ALGORITHMS FOR INFORMATION RETRIEVAL AND INTELLIGENT WEB

Multimedia and Multimodal Information Retrieval

Dr. C P Chavan

Department of Computer Science Engineering

ALGORITHMS FOR INFORMATION RETRIEVAL AND INTELLIGENT WEB

PES UNIVERSITY

Multimedia and Multimodal Information Retrieval - Introduction, Requirements, Applications, and Challenges

Introduction

- With emergence of Web 2.0 and lots of content available today, Web, mobile access, and digital television has boosted the production and consumption of audio-visual materials, making the Web a truly multimedia platform.
- This Unit will give us a concise overview of Multimedia Information Retrieval (MIR), the long-standing discipline at the base of audio-visual search engines, and connects the research challenges in this area to the objectives and research goals of Search Computing.
- The grand challenge of MIR is bridging the gap between queries and content: the former are either expressed by keywords, like in text search engines, or, by extension, with non-textual samples (e.g., an image or a piece of music). Unlike in text search engines, where the query has the same format of content and can be matched almost directly to it, query processing in MIR must fill an enormous gap

Introduction

- An MIR system can be seen as an infrastructure for governing two main processes:
  - the content process deals with metadata
  - the query process deals with user's query
- The link between the content process and the query process is represented by metadata, which encode the knowledge that the MIR system is able to extract from the media assets, index and use for answering queries.

Requirements

The requirements of an MIR system go beyond the problems faced in classical text based IR.

- **Opacity of Content**
- **Query Formulation Paradigm**
- **Relevance Computation**

Requirements

Opacity of Content

- When we consider a classic case of text IR, the query and content use the same medium. (Transparency of Content).
- But, MIR content is opaque.
- The knowledge necessary to verify if an item is relevant to a user's query is deeply embedded in it and must be extracted by means of a complex pre-processing.
- Example : Extracting speech transcriptions from a video.

Requirements

Query Formulation Paradigm

- In traditional engines, not only keywords but queries can be in the form of an analogy.
- Thus, content sample is similar to whatever is user searching for.
- But in case of MIR, content samples could be used as queries and those could be images, a piece of audio or a video fragment.

Requirements

Relevance Computation

- In text IR, relevance of documents to the user's query is computed as similarity scores. Which involves comparison between vectors of words in the query and document.
- In MIR, comparison has to be done on a wider variety of features – not only the medium of query and the content but also their

Applications

MIR applications requirements have been extensively addressed in the last three decades, both in the industrial and academic fields :

- **Architecture, Real estate and Design**
- **Broadcast media selection**
- **Cultural service**
- **Digital libraries**
- **Education**
- **Entertainment**
- **Journalism**
- **Investigation**
- **Multimedia directory**
- **Multimedia editing**
- **Remote sensing**
- **Social**
- **Surveillance**

Challenges

MIR systems operate on a very heterogeneous spectrum of content, ranging from home-made content created by users (like vlogs, reviews, etc) to high value premium productions (like movies, trailers, webseries).

The challenges in the design of an MIR solution, by following the lifecycle of multimedia content is as follows:

- Entrance into the system (**Acquisition)**
- Preparation of analysis (**Normalization)**
- Extraction of metadata for building a search engine
- Constructing indexes **(Indexing)**
- Processing of a query (**Querying**)
- Presentation of results **(Browsing)**

Challenges

Content Acquisition

- Multimedia content can be acquired in a way similar to document acquisition:
  - By crawling the Web or local media repositories.
  - By user's contribution or syndicated contribution from content aggregators.
    - Directly from production devices directly connected to the system

- Just like different sources, the size of media files also varies make the content ingestion task more complicated, e.g., because the probability of download failures increases, the cost of storing duplicates or near duplicates becomes less affordable, and the presence of DRM issues on the downloaded content is more frequent.

