Web Information Retrieval

Multimedia Search

SOSE2023

Frank Hopfgartner, Stefania Zourlidou
Institute for Web Science and Technologies

Recapitulation

- HCI
  - Query Specification
  - Query Reformulation
  - Retrieval Results Display
  - Visualizing Research Results
  - Design & Evaluation

Objectives of this Lecture

universität koblenz
weiter:denken

- Multimedia Information Retrieval (MIR)
  - Motivations and challenges of multimedia search
  - MIR architecture
  - MIR metadata
  - MIR content processing
  - Research projects and commercial systems

universität koblenz
weiter:denken

1. Motivations and Challenges of Multimedia Search

Multimedia Information Retrieval (MIR)

universität koblenz
weiter:denken

- It is a field of study that focuses on the development and application of techniques and algorithms to effectively search, retrieve, and analyze multimedia data, such as images, videos, and audio
- It aims to bridge the semantic gap between low-level visual or auditory features and high-level semantics, enabling users to find relevant multimedia content based on their information needs
- It encompasses various aspects, including content-based retrieval, cross-modal retrieval, metadata extraction, indexing, relevance feedback, and user interaction
- It plays a crucial role in harnessing the vast amount of multimedia data available today, facilitating efficient search, organization, and utilization of multimedia content across diverse domains and applications

MIR vs. IR

- IR
  - focuses on textual information (web pages, articles, or books)
  - relies on text-based indexing (keyword indexing, inverted indexes, and relevance ranking algorithms based on textual content)
  - emphasizes textual relevance: relevance assessment in IR is primarily based on the matching of keywords and textual similarity between query and document

MIR vs. IR

- MIR
  - handles various types of media (e.g. images, videos, audio)
  - explores content-based retrieval: MIR focuses on extracting features from multimedia data, such as color, texture, shape, or audio patterns, to enable content-based search and retrieval
  - addresses the semantic gap: MIR aims to bridge the gap between low-level visual or auditory features and high-level semantics, enabling retrieval based on the content itself
  - incorporates cross-modal retrieval: MIR includes techniques for searching across different modalities, such as finding images based on textual queries or retrieving audio based on visual inputs

Motivations

- The growth of digital content has reached impressive rates in the last decade
- The convergence of the fixed-network Web, mobile access, and digital television has boosted the production and consumption of audio–visual mater
- This trend challenges search due to
  - the more complex nature of multimedia with respect to text in all the phases of the search process
    - from the expression of the user's information need
    - to the indexing of content and the processing of queries by search engines

Some examples

- Finding the title and author of a song recorded with one's mobile in a crowded disco
- Locating news clips containing interviews to President Obama and accessing the exact point where the Health Insurance Reform is discussed
- Finding a song matching in mood the images to be placed in a slideshow

These are only a few examples of what MIR is about: satisfying a user's information need that spans across multiple media, which can itself be expressed using more than one medium.

Requirements

- Opacity of Content: 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 preprocessing task (e.g., extracting speech text from a video)
- Query Formulation Paradigm: The user's information need can be formulated not only by means of keywords, as in traditional search engines, but also by analogy, e.g., by providing a sample of content "similar" to what the user is searching for
- Relevance Computation: In MIR, the comparison must be done on a wide variety of features, characteristic not only of the specific medium in which the content and the query are expressed, but even of the application domain (e.g., two audio files can be deemed similar in a music similarity search context, but dissimilar in a topic-based search application)

MIR applications

- Digital libraries
  - mage catalogues, musical dictionaries, biomedical imaging catalogues, film, video, and radio archives
- E-commerce
  - personalized advertising, on-line catalogues, directories of e-shops
- Education
  - repositories of multimedia courses, multimedia search
- Journalism
  - searching speeches of a certain politician using his name/voice/face
- Social uses
  - dating services, podcasts

Challenges

universität koblenz
weiter:denken

- Heterogeneity and complexity of multimedia data
- Semantic gap between low-level features and high-level semantics
- Scalability and efficiency of retrieval systems

- More specifically…

Challenge 1: content acquisition

- In MIR content is acquired from many sources and in multiple ways:
  - from media-production devices (scanners, digital cameras, smartphones, etc.)
  - by crawling the Web or local repositories
  - by the user’s contribution
  - by syndicated contribution from content aggregators
  - via broadcast capture (from air/cable/satellite broadcast, IPTV, Internet TV multicast, etc.)

