the database.
- Instead of that a set of information called as features is extracted and maintained for each object.
- When the user submit query, the features of the query object are matched with respect to the feature of the object stored in the database and only the objects that are more similar to the query are returned to the user.
- While designing any multimedia query language following three aspects should be considered.
- Interfaces to be provided to the user to enter his/her queries
- Conditions on multimedia objects can be specified in the user request.
- The impact of uncertainty, proximity and weights on the design of query language.

4.9.1 Request Specification

To query multimedia objects, two different interfaces are presented to the user –

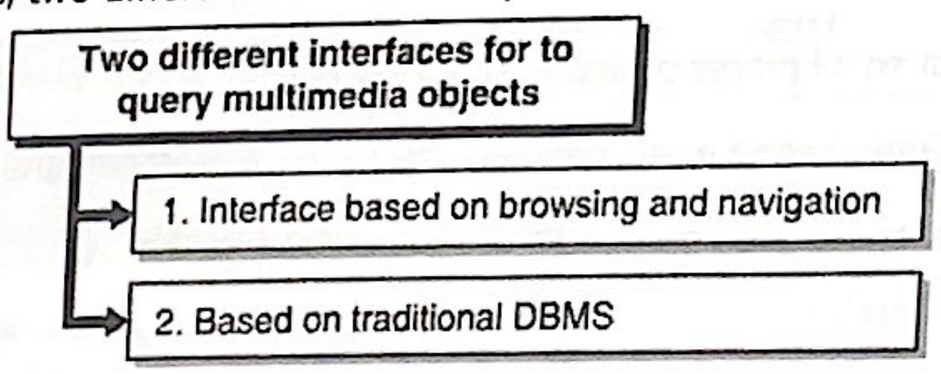

Fig. 4.9.1 : Interfaces for to query multimedia objects

1. Interface based on browsing and navigation

- Users can browse and navigate inside the structure of multimedia objects to locate the desired object. This approach is used in CAD/CAM/CASE environment due to the complex structure of the objects under consideration.
- But navigation is not always the best way to find multimedia objects because they may be heavily time consuming when the object desired is deeply nested.

2. Based on traditional DBMS

- Specify the conditions that objects of interest must satisfy.
- Queries can be specified in two different ways :

i. Traditional database context.

User can enter query by using a specific query language.

ii. Query by example

- When there is image and audio data then this way is preferred. In this approach queries are specified by using actual data inside a visual environment. The user provides the system with an object example that is then used to retrieve all the stored objects similar to the given one.
- This approach requires the use of GUI environment where user can pick examples and compose the query objects. To pick examples, the system must supply some domain i.e. for each object feature a set of typical values.
- Example of this type is "Retrieve all houses of the same shape and different color".

4.9.2 Conditions on Multimedia Data

University Questions

Q. Explain the query predicates of multimedia query language in detail. **SPPU : May 14, 8 Marks**

Q. What do you understand by multimedia query language? Explain various query predictors. **SPPU : May 17, May 19, 8/9 Marks**

- Multimedia query language should provide predicates for expressing conditions on the attribute, the content and the structure of multimedia objects.
- Query predicate can be classified as shown in Fig. 4.9.2.

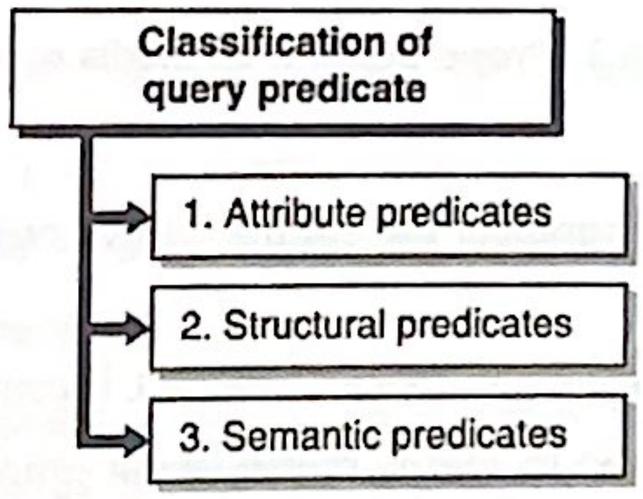

Fig. 4.9.2 : Classification of query predicate

1. Attribute predicates

- Concern the structured content attribute of the multimedia objects.
- These are predicates against traditional attribute. i.e. attribute for which an exact value is supplied for each object.
- Examples of attribute are the speaker of an audio object, the size of an object and the type of an object.
- By querying these predicates the system applies exact-match retrieval.

