Metadata:
Automated generation

CS 431 - March 16, 2005
Carl Lagoze - Cornell University
Acknowledgement

• Liz Liddy (Syracuse)
• Judith Klavans (U. Maryland)
• IVia Project
What we’ve established so far

• In some cases metadata is important
  - Non-textual objects, especially data
  - Not just search (browse, similarity, etc.)
  - Intranets, specialized searching
  - Deep web

• Human-generated metadata is problematic
  - Expensive when professionally done
  - Flakey or malicious when non-professionally done
How much can automation help?

• **Trivial approaches**
  - Page scraping and trivial parsing

• **Non-trivial approaches**
  - Natural Language Processing
  - Machine Learning
    • Naïve Bayes
    • Support Vector Machines
    • Logistic Regression
DC-dot

- Heuristic parsing of HTML pages to produce embedded Dublin Core Metadata
- http://www.ukoln.ac.uk/metadata/dcdot/
Breaking the MetaData Generation Bottleneck

- Syracuse University, U. Washington - Automatic Metadata Generation for course-oriented materials

- **Goal**: Demonstrate feasibility of high-quality automatically-generated metadata for digital libraries through Natural Language Processing

- **Data**: Full-text resources from ERIC and the Eisenhower National Clearinghouse on Science & Mathematics

- **Metadata Schema**: Dublin Core + Gateway for Educational Materials (GEM) Schema
# Metadata Schema Elements

<table>
<thead>
<tr>
<th>Dublin Core Metadata Elements</th>
<th>GEM Metadata Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Contributor</td>
<td>• Audience</td>
</tr>
<tr>
<td>• Coverage</td>
<td>• Cataloging</td>
</tr>
<tr>
<td>• Creator</td>
<td>• Duration</td>
</tr>
<tr>
<td>• Date</td>
<td>• Essential Resources</td>
</tr>
<tr>
<td>• Description</td>
<td>• Grade</td>
</tr>
<tr>
<td>• Format</td>
<td>• Pedagogy</td>
</tr>
<tr>
<td>• Identifier</td>
<td>• Quality</td>
</tr>
<tr>
<td>• Language</td>
<td>• Standards</td>
</tr>
<tr>
<td>• Publisher</td>
<td></td>
</tr>
<tr>
<td>• Relation</td>
<td></td>
</tr>
<tr>
<td>• Rights</td>
<td></td>
</tr>
<tr>
<td>• Source</td>
<td></td>
</tr>
<tr>
<td>• Subject</td>
<td></td>
</tr>
<tr>
<td>• Title</td>
<td></td>
</tr>
<tr>
<td>• Type</td>
<td></td>
</tr>
</tbody>
</table>
Method: Information Extraction

• Natural Language Processing
  – Technology which enables a system to accomplish human-like understanding of document contents
  – Extracts both explicit and implicit meaning

• Sublanguage Analysis
  – Utilizes domain and genre-specific regularities vs. full-fledged linguistic analysis

• Discourse Model Development
  – Extractions specialized for communication goals of document type and activities under discussion
Information Extraction

Types of Features:

- Non-linguistic
  - Length of document
  - HTML and XML tags
- Linguistic
  - Root forms of words
  - Part-of-speech tags
  - Phrases (Noun, Verb, Proper Noun, Numeric Concept)
  - Categories (Proper Name & Numeric Concept)
  - Concepts (sense disambiguated words / phrases)
  - Semantic Relations
  - Discourse Level Components
Sample Lesson Plan

Stream Channel Erosion Activity

Student/Teacher Background:
Rivers and streams form the channels in which they flow. A river channel is formed by the quantity of water and debris that is carried by the water in it. The water carves and maintains the conduit containing it. Thus, the channel is self-adjusting. If the volume of water, or amount of debris is changed, the channel adjusts to the new set of conditions.

Student Objectives:
The student will discuss stream sedimentation that occurred in the Grand Canyon as a result of the controlled release from Glen Canyon Dam.

...
NLP Processing of Lesson Plan

Input:
The student will discuss stream sedimentation that occurred in the Grand Canyon as a result of the controlled release from Glen Canyon Dam.

Morphological Analysis:
The student will discuss stream sedimentation that occurred in the Grand Canyon as a result of the controlled release from Glen Canyon Dam.

Lexical Analysis:
The student will discuss stream sedimentation that occurred in the Grand Canyon as a result of the controlled release from Glen Canyon Dam.
Syntactic Analysis - Phrase Identification:

The student will discuss stream sedimentation that occurred in the Grand Canyon as a result of the controlled release from Glen Canyon Dam.