Challenge 1: content acquisition & metadata

- Metadata are textual descriptions that accompany a content element; they can range in quantity and quality, from no description (e.g., Web cam content) to multilingual data (e.g., closed captions and production metadata of motion pictures).
- Some examples of where metadata can be found
  1. embedded within content (e.g., closed captions)
  2. in surrounding Web pages or links (HTML content, link anchors, etc.)
  3. in domain-specific databases (e.g., IMDb for feature films)

Challenge 2: content normalization

- In conventional text-based search engines
  - the context undergoes a series of operations to make it suitable for indexing (parsing, tokenization, lemmatization, and stemming)
  - the index consists of words
- Multimedia content
  - requires a more advanced preprocessing phase due to the nature of the elements that need to be indexed, referred to as "features" or "annotations"
  - these elements are a combination of numerical and textual metadata that must be extracted from the raw content using sophisticated algorithms

Challenge 3: content indexing

- Multimedia content cannot be indexed as is, but features must be extracted from it
- Such features must be both sufficiently representative of the content and compact to optimize storage and retrieval
- Features can be grouped into two categories
  - Low-Level Features: Concisely describe physical or perceptual properties of a media element (the color or edge histogram of an image)
  - High-Level Features: Domain concepts characterizing the content (extracted objects and their properties, geographical references, content categorizations, etc.)
- Feature detection may even require a change of medium with respect to the original file, as in the case of speech-to-text transcription

Challenge 4: content querying

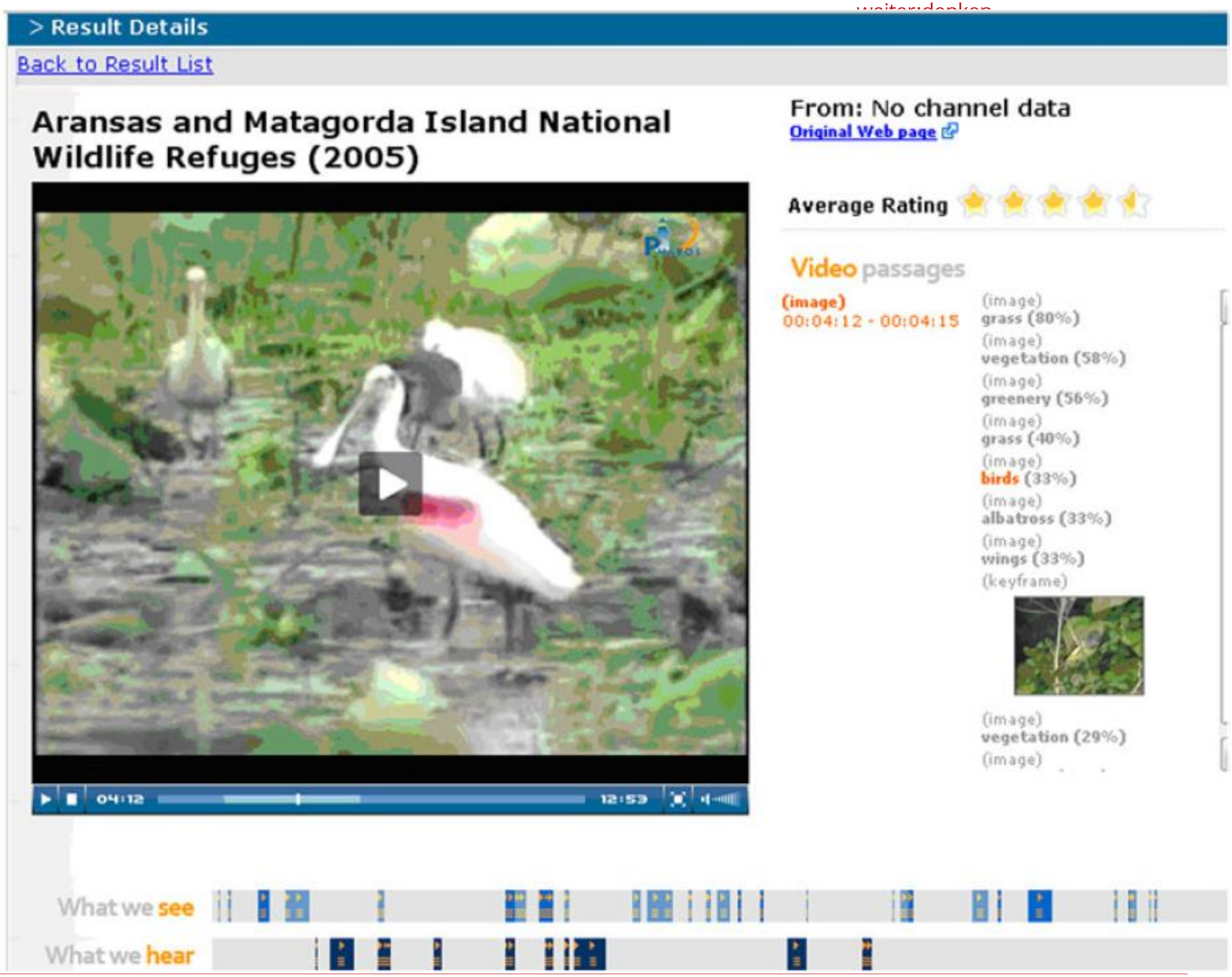
- In MIR, the expression of the user's information need allows for alternative query representation formats and matching semantics
- Examples of queries can be
  - Textual: One or more keywords, to be matched against textual metadata extracted from multimedia content
  - Monomedia: A content sample in 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)
  - Multimedia: 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

