A Metadata Generation System for Scanned Scientific Volumes

Xiaonan Lu¹, Brewster Kahle³, James Z. Wang¹,², and C. Lee Giles¹,²

¹Department of Computer Science and Engineering
²College of Information Sciences and Technology
The Pennsylvania State University, University Park, PA
³Internet Archive, San Francisco, CA

xlu@cse.psu.edu, brewster@archive.org, jwang@ist.psu.edu, giles@ist.psu.edu
Background: Biodiversity Heritage Library at the Internet Archive

- Over 400,000 scanned items in text archive (www.archive.org)
  - Batch scanned, OCRed, Historical
  - No efficient navigation of content: Complex objects, Little or no metadata available
Related work: Automatic metadata extraction (a version of the NLP problem)

- Long standing problem
  - Many data types: text, images, audio, etc.
  - Existing systems: CiteSeerX, Scholar, Rexa, etc.

- References
  - Descriptive metadata generation
    - Scientific digital libraries: research paper metadata extraction [Giuffrida, etc. 2000] [Han, etc. 2003] [Mao, etc. 2004]
    - Dublin Core + GEM metadata extraction [Yilmazel, etc. 2004]
    - NLM metadata generation [Mao, etc. 2000]
    - Metadata Research Center, the School of Information and Library (SILS), UNC
  - Structural metadata generation
    - Personalized content access [Dushay, 2002]
    - Page classification [Cesarini, 2001]
    - Speech transcript generation [Liu, etc., 2005]
Our document analysis project

☑ Goal
  ☑ Analyze scanned and OCR'd documents
    ☑ Volumes of Journals, books, ...
  ☑ Generate metadata for browsing, retrieval, and presentation of content
    ☑ Structural metadata
    ☑ Descriptive metadata

☑ Collaborators
  ☑ Internet Archive:
    ☑ Brewster Kahle, Hank Bromley, ...
  ☑ Smithsonian Institute:
    ☑ Tom Gannett, Martin Kalfatovic, Suzanne Pilsk, and Conrad Ziyad
  ☑ Missouri Botanical Garden:
    ☑ Chris Freeland
  ☑ Penn State:
    ☑ Xiaonan Lu, Dr. James Z. Wang, and Dr. C. Lee Giles

7/24/2008 The Pennsylvania State University
Problem statement

- **Scope:** scanned scientific volumes
  - unit: volume (200~1000 pages/volume)
  - multiple journal issues, many articles
- **Problem:** automatic generation of metadata
  - at volume, issue, and article levels
- **Support content navigation**
- **Applications**
  - Digital libraries, museums, and archives
Challenges

- Diversity in formatting styles from various historical journals
- OCR errors
Our contributions

- Automatic metadata generation software
  - Analyze scanned scientific volumes
  - Generates metadata describing content
  - Support efficient access to content

- A supervised learning based method
  - Label metadata elements
  - Based on style, linguistic, and format features of OCRed text

- Experiments on actual archives
  - Various journals
  - Integration
Experiments at the Internet Archive

Article navigation services in test mode at the Internet Archive
Metadata generation software requirements

- **Input**
  - OCRed text (txt, djvu.xml,...)
  - Manually generated metadata (volume title and page number assertions)
  - PDF file (for comparison and evaluation)

- **Functionality**
  - Parse OCRed text and position information
  - Calculate word, line, paragraph, and page features for metadata tagging
  - Generate structural and descriptive metadata represent in XML

- **Output**
  - Volume, issue information and positions
  - List of articles and their descriptive metadata: title, author, start page, ...
Software components

- Identify and automatically extract metadata
  - At volume and issue level: volume no, volume start and end page, Issue no, issue start and end page, etc.
  - At article level: title, author, and location, etc.
- Use both text and image information

Metadata generation software
What input type to use?

- Available file formats:
  - Pure text, DjVu XML, PDF.

