Department of Informatics
Aristotle University of Thessaloniki
Fall-Winter 2016-2017

Multimedia Database Systems & Multimedia Information Retrieval

Introduction to (Multimedia) Information Retrieval

Outline

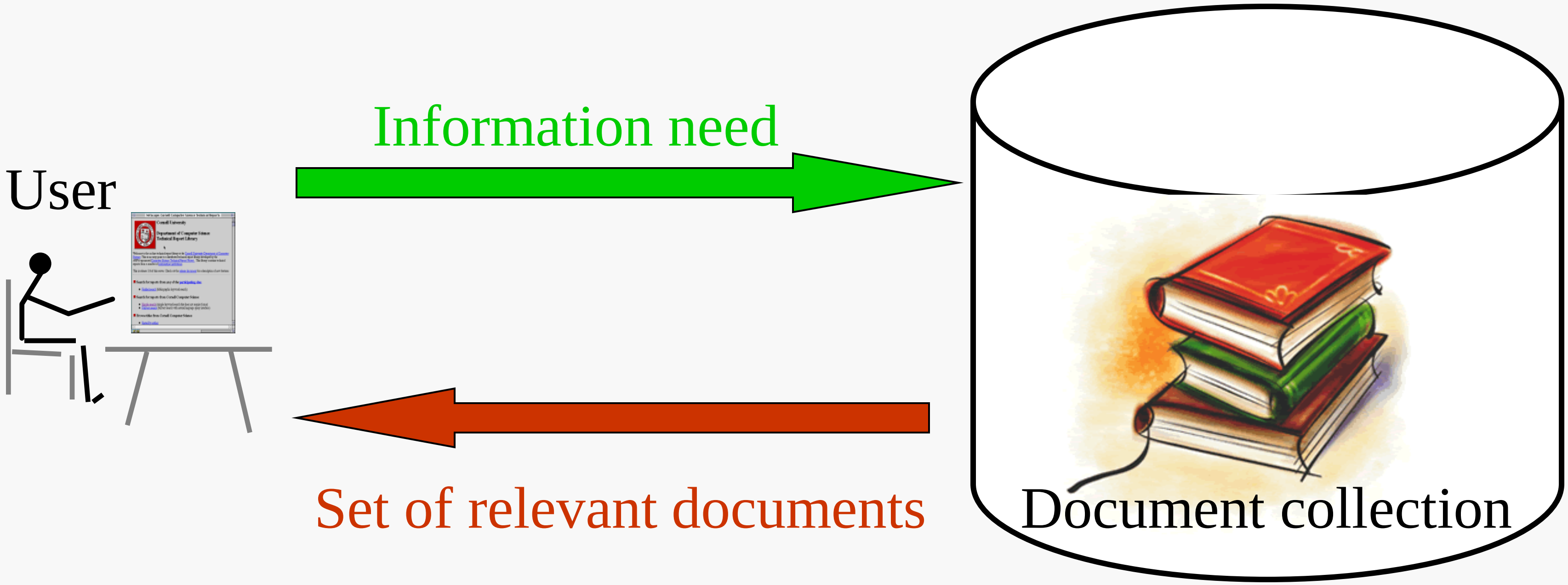
- Introduction to Information Retrieval (IR)
- Multimedia Information Retrieval (MIR) Motivation
- MIR Fundamentals
- MIR Challenges
- Issues in MIR
  - Image retrieval by content
  - Audio retrieval by content
  - Video retrieval by content
  - Indexing and searching
- Conclusions
- Bibliography

Introduction to Information Retrieval

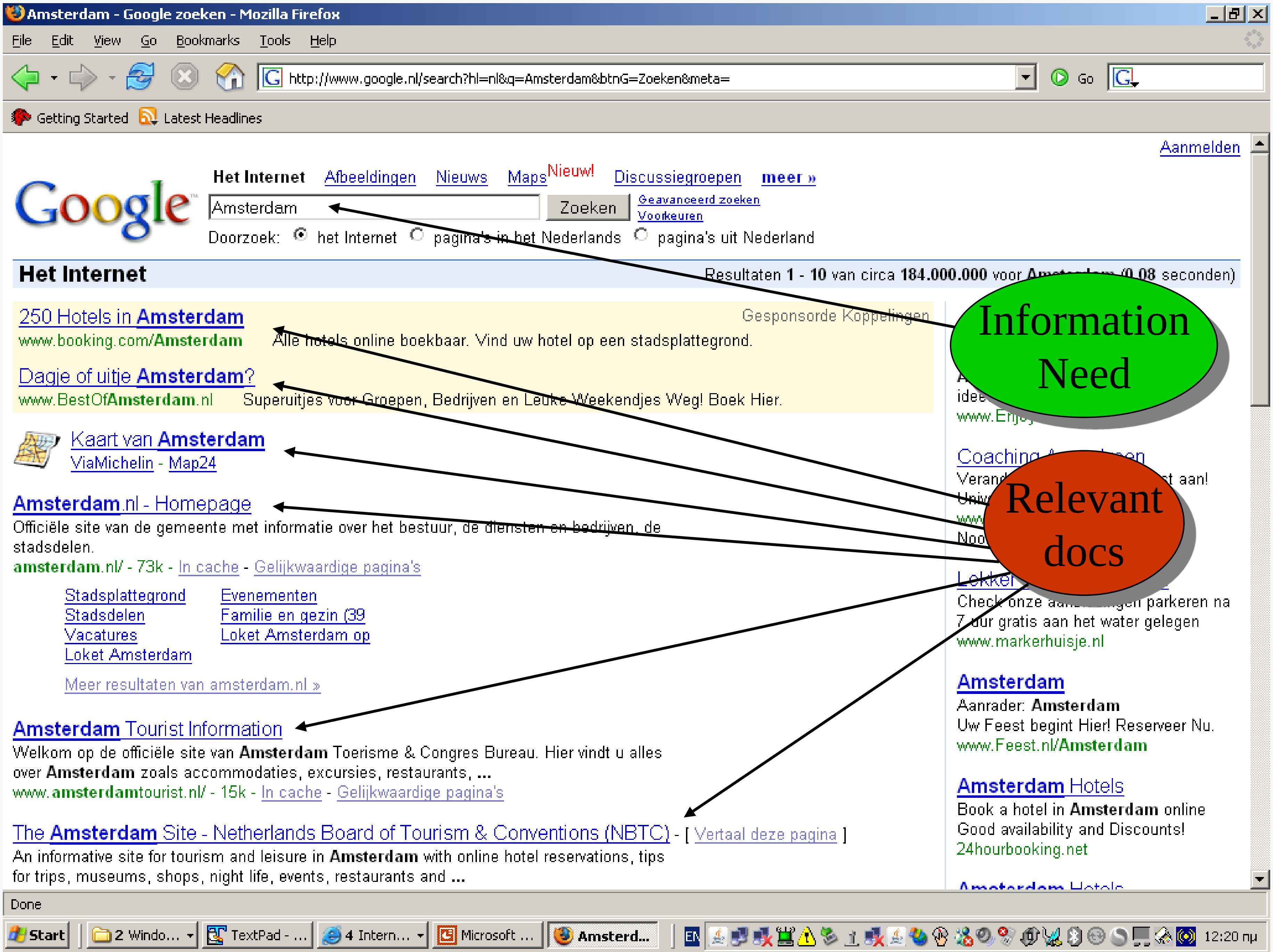
- Information Retrieval (IR) has been an active area of research and development for many years. The area of classic IR studies the representation, storage and processing of text documents.
- The primary target of an IR system is the following: given a collection *D* of documents and a user's information need *IN* determine which documents from *D* are relevant with respect to *IN*.