2. Structural predicates

- Concern the structure of data being considered.
- Predicate are answered by using some form of metadata and information about the database schema.
- Example of this type is "Find all multimedia objects containing at least one image and a video clip".

3. Semantic predicates

- Concern the semantic content of the queried data depending on the features that have been extracted and stored for each multimedia object.
- Example of this type is "Find all the object containing the word OFFICE."
- The query "Find all the blue houses" is a query on the image content. This query can be executed only if color and shape are features that have been previously extracted from the image.
- Current systems support semantic predicates only with respect to specific features, such as the color, the shape, the texture and sometimes the motion.
- The main difference between attribute predicates and semantic predicates is that in semantic predicate an exact match cannot be applied. I.e. there is no guarantee that object retrieved are 100% correct or precise. The result of a query involving semantic predicates is a set of objects which has an associated degree of relevance with respect to the query.

Properties of multimedia objects

The structural and semantic predicates can also refer to following two properties of multimedia objects :

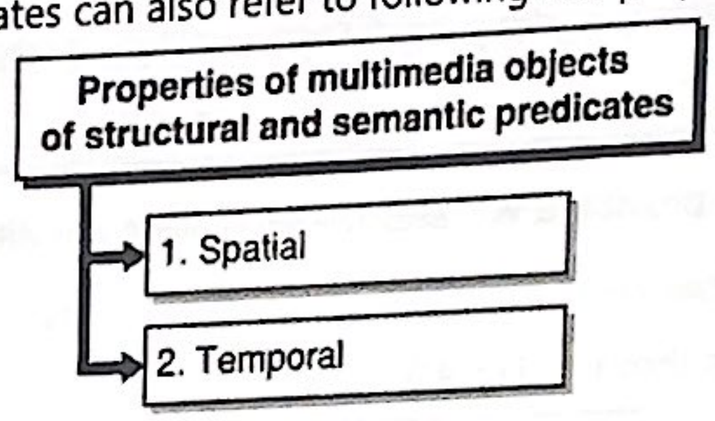

Fig. 4.9.3 : Properties of multimedia objects

1. Spatial

- Spatial semantic predicates specify conditions about the relative positions of a set of objects in an image or a video.
- Examples of spatial semantic predicates are: contain, intersect, is contained in, is adjacent to.
- Spatial structural predicates are used to specify spatial layout properties for the presentation of multimedia objects.
- Spatial structural predicates are also used to impose a condition on the spatial layout of the retrieved objects.
- Example of spatial structural predicate is query of the form "Find all the object containing an image overlapping the associated text.

2. Temporal

- Temporal semantic predicates are related to continuous media like audio and video.
- Temporal semantic predicates all to express temporal relationship among the various frames of a single audio or video.

- Example of temporal semantic predicates is "find all the objects that contain an audio component where the hint of the discussion is first policy, and then economy".
- Temporal structural predicates are used to specify temporal synchronization properties for the presentation of multimedia objects.
- Example of structural temporal predicate is query of form "Find all the objects in which a jingle is played for the duration of an image display".
- The temporal and spatial predicates can be combined to express more articulated requirements. Example is the query of the form "find all the objects in which the logo of a car company is displayed and when it disappears, a graphics showing the increase in the company sales is shown in the same position where the logo was."

4.9.3 Uncertainty, Proximity and Weights in Query Expressions

University Question

Q. Discuss uncertainty, proximity, and weights in query expressions. **SPPU : May 12, 8 Marks**

- One of the aspects in designing a query language is how to specify the degree of relevance of the retrieved objects. And this can be done by-

