Enhancing Metadata Assignment in a Digital Library with Background Knowledge

James Silas Creel
Texas A&M University
Department of Computer Science and Engineering

1. INTRODUCTION

Traditionally, information retrieval experts have addressed the problem of information overload. This pursuit has yielded automated electronic bibliographic card catalogs as well as effective search engines for enormous collections including the World Wide Web. Despite advances in this area, librarians lament the ambiguity or absence of metadata required to effectively communicate content.

In the 20th century, technocrats such as Vanevar Bush and J. C. R. Licklider foresaw a future in which automatic interactive systems aid humans in navigating and understanding the knowledge of science and humanities. Eventually, the Web and Hypertext Transfer Protocol presented a medium for the digitization and dissemination of knowledge in textual and multimedia forms. More recently, the Semantic Web has provided a framework to encode the metadata about such corpora and thereby automate information discovery at the finest levels of semantic granularity. Yet metadata of such scope and quality are seldom realized. Various obstacles inhibit the acquisition, encoding, and sharing of metadata. In the context of institutional repositories, users' ignorance of the technical language of bibliographic cataloging inhibits knowledge capture at the time of document ingestion. Furthermore, the technical languages of digital bibliographic catalogs suffer from the same semantic problems as other formal knowledge representation languages, including context-specific definitions and conflicting representations. The difficulties of large-scale knowledge formalization have been documented by scholars like Marvin Minsky and Lucy Suchman and realized all too well during large-scale projects such as Doug Lenat's Cyc.

However, the institutional repository offers structures to deal with the obstacles to capturing, encoding, and sharing metadata. The institutional repository not only encompasses schemas and standards for encoding information and metadata, but it is also oriented toward the intellectual domains of its host institution. This situation is ripe for knowledge-based approaches to repository management.

The work proposed here aims to enhance knowledge capture by leveraging semantic structures of document encodings, by statistically applying formal domain knowledge to suggest metadata values, and by eliciting knowledge from users about the topics of their documents. The setting for the proposed work is a DSpace digital repository with metadata encodings in the Dublin Core standard.

The proposed work extends the typical model of metadata assignment employed in modern digital libraries of using information extraction to assign values to metadata fields. An inherent problem with interpretation of such values is that they are supposed to signify various things – authors, documents, topics, etc., but extracted data are ambiguous with respect to their significations. Information extraction does not resolve references, but considerable work has been devoted to automatic disambiguation of terms occurring in natural language. However, in the digital library domain, name disambiguation has concentrated on specific types of references (predominantly author names) in isolation. In order to apply name disambiguation techniques to the variety of references appearing in metadata field values, one must enumerate and distinguish the possible referents for each type of name considered. We hypothesize that a formal knowledge base can structure concepts of referents sufficiently well to apply name disambiguation to a variety of metadata values extracted from documents. Crucially, by formally encoding concepts of referents of documents, we may comprehend the relations between the concepts. For example, an autobiography should have an *author* metadata field value signifying the same individual as a *subject* metadata field for that document.

Furthermore, we hypothesize that the joint application of information extraction and name disambiguation techniques to the metadata assignment process can produce more statistically accurate metadata values than either technique alone.

2. RELATED WORK

Automatic metadata assignment has received increased attention in recent literature. Approaches have concentrated on specific types of document templates, metadata fields, collections, and, increasingly, on knowledge-based approaches to document annotation.

2.1 Automatic Metadata Assignment in Digital Libraries

We examine several previous efforts at automatic metadata assignment in digital libraries: the iVia project [4], the paperBase system [8], document code analysis in the It's:learning [*sic*] system [1], a topic-based name disambigua-

tion technique using unsupervised learning [7], and a name disambiguation technique using supervised learning [9].

The iVia virtual library software, a prominent effort to develop automatic metadata assignment and evaluation tools in a digital library, was developed as a supplement to the INFOMINE digital library[1] as documented by Gordon W. Paynter in [4]. Given the URL of an HTML document, iVia assigns title, creator, keyphrase, description, Library of Congress Subject Heading (LCSH), and INFOMINE Category metadata. Specific algorithms, some relying on document structure and others on statistical analysis of text, are employed for each metadata field. In the appendix we examine the specific procedures for assignment of metadata fields in iVia and the evaluation of those assignments. The iVia software has recently seen widespread adoption in the digital library community and is currently managed under the auspices of the DataFountains project[2]. Notably, in the UK, The University of Hull used iVia as the basis of their RepoMMan project[3] to analyze document content to provide suggestions to users inputting metadata.

Emma Tonkin and Henk L. Muller describe the paperBase system which helps users enter metadata for submissions to an archive of preprints [8]. The paperBase system extracts metadata from preprints and chooses keywords for them from a controlled vocabulary using a Bayesian classifier. The metadata populate a web form in a digital repository interface which the user can edit and complete.

Edvardsen et al. describe an approach to "Automatic Metadata Generation (AMG)" that leverages knowledge of the file formatting of common types of digital documents [1]. Their "document code analysis" approach can produce high quality automatic metadata assignments in large collections of heterogeneous documents by incorporating a variety of document visualization and formatting schemes.

Yang Song et al. address the problem of disambiguating author names in the CiteSeer digital library of academic papers [7]. For this purpose, Song et al. employ a two-staged approach to disambiguate author names. In the first stage, a Bayesian model is used to determine probable topics for documents. In the second stage, an unsupervised clustering algorithm, using probable topics as features, clusters documents to distinguish document authors and infers that the authors of a cluster of documents are identical.