Introduction to Information Retrieval

Simple view of the IR process



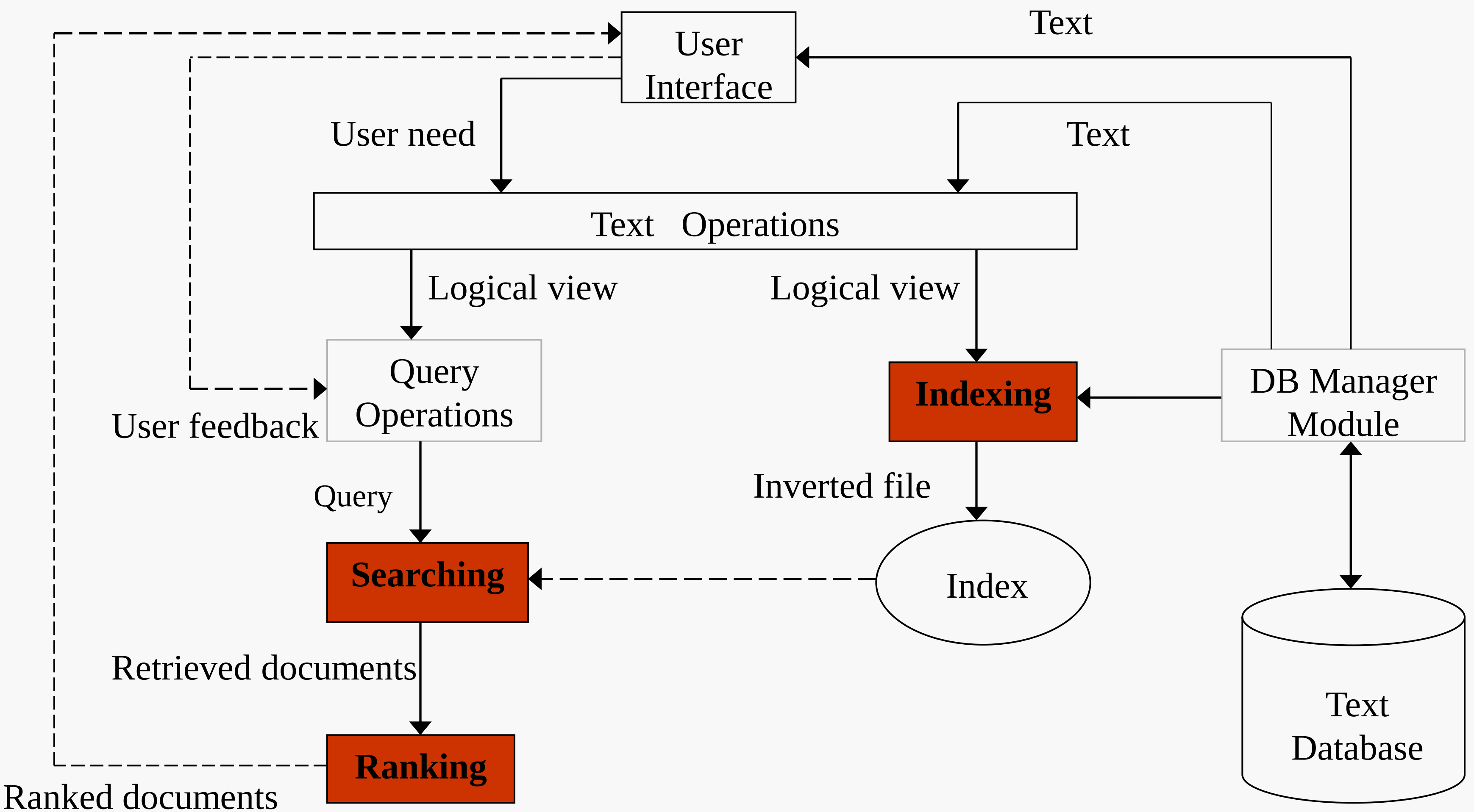
<u>The set of documents in the answer MUST be relevant to the user's information need. Otherwise the IR process results in complete failure.</u>



Amsterdam - Google zoeken - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.google.nl/search?hl=nl&q=Amsterdam&btnG=Zoeken&meta=
Getting Started
Latest Headlines
Aanmelden
Het Internet
Afbeeldingen
Nieuws
Maps
Nieuw!
Discussiegroepen
meer »
Amsterdam
Zoeken
Geavanceerd zoeken
Voorkeuren
Doorzoek: het Internet pagina's in het Nederlands pagina's uit Nederland
Het Internet
Resultaten 1 - 10 van circa 184.000.000 voor Amsterdam (0.08 seconden)
Gesponsorde Koppelingen
250 Hotels in Amsterdam
www.booking.com/Amsterdam
Alle hotels online boekbaar. Vind uw hotel op een stadsplattegrond.
Dagje of uitje Amsterdam?
www.BestOfAmsterdam.nl
Superuitjes voor Groepen, Bedrijven en Leuke Weekendjes Weg! Boek Hier.
Kaart van Amsterdam
ViaMichelin - Map24
Amsterdam.nl - Homepage
Officiële site van de gemeente met informatie over het bestuur, de diensten en bedrijven, de stadsdelen.
amsterdam.nl/ - 73k - In cache - Gelijkwaardige pagina's
Stadsplattegrond
Stadsdelen
Vacatures
Loket Amsterdam
Evenementen
Familie en gezin (39
Loket Amsterdam op
Meer resultaten van amsterdam.nl »
Amsterdam Tourist Information
Welkom op de officiële site van Amsterdam Toerisme & Congres Bureau. Hier vindt u alles over Amsterdam zoals accommodaties, excursies, restaurants, ...
www.amsterdamtourist.nl/ - 15k - In cache - Gelijkwaardige pagina's
The Amsterdam Site - Netherlands Board of Tourism & Conventions (NBTC) - [Vertaal deze pagina]
An informative site for tourism and leisure in Amsterdam with online hotel reservations, tips for trips, museums, shops, night life, events, restaurants and ...
Information Need
Relevant docs
Check onze aanbiedingen parkeren na 7 uur gratis aan het water gelegen
www.markerhuisje.nl
Amsterdam
Aanrader: Amsterdam
Uw Feest begint Hier! Reserveer Nu.
www.Feest.nl/Amsterdam
Amsterdam Hotels
Book a hotel in Amsterdam online
Good availability and Discounts!
24hourbooking.net
Done
Start
TextPad - ...
Microsoft ...
Amsterd...
12:20 πμ