Challenges

- The ability of the content ingestion subsystem to maintain or even improve the intrinsic quality of the downloaded digital assets—for example, by obtaining them at the best resolution feasible given the bandwidth constraints and preserving all the available metadata associated with them—is crucial.

- Building scalable and intelligent content acquisition systems, which could ingest content exploiting different communication protocols and acquisition devices, decide the optimal resolution in case alternative representations are available, detect and discard duplicates as early as possible, respect DRM issues, and enrich the raw media asset with the maximum amount of metadata that could be found inside or around it.

Challenges

PES UNIVERSITY

Content Normalization

- Multimedia content needs a more sophisticated preprocessing phase, because the elements to be indexed (called “features” or “annotations”) are numerical and textual metadata that need to be extracted from raw content by means of complex algorithms.
- Due to the variety of multimedia encoding formats, prior to processing content for metadata extraction, it is necessary to submit it to a normalization step, with a twofold purpose:
    - translating the source media items represented in different native formats into a common representation format. (e.g. MPEG4 for video files)
    - producing alternative variants of native content items, e.g., to provide freebies (free sample copies) of copyrighted elements or low resolution copies for distribution on mobile or low-bandwidth delivery channels. (e.g. making a 3GP version of video files)

Challenges

Content Indexing and Analysis

- After the normalization step, a multimedia collection has to be processed in order to make the knowledge embedded in it available for querying, which requires building the internal indexes of the search engine.
- Indexes are a concise representation of the content of an object collection, constructed out of the features extracted from it; the features used to build the indexes must be both sufficiently representative of the content and compact to optimize storage and retrieval.
- Features are traditionally grouped into two categories:
  - **Low level features**: concisely describe physical or perceptual properties of a media element (e.g., the colour or edge histogram of an image).
  - **High level features**: domain concepts characterizing the content (e.g., extracted objects and their properties, geographical references, etc.).

Challenges

- As in text, where the retrieved keywords can be highlighted in the source document, also in MIR there is the need of locating the occurrences of matches between the user's query and the content. Such requirement implies that features must be extracted from a time continuous medium, and that the coordinates in space and time of their occurrence must be extracted as well
- Content analysis and indexing are the prominent research problem of MIR, as the quality of the search engine depends on the precision at which the extracted metadata describe the content of a media asset.

Challenges

PES UNIVERSITY

Content Querying

- The expression of the user's information need allows for alternative query representation formats and matching semantics.
- Textual: one or more keywords, to be matched against textual metadata extracted from multimedia content.
- Mono-media: a content sample in a single media (e.g., an image, a piece of audio) to be matched against an item of the same kind (e.g., query by music or image similarity, query by humming) or of a different medium (e.g., finding the movies whose soundtrack is similar to an input audio file).
- Multi-media: a content sample in a composite medium, e.g., a video file to be matched using audio similarity, image similarity, or a combination of both.

Challenges

- Another implication of non-textual queries is the need for the MIR architecture to coordinate query processing across multiple dedicated search engines.

- The grand challenge of MIR query processing is in part the same as for textual IR: retrieving the media objects more relevant to the user's query with high precision and recall. MIR adds the specific problem of content-based queries, which demand suitable architectures for analysing a query content sample on the fly and matching its features to those stored in the indexes

Challenges

PES UNIVERSITY

Content Browsing

- Unlike data retrieval queries (such as SQL or XPATH queries), IR queries are approximate and thus results are presented in order of relevance, and often in a number that exceeds the user's possibility of selection.
- The interface must also permit users to quickly inspect continuous media and locate the exact point where a match has occurred. This can be done in many ways, e.g., by means of annotated time bars that permit one to jump into a video where a match occurs, with VCR-like commands, and so on.
- The challenge of MIR interfaces is devising effective renditions (visual, but also aural) that could convey both the global characteristics of the result set (e.g., the similarity distribution across a result collection) and the local features of an individual result item that justify the query match.

