Multimedia Indexing

91.550: Topics in Information Retrieval, Search
And Autonomous Agents
Dept. of Computer Science, UML
Aminata Kane

Outline

- Text and Multimedia Languages and Properties
  - Documents
    - Metadata
    - Text
      - Introduction
      - Metadata
      - Text
    - Similarity models
  - Markup languages
    - SGML
    - HTML
    - XML
  - Multimedia
    - Format
    - Textual images
    - Graphic and virtual reality
    - HyTime
    - Taxonomy of web languages
- Non-text Queries
- Music retrieval
Introduction

A document is a single unit of information, could be logical (research article) or physical (web page).

Components of a document:
- Syntax and structure: decided by the application
- Semantics: specified by the author
- Presentation style: how it is displayed or printed
- Metadata: information about itself

Syntax (languages) of a document:
- Can express structure, presentation style, semantics
- Usually proprietary and specific to languages
- There are open and generic languages
  - Can be interchanged between applications
  - Flexible
- Trends: favor the languages which both human and computer can read
  - e.g., SGML (standard generalized markup language)

Introduction (2)

Formatting style of document

Most documents have a particular formatting style.

Types:
- External formatting (new applications)
  - Separate the style from the information
- Embedded style
  - e.g., TeX, Rich Text Format (RTF)

Styles is how it is seen on the screen and can be defined by:
- Document authors or
- Reader
**Metadata**

- **Data about data**
  - Author, the date of publication, source, document length, document genre

- **Applications of metadata**
  - Cataloging: e.g., BibTeX format
  - Content rating: e.g., for protecting children
  - Intellectual property rights
  - Authentication: e.g., digital signatures
  - Privacy levels: who can or who should not have access
  - E-commerce
  - Description of non-textual data: e.g., keywords describing an image

- **Web metadata**
  - RDF: Resource Description Framework
    - Allows description of Web resources (nodes and attached attribute/value pairs)
    - Nodes: any web resources, e.g., URI, URL
    - Attributes: node properties. metadata can be used

---

**Markup Languages**

- **Markup**
  - Extra textual syntax to describe formatting, structure information, semantics, attributes
    - The formatting commands of TeX is markup but less structured than the usual ones.

- **Examples**
  - SGML (standard generalized markup language)
  - XML (extensible markup language): subset of SGML
  - HTML (HyperText markup language)
SGML

- Metalanguage developed by IBM for tagging text
- Each instance includes a document type declaration (DTD)
  - Describes and names the piece
  - Defines the relationship in between pieces
  - Does not define semantics
    - e.g., two tags have the same name but different meaning in different applications
- DTD notations
  - `<tagname>` … `</tagname>`
  - ‘ELEMENT’: defines element
  - ‘-‘: mandatory, ‘O’: optional
  - Regular expression
    - ‘,’: concatenation, ‘|’ or, ‘*:0 or 1), ‘+’: (0 or more), ‘+’
  - ‘PCDATA’: ASCII characters, ‘NDATA’: binary data, ‘EMPTY’
  - ‘ATTLIST’: possible attributes

SGML - DTD Example for mail

```xml
<!DOCTYPE e-mail SYSTEM "e-mail.dtd">
<?prolog>
<e-mail id="04001998">
  <sender> Pablo Beruda </sender>
  <address> Faberico Garcia Lorca </address>
  <subject> Pictures of my house in Isla Negra </subject>
  <to> Gabriel Garcia Marques </to>
  <prolog>
    <par>
      As promised in my previous letter, I am sending two digital
      pictures to show you my house and the splendid view
      of the Pacific Ocean from my bedroom (photograph from ref=I1998).
    </par>
    <image id="I1998" "photol.gif"/>
    <image id="I2000" "photo2.jpg"/>
  </par>
  <par>
    Regards from the South, Pablo.
  </par>
</e-mail>
```
SGML - Output