Introduction to Information Retrieval

The IR process in detail



Information Retrieval vs Data Retrieval

	DR	**IR**
Matching	exact	partial, best
Items wanted	matching	relevant
Queries	precise	imprecise
Information	data, numeric	natural lang.
Query language	SQL	natural lang. (e.g., keywords)

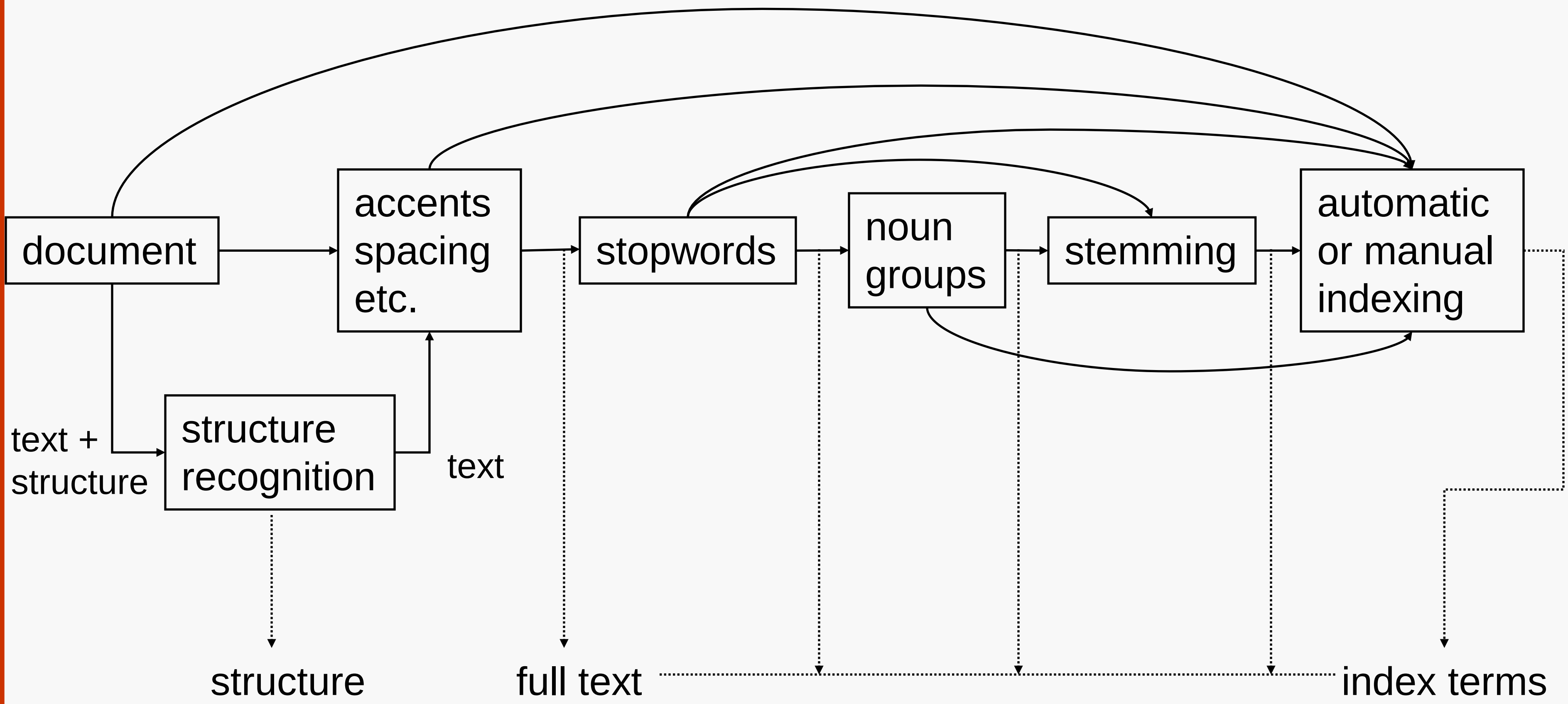
IR is supported by IR Systems

DR is supported by Database Systems

Document representation

The first important issue is how to represent the document collection. Usually, we assume that each document is a collection of words (terms). Some of the terms are eliminated since they are considered conceptually unimportant (e.g., the term “the”). As another preprocessing step we may consider stemming (e.g., planets→planet).

Document representation



Example of a document collection:

D_1: the Halley comet is here

D_2: a comet is not a planet

D_3: planet Earth is smaller than planet Jupiter

Query example: I need information about Halley comet

Question: how to process this query?

The query processing technique used depends on the following factors:

- the indexing scheme used, and
- the retrieval model supported.

Popular indexing schemes: inverted index, signature index, etc.

Popular retrieval models: boolean, vector, probabilistic, etc.

Inverted index example

For each term in the collection we record the total number of occurrences as well as the term position in each document

lexicon	posting lists
the	1, (*D*1, 1)
Halley	1, (*D*1, 2)
comet	2, (*D*1, 3), (*D*2, 2)
is	3, (*D*1, 4), (*D*2, 3), (*D*3, 3)
here	1, (*D*1, 4)
a	2, (*D*2, 1), (*D*2, 5)
not	1, (*D*2, 4)
planet	2, (*D*2, 6), (*D*3, 1, 6)
Earth	1, (*D*3, 2)
smaller	1, (*D*3, 4)
than	1, (*D*3, 5)
Jupiter	1, (*D*3, 6)

Collection

*D*1: the Halley comet is here

*D*2: a comet is not a planet

*D*3: planet Earth is smaller than planet Jupiter

Boolean retrieval model

- Each document in the collection is either relevant or irrelevant (on-off decision).
- Moreover, each query term is either present or absent in a document.
- A document will be part of the answer if it satisfies the query constraints.
- Queries are formed by using the query terms with logical operators AND, OR and NOT.

