Automatic Metadata Generation for Scanned Scientific Volumes

Xiaonan Lu[1] and Brewster Kahle[2]
[1]Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA
[2]Internet Archive, San Francisco, CA

ABSTRACT
Large scale digitization projects have been conducted at the Internet Archive digital library to preserve cultural artifacts and to provide permanent access. The increasing amount of digitized resources requires advanced tools and methods that will efficiently analyze and manage digitized resources. In this position paper, we identify several issues related to scanned books projects, present our initial work on automatic metadata generation for scanned scientific journals, and suggest potential future actions.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *Collection, Systems issues;* H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms
Algorithms, Experimentation, Verification

Keywords
Metadata Generation, Scanned Journal

1. INTRODUCTION
Large scale digitization projects are underway at digital libraries to preserve cultural artifacts and to provide open web access. The Internet Archive has scanned and preserved seven collections with a total of more than 400,000 items in the text archive only. As an ongoing effort, ten major natural history museum libraries, botanical libraries, and research institutions have joined to form the Biodiversity Heritage Library (BHL). The BHL partners will digitize the published literature of biodiversity held in their respective collections and provide basic and important content for immediate research and for multiple bioinformatics initiatives [1].

In digital libraries, Digitalized resources are often compound objects consisting of a large number of scanned images of pages, OCRed text, and viewable PDF files generated by automatic scanning and recognition processes. In contrast, there is usually only a very limited amount of manually-generated metadata available to describe the structure and content of digitized resources. Thus, digital libraries need automated tools to generate metadata for digitized resources, to describe the intellectual content of compound objects and to connect different components [2].

We have worked on scanned books projects with original documents coming from various libraries and institutes. We aim to create on-line systems that advocate user engagement in scanned books. We are mostly interested in open source techniques that will potentially benefit the effort of scanning, preserving, and presenting historical books.

In this position paper, we present several problems we have observed while conducting scanned books projects at the Internet Archive: such as the choice of the metadata schema for scanned books, the OCR software for historical documents, and the tools that can automatically process large scale scanned books and help content based retrieval of scanned resources. We will also present our initial work on automatic metadata generation for bound volumes of scientific publications. Finally, we talk about potential future action directions.

BooksOnline'08, October 30, 2008, Napa Valley, California, USA.

2. ISSUES RELATED TO SCANNED BOOKS PROJECTS

2.1 OCR Performance
Based on our observations of OCRed text on scanned historical publications we have created, accuracy of the OCR process is relatively low due to several reasons. One reason is that printed text produced by old technologies, such as typewriter, may cause lower OCR accuracy. Besides, the mixture of text and symbols in different languages brings extra difficulties. For example, the Roman numerals are frequently used as index and person names may appear in different languages as the one of the main text. Finally, ancient variations of words no longer used these days cause OCR errors.

2.2 Metadata Schema
There are many existing metadata schemas used in various digital libraries for encoding descriptive, administrative, and structural metadata regarding objects within digital libraries. During the process of developing, maintaining, and upgrading digital libraries, we face important issues related to metadata schema, such as the choice of right metadata schemas for different types of libraries, the transform from one schema to another schema, the domain knowledge, and the compatibility issues.

2.3 Automatic Metadata Generation
We attempt to tackle the problem of automatic metadata generation for scanned books. Our current goal is to develop tools which can automatically generate structural and descriptive

metadata for digitized volumes of scientific journals. These resources usually contain rich content, including the hierarchy of issues, articles and various types of pages. The generated metadata will facilitate various web-based content access functionalities, such as content navigation and article search.

3. MACHINE LEARNING BASED METADATA GENERATION

We have developed an automatic metadata generation system which has been integrated into the Internet Archive for testing. The metadata generation system extracts metadata from volumes of printed journals. Specifically, the system takes OCRed text and limited human-generated metadata of a volume of journal as input, identifies different types of content within the volume, generates multi-level metadata, and then outputs the metadata in XML form. The automatically-generated metadata consists of descriptive information about the collection of published articles within the volume as well as correspondence between scanned images of pages, pages within viewable PDF file, and pages within the original volume. These metadata is used to support internal navigation functionalities.

We use a machine learning based approach to generate article level metadata. The occurrence of an article, i.e., the start line of an article is detected by classification of text lines based on line features.

3.1 Feature Extraction

The metadata generation system takes the DjVu XML file of a scanned volume as input and parses the hierarchy of objects contained within the file. The system calculates features for every word, line, paragraph, and page of the OCRed text.

In order to identify articles and extract descriptive metadata for articles, we choose the text line as the unit for feature extraction because the article title and the author information usually occupy one or a few consecutive lines. Additionally, text within the same line usually has the same style. Thus, line features are designed to estimate properties of OCRed text within a line, which can be calculated based on OCRed text and bounding box information in the DjVu XML file. Specifically, we have designed style features, semantic and linguistic features, structure and context features, and font features to estimate properties of text lines.

3.2 Metadata Generation

A supervised-learning based method is applied to detect articles. In the learning and classification process, text line is the basic unit. Various types of features have been extracted for every text line. And, every line in the scanned volume corresponds to one vector of feature values. In order to train the machine, we manually label every line in the training set, i.e., determine if it is the start of an article. The set of training data, including line features and labels, are fed into the learning module to create the learned model. Finally, the classifier takes feature vectors of unlabeled text lines and the previously trained model as input and output the class label of the text lines, i.e., tells if each text line starts an article.

After the start line of an article is detected, the article tile and the author information are generated by analyzing the article start line and limited following lines. Other article level metadata elements, such as the volume and issue number, are generated based on volume and issue level metadata.

3.3 Experiments

We randomly selected scanned volumes from the real world data set, manually checked several thousands of pages, and compared the automatically-generated metadata with the ground-truth metadata. The performance of the automatic metadata generation system demonstrates its usage for real world usage.

4. CONCLUSTION AND FUTURE DIRECTIONS

4.1 Conclusion

We have identified several problems related to our scanned books projects at the Internet Archive: including the metadata schema, the OCR process, and the automatic metadata generation. We presented our effort on the automatic metadata generation system which aims to achieve efficient and user-friendly access to scanned resources.

4.2 Future Directions

In the future, we plan to continue our work on automatic metadata generation system by integrating analysis of table of contents information. We would also work on selecting a metadata schema for our scanned books and other projects. Finally, we are very interested in universal OCR softwares.

5. REFERENCES

[1] Biodiversity Heritage Library: *http://www.biodiversitylibrary.org*.

[2] R. Wendler. LDI Update: Metadata in the Library. *Library Notes*, no. 1286 (July/August): 4-5, 1999.

[3] Xiaonan Lu, Brewster Kahle, James Z. Wang and C. Lee Giles, ``A Metadata Generation System for Scanned Scientific Volumes," *Proceedings of the ACM and IEEE Joint Conference on Digital Libraries***,** pp. 167-176, 2008.