Leveraging Existing Bibliographic Metadata to Improve Automatic Document Identification in Web Archives

The University of North Texas (UNT) Libraries in partnership with the University of Illinois Chicago (UIC) Computer Science Department are seeking IMLS support for an applied research grant that aligns with the agency goals to advance collections stewardship and access (#3.1), and with the long-term goal to improve access to digital resources housed in web archives. This goal aligns with the NLG Program's Objective #3.2 to support innovative approaches to digital collection management. This research project will build on findings from a previously funded IMLS research grant (LG-71-17-0202-17) that was a first effort in training machine models to help identify high-value documents and publications within web archives. This proposal seeks to incorporate existing bibliographic metadata related to state government documents collections to better train machine models and allow for a reduction in human effort, as the process is still time consuming and requires highly-trained content curators. The project team includes PI Mark Phillips, Associate Dean for Digital Libraries at the UNT Libraries, and Co-PI Cornelia Caragea, Associate Professor in the UIC Computer Science department. Two graduate students will assist in the project. The project team will partner with the Library of Michigan and the Internet Archive's Archive-It service as data sources to further test this new approach in building machine models for this task. Finally, an advisory committee of professionals from collecting institutions and machine learning researchers will provide guidance and advice for the project. We respectfully request $396,045 in support.

Project Justification

Web archives have continued to gain popularity in collecting institutions like libraries, archives, and museums. These harvested websites serve a wide range of needs including the preservation of important publications and documents for an institution's holdings. While accessible through the web archive itself, these materials are often not identified or described at the item level, preventing users from finding resources and institutions from maintaining complete collections. As more web content is collected, identifying resources within a web archive that meet an existing collection development scope becomes costlier and more daunting.

In 2017, the UNT Libraries and the UIC Computer Science department received IMLS support under the National Digital Platform category for a research project to evaluate the use of machine learning algorithms to identify and extract publications contained in existing web archives as a way of surfacing these documents. This research was remarkably successful, resulting in: [1] the development of gold-standard, manually-labeled datasets from three different web archive domains, an institutional repository from a web archive of a university domain, state publications from a web archive of a state government, and technical reports from a large federal agency; [2] the design of supervised machine learning models that accurately classify PDF documents from web archives against the scope for a given collection. *Despite this success, we identified two major challenges that serve as the research questions for this project.* First, how can large amounts of labeled data be generated for supervised approaches with less intensive human effort, which is often impractical? Second, how will models generalize under distribution or vocabulary shifts (e.g., on data from one state to another, or from one collection type/scope to another), when no labeled datasets are available in the target domain?

We are now seeking IMLS support for a research project to address these challenges through a two-phase plan: [1] explore the use of available bibliographic metadata (representing decades of work by librarians within the defined collection scopes) together with unlabeled data to create large machine-learning training sets while reducing human effort; and [2] focus on the design of unsupervised domain adaptation techniques, i.e., training on one domain and testing on another, to address model generalization under distribution or vocabulary shifts. As a by-product of this research, we will generate test datasets from curators at the Library of Michigan and their web-archived content via the Internet Archives' Archive-It service that will serve as evaluation test sets for in-domain and out-of-domain scenarios.

Project Work Plan

Our research project tests the assumption that we will be able to successfully incorporate existing bibliographic metadata and catalog records to aid automated classification algorithms in accurately extracting documents and publications that meet collection development policies for a variety of organizations. We will focus on state government documents held in web archives at the UNT Libraries (for Texas) and at the Library of Michigan (for Michigan). We will use existing bibliographic metadata at these institutions, which describe and implicitly encode collection decisions from decades of manual curation efforts.

During the first phase, the team will investigate distant supervision by mapping the bibliographic metadata onto existing datasets--annotated during our previous/completed research--to understand which metadata fields are highly correlated with the positive vs. negative class and exploit these correlations to automatically create large model training sets. Additionally, the team will treat the task as positive-unlabeled learning, in which an incomplete set of positive examples is available as well as a set of unlabeled examples (some positive and others negative). We will adapt the learning algorithm so that training examples have individual weights, i.e., positive examples are given unit weight and unlabeled examples are duplicated where one copy of each unlabeled example is made positive with some weight w and the other copy is made negative with weight 1-w. By reweighting the importance of these training examples, we aim to model the uncertainty in the negative examples. In the second phase, we will focus on the design of unsupervised domain adaptation models to understand how well models trained in one domain (Texas state publications) can be adapted to another domain (Michigan state documents).

An advisory board comprising both collection managers and machine-learning professionals will offer guidance and review documentation and methods.

Project Results

Our primary goal is to test solutions outlined here in response to the challenges identified in our previous research. We hope that leveraging bibliographic metadata created over time will reduce current human effort while improving precision and recall of automated extraction of “in-scope” resources from web archives to better meet collection stewardship and user access needs.

Deliverables will include: [1] approaches that incorporate existing knowledge from decades of effort encoded as metadata records in library catalogs to create large training sets using distant supervision and positive-unlabeled data. These datasets will serve as benchmarks for training and testing *robust* machine learning algorithms to identify and extract materials from web archives; [2] accurate unsupervised domain adaptation models and workflows to evaluate model generalization under distributional / vocabulary shifts; [3] annotated test collections from a different state (Michigan) and different collection types to evaluate the model generalization and transferability from one domain to another; [4] a white paper describing the current status of web archiving efforts to preserve state government documents collections. The project team will share all code and documentation via GitHub repositories maintained by the UNT Libraries (https://github.com/unt-libraries/); published output will be available in the UNT Scholarly Works Repository and UNT Data Repository.

Budget

We respectfully request $396,045 in IMLS funds: $178,308 in salaries (Phillips, 0.96 person months per year; Caragea, 1 person months per year; 2 graduate research assistants each at 6 person months per year), $35,178 in fringe benefits, $36,627 in tuition reimbursement for the graduate students; $8,000 for travel to digital library and machine learning conferences, $2,000 for publication fees and $135,932 in indirect costs at federally negotiated indirect cost rates for the University of North Texas and the University of Illinois at Chicago (41.5% and 59.5% respectively)..