

Metadata Extraction and Quality Evaluation

The Paper

- *Developing Practical Automatic Metadata Assignment and Evaluation Tools for Internet Resources*
 - Gordon Paynter, 2005
 - (In the proceedings of the JCDL 2005)

Participants in iVia, NSDL iVia and Data Fountains:

- **IMLS**
- **Library of the University of California, Riverside:**
Ruth Jackson, Steve Mitchell, Johannes Ruscheinski, Paul Vander Griend, Walter Howard, Jason Scheirer
- **NSF NSDL Core Integration, Cornell University:**
Carl Lagoze and John Saylor
- **Computer Science Department, Cornell University:**
Rich Caruana and Thorsten Joachims
- **Computer Science Department, University of Massachusetts:**
Andrew McCallum
- **University of California, Riverside, Computing and Communications:**
Charles Rowley, Jerry Keith and Tim Paul
- **Indian Institute of Technology, Bombay:**
Soumen Chakrabarti

The Project

iVia/DataFountains

- Open source web crawler (everything released under GPL/LGPL)
- Generates quality metadata from web resources
- Extracts information from natural language in web resources
- Generates and extracts data for controlled/limited vocabulary schema (LCSH)
- Resource discovery
- Focused and unfocused crawls

Metadata Extraction and Generation

Extraction v. Generation

Extraction

- Taking information that is in the document: titles, provided keywords, etc
- Sometimes obvious
- Sometimes not (summaries, finding authors without Dublin core, etc.)

Generation

- What's not there: general topic, keywords (if not provided), etc.
- One important family of generation is classification

The Easy Part:

Extraction: What's Explicitly There

Dublin Core Elements

- Extracting Dublin Core elements from HTML is easy:
 - `<meta name="DC.Author" content="William Shakespeare" />`
 - `<meta name="DC.Title" content="A Midsummer Night's Dream" />`

Unfortunately

Dublin Core is nice but it doesn't have everything we want

We want a lot more metadata fields than DC has defined

Very few sites on the internet actually *have* DC data

DC data is often misleading, wrong or incomplete

The Hard Part:

Generation and Classification
(what's not there)

Extraction (When the answers aren't
obvious)

Things Working On Our Side

- Large amounts of training data
- Collaborators who are leaders in their fields
- Clever algorithms and strategies
 - SVM, Logistic Regression, HITS/PageRank, PhraseRate
- Large amounts of computing resources
- Implicit structure in web pages

Some Fields

Titles

- Title meta tags
- Dublin Core titles
- Page title in `<title>` tag
- `<h1>` header tags
- The words in the first 50 characters of the page

Creator

- Dublin Core
- Named Entity Extraction (very early versions, not yet out of testing)

Keyphrases

- Keyword[s] tags in <meta>
- PhraseRate
 - Creates short (3-5 word) keyphrases using the structure of the page as a guide
- Combines the two, giving PhraseRate generated phrases higher weight (spamming)

LCC/LCSH Headings

- Purely statistical classification
- Naïve Bayes, Logistic Regression and Pachinko
 - SVM in testing

Extraction and Generation/Classification

Different approaches for evaluating
each

Evaluating Extraction

- (Near-) Purely algorithmic
- Iterative: if we aren't getting good data, we need to go back to programming something better
- Overlap and subjectivity: very seldom are there absolutely wrong and right answers

Evaluating Generated Metadata

Assignment and Recall

- Assignment accuracy: percentage of correct subjects assigned

$$\frac{\textit{number_of_correct_assignments}}{\textit{total_number_of_assignments}}$$

- Recall accuracy: percentage of correct subjects returned

$$\frac{\textit{number_of_correct_assignments}}{\textit{total_number_of_expert_assigned_values}}$$

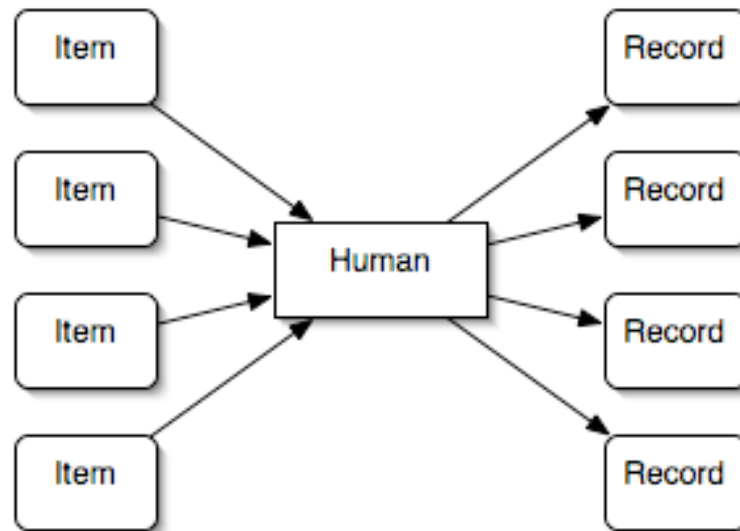
A problem with these metrics

- If you assign 100% of subjects to each document, you have 100% accuracy in assignment since you have assigned it all the subjects it belongs to (and then some)
- If you return 100% of subjects for a document, you have 100% recall accuracy since every correct subject is there (and then some)
- Introduce a score performance penalty for incorrect assignment (items that shouldn't be there)