Challenge 5: content browsing

- In MIR applications, understanding if a content element is relevant has additional challenges. For example
  - summarizing a video may be done in several alternative ways: by means of textual metadata, with a selection of key frames, with a preview (e.g., the first 10 seconds), or even by means of another correlated item (the free trailer of a copyrighted feature film)
  - 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
    - 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

Challenge 5: content browsing

- The PHAROS multimedia search platform: accessing video results of a query
- Two time bars (labelled “what we hear”, “what we see”) allow one to locate the instant where the matches for a query occurring the video frames and in the audio, inspect the metadata that support the match, and jump directly to the point of interest

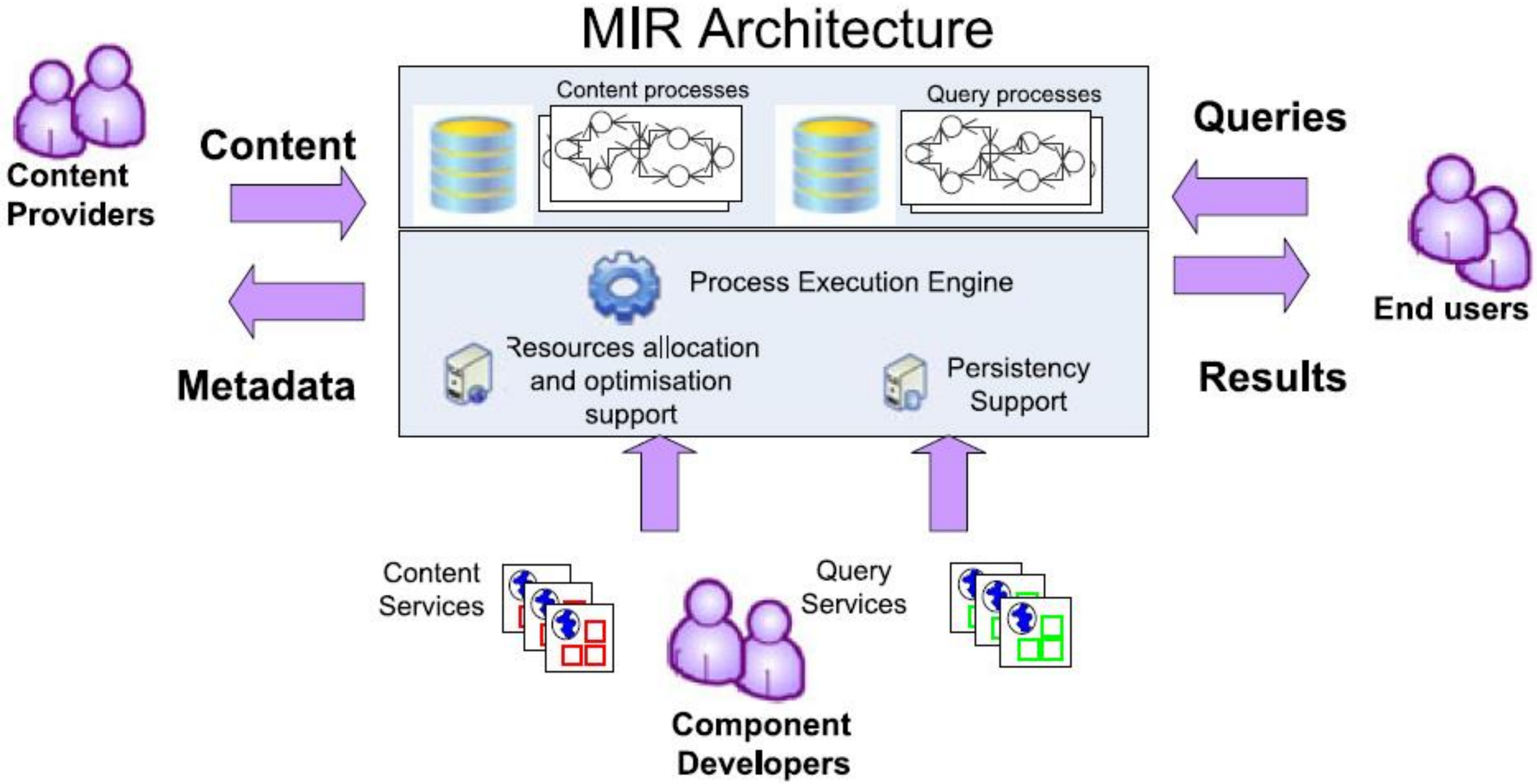


2. MIR architecture

MIR architecture overview

- Components and their roles in the MIR architecture
- Interaction between components
- Benefits of using an architecture for MIR