Semantic Analysis Phase 1 - Proper Name Interpretation:

The student will discuss stream sedimentation that occurred in the Grand Canyon as a result of the controlled release from Glen Canyon Dam.
NLP Processing of Lesson Plan (cont’d)

Semantic Analysis Phase 2 - Event & Role Extraction

Teaching event: discuss
actor: student
topic: stream sedimentation

event: stream sedimentation
location: Grand Canyon
cause: controlled release
HTML Document

MetaExtract

HTML Converter

eQuery Extraction Module

PreProcessor Tf/idf

Metadata Retrieval Module

Configuration

Cataloger
Catalog Date
Rights
Publisher
Format
Language
Resource Type

Creator
Grade/Level
Duration
Date
Pedagogy
Audience
Standard

Title
Description
Essential Resources
Relation

Keywords

Output Gathering Program

HTML Document with Metadata
<table>
<thead>
<tr>
<th><strong>Automatically Generated Metadata</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title:</strong> Grand Canyon: Flood! - Stream Channel Erosion Activity</td>
</tr>
<tr>
<td><strong>Grade Levels:</strong> 6, 7, 8</td>
</tr>
<tr>
<td><strong>GEM Subjects:</strong> Science--Geology, Mathematics--Geometry, Mathematics--Measurement</td>
</tr>
<tr>
<td><strong>Keywords:</strong> Proper Names: Colorado River (river), Grand Canyon (geography / location), Glen Canyon Dam (buildings&amp;structures)</td>
</tr>
<tr>
<td><strong>Subject Keywords:</strong> channels, clayboard, conduit, controlled_release, cookie_sheet, cup, dam, flow_volume, hold, paper_towel, pencil, reservoir, rivers, roasting_pan, sand, sediment, streams, water,</td>
</tr>
</tbody>
</table>
Automatically Generated Metadata (cont’d)

Pedagogy: Collaborative learning
         Hands on learning
Tool For: Teachers
Resource Type: Lesson Plan
Format: text/HTML
Placed Online: 1998-09-02
Name: PBS Online
Role: onlineProvider
Homepage: http://www.pbs.org
Project CLiMB Computational Linguistics for Metadata Building

- Columbia University 2001-2004
- Extract metadata from associated scholarly texts
- Use machine generation to assist expert catalogers
Problems in Image Access

- Cataloging digital images
- Traditional approach:
  - manual expertise
    - labor intensive
    - Expensive
    - General catalogue
      records not useful
      for discovery
- Can automated techniques help?
  - Using expert input
  - Understanding contextual information
  - Enhancing existing records
September 14, 1908, the basis of the Greenes' final design had been worked out. It featured a radically informal, V-shaped plan (that maintained the original angled porch) and interior volumes of various heights, all under a constantly changing roofline that echoed the rise and fall of the mountains behind it. The chimneys and foundation would be constructed of the sandstone boulders that comprised the local geology, and the exterior of the house would be sheathed in stained split-redwood shakes. —Edward R. Bosley. Greene & Greene. London : Phaidon, 2000. p. 127
Chinese Paper Gods
Anne S. Goodrich Collection

C.V. Starr East Asian Library,
Columbia University
Pan-hu chih-shen
God of tigers
Alex Katz
American, born 1927
Six Women, 1975
Oil on canvas
114 x 282 in.
Alex Katz has developed a remarkable hybrid art that combines the aggressive scale and grandeur of modern abstract painting with a chic, impersonal realism. During the 1950s and 1960s—decades dominated by various modes of abstraction—Katz stubbornly upheld the validity of figurative painting. In major, mature works such as Six Women, the artist distances himself from his subject. Space is flattened, as are the personalities of the women, their features simplified and idealized: Katz’s models are as fetching and vacuous as cover girls. The artist paints them with the authority and license of a master craftsman, but his brush conveys little emotion or personality. In contrast to the turbulent paint effects favored by the abstract expressionist artists, Katz pacifies the surface of his picture. Through the virtuosic technique of painting wet-on-wet, he achieves a level and unifying smoothness. He further “cools” the image by adopting the casually cropped composition and overpowering size and indifference of a highway billboard or big-screen movie.