Techniques for Content Processing

- **Transformation**: This kind of operation converts the format of media items, for making the subsequent analysis steps more efficient or effective. For instance, a video transformer can modify an MPEG2 movie file to a format more suitable for the adopted analysis technologies (e.g., MPEG); likewise, an audio converter can transform music tracks encoded in MP3 to WAV, to eliminate compression and make content analysis simpler and more accurate.

- **Feature Extraction:** calculates low-level representations of media contents, i.e. feature vectors, in order to derive a compact, yet descriptive, representation of a pattern of interest. Such representation can be used to enable content based search, or as input for classification tasks. Examples of visual features for images are colour, texture, shape, etc.; examples of aural features for music contents are loudness, pitch, tone (brightness and bandwidth).

Techniques for Content Processing

- **Classification**: assigns conceptual labels to content elements by analyzing their raw features; the techniques required to perform this operations are commonly known as machine learning. For instance, an image classifier can assign to image files annotations expressing the subject of the pictures (e.g., mountains, city, sky, sea, people, etc.), while an audio file can be analyzed in order to discriminate segments containing speech from the ones containing music.

Techniques for Content Processing

PES UNIVERSITY

Techniques for Audio Analysis

Techniques	Purpose
Audio Segmentation	split audio track according to the nature of its content For example: a file can be segment according to the presence of noise, music, speech, etc.
Audio event identification	identify the presence of events like gunshots and scream in an audio track.
Music genre identification	to identify the genre (e.g., rock, pop, jazz, etc.) or the mood of a song
Speech recognition	to convert words spoken in an audio file into text. Speech recognition is often associated with Speaker identification, that is to assign an input speech signal to one person of a known group

Techniques for Content Processing

PES UNIVERSITY

Techniques for Image Analysis

Techniques	Purpose
Semantic Concept Extraction	the process of associating high-level concepts (like sky, ground, water, buildings, etc.) to pictures.
Optical character recognition (OCR)	to translate images of handwritten, typewritten or printed text into an editable text.
Face recognition and identification	to recognize the presence of a human face in an image, possibly identifying its owner.
Object detection and identification	to detect and possibly identify the presence of a known object in the picture.

Techniques for Content Processing

PES UNIVERSITY

Techniques for Video Analysis

Techniques	Purpose
Scene detection	detection of scenes in a video clip; a scene is one of the subdivisions of a play in which the setting is fixed, or that presents continuous action in one place
Video text detection and segmentation	to detect and segment text in videos in order to apply image OCR techniques.
Video summarization	to create a shorter version of a video by picking important segments from the original.
Shot detection	detection of transitions between shots. Performed by means of Keyframe segmentation algorithms that segment a video track according to the key frames produced by the compression algorithm.

Techniques for Content Processing

Having come across a variety of these methods, how do we leverage them for an effective multimodal analysis?

The answer lies in arbitrary combinations of transformation, feature extraction techniques and classification operation that result in several analysis algorithms.

The techniques shown in the previous tables can be used in isolation, to extract different features from an item. Since the corresponding algorithms are probabilistic, each extracted feature is associated with a confidence value that denotes the probability that item X contain feature Y. To increase the confidence in the detection, different analysis techniques can be used jointly to reinforce each other.

Examples of MIR Query Languages

- The last two decades have witnessed to a lot of efforts in the definition of more expressive and structured query languages, designed specifically for multimedia retrieval.
- **POQL** : is a general purpose query language for object oriented multimedia databases exposing arbitrary data schema.
- **MuSQL**: is a music structured query language, composed of a schema definition sub-language and a data manipulation sub-language.
- One of the latest attempts in providing a unified language for MIR is represented by the MPEG Query Format (MPQF).

