Multimedia Information Retrieval:

Or “This stuff is really hard... really!”
Matt Earp / Kid Kameleon
Setlist

• Review of Multimedia Information Organization
• MMIR in Music - Matt
• MMIR in Video and Photo - Yiming
Classical Search

- The user translates an information need or question(s) into a QUERY

- The query expresses the information need in a format or as a set of descriptive features that the system can handle

- The processable representation of these features make up the INDEX or INDICES

- The system matches the descriptive features in the query against the features that describe the "documents" or "information objects" (or pointers to them) stored by the system

- Items are retrieved when the degree of the match exceeds some measure of similarity (which might be "exact match" for some queries or systems)

- The system presents the retrieved items according to the measure of similarity
The Tradeoff: Organization \{and,or,vs\} Retrieval

- A recurrent theme in this course has been the tradeoff between organization and retrieval
- The more effort put into organizing information, the more effectively it can be retrieved
  - Using controlled vocabularies in metadata to describe instances
  - Developing "ontologically-grounded" categories to organize instances
- The more effort we put into retrieving information, the less it needs to be organized first
  - Crawling the web to create indexes and analyze link structures to organize information so that it can be found
  - Analyzing the content of text documents to compute their similarity on the basis of lexical and conceptual co-occurrence
Why “Describing Text” is Relatively Straightforward

• Most of the concepts and techniques that authors or other people might use for "describing things" were designed for text information

• Many techniques for extracting text descriptions from the information being described can automated because the content is text

• "Text is text" - the text content of information objects can vary in formats, fonts, etc. -- but the alphabet defines equivalence classes for these different representations

• The internal structure of text information and collections is explicit, which enables useful descriptions to be assigned at different granularity
Why Describing Multimedia (with Text) is Challenging

• There’s no standard “alphabet”
• Can vary in format, resolution, etc and still be the “same” object (or not...)
• Features can be implicit, hard to detect, or be sensorily ambiguous
• Multimedia Content description are not explicit in the content and are often more emotional ambiguous, subjective, or generative
• Temporal issues of Multimedia can create the need for new types of descriptors
• Automated techniques for assigning text descriptors to content can’t capture all the semantic complexity in multimedia content
Two methods of Description for MM

1. Add machine/human readable text (metadata) that can be queried -- Text Query

2. Analyze the content to determine meaningful attributes, then search based on similarity to an example -- Query by Example
Adding Text: A “Simple” Example

- Find information about: *Endtroducing* by DJ Shadow

- Look in a variety of places and see what’s displayed: Allmusic.com, Discogs.com, Amazon.com, Last.fm, iTunes ...

- Wikipedia entry includes the sample list

- Ultimately, it’s not so simple!
The Sensory Gap

- What are these pictures about?
The Sensory Gap 2

- The gap between an object and a computer's ability to sense and describe the object
- An infinite number of different "signals" or representations can be produced by the same object, depending on its distance, orientation, illumination, etc.
- Likewise, different objects can produce similar "signals"
- Human perception is remarkable in its ability to recognize when different "signal patterns" are the same object or when similar patterns are different ones
- But the "similarity" problem is much more difficult for computers
The Semantic Gap

- Instruments, devices, sensors and so on encode data in formats that are optimized for efficient capture, storage, decoding, or other criteria.

- As a result, the content/representation/encoding/material of the object is semantically opaque, and can't be (easily) processed to understand what the object "means".

- So there is a gap between the semantic descriptions that people assign to objects and the descriptions that can be assigned by computers or other automated mechanisms.

- That's why multimedia needs metadata if we are going to use text-based search!
Music Examples

- Trade off between text search and query-by-example
- Trade off between a PIM system that may be great for a very limited domain but may not have extensibility
Annotatable Audio

• Tom Coates created the prototype system for BBC Radio and Music Interactive in 2005 with several others

• The idea was to mark up/make semantically useful radio broadcasts from the BBCs 80 year history

• Based User generated transcription, tagging, and annotation of sound files, although the system was never put into place

• Example videos: playing media, editing media

• More on Tom Coats here: http://www.plasticbag.org
Query By Humming

• Bryan Pardo (of Northwestern) creates the VocalSearch system

• The goal is to have the query object (a hum) match as closely as it can to the corpus (music)

• Pardo has gone on to do serious number-crunching work around music element extraction and “source separation”

• More here: http://www.cs.northwestern.edu/~pardo/
Music Retrieval: Internet Radio

- Net radio is still struggling to find the balance between traditional “terrestrial” radio and a personal music collection
- Different models play with various sorts of recommendation systems to bring you the music you “want” to hear
- Many systems are constrained by national and international laws around Copyright
- No one has quite gotten it “right” yet (in my opinion), but some do a good job
Last FM

- World’s “largest music platform” (whatever that means), based in the UK
- An Internet Radio station based around community
- Takes as its central premise that people want to hear stuff that the people LIKE them want to hear
- [http://www.Last.fm](http://www.Last.fm)
Pandora

- Less focused on some of the community aspects of music, more about producing the “right” songs
- The Musical Genome Project (started in 2000) is its computational foundation
- Extracts vital “musical DNA” out of the song (things like pitch, timbre, “shoutiness”, etc) and uses that to come up with what it will play you next
- Some elements of the site (like only being able to skip 6 songs an hour) are dictated by law.
- http://www.Pandora.com
Why pundits don’t know how to talk about this stuff