Example queries:

Halley AND comet

Comet OR planet

Comet AND NOT planet

Vector-space model

- Each document is represented as a vector in the *T*-dimensional space, where *T* is the total number of terms used to represent the document collection.
- For each pair (*ti*,*dj*) where *ti* is the *i*-th term and *dj* is the *j*-th document there is a value *wi,j* expressing the weight (or the importance) of term *ti* in the document *dj*.

Question 1: how are these weights calculated?

Question 2: how can we determine the similarity of a document with respect to a query?

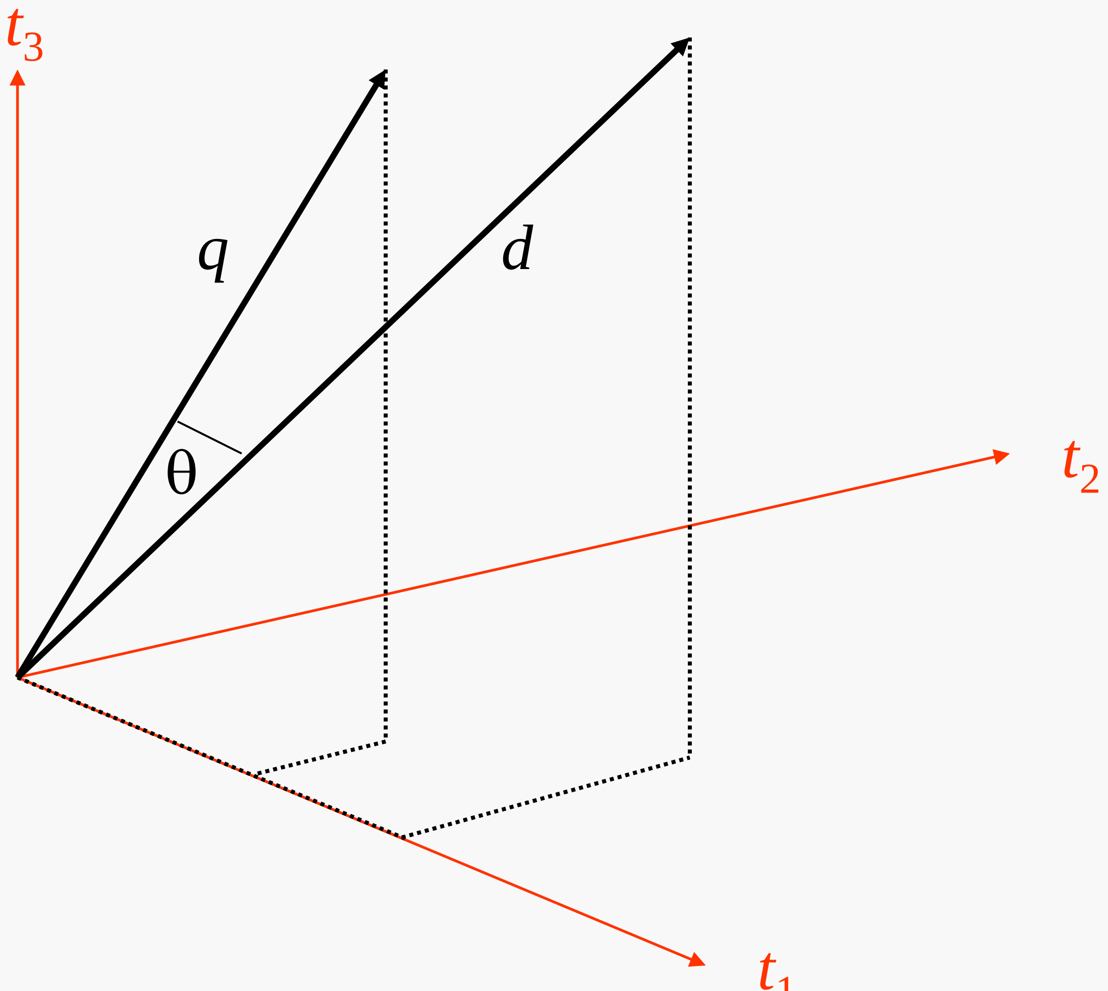
Introduction to Information Retrieval

Weight calculation: We take into account the number of occurrences of a term in a document and the number of documents containing a specific term.

Similarity calculation: Both the query and each of the documents are represented as vectors in a multidimensional space. The similarity is expressed by applying a function, e.g. cosine similarity.

$$\cos(\theta) = \frac{\boldsymbol{x}_1 \cdot \boldsymbol{x}_2}{|\boldsymbol{x}_1|\,|\boldsymbol{x}_2|}$$

Cosine similarity example



Efficiency and Effectiveness

The performance of an IR system is measured by two different factors.

- the efficiency of the system is the potential to answer queries fast,
- the effectiveness measures the quality of the results returned.

Both are very important and there is a clear trade-off between them. In many cases, we sacrifice effectiveness for efficiency and vise versa. Decisions depend heavily on the application.

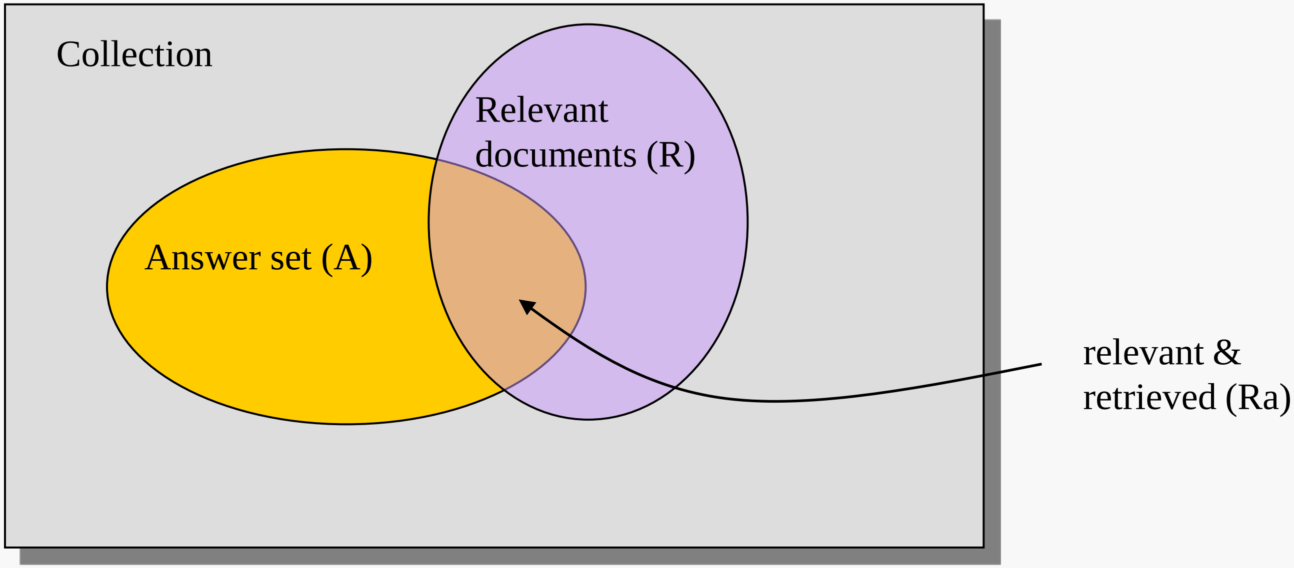
Efficiency and Effectiveness

The efficiency of the IR system depends heavily on the access methods used to answer the query.