- No specification about output style
  - Often output specification is associated with the SGML documents
- Output specification standards
  - DSSSL (Document Style Semantic Specification Language)
  - FOSI (Formatted Output Specification Instance)
- TEI (Text Encoding Initiative)
  - Cooperative project. Started 1987
  - Guidelines for the preparation and interchange of electronic texts
  - Provides document format through SGML DTDs
    - TEI Light
      - Similar to latex style file

HTML

- HyperText Markup Language
  - Created in 1992
  - Version 4.0 released in 1997
    - Style sheets, internationalization, frames, richer tables and forms
    - Being extended for some issues like math formulas
- DHTML (Dynamic HTML)
  - HTML page that has program (e.g., Javascript) inside
- CSS (Cascade Style Sheet)
  - Presentation style associated with HTML
  - Modestly supported by current browsers
    - Not clear who (creator or reader) should define presentation
**Characteristics**
- Extensible Markup Language
- Subset of SGML
- Allows human-readable and machine-readable semantic markup

**Syntax**
- More rigid syntax than in HTML
  - Ending tags cannot be omitted
  - Distinguishes lower and upper case
  - Attribute values must be between quotes
- RMD (Required Markup Declaration)
  - RMD="NONE": no DTD
  - RMD="INTERNAL": DTD is inside
  - RMD="ALL": use of external source for DTD. default

---

```xml
<?xml version="1.0" encoding="iso-8859-1"?>
<e-mail id="94108rby" date_sent="02101998">
    <prolog>
        <sender> Pablo Neruda </sender>
        <address> Federico Garcia Lorca </address>
        <address> Ernest Hemingway </address>
        <subject> Pictures of my house in Isla Negra </subject>
        <Cc> Gabriel Garcia Marquez </Cc>
    </prolog>
    <content>
        <par>
            As promised in my previous letter, I am sending two digital pictures to show you my house and the splendid view of the Pacific Ocean from my bedroom [photo <img id="F1" ref="photo1.gif"/>].</par>
        <par>
            Regards from the South, Pablo.</par>
    </content>
</e-mail>
```
XML (3)

- Can define new tags, complex (nested) structures
- DTD is not required
- XSL (Extensible Style Sheet Language)
  - Like Cascade Style Sheet in HTML
  - e.g., extract a table from XML document
  - Can transform XML documents to HTML and CSS
- Recent extensions
  - MathML (mathematical Markup Language)
  - SMIL (Synchronized Multimedia Integration Language)
    - Schedule multimedia presentation
  - RDF (Resource Description Format)
    - Metadata information for XML
- Directions
  - Parseable, hierarchical object model

Multimedia Languages and Properties

- Properties
  - Text, Audio, static images, dynamic objects (animations, video…)
  - Different in volume, format, processing requirements (temporal dimension for audio and video)
- Formats
  - Image formats
  - Audio formats
  - Animation or moving images
- Textual Images
- Graphic and virtual reality
  - CGM (Computer Graphics Metafile)
  - VRML
- Taxonomy of Web Languages
Image Formats

- Direct representation of pixel based (bit-map, simplest)
  - XBM, BMP, PCX
  - Consume too much space (single screen frame = 1 Mb)
- Compression format
  - GIF (CompuServe's Graphic Interface Format)
  - Good for black and white pictures and small # of colors
- Lossy compression format (Eliminating parts)
  - JPEG (Joint Photographic Experts Group)
  - Eliminate part of image that have less impact on human eye
  - Uncompressing a compressed does not give the original
- TIFF (Tagged Image File Format)
  - Used to exchange in different applications and computer platforms
  - Has file for metadata, supports compression

Image Formats (2)

- TGA (Targa Image File)
  - Associated with video game boards
- JBIG:
  - Bi image level format for fax
- WSG:
  - Highly accurate and compressed format for fingerprint and satellite
- PNG (Portable Network Graphics):
  - Bit-mapped image for the Internet
Audio and Dynamic Image Formats

- **Audio formats**
  - For small pieces of digital audio
    - AU, MIDI, WAVE
  - For audio libraries
    - RealAudio, CD format