- Choose DjVu XML
  - Full OCRRed text
  - Physical structures
  - Bounding box

```
\begin{verbatim}
<<PARAGRAPH>
By E. A. SCHWARZ.
\end{verbatim}
```

7/24/2008 The Pennsylvania State University
Investigate page and line features

- Page and line feature can determine
  - Special pages (title pages, empty pages)
  - Start of articles
  - Link page number with page image files
Automatically generate page and line features

- Parse human generated metadata
  - Volume title, creation date, etc.
  - Page number and page image correspondences
- Parse paragraph, line, word contained within DjVu XML
  - Calculate page features (# of lines, words, special patterns etc.)
  - Calculate line features (# of words, alignment, average size, position, etc.)
Line features

- **Style features**
  - Capital, alignment

- **Linguistic features**
  - Entity name, special words

- **Context features**
  - Position, space between neighboring lines

- **Font features**
  - Word and character height and width

<table>
<thead>
<tr>
<th>Feature ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Capital mode</td>
</tr>
<tr>
<td>2</td>
<td>Average word height</td>
</tr>
<tr>
<td>3</td>
<td>Average character width</td>
</tr>
<tr>
<td>4</td>
<td>Normalized word height</td>
</tr>
<tr>
<td>5</td>
<td>Normalized character width</td>
</tr>
<tr>
<td>6</td>
<td>Alignment</td>
</tr>
<tr>
<td>7</td>
<td>Line position</td>
</tr>
<tr>
<td>8</td>
<td>Line ID</td>
</tr>
<tr>
<td>9</td>
<td>Distance to previous line</td>
</tr>
<tr>
<td>10</td>
<td>Distance to next line</td>
</tr>
<tr>
<td>11</td>
<td># of words</td>
</tr>
<tr>
<td>12</td>
<td>Person name</td>
</tr>
<tr>
<td>13</td>
<td>Special words</td>
</tr>
<tr>
<td>14</td>
<td>Paragraph begin</td>
</tr>
</tbody>
</table>
Calculate line features

- Avg. word height (m: # of words)
  \[ \frac{\sum_{i=1}^{m} (y_1^i - y_2^i)}{m} \]

- Avg. letter width (n: # of letters within a word)
  \[ \frac{\sum_{i=1}^{m} (x_2^i - x_1^i)}{\sum_{i=1}^{m} n_i} \]

- Normalized word height
  \[ \frac{\sum_{j=1}^{m_i} (y_1^{i,j} - y_2^{i,j})}{m_i} \]
  \[ \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} (y_1^{i,j} - y_2^{i,j})}{m_i} \]
  \[ \frac{n}{\sum_{i=1}^{n} \sum_{j=1}^{m_i} k_{i,j}} \]

- Normalized letter width
  \[ \frac{\sum_{j=1}^{m_i} (x_2^{i,j} - x_1^{i,j})}{\sum_{i=1}^{m_i} k_{i,j}} \]
  \[ \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} (x_2^{i,j} - x_1^{i,j})}{n} \]
Volume and issue level: title page detection

• Combine page number information and special patterns (key words like “Volume”, “issue”, etc.)

• Use Levenshtein distance to tolerate OCR error
  – Definition: minimum number of operations (insertion, deletion, substitution) needed to transform. \( L(\text{“Volume”}, \text{“Volme”}) = 1 \)
  – Why? : measure the amount of difference, widely used in spell checkers
  – Other distance metrics: hamming distance (only for equal length)
Article level: find article start lines

- Integrate line features
  - Style, linguistic, context and font features
- Supervised (SVM) learning based classification
  - Classify lines into title begin, title end, and others
Generate article metadata items

• Title and author
  – Title: between title begin and title end
  – Author follows title end, using keyword match and person name detection

• Volume and issue
  – Use metadata input and rule-based title page detection

• Start page number, PDF page number, page leaf num
  – Based on metadata, and collection of “objects” in the raw OCRed text

• Start page index, start page image file
  – Recorded page properties

• End page
  – Before the next detected article
Experimental results

- Implementation
  - Linux platform
  - Java DOM, Perl, and SVM Light
- Data sets
  - Journals contributed by the Smithsonian Institute based on the needs of the Biodiversity Heritage Library (BHL) project
  - Digitized and hosted by the Internet Archive for public access
  - Sample journals as shown in table:

<table>
<thead>
<tr>
<th>Volume title</th>
<th>Century</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proceedings of the Entomological Society of Washington</td>
<td>19th</td>
</tr>
<tr>
<td>Proceedings of the Biological Society of Washington</td>
<td>19th</td>
</tr>
<tr>
<td>Journal of Natural Philosophy, Chemistry &amp; the Arts</td>
<td>19th</td>
</tr>
<tr>
<td>Magazine of Natural History and Journal of Zoology, Botany,</td>
<td>19th~19th</td>
</tr>
<tr>
<td>Mineralogy, Geology and Meteorology</td>
<td></td>
</tr>
</tbody>
</table>
Performance metrics

• Standard metrics
  – Precision: # of correctly labeled articles / # of labeled articles
  – Recall: # of correctly labeled articles / # of articles within the volume