MIR architecture overview



Reference architecture of a MIR system

MIR architecture

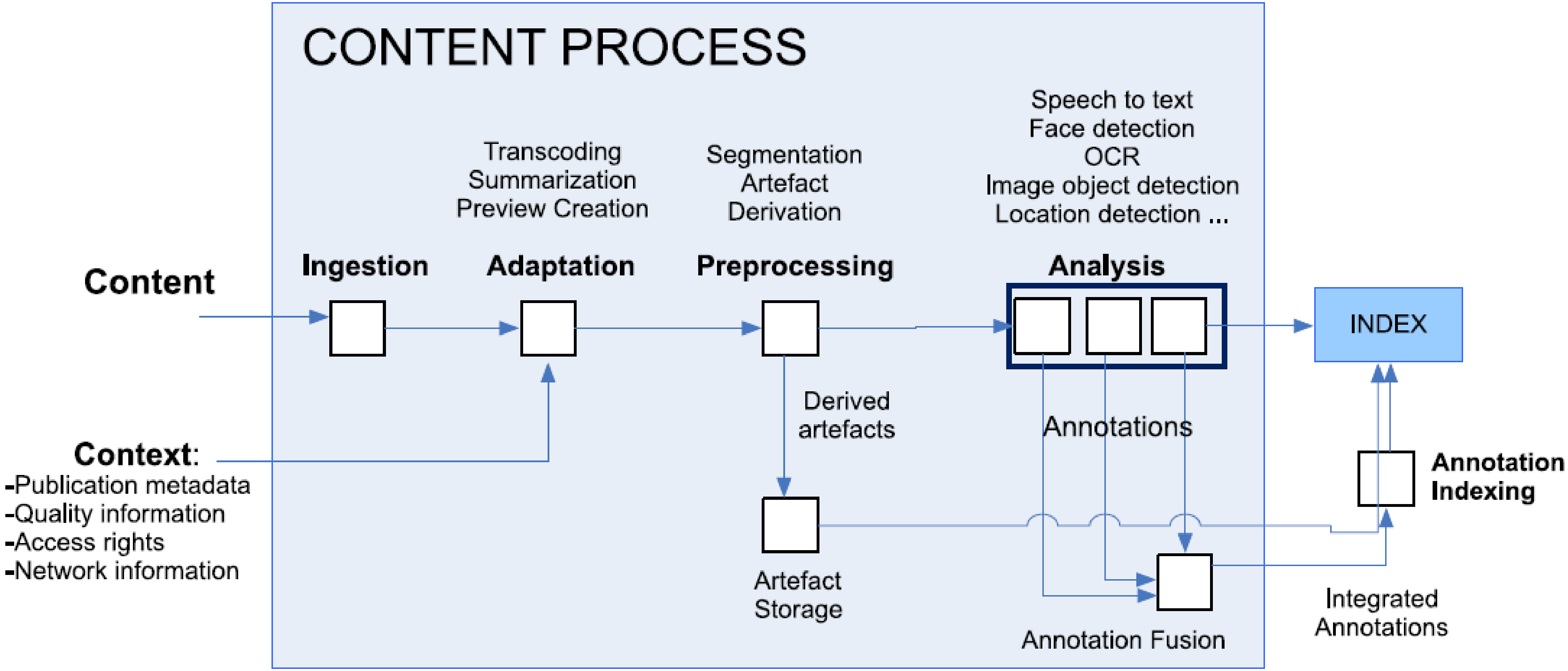
- It can be described in very general terms as a platform for composing, verifying, and executing search processes, defined as complex workflows made of atomic blocks, called search services
- The most important categories of MIR processes are
  - content processes, which have the objective of acquiring multimedia content from external sources (the user, a video portal, a TV channel) and extracting features from it; and
  - the query processes, which have the objective of acquiring a user's information need and computing the best possible answer to it
- The most important categories of search services are
  - content services, which embody functionality relevant to content acquisition, analysis, enrichment, and adaptation
  - query services, which implements all the steps for answering a query and computing the ranked list of results

Content Process

The typical steps of a content process comprise

- the ingestion of the content into the platform (via crawling, upload)
- the adaptation of the content (e.g., the transcoding to an internal format)
- its preprocessing (e.g, video segmentation, the extraction of derivatives, such as thumbnails and summaries)
- the proper analysis, which is a possibly complex workflow of analysis operators, each extracting a given low-level feature of high-level annotation
  - annotations derived independently can be integrated by a fusion step, to reinforce the confidence in the extracted knowledge, and then indexed in the annotation indexing step

Content Process



Example of a MIR content process

Query Process

- The input of the query process is an information need, which can be a keyword and/or a non-textual element, such as a content sample
- A collateral source of input is the query context, which expresses additional circumstances, often implicit, about the information need
  - Well-known examples of query context are user preferences, past user queries and their responses, access device, location, access rights, and so on
- The query is first acquired, which may require the multimedia part to be subjected to a content processing step for transforming the multimedia object into a tractable representation (such as a feature vector or a set of high-level annotations)