The effectiveness, on the other hand, depends on the retrieval model and the query processing mechanism used to answer the query.

Important: Two DB systems will provide the same results for the same queries on the same data. However, two IR systems will generally give different results for the same queries on the same data.

Effectiveness measures

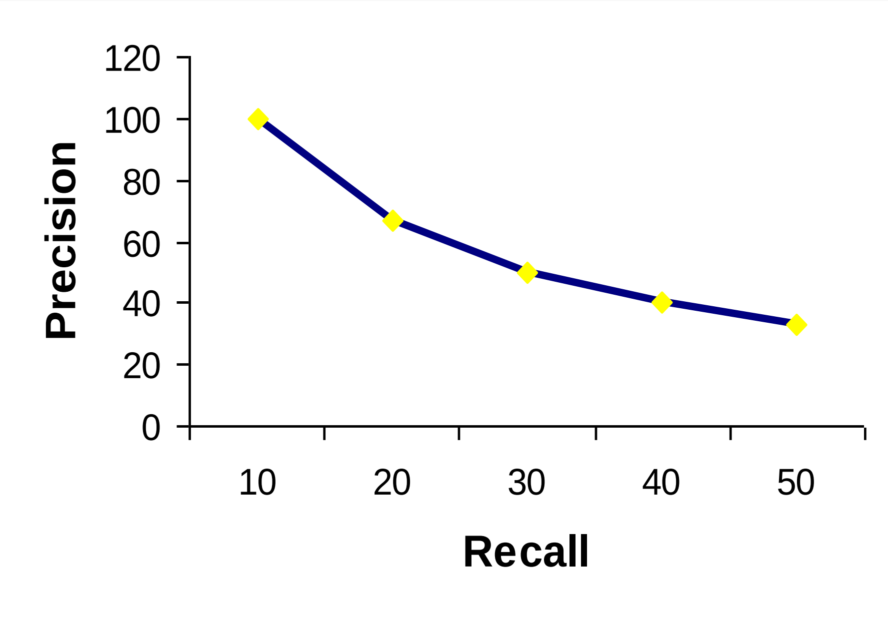


Recall = |Ra| / |R|

Precision = |Ra| / |A|

Recall-Precision example

$$R_q = \left\{ \begin{matrix} d_3, d_5, d_9, d_{25}, d_{39}, \\ d_{44}, d_{56}, d_{71}, d_{89}, d_{123} \end{matrix} \right\}$$



Rank	Doc	Rel	Recall	Precision
0			0 %	0 %
1	d_{123}	✔	10 %	100 %
2	d_{84}		10 %	50 %
3	d_{56}	✔	20 %	67 %
4	d_{6}		20 %	50 %
5	d_{84}		20 %	40 %
6	d_{9}	✔	30 %	50 %
7	d_{511}		30 %	43 %
8	d_{129}		30 %	38 %
9	d_{187}		30 %	33 %
10	d_{25}	✔	40 %	40 %
11	d_{38}		40 %	36 %
12	d_{48}		40 %	33 %
13	d_{250}		40 %	31 %
14	d_{113}		40 %	29 %
15	d_{3}	✔	50 %	33 %

MIR Motivation

Large volumes of data world-wide are not only based on text:

- Satellite images (oil spill), deep space images (NASA)
- Medical images (X-rays, MRI scans)
- Music files (mp3, MIDI)
- Video archives (youtube)
- Time series (earthquake measurements)

Question: how can we organize this data to search for information?

E.g., *Give me music files that sound like the file "query.mp3"*

Give me images that look like the image "query.jpg"

MIR Motivation

One of the approaches used to handle multimedia objects is to exploit research performed in classic IR.

Each multimedia object is annotated by using free-text or controlled vocabulary.

Similarity between two objects is determined as the similarity between their textual description.

MIR Challenges

- Multimedia objects are usually large in size.
- Objects do not have a common representation (e.g., an image is totally different than a music file).
- Similarity between two objects is subjective and therefore objectivity emerges.
- Indexing schemes are required to speed up search, to avoid scanning the whole collection.
- The proposed techniques must be effective (achieve high recall and high precision if possible).

MIR Fundamentals

In MIR, the user information need is expressed by an object *Q* (in classic IR, *Q* is a set of keywords). *Q* may be an image, a video segment, an audio file. The MIR system should determine objects that are similar to *Q*.

Since the notion of similarity is rather subjective, we must have a function $S(Q,X)$, where *Q* is the query object and *X* is an object in the collection. The value of $S(Q,X)$ expresses the degree of similarity between *Q* and *X*.

MIR Fundamentals

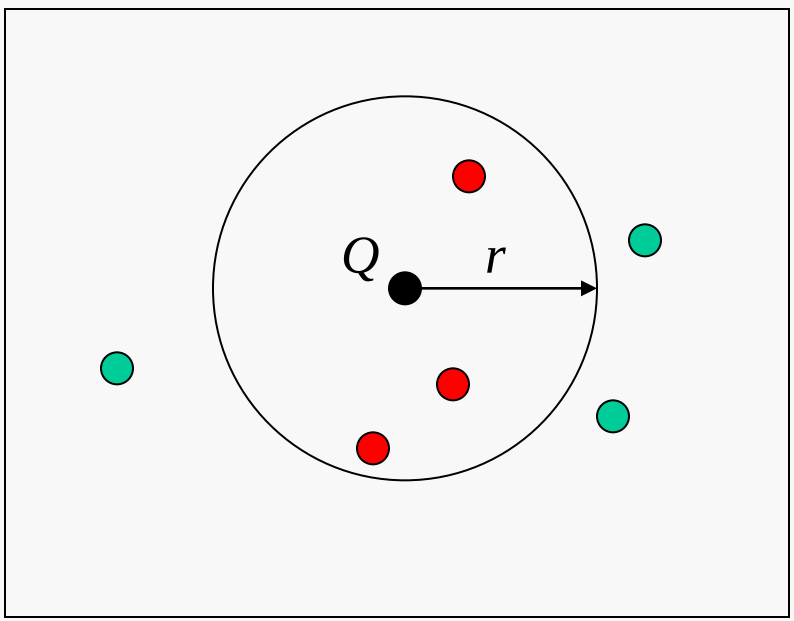
Queries posed to an MIR system are called similarity queries, because the aim is to detect similar objects with respect to a given query object. Exact match is not very common in multimedia data.