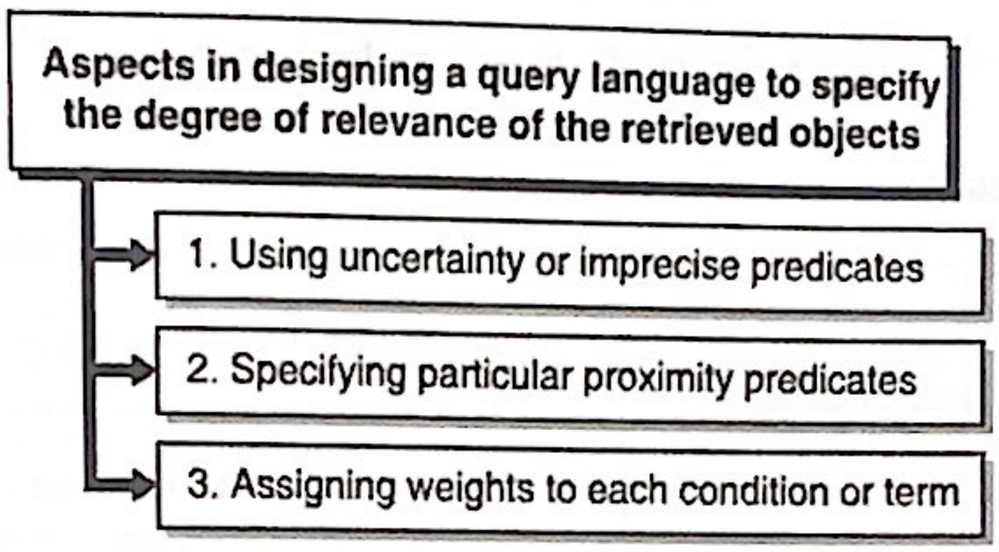

Fig. 4.9.4

1. Using uncertainty or imprecise predicates

- Some imprecise terms and predicates such as normal, unacceptable, typical are used.
- Each of these terms represents a set of possible acceptable values and not the precise values. And these possible acceptable values are with respect to which the attribute or the features has to be matched

2. Specifying particular proximity predicates

- The predicate does not represent a precise relationship between objects and values or between attribute/feature and values. Instead of that the relationship is based on the computation of a semantic distance between the query object and the stored objects on the basis of the extracted features.
- Example of proximity predicate is nearest object search, in this the user requests all the objects which are closest or within a certain distance of a given object.

3. Assigning weights to each condition or term

- Weight specifies the degree of precision by which a condition must be verified by an object.
- Example of this type is query of the form "Find all the objects containing an image representing a screen and a keyboard".

- The use of imprecise terms and relationship, the use of weights allows the user to drive the similarity-based selection of relevant objects.
- The corresponding query is executed by assigning some importance and preference values to each predicate and term. Then the object are retrieved and presented to the user as an ordered list.
- This ordering is given by a score associated with each object, giving a measure of the matching degree between the object and the query. And the computation of score is based on probabilistic models using the preference values assigned to each predicate.

4.9.4 Query Languages to Support Retrieval of Multimedia Objects

- In this section we will see some query languages supporting retrieval of multimedia objects.
- There are two query languages for retrieval of multimedia objects as shown in Fig. 4.9.5.

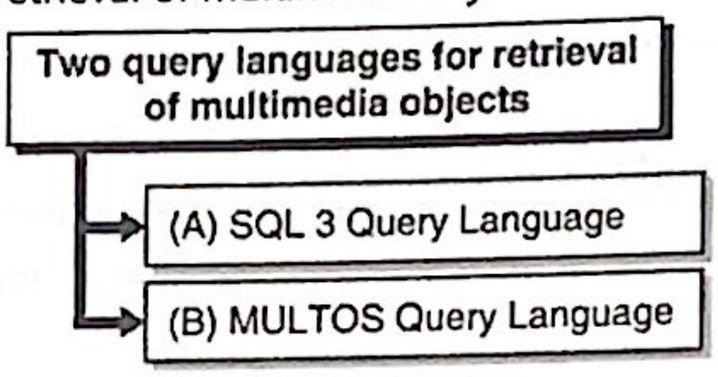

Fig. 4.9.5 : Query language for retrieval

4.9.4 (A) SQL 3 Query Language

University Question

Q. Write short notes on : SQL3 query language. **SPPU : Dec. 14, 4 Marks**

- In this section we will see the facilities provided by the standard SQL 3 to support retrieval of multimedia objects.
- Due to the extensibility to deal with complex objects make SQL 3 suitable for modelling multimedia data.
- From the query language point of view there are major improvements of SQL 3 with respect to SQL - 92 as :