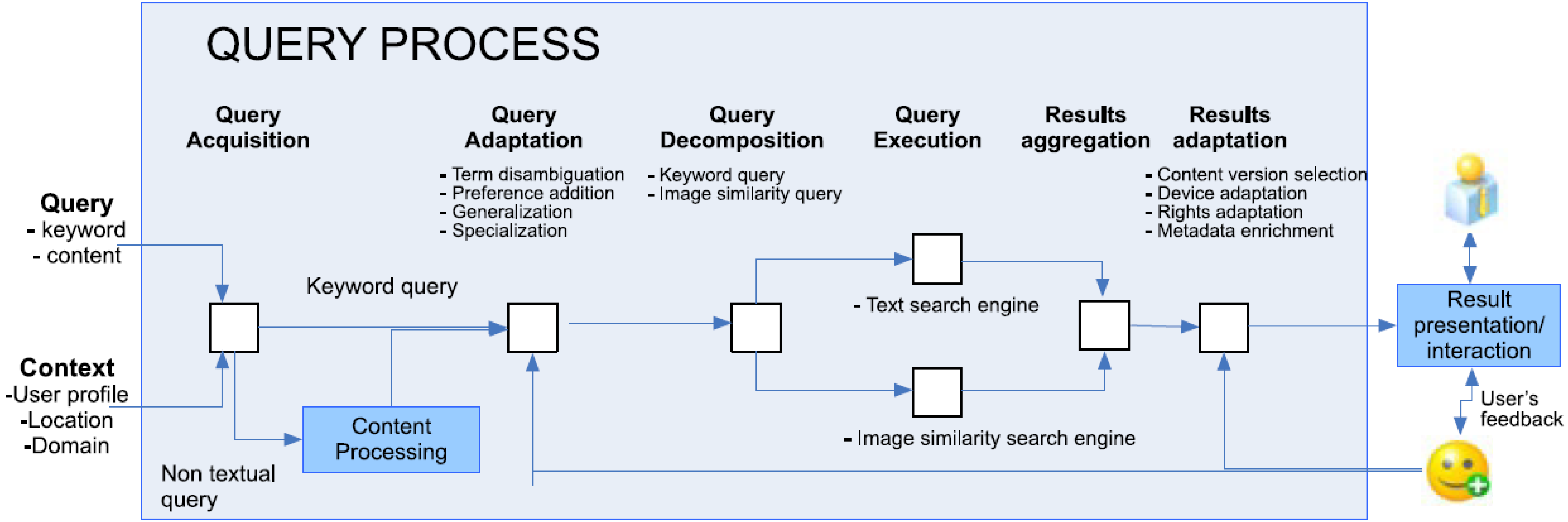
Query Process

- The input of the query process is an information need, which can be a keyword and/or a non-textual element, such as a content sample
- A collateral source of input is the query context, which expresses additional circumstances, often implicit, about the information need
  - Well-known examples of query context are user preferences, past user queries and their responses, access device, location, access rights, and so on
- The query is first acquired, which may require the multimedia part to be subjected to a content processing step for transforming the multimedia object into a tractable representation (such as a feature vector or a set of high-level annotations)

Query Process

- Next adaptation is performed (for example, by disambiguating terms based on user's preferences or on context information). In this stage, the query context is used to adapt the query process, e.g., to expand the original information need of the user with additional keywords reflecting her preferences, or disambiguating a query term based on the application domain where the query process is embedded
- Query decomposition fragments the query into its constituents (e.g, the original keyword plus feature vectors extracted from the sample image), so as to enable query execution by the appropriate search engines
- The output of the query process is a result set, which contains information on the retrieved objects that match the input query. The description of the objects in the result set can be enriched with metadata coming from sources external to the MIR platform, e.g., additional metadata on a movie taken from IMDb

Query Process



Example of a MIR query process

Query Process

Queries are classified as

- monomodal, if they are represented in a single medium (a text keyword, a music fragment, an image)
- multimodal, if they are represented in more than one medium (a keyword AND an image).

Queries can also be classified as

- mono-domain, if they are addressed to a single search engine, e.g., a general-purpose image search engine like Google Images or a special-purpose search service like Empora garments search