There are two basic types of similarity queries:

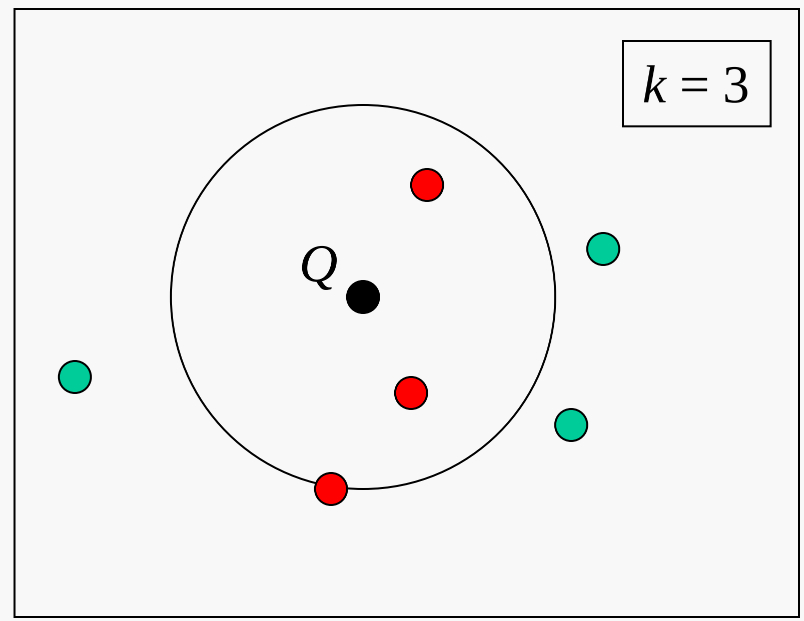
- A range query is defined by a query object *Q* and a distance *r* and the answer is composed of all objects *X* satisfying $S(Q,X) <= r$.
- A *k*-nearest-neighbor query is defined by an object *Q* and an integer *k* and the answer is composed of the *k* objects that are closer to *Q* than any other object.

MIR Fundamentals

Similarity queries in 2-D Euclidean space



range query



k-NN query

MIR Fundamentals

Given a collection of multimedia objects, the ranking function *S*( ), the type of query (range or *k*-NN) and the query object *Q*, the brute-force method to answer the query is:

<u>Brute-Force Query Processing</u>

[Step1] Select the next object *X* from the collection

[Step2] Test if *X* satisfies the query constraints

[Step 3] If YES then report *X* as part of the answer

[Step 4] GOTO Step 1

Problems with the brute-force method

- The whole collection is being accessed, increasing computational as well as I/O costs.
- The complexity of the processing algorithm is independent of the query (i.e., O(n) objects will be scanned).
- The calculation of the function S( ) is usually time consuming and S( ) is evaluated for ALL objects, the overall running time increases.
- Objects are being processed in their raw form without any intermediate representation. Since multimedia objects are usually large in size, memory problems arise.

MIR Fundamentals

Multimedia objects are rich in content. To enable efficient query processing, objects are usually transformed to another more convenient representation.

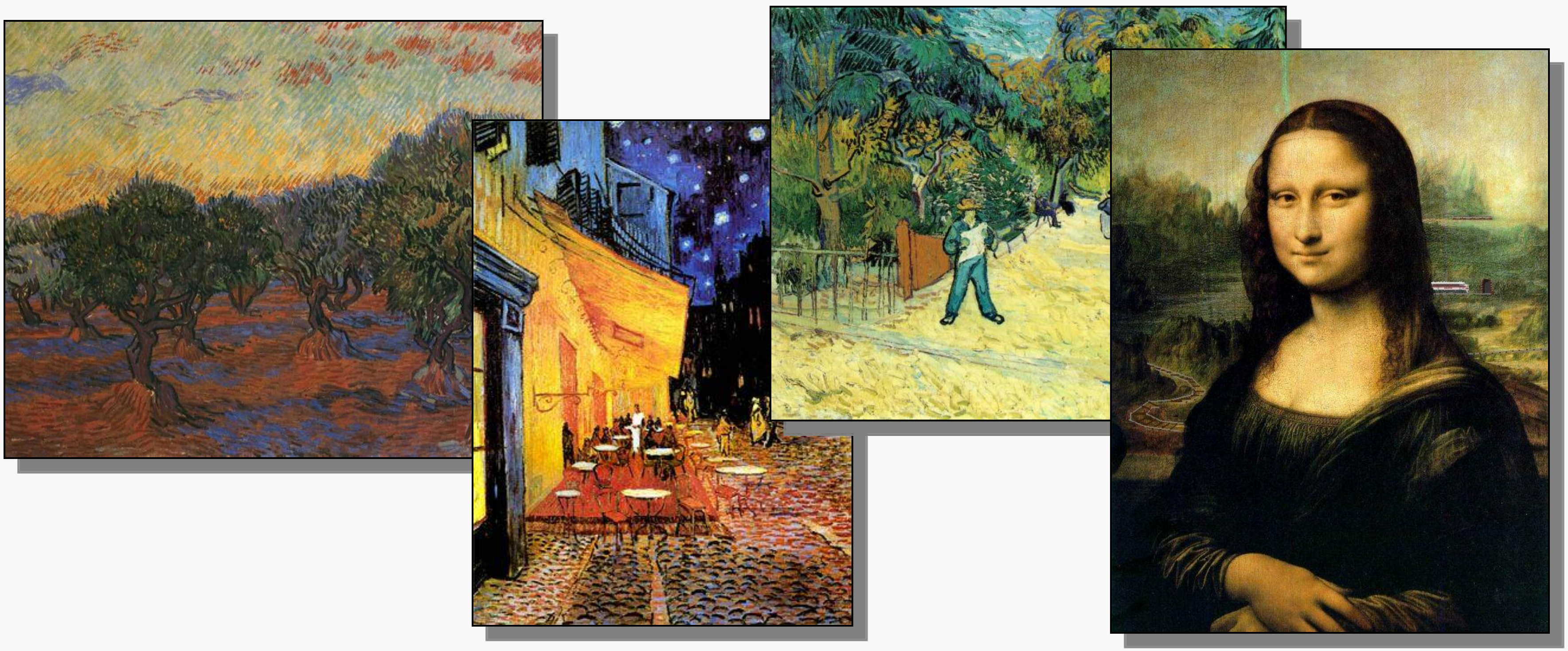
Each object X in the original collection is transformed to another object $T(X)$ which has a simpler representation than X.

