

Creative Computing II

Christophe Rhodes
c.rhodes@gold.ac.uk

Autumn 2009, Tuesdays, 10:00–15:00
Winter 2010, Tuesdays, 13:30–17:30

Multimedia Information Retrieval

Information Retrieval:

- ▶ name given to general field of retrieving information in response to a **query**;
- ▶ Van Rijsbergen, C.J. *Information Retrieval*.

Nature of query and retrieval:

- ▶ query: specificity scale;
- ▶ retrieval: only exact, or also approximate matches;
- ▶ domain and range of retrieval process.

Multimedia Information Retrieval

Specificity

How **specific** is the query? Examples:

- ▶ 'the fourth song on the *Sgt Pepper* album';
- ▶ '*Hallelujah*';
- ▶ 'something bluesy';
- ▶ 'a track I would like to listen to now'.

[figure]

Multimedia Information Retrieval

Exactness

How stringent are our requirements on the retrieval?

- ▶ 'the fourth song on the *Sgt Pepper* album';
- ▶ '*Hallelujah*';
- ▶ 'something bluesy';
- ▶ 'a track I would like to listen to now'.

Are only **exact** matches acceptable, or are **approximate** matches good enough too?

Multimedia Information Retrieval

Domain and Range

What are we trying to retrieve information from?

- ▶ database (**corpus**) of materials?
- ▶ album?
- ▶ track?
- ▶ 100ms of audio?

What information are we trying to retrieve?

- ▶ album?
- ▶ track?
- ▶ musical information (e.g. key)?
- ▶ metadata?

Multimedia Information Retrieval

Fingerprinting

Problem statement: produce a unique identifier for a piece of multimedia that is (reasonably) invariant to distortion.

media item \rightarrow perceptual hash

item \approx item' \implies hash = hash'

Purposes:

- ▶ duplicate identification;
- ▶ rights management;
- ▶ directed sales.

e.g. Shazam, pHash.

Multimedia Information Retrieval

Text-based search

Problem statement: given some textual metadata, retrieve media items to which that metadata applies.

textual metadata \rightarrow media item*

Purposes:

- ▶ content delivery;
- ▶ personal collection organization;
- ▶ media navigation and discovery.

Multimedia Information Retrieval

Similarity

Problem statement: find items from a database which are 'similar' in some way to a query.

media item \rightarrow media item*

item \approx item'

Purposes:

- ▶ media discovery;
- ▶ rights management.

Multimedia Information Retrieval

Similarity

Problem statement: find items from a database which are 'similar' in some way to a query.

media item \rightarrow media item*

item \approx item'

Purposes:

- ▶ media discovery;
- ▶ rights management.

What does 'similar' mean?

- ▶ cover song?
- ▶ remix? mashup?
- ▶ same key? same genre? style?
- ▶ same structure? same artist?

Different things to different people (or at different times).

Multimedia Information Retrieval

Textual Features

- ▶ textual metadata;
- ▶ collaborative filtering.

Multimedia Information Retrieval

Textual Features

- ▶ textual metadata;
- ▶ collaborative filtering.

Textual feature treatment techniques:

- ▶ **stopword** or **noise word** removal;
- ▶ **stemming**;
- ▶ distance measures.

Multimedia Information Retrieval

Textual Features

- ▶ textual metadata;
- ▶ collaborative filtering.

Textual feature treatment techniques:

- ▶ **stopword** or **noise word** removal;
- ▶ **stemming**;
- ▶ distance measures.

Common search strategy:

- ▶ term-frequency–inverse-document-frequency

Multimedia Information Retrieval

Textual Features

stopword removal:

- ▶ some common words are not useful in an index.
- ▶ e.g. 'the', 'a', 'but', 'who', 'I'
- ▶ these words are typically removed prior to construction of an index (or ignored in distance measures if there is no index).

Stemming:

- ▶ many words come in different forms
- ▶ verbs: conjugation;
- ▶ nouns: pluralization;
- ▶ adverb / adjective duality.