- multi-domain, if they exploit different independent search services, e.g., a face search service like Facesaerch and a video search service like Blinkx

3. MIR metadata

MIR metadata

- The current state of the practice in content management presents a number of metadata vocabularies dealing with the description of multimedia content
  - we list a set of common vocabularies currently adopted in MIR applications
  - the list is not intended to be complete, but rather to show how the portability, completeness, and extensibility of a metadata format can affect its usage in MIR systems

MPEG-7

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

MXF (Material eXchange Format)

- It is an open file format, aimed at the interchange of audio–visual material, along with associated data and metadata, for various applications used in the television production chain
- MXF metadata address both high-level and administrative information, like the file structure, keywords 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

Exchangeable Image File Format (Exif)

- It is 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 (height, width, resolution), a preview thumbnail, etc
- By being embedded in picture raw content, 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

ID3

- It is a tagging system that enriches audio files by embedding metadata information
- It 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)
- It is a worldwide standard for audio metadata, adopted in a wide set of applications and hardware devices

3. MIR content processing

MIR content processing

- Content processing is the activity performed over a content item with the aim of creating a representation suitable for indexing and retrieval purposes.
- The way contents are processed is application dependent, as it relates to the nature of the processed items.
- IR systems typically deal with textual contents. MIR systems are not an exception, as information is often represented in a textual format. Therefore, textual processing is a common activity in MIR applications, and it exploits the same standard operations for text analysis already discussed for textual IR.