- **Dynamic images**
  - MPEG (Moving Picture Expert Group)
    - Related to JPEG
    - Codes the changes with respect to a given referential (image of base)
  - Number of frames is important,
  - Includes audio (like AVI and QuickTime)
  - AVI, QuickTime
    - Developed by Apple

Textual Images

- **Images of document**
  - Obtained by scanning the documents (mostly for archiving purposes, efficient for retrieval purposes)
  - Further compression can be done by extracting the different text symbols or marks from the textual image
  - Non-textual information such as logos or signatures can be extracted using the segmentation method
Graphics and Virtual Reality

- CGM (Computer Graphics Metafile)
  - Two dimensional data interchange of structural graphical objects and associated attributes.
  - Can represent vector graphics, raster graphics, text.
  - Specifies which element are allowed to occur in which position of a metafile
- VRML (Virtual Reality Modeling Language)
  - Describes interactive 3D objects
  - Subset of Silicon Graphics Open Inventor file format
  - Used in engineering, scientific visualization, entertainment, educational titles, Web pages

Time format

- HyTime(hypermedia time based Structuring language
  - Standard (ISO/IEC 10744)
  - For multimedia documents markup
  - Applications are SDML(Standard Music Description Language) and MID(Metafile for interactive Document)
  - Concept based on
    - Complex locating of document object
    - Hyperlinks between document objects
    - Numeric measured association between document objects
Taxonomy of Web Languages

- Solid line: instance
- Dashed line: derived

Overview of System Components and Process Flow
Non-text Queries (VisualSeek)

Developed by the Image and AdvanceTelevision Lab, Columbia University, NY. It presents among other things:

- automated extraction of localized regions and features
- efficient representation of features
- preservation of spatial properties
- extraction from compressed data
- fast indexing and retrieval
Color set-back projection technique

- selection of a color set
- given image $I$, an image $B$ is generated according to

$$B[x, y] = \max_j a[k, j]c[j], \quad k \in \{0, \ldots, 165\}$$

With $k$, index of the color of the pixel $(x, y)$ in $I$ and $a[k,j]$ similarity between the 2 colors
Diagram used first Image color used

VisualSeek variant (VideoQ)
Generating the candidate video shot list for a single query
Music Retrieval Techniques

- Brian Whitman: http://www.media.mit.edu/~bwhitman Dan Ellis' Rosa lab http://www.ctr.columbia.edu/~dpwe/LabROSA/
- Jonathan Foote Http://www.fxpal.com/people/foote/ (ACM multimedia 98)
This link presents a great way to retrieve music:

http://www.fxpal.com/people/foote/musicdoc1.html

- Some things to search for:
  - Piano music
  - Grunge rock
  - Acoustic guitar
  - Reggae
  - Jazz
  - Medieval plainsong

Example(2)

1.000 Paulistane+SolidLove
0.335 Paulistane+WildFlower
0.886 WarmEvenings+SummerNight
0.835 Paulistane+NotesCancoes
0.830 ShimaShoka+Rasas
0.779 Paulistane+WhoKnows
0.715 WarmEvenings+HeLovesAndSheLoves
0.711 Belaflek+UpAndRunning
0.699 ShimaShoka+MeroViglaco
0.687 Belaflek+TheGreatCircleRoute
0.673 Paulistane+SambaDoAria
0.658 SoftSounds+Barefoot
0.645 ShimaShoka+PrestoVH
0.639 Paulistane+Luminosos
0.615 DukePearson+Gaslight

Search for similar files  Play selected file  Reset
Paulistana+SolnLove (referential )
- Distance = 1.000
- Sound = 

Paulistana+Wildflower (2nd in list)
- Distance = 0.935
- Sound = 

Belafleck+TheGreatCircleRoute (10th)
- Distance = 0.687
- Sound = 

Chemical brothers+songToTheSiren (last)
- Distance = 0.000
- Sound =
**How does it work?**

- Works by calculating the "distance" between the selected file and all other files.
- "closer," and hence more similar files are nearer to the top.
- Parameterizations into Mel frequency cepstral coefficient.
- Tree construction.
- Tree Structured Quantization of vectors.
- Template generation.
- Comparaison by computing metric distance.