Evaluation and Generation

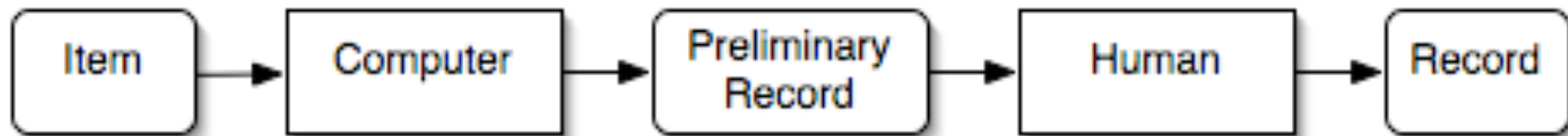
Human Generation

- Have a human assign all metadata to every record
- Expensive: reviewing hundreds of thousands of records will have large costs in expert time and money



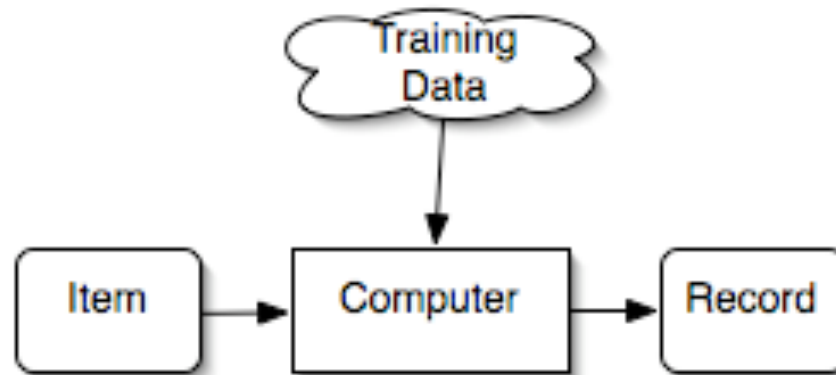
Computer-Assisted Human Generation

- Take a random record and assign metadata using automatic assigners
- Have an expert rate the appropriateness of the data and tweak the assigners to better fit the expert's suggestions
- Still expensive: you want to minimize the amount of human interaction



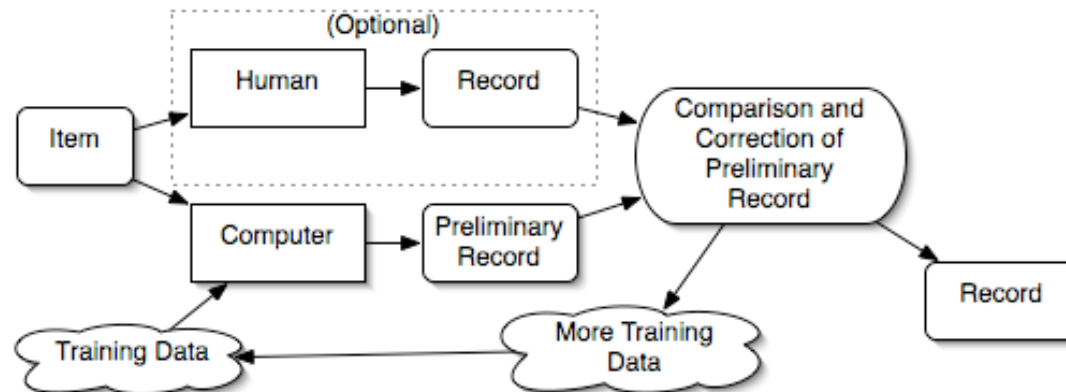
Automatic Generation

- Train metadata classifiers on INFOMINE records
- Take a random expert-created record in INFOMINE
- Run our metadata extraction/generation/classification tools on it



Another, iterative generation/evaluation method

- Continuously compare our (large-volume) automatically created records with our (small-volume) expert-created records as new INFOMINE records become available
- Operate in chronological order: newest expert records take precedence over older record
- Correlate how well our machine classifies records v. the human data
- Adjust automatic classifiers accordingly



Advantages to this approach

- We still have human-assigned data
- Humans don't need to guide the automatic assigners -- the automatic assigners shadow the human data and use it to continually train classifiers
- As humans change how they classify items diachronically, computer picks up on trends and follows

Future Directions

- More streamlined human to computer classifier integration and editing workflow, including visualization
- Instilling trust in users over metadata generation
- Using human data to pick better classifiers from sets of classifiers (ensemble classification instead of just beating on a single classifier until it fits)
- More sophisticated named entity extraction (CRFs)
- There is always room for improvement in classifiers and extractors

Questions