MIR content processing

- MIR systems, with respect to purely textual IR systems, must also process audio, video, or images in order to produce annotations, which requires specialized operations
- Mono-annotation analysis is defined as an analysis operation where a single combination of file type and content type (e.g., the audio track of a video file) is represented by a single annotations set
- Multi-annotation analysis provides multiple view-points over the same content, in order to produce more descriptive annotations: for instance, the audio track of a movie can be analyzed first to identify the speaking actors and then to segment it according to speakers' turns

Multiple annotations

- Multiple annotations can be considered separately, as independent descriptions of the analyzed content, or jointly, in order, for instance, to raise the overall confidence on the produced metadata
- In the former case, the annotations associated with the managed contents are defined as monomodal; otherwise, we talk about multimodal annotations

Multiple annotations: an example

Suppose a movie file

- the fact that in a single scene both the face and the voice of a person are identified as belonging to an actor “X” can be considered as a correlated event, in order to describe the scene as “scene where actor X appears” with a high confidence.
- Multimodality is typically achieved by means of annotation fusion techniques
  - media processing operations are probabilistic processes, where the result is characterized by a confidence value
  - multiple features extracted from media data can be fused to yield more robust classification detection,
    - e.g., multiple content segmentation techniques (e.g., shot detection and speaker’s turn segmentation) can be combined in order to achieve better video splitting
    - voice identification and face identification techniques can be fused in order to obtain better person identification

Multimedia processing operations

- Transformation
  - To convert the format of media items
- Feature extraction
  - To calculate low-level representations of media contents
- Classification
  - To extract and assign conceptual labels to content elements by analyzing their raw representations; the techniques required to perform this operation are commonly known as machine learning

4. Research projects and commercial systems

Research projects

- PHAROS (Platform for searcH of Audiovisual Resources across Online Spaces)
- THESEUS is a German research program aimed at developing a new Internet-based infrastructure to better exploit the knowledge available on the Internet
- ALVIS is an FP6 founded open source project aiming at developing open source semantic search engines in P2P-distributed architecture
- I-SEARCH: a EU-funded project has produced a unified framework for multimodal content indexing, sharing, search and retrieval, able to handle specific types of multimedia and multimodal content (text, 2D image, sketch, video, 3D objects and audio) alongside with real-world information, which can be used as queries for retrieving correlated content of any of the aforementioned types

Commercial systems

- Midomi applies audio processing technologies to offer a music search engine. The interface allows users to upload voice recordings of songs and then query such music files by humming or whistling
- Shazam, a commercial music search engine that enables users to identify tunes using their mobile phone. They can record a sample of a few seconds from any source (even with bad sound quality), and the system returns the identified song with the necessary details: artist, title, album, etc.
- Voxalead is an audio search technology demonstrator implemented by Exalead to search in TV news, radio news, and VOD programs by content. The system uses a speech-to-text transcription module and transcribes political speeches in several languages
- Blinkx is a search engine supporting keyword queries on both videos and audio streams. Blinkx, like Voxalead, incorporates speech recognition to match the text query to the video or audio speech content

universität koblenz
weiter:denken

5. Summary

Summary

- Motivations and challenges
- MIR architecture
- MIR metadata
- MIR content processing
- Examples of MIR research projects and commercial systems

References

[1] https://olat.vcrp.de/auth/RepositoryEntry/4071063853

[2] https://nlp.stanford.edu/IR-book/information-retrieval-book.html C. D. Manning, P. Raghavan and H. Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

[3] S. Ceri, A. Bozzon, M. Brambilla, E. Fraternali, S. Quarteroni, Web Information Retrieval, Springer-Verlag Berlin Heidelberg, 2013.

- Chapter 13, Multimedia Search, available in OLAT.