Examples of Commercial MIR System

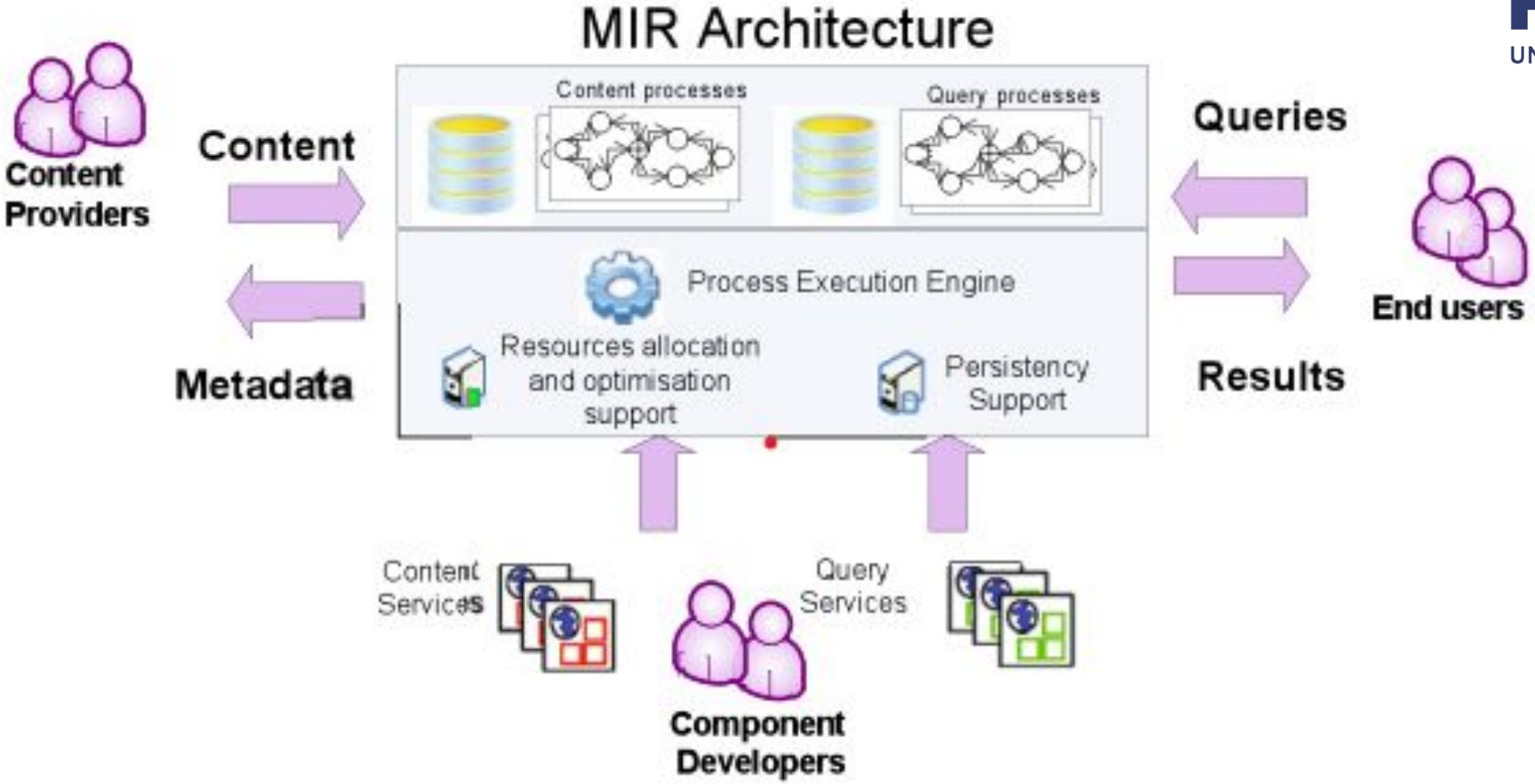
- **PHAROS** – Proposed by Italy and implemented collaboratively with other European countries.
- **VITALAS** – European commission information society
- **THESEUS** – Germany
- **Quaero, Voxalead** – France
- **Midomi (SoundHound), Google Images and Microsoft Bing, Tiltomo** - USA
- **Shazam** – UK (now owned by Apple)
- **SAPIR** – France, Italy, Israel
- **Blinkx** – UK (now known as RhythmOne and based in USA)