In Six Women, Katz portrays a gathering of young friends at his Soho loft. The apparent informality of the scene is deceptive. It is, in fact, carefully staged. Note the three pairs of figures: the foreground couple face each other; the middle ground pair alternately look out and into the picture; and the pair in the background stand at matching oblique angles. The artist also arranges the women into two conversational triangles. Katz studied each model separately, then artfully fit the models into the picture. The image suggests an actual event, but the only true event is the play of light. From the open windows, a cordial afternoon sunlight saturates the space, accenting the features of each woman.

http://ncartmuseum.org/collections/offviewcaptions.shtml#alex
Segmentation

• Determination of relevant segment
• Difficult for Greene & Greene
  - The exact text related to a given image is difficult to determine
  - Use of TOI to find this text
• Easy for Chinese Paper Gods and for various art collections
• Decision: set initial values manually and explore automatic techniques
Text Analysis and Filtering

1. Divide text into words and phrases
2. Gather features for each word and phrase
   • E.g. Is it in the AAT? Is it very frequent?
3. Develop formulae using this information
4. Use formulae to rank for usefulness as potential metadata
What Features do we Track?

• Lexical features
  - Proper noun, common noun
• Relevancy to domain
  - Text Object Identifier (TOI)
  - Presence in the Art & Architecture Thesaurus
  - Presence in the back-of-book index
• Statistical observations
  - Frequency in the text
  - Frequency across a larger set of texts, within and outside the domain
Techniques for Filtering

1. Take an initial guess
   - Collect input from users
   - Alter formulae based on feedback

2. Use automatic techniques to guess (machine-learning)
   - Collect input from users
   - Run programs to make predictions based on given opinions (Bayesian networks, classifiers, decision trees)

3. The CLiMB approach: Use both techniques!
Initial Manual Filter

- Increase score if proper noun;
- Decrease score if very frequent in Brown corpus;
- Increase score if frequent in back-of-book indexes;
- Increase score if particularly frequent in domain specific texts;
- Increase score if present in authority lists
Early Results

Cordelia Culbertson
Greene
James Culbertson
James A. Culbertson
house
special furnishings Charles
Cordelia A. Culbertson house
Blacker house
Tichenor house
bedrooms
Greene furniture
Pacific Coast Architect
Culbertson residence
single-story elevation
Georgia O'Keeffe (American, 1887-1986)

Cebolla Church, 1945

Oil on canvas, 20 1/16 x 36 1/4 in. (51.1 x 92.0 cm.)
Purchased with funds from the North Carolina Art Society (Robert F. Phifer Bequest), in honor of Joseph C. Sloane, 72.18.1

North Carolina Museum of Art

MARC format

100 O'Keeffe, Georgia, ≠d 1887 -1986.
245 Cebolla church ≠ h [slide] / ≠ c Georgia O'Keeffe.
260 ≠c2003
300 1 slide : ≠ b col.
500 Object date: 1945.
500 Oil on canvas.
500 20 x 36 in.
535 North Carolina Museum of Art ≠ b Raleigh, N.C.
650 Painting, American ≠ y 20th century.
650 Women artist ≠ z United States
650 Church buildings in art.
Driving through the New Mexican highlands near her home, Georgia O'Keeffe would often pass through the village of Cebolla with its rude adobe Church of Santo Niño. The artist was moved by the poignancy of the little building: its sagging, sun-bleached walls and rusted tin roof seemed so typical of the difficult life of the people.

When O'Keeffe came to paint the church she addressed it directly, emphasizing its isolation and stark simplicity. Literally formed out of the earth, the building affirms the permanence and the hard, defiant patience of the people. For O'Keeffe, it symbolized human endurance and aspiration. "I have always thought it one of my very good pictures", she wrote, "though its message is not as pleasant as many others".

And the question remains: What is that in the window?
Data Fountains

• fully-automated collection aggregation and metadata generation
• semi-automated approaches that strongly involve and amplify the efforts of collection experts
• U.C. Riverside
Architecture overview of DF

- Seed Set Generator
- Nalanda iVia Focused Crawler
- Distiller
- Metadata Exporter
Seed Set Generator

- Seed sets are sets of URL’s that define a topic of interest
- Seed sets can be supplied in various formats by a client (e.g. simple text file with a list of URL’s)
- Typically need around 200 highly topic-specific URL’s
- Problem: most users would come up with only a few dozen
- Solution: scout module uses a search engine such as Google to fatten up the user-provided initial set
Nalanda iVia Focused Crawler

- Primarily developed by Dr. Soumen Chakrabarti (IIT Bombay), a leading crawler researcher
- Sophisticated focused crawler using document classification methods and Web graph analysis techniques to stay on topic
- Supports user interaction via URL pattern blacklisting etc
- Uses a classifier to prioritize links that should be followed
- Returns a list of URL’s likely to be on the initial seed set topic
Distiller

