6 – information retrieval

Based on the *Amsterdam Drugport* scenario we can identify the need for information retrieval capabilities for images, text documents, audio and video. Before we delve intoany of the media types, we discuss an information system architecture that allows for a uniform approach to a variety of media types. After discussing information retrieval and content annotation for the various media types, we elaborate on the uniform approach by defining the notion of *media abstraction*.

information retrieval

- multimedia scenarios
- information system architecture
- images
- documents
- audio, video
- media abstractions

You may leave out the audio media type, since the material discussed there is not immediately relavant to the scenario sketched. The section on feature extraction might be discussed seperately or in combination with the audio 'research directions'.

0.1 scenarios

Multimedia is not only for entertainment. Many human activities, for example medical diagnosis or scientific research, make use of multimedia information. To get an idea about what is involved in multimedia information retrieval look at the following scenario, adapted from [MMDBMS],

Amsterdam Drugport

> *Amsterdam is an international centre of traffic and trade. It is renowned for its culture and liberal attitude, and attracts tourists from various ages, including young tourists that are attracted by the availability of soft drugs. Soft drugs may be obtained at so-called coffeeshops, and the possession of limited amounts of soft drugs is being tolerated by the authories.*
>
> *The European Community, however, has expressed their concern that Amsterdam is the centre of an international criminal drug operation. Combining national and international police units, a team is formed to start an exhaustive investigation, under the code name* Amsterdam Drugport.

Now, without bothering ourselves with all the logistics of such an operation, we may establish what sorts of information will be gathered during the investigation, and what support for (multimedia) storage and (multimedia) information retrieval must be available.

Information can come from a variety of sources. Some types of information may be gathered continuously, for example by video cameras monitoring parking lots, or banks. Some information is already available, for example photographs in a (legacy database) police archive. Also of relevance may be information about financial transactions, as stored in the database of a bank, or geographic information, to get insight in possible dug traffic routes.

From a perspective of information storage our informatio (data) include the following media types: images, from photos; video, from surveillance; audio, from interviews and phone tracks; documents,from forensic research and reports; handwriting, from notes and sketches; and structured data, from for example bank transactions.

We have to find a way to store all these data by developing a suitable multimedia information system architecture, as discussed in chapter 6. More importantly, however, we must provide access to the data (or the information space, if you will) so that the actual police investigation is effectively supported. So, what kind of queries can we expect? For example, to find out more about a murder which seems to be related to the drugs operation.

retrieval

- *image query* – all images with this person
- *audio query* – identity of speaker
- *text query* – all transactions with BANK Inc.
- *video query* – all segments with victim
- *complex queries* – convicted murderers with BANK transactions
- *heterogeneous queries* – photograph + murderer + transaction
- *complex heterogeneous queries* – *in contact with* + murderer + transaction

Apparently, we might have simple queries on each of the media types, for example to detect the identity of a voice on a telephone wiretap. But we may also have more complex queries, establishing for example the likelihood that a murderer known by the police is involved, or even *heterogeneous* queries (as they are called in [MMDBMS]), that establish a relation between information coming from multiple information sources. An example of the latter could be, *did the person on this photo have any transactions with that bank in the last three months*, or nore complex, *give me all the persons that have been in contact with the victim (as recorded on audio phonetaps, photographs, and video surveillance tapes) that have had transactions with that particular bank.*

I believe you'll have the picture by now. So what we are about to do is to investigate how querying on this variety of media types, that is images, text, audio and video, might be realized.

research directions– *information retrieval models*

Information retrieval research has quite a long history, with a focus on indexing text and developing efficient search algorithms. Nowadays, partly due to the wide-spread use of the web, research in information retrieval includes modeling, classification and clustering, system architectures, user interfaces, information visualisation, filtering, descriptive languages, etcetera. See [IR].

Information retrieval, according to [IR], deals with the representation, storage, organisation of, and access to information items. To see what is involved, imagine that we have a (user) query like:

> *find me the pages containg information on ...*

Then the goal of the information retrieval system is to retrieve information that is useful or relevant to the user, in other words: *information that satisfies the user's information need.*

Given an information repository, which may consist of web pages but also multimedia objects, the information retrieval system must extract syntactic and semantic information from these (information) items and use this to match the user's information need.

Effective information retrieval is determined by, on the one hand, the *user task* and, on the other hand, the *logical view* of the documents or media objects that constitute the information repository. As user tasks, we may distinguish between *retrieval* (by query) and *browsing* (by navigation). To obtain the relevant information in retrieval we generally apply *filtering*, which may also be regarded as a ranking based on the attributes considered most relevant.