ALGORITHMS FOR INFORMATION RETRIEVAL AND INTELLIGENT WEB

PES UNIVERSITY

Multimedia and Multimodal Information Retrieval - Architecture, Metadata

MIR Architecture



MIR Architecture

- The architecture of a MIR system can be described as a platform for composing, verifying, and executing search processes, defined as complex workflows made of atomic blocks, called **search services**.

- A **search service** is a wrapper for any software component that embodies functionality relevant to a MIR solution

- At the core of the architecture there is a **Process Execution Engine** which is a runtime environment, optimized for the scalable enactment of data-intensive and computation-intensive workflows made of search services.

MIR Architecture

- The most important categories of MIR workflows are
  - **Content Processes**, which have the objective of acquiring multimedia content from external sources (e.g. from the user or a from video portal) and extracting features from it.
  - **Query Processes**, which have the objective of acquiring a user's information need and computing the best possible answer to it.

- . Accordingly, the most important categories of search services are
  - **content services**, which embody functionality relevant to content acquisition, analysis, enrichment, and adaptation; and
  - **query services**, which implements all the steps for answering a query and computing the ranked list of results.

MIR Architecture

Examples of content services can be:

- algorithms for extracting knowledge from media elements,
- transducers for modifying the encoding format of media files

Examples of query services are:

- query disambiguation services for inferring the meaning of ambiguous information needs, or
- social network analysis services for inferring the preferences of a user and
- personalizing the results of a user's query.

MIR Architecture - Content Process

- A content process aims at gathering multimedia content and at elaborating it to make it ready for information retrieval. A MIR platform may host multiple content processes, as required for elaborating content of different nature, in different domains, for different access devices, for different business goals, etc.
- The input to the process is twofold:
  - Multimedia Content (image, audio, video).
  - Information about the content, which may include
    - **publication metadata** (HTML, podcast, RSS, MediaRSS, MPEG7, etc.),
    - **quality information** (encoding, user's rating, owner's ratings, classification data),
    - **access rights** (DRM data, licensing, user's subscriptions), and
    - **network information** (type and capacity of the link between the MIR platform and the content source site - local disk-based, remote – LAN, WAN