The transformation used depends on the type of multimedia objects. Therefore, different transformations are used for images, audio files and videos.

The transformation process is related to feature extraction. Features are important object characteristics that have large discriminating power (can differentiate one object from another).

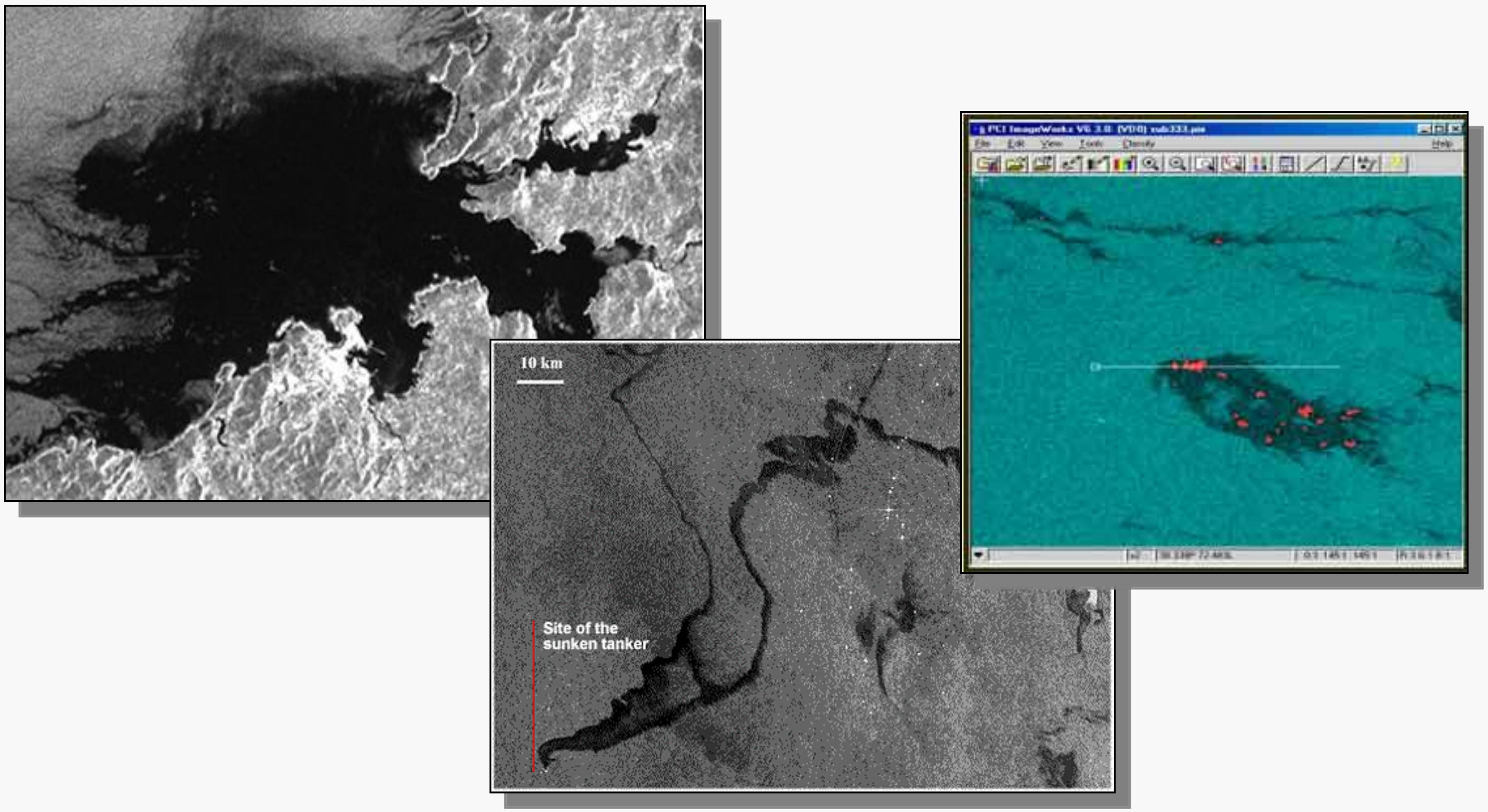
MIR Fundamentals

Image Retrieval: paintings could be searched by artists, genre, style, color etc.