Identify the **stem** of the word, so that all variants are findable. (cf. **Porter stemmer**)

Multimedia Information Retrieval

Textual Distance Measures

Identity measure:

- ▶ if the two words are exactly the same, their distance is 0;
- ▶ otherwise, the distance between them is 1.

$$d(\text{choose}, \text{choose}) = 0$$

$$d(\text{choose}, \text{chives}) = 1$$

Multimedia Information Retrieval

Textual Distance Measures

Identity measure:

- ▶ if the two words are exactly the same, their distance is 0;
- ▶ otherwise, the distance between them is 1.

$$d(\text{choose}, \text{choose}) = 0$$

$$d(\text{choose}, \text{chives}) = 1$$

This distance measure is too specific:

- ▶ usually want to have some tolerance (e.g. for misspellings)
- ▶ all non-identity pairs at the *same* distance.

$$d(\text{professor}, \text{professor}) = 1$$

$$d(\text{professor}, \text{cabbage}) = 1$$

Multimedia Information Retrieval

Textual Distance Measures

Hamming distance:

- ▶ if the two words have the same length, then their Hamming distance is the number of positions at which they differ;
- ▶ if the two words have different lengths, the Hamming distance is undefined.

$$d(\text{choose}, \text{choose}) = 0$$

$$d(\text{choose}, \text{chives}) = 4$$

Multimedia Information Retrieval

Textual Distance Measures

Hamming distance:

- ▶ if the two words have the same length, then their Hamming distance is the number of positions at which they differ;
- ▶ if the two words have different lengths, the Hamming distance is undefined.

$$d(\text{choose}, \text{choose}) = 0$$

$$d(\text{choose}, \text{chives}) = 4$$

This distance measure is not ideal for natural language:

- ▶ many misspellings change a word's length; doesn't model common ways of making mistakes.
- ▶ (useful in other contexts: particularly bit strings)

$$d(\text{professor}, \text{professor}) = \perp$$

$$d(\text{professors}, \text{professor}) = 5$$

Multimedia Information Retrieval

Textual Distance Measures

Levenshtein distance:

- ▶ define a set of permitted operations and associated costs:
 - ▶ insert;
 - ▶ delete;
 - ▶ substitute;
- ▶ Levenshtein distance between two words is the minimum cost to transform one word into another.

$$d(\text{choose}, \text{choose}) = 0$$

$$d(\text{choose}, \text{chives}) = 2s + d + i$$

- ▶ Often an appropriate measure to use for comparing words;
- ▶ Models ways of making mistakes;
- ▶ Can be computed in $O(N^2)$ time for words of length N .

Multimedia Information Retrieval

Textual Document Retrieval

Term-Frequency–Inverse-Document-Frequency (**tf-idf**):

- ▶ intuition:
 - ▶ term frequency: the more often a term is in a document, the more relevant it is;
 - ▶ inverse document frequency: the more documents a term is in, the less discriminating it is;
- ▶ Therefore, maximize a measure combining the term frequency and the inverse document frequency.

Multimedia Information Retrieval

Textual Document Retrieval

Term-Frequency–Inverse-Document-Frequency (**tf-idf**):

- ▶ intuition:
 - ▶ term frequency: the more often a term is in a document, the more relevant it is;
 - ▶ inverse document frequency: the more documents a term is in, the less discriminating it is;
- ▶ Therefore, maximize a measure combining the term frequency and the inverse document frequency.
- ▶ $tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}$

Multimedia Information Retrieval

Textual Document Retrieval

Term-Frequency–Inverse-Document-Frequency (**tf-idf**):

- ▶ intuition:
 - ▶ term frequency: the more often a term is in a document, the more relevant it is;
 - ▶ inverse document frequency: the more documents a term is in, the less discriminating it is;
- ▶ Therefore, maximize a measure combining the term frequency and the inverse document frequency.

$$\text{tf}_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}$$

$$\text{idf}_i = \log \frac{|D|}{|d_j: n_{ij} > 0|}$$