- Attempts to rank URL’s returned by the NIFC according to their relevance to the client-provided topic

- Uses improved Kleinberg-like Web graph analysis to assign hub and authority values to each URL

- Returns scores for each provided URL
Metadata Exporter

- Final stage of DF
- Provides clients with convenient data formats to incorporate the best on-topic URL’s into their own databases
- Allows different amounts/quality of metadata to be exported based on the client’s selected service model
- Supports various export types and file formats (simple URL lists, delimiter-separated file formats, XML file formats, MARC records and export via OAI-PMH)
Classification: Example Subject Categories

- **LCC**: Library of Congress Categories
- **LCSH**: Library of Congress Subject Headings
- **INFOMINE** Subject Categories
  - Biological, Agricultural, and Medical Sciences
  - Business and Economics
  - Cultural Diversity
  - Electronic Journals
  - Government Info
  - Maps and Geographical Information Systems
  - Physical Sciences, Engineering, and Mathematics
  - Social Sciences and Humanities
  - Visual and Performing Arts
Example

The Korea Rice Genome Database

The Korea Rice Genome Database WWW Server is maintained by the Dept of Biological Science, Myongji University as a part of Korea Rice Genome Research Program supported by the National Institute of Agricultural Science and Technology (NIAST), Rural Development Administration (RDA) under the direction of Science and Technology Policy Institute (STEPI).

Current Status of Rice Genome Sequencing by Internation Consortium

Rice EST Database

Search for Rice EST Information

The Rice EstDB registered in GENBANK is updated every other month and EST sequence is analyzed by blastx. The homology is analyzed by blastx ver 2.0.

You can download compressed form of the latest version of about 100,015 Rice cDNA sequences containing about 800 complete coding sequences in the following format as of Aug. 2002.

RiceEST promoted.gz: Fasta format of all GenBank Rice ESTs
RiceEST gbk.gz: GenBank format of all GenBank Rice ESTs
completeCD fasta Z: Fasta format of all Rice complete coding sequences
completeCD gb.Z: GenBank format of all Rice complete coding sequence

Rice Map

Search for Rice Map Information

Academia Sinica Map
Cornell Map
Japan NIAR-STAFF Tsukuba Map
Korea Map - NIAST, RDA
Morphological Map
Example: Korea Rice Genome Database

- Is it about…
  - Geography?
  - Agriculture?
  - Genetics?

- Which **INFOMINE** category do we put it in?
  - Biological, Agricultural, and Medical Sciences

- Pretty obvious, right?
  - For *humans*, yes. But how do we automate it?
Automating Document Classification

• We need a way to measure *document similarity*

• Each document is basically just a list of words, so we can count how frequently each word appears in it

• These word frequencies are one of many possible *document attributes*

• Document similarity is mathematically defined in terms of document attributes
Automating Document Classification

- The previous slide contains 51 words
  - *document* 6
  - *word, of* 3 each
  - *we, a, in, is, each* 2 each
  - All other words 1 each

- Note that we consider words such as *word* and *words* to be the same

- We also don’t care about capitalization

- In general, we’d also ignore non-descriptive words such as *we, a, of, the, and so on*
Automating Document Classification

- Not an easy task
  - The distribution of words shows that the slide in question is not very rich in content
    - The most frequent word (*document*) is not very descriptive
    - The most descriptive word (*classification*) does not appear very frequently in the slide
  - How descriptive and how frequent a word should be depends on the category

- The task is easier when:
  - we have a large number of content-rich documents
  - categories are characterized by very specific words which don’t appear very frequently in other categories
Automating Document Classification

- Two documents sharing a large number of category-specific words are considered to be very similar to each other.

- Document similarity can thus be quantified and computed automatically.

- Documents can then be ranked by their similarity to each other.

- A large group of documents that are all very similar to each other can then be considered to *define* the category they belong to (the set of all such groups is called the *Training Corpus*).

- One way to classify a document is then to put it in the same category as that of the training document that it’s most similar to.
Automating Document Classification

- The classification method just described is known as the *Nearest Neighbor* method.

- There are other methods, which may be more suited for the classification of documents from the Internet:
  - *Naïve Bayes*
  - *Support Vector Machine (SVM)*
  - *Logistic Regression*

- *Infomine* uses a flexible approach – supporting all of these methods – in an attempt to produce highly-accurate classifications.