The logical view of text documents generally amounts to a set of index terms characterizing theb document. To find relevant index terms, we may apply operations to the document, such as the elimination of stop words or text stemming. As you may easily see, full text provides the most complete logical view, whereas a small set of categories provides the most concise logical view. Generally, the user task will determine whether semantic richness or efficiency of search will be considered as more important when deciding on the obvious tradeoffs involved.

information retrieval models In [IR], a great variety of information retrieval models is described. For your understanding, an information retrieval model makes explicit how index terms are represented and how the index terms characterizing an information item are matched with a query.

When we limit ourselves to the classic models for search and filtering, we may distinguish between:

information retrieval models

- boolean or set-theoretic models
- vector or algebraic models
- probabilistic models

Boolean models typically allow for *yes/no* answers only. The have a set-theoretic basis, and include models based on fuzzy logic, which allow for somewhat more refined answers.

Vector models use algebraic operations on vectors of attribute terms to determine possible matches. The attributes that make up a vector must in principle be orthogonal. Attributes may be given a weight, or even be ignored. Much research has been done on how to find an optimal selection of attributes for a given information repository.

Probabilistic models include general inference networks, and belief networks based on Bayesan logic.

Although it is somewhat premature to compare these models with respect to their effectiveness in actual information etrieval tasks, there is, according to [IR], a general consensus that vector models will outperform the probabilistic models on general collections of text documents. How they will perform for arbitrary collections of multimedia objects might be an altogether different question!

Nevertheless, in the sections to follow we will focus primarily on generalized vector representations of multimedia objects. So, let's conclude with listing the advantages of vector models.

vector models

- attribute term weighting scheme improves performance
- partial matching strategy allows retrieval of approximate material
- metric distance allows for sorting according to degree of similarity

Reading the following sections, you will come to understand how to adopt an attribute weighting scheme, how to apply partial matching and how to define a suitable distance metric.

So, let me finish with posing a research issue: *How can you improve a particular information retrieval model or matching scheme by using a suitable method of knowledge representation and reasoning?* To give you a point of departure, look at the logic-based multimedia information retrieval system proposed in [Dolores].

0.2 architectural issues

The notion of *multimedia information system* is sufficiently generic to allow for a variety of realizations. Let's have a look at the issues involved.

As concerns the database (that is the storage and rerieval facilities), we may have to deal with homegrown solution, commercial third party databases or (even) legacy sources. To make things worse, we will usually want to deploy a combination of these.

With respect to the information architecture, we may wish for a common format (which unifies the various media types), but in practice we will often have to work with the native formats or be satisfied with a hybrid information architecture that uses both media abstractions and native media types such as images and video.

The notion of media abstraction, introduced in [MMDBMS], allows for uniform indexes over the multimedia information stored, and (as we will discuss in the next section) for query relaxation by employing hierarchical and equivalence relations.

Summarizing, for content organisation (which basically is the information architecture) we have the following options:

content organisation

- *autonomy* – index per media type
- *uniformity* – unified index
- *hybrid* – media indexes + unified index

In [MMDBMS], a clear preference is stated for a uniform approach, as expressed in the *Principle of Uniformity*:

Principle of Uniformity

> *... from a semantical point of view the* content *of a multimedia source is independent of the source itself, so we may use statements as meta data to provide a description of media objects.*

Naturally, there are some tradeoffs. In summary, [MMDBMS] claims that: metadata can be stored using standard relational and OO structures, and that manipulating metadata is easy, and moreover that feature extraction is straightforward. ¡/lp¿ Now consider, is feature extraction really so straightforward as suggested here? I would believe not. Certainly, media types can be processed and analysis algorithms can be executed. But will this result in meaningful annotations? Given the current state of the art, hardly so!

research directions– *the information retrieval cycle*

When considering an information system, we may proceed from a simple generic software architecture, consisting of:

software architecture

- a database of media object, supporting
- operations on media objects, and offering
- logical views on media objects

However, such a database-centered notion of information system seems not to do justice to the actual support and information system must provide when considering the full information retrieval cycle:

information retrieval cycle

1. specification of the user's information need
2. translation into query operations
3. search and retrieval of media objects
4. ranking according to likelihood or relevance
5. presentation of results and user feedback
6. resulting in a possibly modified query

When we look at older day information retrieval applications in libraries, we see more or less the automation of card catalogs, with search functionality for keywords and headings. Modern day versions of these systems, however, offer graphical userinterfaces, electronic forms and hypertext features.

When we look at the web and how it may support digital libraries, we see some dramatic changes with respect to the card catalogue type of applications.

We can now have access to a variety of sources of information, at low cost, including geographically distributed resources, due to improved networking. And, everybody is free to make information available, and what is worse, everybody seems to be doing so. Hence, the web is a continuously growing repository of information of a (very) heterogeneous kind.

Considering the web as an information retrieval system we may observe, following [IR], that:

- despite high interactivity, access is difficult;
- quick response is and will remain important!

So, we need better (user-centered) retrieval strategies to support the full information retrieval cycle. Let me (again) mention someof the relevant (research) topics:*user interfaces, information visualisation, user-profiling and navigation.*