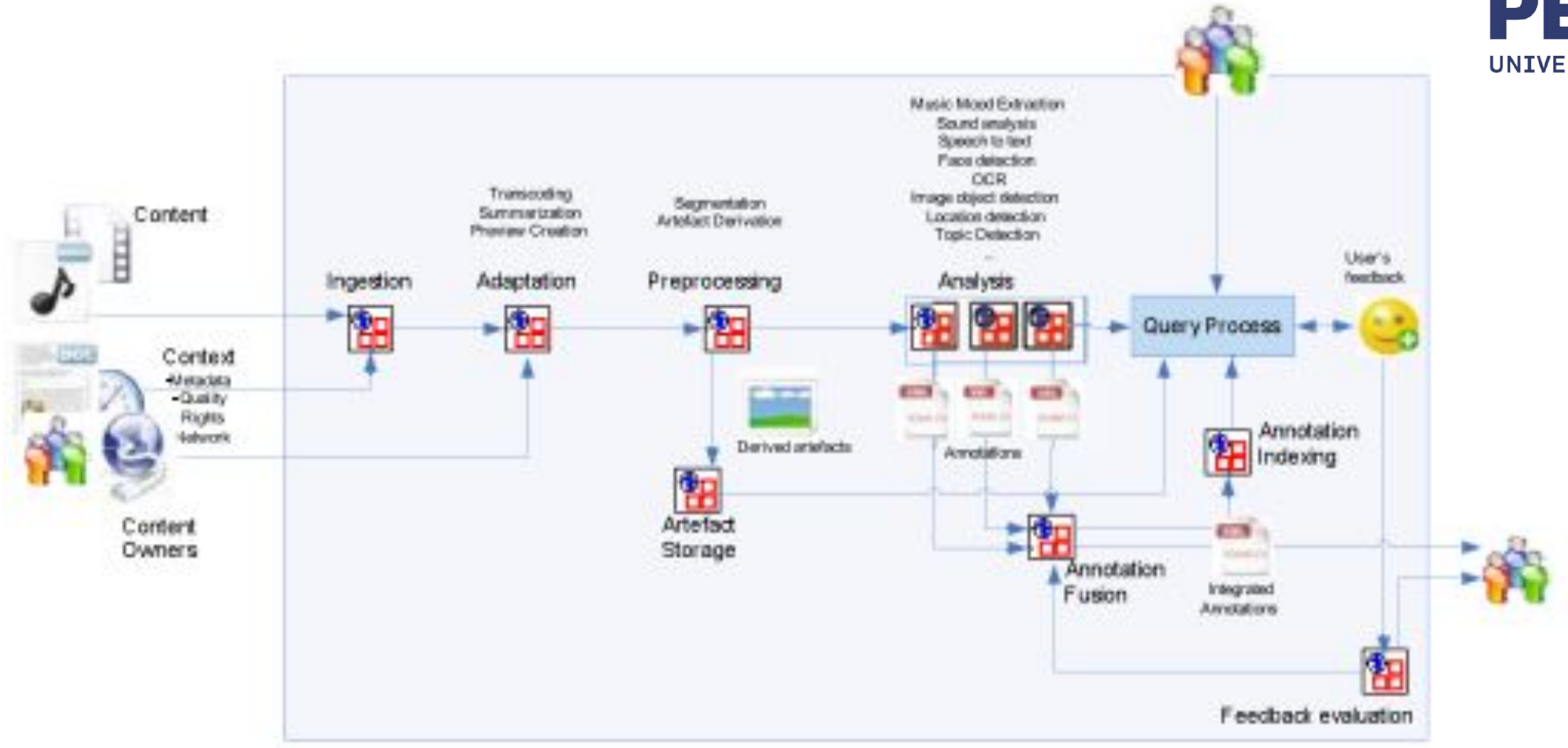
MIR Architecture - Content Process

- The output of the process is the textual representation of the metadata that capture the knowledge automatically extracted from the multimedia content via content processing operations.
- The calculated metadata are integrated with the metadata gathered by the content acquisition system, which are typically added to the content manually by the owner or by the Web users.

MIR Architecture - Content Process

- A content process can be designed so as to dynamically adapt to the external context as follows by analyzing:
  - content metadata
  - the access rights metadata
  - the geographical region where the content comes from
  - the content delivery modality

MIR Architecture - Content Process



MIR Architecture - Query Process

- A MIR Query process accepts in input information need and formulates the best possible answer from the content indexed in the MIR platform.
- The input of the query process is an information need, which can be a keyword or a content sample. The output is a result set, which contains information on the objects (typically content elements) that match the input query.
- A collateral source of input is the query context, which expresses additional circumstances about the information need, often implicit. Well-known examples of query context are: user preferences, past users' queries and their responses, access device, location, access rights, and so on.

MIR Architecture - Query Process

- Queries are classified as mono-modal, if they are represented in a single medium (e.g., a text keyword, a music fragment, an image) or multi-modal, if they are represented in more than on medium (e.g., a keyword AND an image).
- As for Search Computing, also in MIR queries can be classified as mono-domain, if they are addressed to a single search engine or multi-domain, if they target different, independent search services

MIR Architecture - Query Process

- Examples of queries in a MIR system

	Mono Domain	Multi Domain
Mono Modal	Find all the results that match a given keyword Find all the images similar to a given image	Find theatres playing movies acted by an actor having the voice similar to a given one
Multi Modal	Find all videos that contain a given keyword and that contain a person with a face similar to a given one.	Find all CDs in Amazon with a cover similar to a given image.