**Template construction**

Diagram showing the process from waveform to histogram counts:
- Waveform
- Window
- Compute MFCCs
- MFCCs
- Quantize via Tree
- Accumulate Histogram Counts
Audio parameterizations into Mel frequency cepstral coefficient

- Audio waveform will first be sampled at 16 kHz, and so transformed into a sequence of 13-dimensional feature vectors (12 MFCC coefficients plus energy)
  - First, the audio is Hamming-windowed in overlapping steps
  - For each window, the log of the power spectrum is computed using a discrete Fourier transform (DFT).
  - The resulting log are weighted by a non-linear map of the frequency scale (for example it emphasizes mid-frequency bands in proportion to their perceptual importance) to form a spectrum
  - The resulting spectrum will be transformed into cepstral coefficients using another discrete fourier transform
Tree Construction

- Each decision in the tree involves comparing the vector with a fixed threshold
  - All music from each artist was considered an individual class, and the tree was automatically trained to separate each class.

- Best MMI (Maximum Mutual Information) split I(X;C)
  - \( I(X;C) = H(C) - H(C|X) \)
  - with \( H \) being the binary entropy function
    - \( H_2(x) = -\log_2(x) - (1-x) \log_2(1-x) \)

- Set of N training Vectors are divided into \( X = \{X_a, X_b\} \)
  - \( X_a = x_d \geq t_d \)
  - \( X_b = x_d \leq t_d \)

- For the demonstration, a tree with 60 bins (leaves) was constructed from the demonstration data

Tree Structured Quantization

- We "quantize" each vector using the specially-designed quantization tree (recursively divides the vector space into bins, each of which corresponds to a leaf of the tree)
  - First the tree will map the 13-dimensional vectors in the vector space \( R^{13} \) into a finite set of vectors \( Y = \{y_i: i = 1, 2, \ldots, N\} \)
  - Associated with each \( y_i \) is a nearest neighbor region called Voronoi region (Given a set of vectors in a space, the Voronoi Region of a particular vector is defined as the set of all points in for which is the nearest vector) and it is defined by:
    - \( V_i = \{x \in R^k : \|x-y_i\| \leq \|x-y_j\| \text{ for all } j \neq i\} \)
  - The set of Voronoi regions partition the entire space \( R^k \) such that for all \( i \neq j \):
    - \( \bigcup_{i=1}^N V_i = R^k \)
    - \( \bigcap_{i=1}^N V_i = \emptyset \)
Codewords in 2-dimensional space.

Tree Structure Quantization (2)

- The representative codeword is supposed to be the closest in Euclidean distance from the input vector.
- The Euclidean distance is defined by:

\[ d(x, y_i) = \sqrt{\sum_{j=1}^{k} (x_j - y_{ij})^2} \]

With \( x_j \) is the \( j \)th component of the input vector, and \( y_{ij} \) is the \( j \)th component of the codeword \( y_i \).
The Encoder and decoder in a vector Quantizer

Template Generation

- The way the vectors are distributed into each bin characterize the audio
- Counting how many vectors fall into each bin yields a histogram template that is used in the distance measure
- sorted by magnitude to produce a ranked list like other search engines
Distance metric

- acoustic similarity is measured using:
  - Euclidian distance
    \[ D_E^2(p, q) = \sum_{i=1}^{N} [p(i) - q(i)]^2 \]
    It treats the histograms as vectors in N-dimensional space, and computes the Euclidean distance between them.
  - Cosine distance
    \[ D_C(p, q) = \frac{\sum_{i=1}^{N} p(i)q(i)}{\sqrt{\sum_{i=1}^{N} p(i)^2 \times \sum_{i=1}^{N} q(i)^2}} \]
References

- http://www2.rgu.ac.uk/~sim/reports/gallery1.htm
- http://www.dlib.org/dlib/february97/columbia/02chang.html
- Modern Information Retrieval (Yates - Nieto) chap. 6, 11, 12
- Image Databases