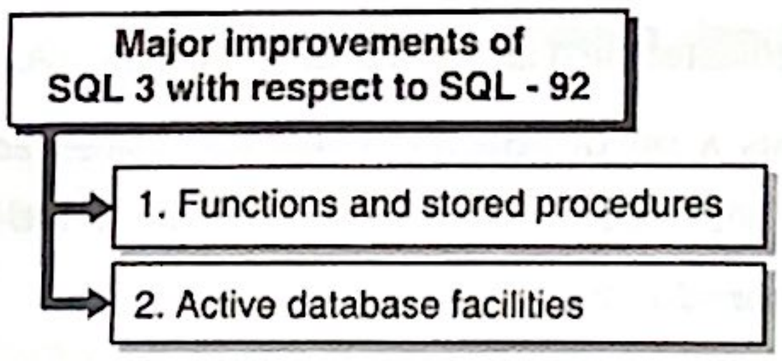

Fig. 4.9.6 : Improvement of SQL3 with respect to SQL-92

1. Functions and stored procedures

- SQL 3 allows the user to integrate external functionalities with data manipulation.
- As SQL 3 allows functions of an external library to be introduced into database system as external functions.
- Such external functions can be either implemented by using an external language or directly implemented by SQL 3.
- If external function is implemented by external language then SQL 3 only specifies which the language is and where the function can be found.

- By this way the impedance mismatch between two different programming languages and type systems is avoided.
- But this approach requires an extension of SQL with imperative programming languages constructs.

2. Active database facilities

- Database are able to react to some system or user-dependent events by executing specific actions, using active rules supported by SQL 3.
- Active rules or triggers are very useful to enforce integrity constraints.
- The ability to deal with external functions and user-defined data types enables the language to deal with objects with a complex structure as multimedia objects.
- Without this characteristic, it is not possible to deal with BLOB (Binary Large OBject) as it reduces the view of multimedia data to single large uninterested data values, which are not a liquate for the rich semantics of multimedia data.
- Spatial and temporal constraints can be enforced by using triggers. And it also preserves the database consistency.
- SQL 3 also allows to model multimedia objects in the framework of a well understood technology.

Limitations of SQL 3

- Does not provide retrieval support and optimization.
- NO information retrieval techniques are integrated into the SQL 3 query processor.
- The ability to perform consent-based search is application dependent.
- Objects are not ranked and therefore are returned the application as a unique set.
- Specialized indexing techniques can be used but they are not transparent to the user.

4.9.4(B) MULTOS Query Language

- The development of the MULTOS query language has been driven by then requirements such that it should be possible to easily navigate through the document structure using path-names.
- Path name identifies either one component or several components.
- If path name identifies only one component then it is termed as total.
- If path name identifies more than one component then it is termed as partial.
- Path names are similar to object oriented path expressions but queries both on the content and on document structure must be supported.
- Query predicates on complex components must be supported. In this case, the predicate applies to all the document subcomponents that have a type compatible with the type required by the query.
- This possibility is very useful when a user does not recall the structure of a complex component.

- In general the form of MULTOS query is as follows :

FIND DOCUMENTS VERSION version-clause

SCOPE scope - clause

TYPE type - clause

WHERE condition - clause

WITH component

Where,

- The version-clause specifies which versions of the documents should be considered by the query.
- The scope-clause restricts the query to a particular set of documents.
- This set of documents is either a user-defined document collection or a set of documents retrieved by a previous query.
- The type-clause allows the restriction of a query to documents belonging to prespecified set of types.
- The condition expressed by the condition clause is applicable to the document belonging to these types and their subtypes.
- When no type is specified, the query is applied to all the document types.
- The condition-clause is a Boolean combination of simple conditions i.e. predicates on documents components.
- Predicates are expressed on conceptual components of documents conceptual components are referenced by pet-names the general form of a predicate is – component restriction.
- Where, component is a path-name and restriction is an operator followed by an expression.
- The with-clause allows one to express structural predicates.
- Component is a path-name and the clause looks for all documents structuring containing such a component.
- Different types to conditions can be specified in order to query different types of media.
- MULTOS supports three main classes of predicates as shown in Fig. 4.9.7.

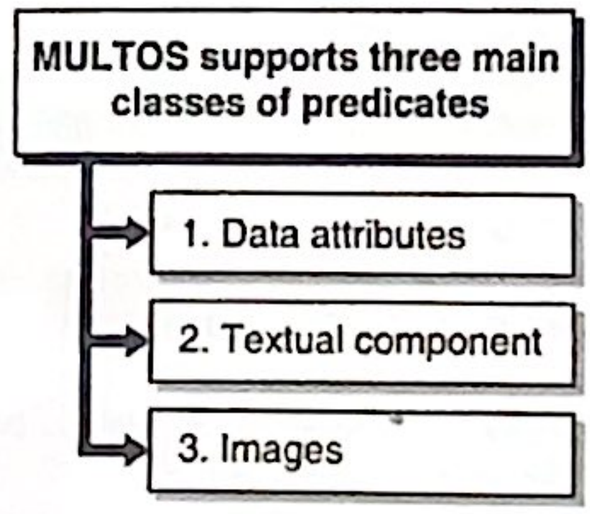

Fig. 4.9.7 : Classes of predicate

1. Data attributes

An exact match search is performed.