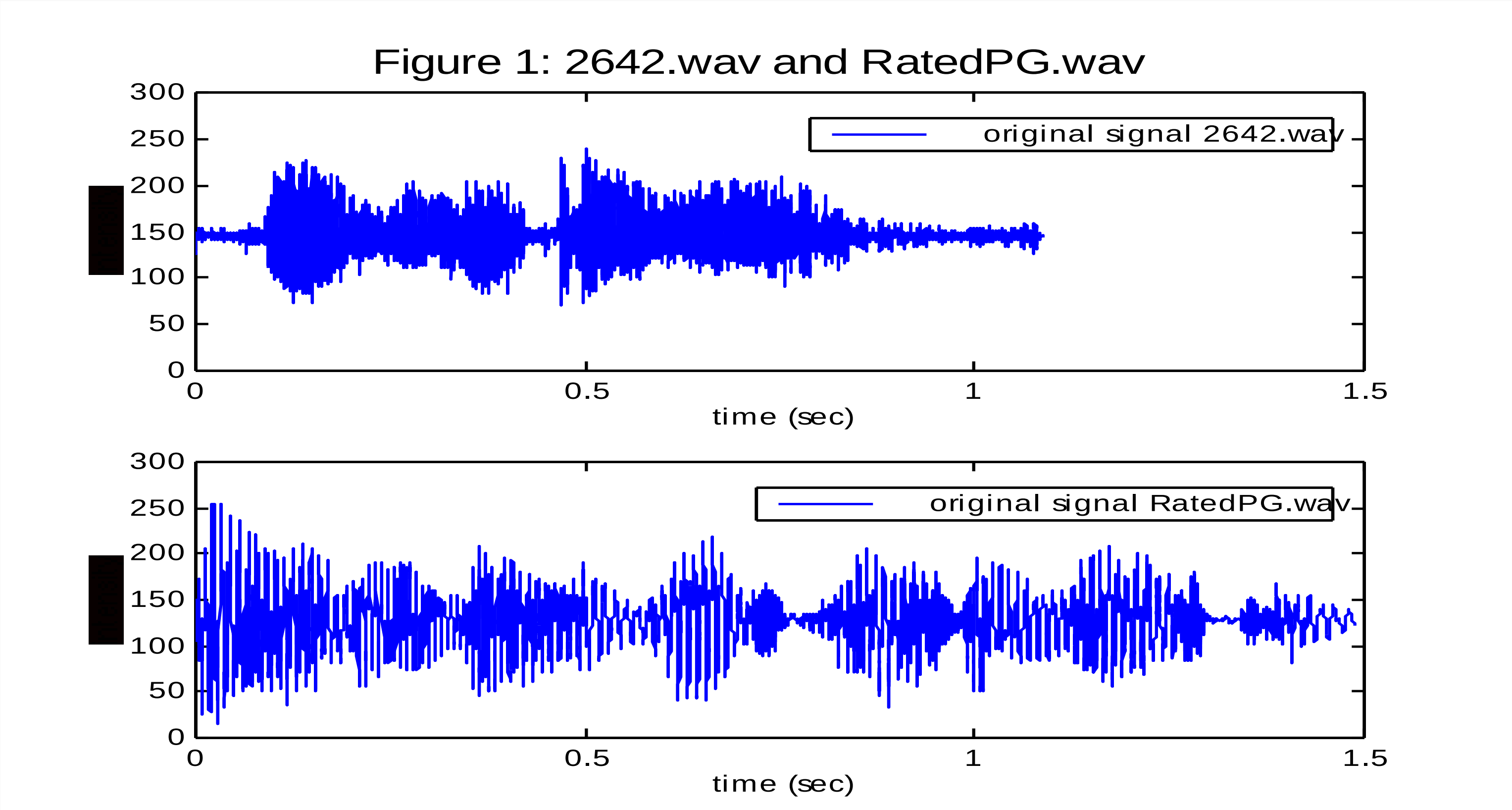


MIR Fundamentals

Satellite images – for analysis/prediction

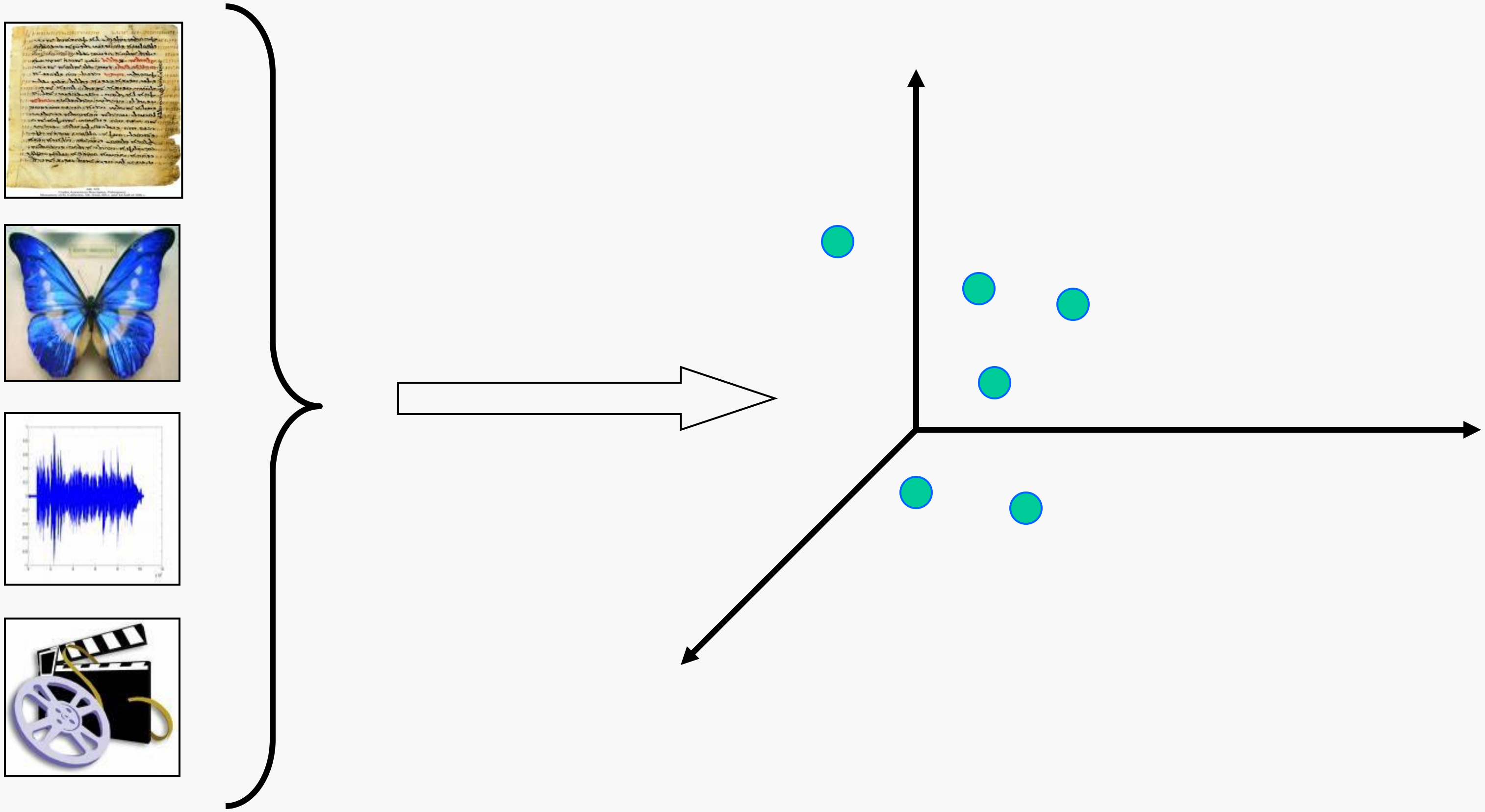


Audio Retrieval by content: e.g, music information retrieval.



MIR Fundamentals

Each multimedia object (text,image,audio,video) is represented as a point (or set of points) in a multidimensional space.



Conclusions

What is MIR?

- MIR focuses on representation, organization and searching of multimedia collections.

Why MIR?

- Large volumes of data are stored as images, audio and video files.
- Searching these collections is difficult.
- Queries involving complex objects can not be adequately described by keywords.

Bibliography

- R. Baeza-Yates and B. Ribeiro-Neto. "*Modern Information Retrieval*". Addison Wesley, 1999.
- C. Faloutsos: "*Searching Multimedia Databases by Content*", Kluwer Academic Publishers, 1996.
- B. Furht (Ed): "*Handbook of Multimedia Computing*", CRC Press, 1999.
- O. Marques and B. Furht: "*Content-Based Image and Video Retrieval*", Kluwer Academic Publishers, 2002.