Metadata

- Metadata are textual descriptions that accompany a content element; they can range in quantity and quality, from no description (e.g., Webcam content) to multilingual data (e.g., closed captions and production metadata of motion pictures).
- Metadata can be found:
  - Embedded within content (e.g., video close captions or Exchangeable image file format (EXIF) data embedded in images).
  - In surrounding Web pages or links (e.g., HTML content, link anchors, etc.).
  - In domain-specific databases (e.g., IMDB for feature films).
  - In ontologies (e.g., like those listed in the DAML Ontology Library)

Metadata

- The current state of the practice in content management presents a number of metadata vocabularies dealing with the description of multimedia content.
- Many vocabularies allow the description of high-level (e.g., title, description) or low-level features (e.g., colour histogram, file format), while some enable the representation of administrative information (e.g., copyright management, authors, date).
- In a MIR system, the adoption of a specific metadata vocabulary depends on its intended usage, especially for what concerns the type of content to describe.

Metadata

Main metadata format categories:

MPEG7

- an XML vocabulary that represents the attempt from ISO to standardize a core set of audio-visual features and structures of descriptors and their (spatial/temporal) relationships.
- MPEG-7 results in an elaborate and complex standard that merges both high-level and low-level features, with multiple ways of structuring annotations.
- MPEG7 is also extensible, so to allow the definition of application-based or domain-based metadata.

Metadata

PES UNIVERSITY

Dublin Core

- a 15-element metadata vocabulary (created by domain experts in the field of digital libraries) intended to facilitate discovery of electronic resources, with no fundamental restriction on the resource type.
- Dublin Core holds just a small set of high-level metadata and relations (e.g. title, creator, language, etc…), but its simplicity made it a common annotation scheme across different domains.
- It can be encoded using different concrete syntaxes, e.g., in plain text, XML or RDF.

Metadata

MXF (Material Exchange Format)

- is an open file format that wraps video, audio, and other bit streams (called "essences"), aimed at the interchange of audiovisual material, along with associated data and metadata, in devices ranging from cameras and video recorders to computer systems for various applications used in the television production chain.
- MXF metadata address both high-level and administrative information, like the file structure, key words or titles, subtitles, editing notes, location, etc.
- Though it offers a complete vocabulary, MXF has been intended primarily as an exchange format for audio and video rather than a description format for metadata storage and retrieval

Metadata

EXIF (Exchangeable Image File Format)

- a vocabulary adopted by digital camera manufacturers to encode high-level metadata like date and time information, the image title and description, the camera settings (e.g., exposure time, flash), the image data structure (e.g., height, width, resolution), a preview thumbnail, etc.
- By being embedded in picture raw contents, EXIF metadata is now a de-facto standard for image management software; to support extensibility, EXIF enables the definition of custom, manufacturer-dependent additional terms.

Metadata

ID3

- a tagging system that enriches audio files by embedding metadata information.
- ID3 includes a big set of high-level (such as title, artist, album, genre) and administrative information (e.g. the license, ownership, recording dates), but a very small set of low-level information (e.g. BPM).
- ID3 is a worldwide standard for audio metadata, adopted in a wide set of applications and hardware devices. However, ID3 vocabulary is fixed, thus hindering its extensibility and usage as format for low-level features

Metadata

- The techniques applied for analysing content are application dependent, and relate both with the nature of the processed items and with the aim of the applications.
- MIR systems deal with more complex media formats, like audio, video and images, and therefore require a more articulated analysis process to produce the metadata needed for indexing.
- The operations that constitute the MIR content process can be roughly classified in three macro categories: **transformation, feature extraction,** and **classification**, based on the stage at which they occur in the analysis process