• Method
  – Select sample journals and manually labeled metadata
  – Manually compare metadata generated by the software with ground-truth
# Performance

<table>
<thead>
<tr>
<th>Sample Journals</th>
<th># of articles</th>
<th># of labeled articles</th>
<th># of correct labels</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>146</td>
<td>141</td>
<td>138</td>
<td>98%</td>
<td>94%</td>
</tr>
<tr>
<td>Sample 2</td>
<td>203</td>
<td>201</td>
<td>134</td>
<td>67%</td>
<td>66%</td>
</tr>
<tr>
<td>Sample 3</td>
<td>180</td>
<td>162</td>
<td>147</td>
<td>91%</td>
<td>82%</td>
</tr>
</tbody>
</table>

Sample 1, 2, 3 corresponds to Proceedings of the Entomological Society of Washington, Journal of Natural Philosophy, Chemistry & the Arts, and Magazine of Natural History and Journal of Zoology, Botany, Mineralogy, Geology and Meteorology, respectively.
Sample failed cases

ACCOUNT OF BOOKS.

A Treatise on Astronomy, in which the Elements of the Science are Cogently deduced in a natural Order, from the Appearances of the Heavens, Astronomy, is an Observer of the Earth's phenomena, demonstrated on Mathematical Principles, and explained by an Application to the various Phenomena. By Oliphant Gregory, Teacher of Mathematics, Cambridge. 8vo. 328 Pages, with Nine Plates. London, sold by Kearsley.


Observations on the Winds and Monsoons, Illustrated with a Chart, Copper engraving, and accompanied with Notes, Geographical and Meteorological. By James Capper, formerly Coloured, and Comptroller-General of the Army and Fortification accounts on the Coast of Coromandel. Quarto, 294 Pages, Dobell.

Accounts of the three last mentioned Works will appear in our next.


This pamphlet contains minute instructions for preparing sugar, syrups, confections, bran, and arrack, from the white beet-root (sugar beet). In order to obtain sugar from the beet-root, the grower directs the roots to be cut into slices, then expels the juice, to boil it down to a strong confidence, and to suffer it to crystallize. The residue, from which the juice has been expressed, may be used for obtaining brandy, but it is better to boil the beet-root to reduce it to a pulp, suffer it to ferment, and then suffer it to distill. In order to distill the spirit of the popular flavour of the root, it will be necessary to distil charcoal powder. Eighty pounds of beet-root yielded eight quarters of brandy. To obtain sugar, the roots are likewise boiled, the juice expressed, and mixed with charcoal powder, evaporated to one-third, then suffered to ferment, and afterwards distilled with the charcoal powder. In order to obtain arrack, the juice is to be evaporated to one-half, then suffered to ferment, and distilled.

ART OF SHAVING.

Undressed with sulphuric acid to prevent oxidation, it yielded a salt of a difficult solubility; but if this acid was added in excess, the salt was readily soluble in water. When igniting it became hard and insoluble in acids. I therefore do not hesitate to believe, that the earth was electric, though not absolutely pure, but mixed with iron.

From these experiments it follows that 200 grms of this hyacinth-like stuff, consist of 100 grms of earth (D):

16 —— aluminium (Al)
12 —— oxide of iron (I).
20 —— zirconia (Zn)
15 —— loss of matter during the operation.

XI.

Philosophical Observations on the Process of common Life.—Art of Shaving.—W. N.

The caprice of fashion, or the modern improvements in habits of the superficial mind, has deprived all the nations of Europe of their hands; and consequently a portion of the goods are always inconsiderable of the small and daily conveniences of life must depend on the facility with which this appendage can be taken off. This subject is frequently treated in conversation; and I am persuaded that many individuals will be glad to know what can be said on the whole respecting it. The fabrication of a good razor depends on to many circumstances: irritations of the various parts of the head, and the hardness of the temper, that are inevitable, after he has exercised his utmost skill, in the last instance, select such instruments as are commonly superior to others, by trial or actual use. I am not aware of any means of choosing a good razor out of a number. All that I can say in this respect is, that a bad razor cannot be easily brought to a fine, or even a moderate edge; for which reason I should prefer that razor which possesses the best edge, and has been skillfully touched upon the hone, that is to say, the razer which, upon being looked...
Future work

• Use domain expert feedback to improve performance
• Create link between “table of contents” and article pages
• Process other historical documents
• Index figures, maps, etc.
  – See for example ArchSeer
  – http://archseer.ist.psu.edu