2. Textual component

By determining all objects containing some specific strings.

3. **Images**

- By specifying conditions on the image content.
- Image predicates allow one to specify conditions on the class to which an image should belong or conditions on the existence of a specified object within an image and on the number of occurrences of an object within an image.
- The following example illustrates the basic features of the MULTOS query language.

Example

- Here we will consider the conceptual structure Generic-letter.
- Consider the following query.

```
FIND DOCUMENT VERSIONS LAST WHERE
    Document . Date > 1/1/1998 AND
    * Product_Description CONTAINS
    ( * Sender . Name = "Olivetti" OR
    * Product_presentation CONTAINS "Olivetti") AND
    "Personal Computer"AND
    (* Address . Country = "Italy" OR
        TEXT CONTAINS "Italy") AND
    WITH * Company_Logo.
```

- According to this query user search for the last version of all documents dated after January 1998. Containing a company logo, having the word "olivetti" either as sender name or in the product presentation (is textural component) with the word 'Personal computer' in the product description section (is another textual component) and with the word 'Italy' either constituting the country in the address or contained in any part of the entire document.
- '*' symbol indicates that the path-name is not complete, that is it could identify more than one component.
- The query language provided by MULTOS also supports the specifications of imprecise queries that can be used when the user has an uncertain knowledge about the consent of the documents he or she is seeking.
- By associating both a preference and an importance value with the attributes in the query uncertainty can be expressed. Such values are then used for ranking the retrieved documents.
- This is as discussed in following example.
- The query is as follows :

```
FIND DOCUMENT VERSION LAST WHERE
(Document . Date BETWEEN (12/31/1998, 1/31/1998) PREFERED
BETWEEN (2/1/1998, 2/15/98) ACCEPTABLE) HIGH AND
(*Sender . Name = "olivetti" OR
*product_Presentation CONTAINS "olvietti") HIGH AND
(*Product_Description CONTAINS "Personal computer") HIGH AND
```

(* Product_Description CONTAINS "good ergonomics") LOW AND

(* Address . Country = "Italy") OR

TEXT CONTAINS "Italy") HIGH AND

WITH * Company_logo HIGH

(IMAGE MATCHES

screen HIGH

keyboard HIGH

AT LEAST 2 floppy_drives LOW) HIGH

- This query finds the last versions of all documents written in January, but possibly even at the beginning of February 1998. Containing a company logo, having the word "Olivetti" either as sender name or in the product presentation with the word "personal computer" in the product description section and with the word "Italy" either constituting the country in the address or contained in any part of the entire document.
- Personal computers are described in the product description section as products having good ergonomics.
- The document should contain the picture of the personal computer. Complete with screen and keyboard, with at least two floppy drives.
- The value 'LOW' associated with the condition on 'good ergonomics' indicates that the user formulating the query is not completely sure about this description of PC but associated value is HIGH.

4.10 Multimedia Information Retrieval : Indexing and Searching

4.10.1 Introduction

- The searching methods will search a database of multimedia object to locate objects that match a query object exactly and so problem is how to design such fast searching methods.
- Object can be two-dimensional colour images gray scale medical images in 2D or 3D, one dimensional time series, digitized voice or music, video clip etc.
- Example of query by content is "in a collection of colour photograph: Find ones with the same colour distribution as a sunset photograph."
- Some specific applications include image databases (like financial, marketing, production time series), scientific databases with vector fields. Audio and video database, DNA databases. Such databases contains queries like, "find all the companies whose stock prices move similarly" or "Find medical X-rays that contain something that has the texture of a tumor."
- Searching for similar patterns in such kind of databases is essential as it helps in predictions. Computer-aided medical diagnosis and teaching, hypothesis testing, data mining and rule discovery.
- The distance of two objects has to be quantified and domain expert will supply such distance function D ( ).
- If two object O_1 and O_2 are given, then the distance i.e. dissimilarity of the two objects is given by D (O_1,O_2)
- For example if the objects are two equal-length time series, then the distance D ( ) can be Euclidean distance i.e. the root of the sum of squared differences.