“There’s so much you can’t get out of the audio,” Lamere said. “The fact that two songs played during your senior prom means a lot to you, yet it doesn’t factor into sound recognition.” Still, computerized analysis of audio can fill in the gaps when artists are so new they simply aren’t being discussed on the Web, he said. In the demonstration, Lamere had Sun’s software “listen” to a Mozart piano sonata, and watched it spit out a list of recommended songs, mainly other classical music.” - Jon Brodkin, Network World, 10/31/07
Echonest

- Creation of Brian Whitman, graduate of the “Music, Mind, and Machine” group under Barry Vercoe at MIT
- “I want to help people find music, and I want artists and labels to find people”
- Done work around the musical DNA concept, machine learning and processing songs for their essential traits
- Does serious number crunching to determine aspects of songs and then recombine them in various ways (“re-synthesis”)
- Also done NLP style textual analysis of record reviews
- Echonest is the (still forthcoming) marriage of the two.
- [http://variogr.am/writings.html](http://variogr.am/writings.html)
Info IR:A (Human) DJ’s view

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Bit Rate</th>
<th>BPM</th>
<th>Time</th>
<th>Artist</th>
<th>Album</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>windfall</td>
<td>320 kbps</td>
<td>140</td>
<td>2:49</td>
<td>Boxcutter</td>
<td>Glyphic</td>
</tr>
<tr>
<td>2</td>
<td>Shallow</td>
<td>192 kbps</td>
<td>141</td>
<td>4:02</td>
<td>Clouds</td>
<td>Clouds EP</td>
</tr>
<tr>
<td>3</td>
<td>Northside Cheng Dub</td>
<td>320 kbps</td>
<td>137</td>
<td>4:38</td>
<td>Blackdown</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Rain</td>
<td>320 kbps</td>
<td>140</td>
<td>6:08</td>
<td>fliPPo</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Untitled</td>
<td>320 kbps</td>
<td>140</td>
<td>6:58</td>
<td>Reform</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Banker</td>
<td>192 kbps</td>
<td>142</td>
<td>4:48</td>
<td>Secret Agent Gel</td>
<td>Valentine's Day Demo</td>
</tr>
<tr>
<td>7</td>
<td>Glock Box</td>
<td>256 kbps</td>
<td>140</td>
<td>5:40</td>
<td>Sky City Rising</td>
<td>Sky City Rising</td>
</tr>
<tr>
<td>8</td>
<td>Why Dem Fight (ft Bunninton Jr)</td>
<td>320 kbps</td>
<td>140</td>
<td>4:26</td>
<td>Sarantis</td>
<td>Subsonic 001</td>
</tr>
<tr>
<td>9</td>
<td>Tuning</td>
<td>256 kbps</td>
<td>141</td>
<td>4:33</td>
<td>Distance</td>
<td>My Demons</td>
</tr>
</tbody>
</table>
Rights and Peer-2-Peer

- MPEG 7 - Multimedia Description Standard
- A way to use XML to include a robust and extensible set of information about the multimedia object (similar to ID3 tags), such as subject, sections, components, and even rights
- MPEG 21 - XML “Rights Expression Language” - hasn’t gained wide implementation yet
- DRM - Digital Rights Management to control media use
- All this can be used in P2P network tracking by companies like BigChampagne (“We are continually incorporating more information from new sources”)
Audio Fingerprinting

- Audio Fingerprinting is the attempt to song based on its “signature”. It differs from Query-By-Humming in that it’s often looking for specific recordings rather that any version of a song.

- It can also be used for song ID (recording a song off the radio for later query) or labels to track sales of their music (“intellectual property”), but often as a precursor to legal action buy the labels (“network scanning”).

- Systems either perform a match for the whole recording on a Table of Content (TOC) look up (Gracenote/CDDB for example) or perform computation to find a specific song (Audible Magic, being used in Myspace)

- Similar work is being done with Video Fingerprinting for YouTube
Audio Fingerprinting

- Audio Fingerprinting is the attempt to song based on its “signature”. It differs from Query-By-Humming in that it’s often looking for specific recordings rather than any version of a song.

- It can also be used for song ID (recording a song off the radio for later query) or labels to track sales of their music (“intellectual property”), but often as a precursor to legal action by the labels (“network scanning”).

- Systems either perform a match for the whole recording on a Table of Content (TOC) look up (Gracenote/CDDB for example) or perform computation to find a specific song (Audible Magic, being used in Myspace).

- Similar work is being done with Video Fingerprinting for YouTube.
The battle continues

- Changing even a small part of a video or sound file (adding a second or two of black for example) can change the blueprint

- So Fingerprinting companies started taking only a small sample of each song

- So users begin altering their entire content enough (using EQ) so it can evade filters

- MP3Tunes/Sideload tries to aggregate ALL free music from across the web in one place (your “music locker”)

- People even put music up on YouTube so it can be found by aggregators
And Now

• More resources on Music IR:
  • http://www.music-ir.org
  • http://www.ismir.net

• Over to Yiming: Picture Query and Synchronous Video