In contrast, the author name disambiguation problem has also been addressed with a supervised learning algorithm by Treeratpituk and Giles [9]. In their approach, a random forest classifier employs a set of similarity profile features applied to documents. The accuracy of the algorithm is shown to compare favorably against a support vector machine (SVM) classifier, another supervised learning algorithm.

2.2 Information Extraction

In the natural language processing and information retrieval literature, *information extraction* denotes the task of inferring pre-characterized data from documents in a certain context. Typical data include dates, locations, or other values to fill slots in a representation of the document's content. The document code analysis techniques of Edvardsen et al. can be seen as a rule- or template-based approach to information extraction in which the rules or templates are encoded by a programmer based on observations and knowledge of the targeted document set. The methods of information extraction are typically corpus-based, meaning that the classifiers are empirically optimized by learning from preclassified natural language examples. One such corpus-based method is the maximum-entropy (maxent) algorithm [5], in which a probability distribution of classifications is learned from a training in such a way that the probability distribution involves no assumptions not warranted by the training set. Thus, the probability distribution maximizes the uncertainty or entropy with respect to the training set. The maxent algorithm is implemented in the OpenNLP natural language processing library[4].

2.3 Knowledge-Based Approaches

Disambiguation of references in documents requires a representation of possible referents and some of their relationships. The creation of an ontology and knowledge base (KB) able to encompass the diversity of document topics and the the noun-phrase varieties of reference is an enormous task that has seen numerous attempts. Here, we consider two ontologies/KBs applied to digital collections: the MESUR project [2] and the Indiana Philosophy Ontology (InPhO) as applied to the Stanford Encyclopedia of Philosophy (SEP)[5] [3].

The MESUR project, funded by Andrew W. Mellon, has a primary goal of studying the relationship of usage-based and citation-based metrics for assigning value to scholarly documents. As part of the MESUR project, Rodriguez et al. have developed an OWL ontology to describe bibliograpic and usage data for a wide range of scholarly artifacts [2]. The project has collected bibliographic and usage data from academic publishers, secondary publishers, and institutional digital library servers that link to documents. The researchers expect to eventually report a dataset of "tens of millions of of bibliographic records, hundreds of millions of citations, and billions of usage events."

Niepert et al. describe the Indiana Philosophy Ontology (InPhO) and its application to the Stanford Encyclopedia of Philosophy (SEP) [3] for purposes of automated management of metadata. The InPhO is an OWL ontology for the domain of philosophy including a *Thinker* sub-ontology of historical figures, an *Idea* sub-ontology of philosophical terms referring to entries' topics, taxonomic relations (like hypernymy and hyponymy) and non-taxonomic relations (like *nationality* and *has-influenced*). The ontology is updated semi-automatically based on statistical methods and authors' feedback when new entries are added to the encyclopedia.

Barry Smith, a major contributor to the Basic Formal Ontology or BFO[6], argues for an ontology based on mereology (the study of composition of wholes and parts) and topology (the study of geometric and spatial relations unaffected by continuous change in shape or size) rather than the set theory of standard model-theoretic semantics [6]. In some cases, this distinction avoids dilemmas whereby a description of objects at one scale or domain expends the vocabulary necessary for a description of another domain. BFO, an OWL-Full ontology, adheres to this philosophy, describ-

[1]http://infomine.ucr.edu/
[2]http://datafountains.ucr.edu/
[3]http://www.hull.ac.uk/esig/repomman
[4]http://maxent.sourceforge.net/
[5]http://plato.stanford.edu
[6]http://www.ifomis.org/bfo/

ing what Aristotle referred to as "substances" like wood or animals which endure self-identically through time (and instantiate the SPAN class in BFO), and "accidents" like greenness or warmth that occur to substances and unfold through time (and instantiate the SNAP class in BFO). As Niepert et al. have remarked, there are multiple ways of decomposing ideas, and few ways to decompose them deeply and comprehensively without inviting controversy [3, p 290]. By imposing uncontroversial restrictions at the highest level of ontology (as with BFO), we hope to avoid painting ourselves into an ontological corner when encoding knowledge about specific domains.

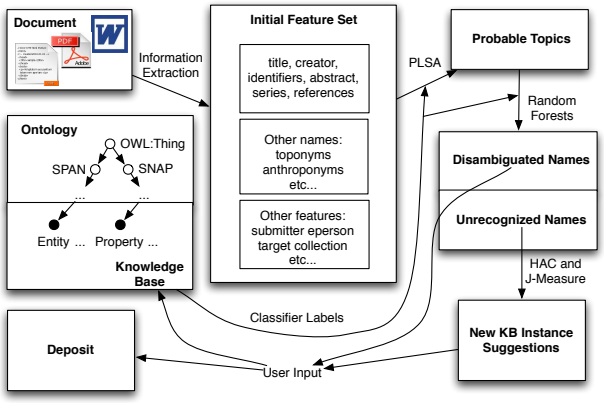
3. APPROACH

A basic difficulty in interpreting documents is that names are ambiguous, taking on different senses in different contexts. We propose a document processing pipeline that takes advantage of contextual information for progressively more refined classifications names in documents. The proposed work aims to automatically suggest Dublin Core (DC) metadata during as a curation task in a digital repository. The setting of the proposed work enables several helpful assumptions regarding the context of documents and user interactions. First, we limit our purview in terms of what content formats the interface can intelligently address: we will target OpenXML, HTML, and PDF documents as these are amenable to document code analysis and together comprise the bulk of textual content in many repositories. Second, we circumscribe the range of metadata to the following fields in the dc schema: creator, title, identifier.*, description.*, relation.references, and subject.*. These metadata fields are characterized in the table below.

dc schema field	In Doc Code?	Ambiguous?
creator	Yes	Yes
title	Yes	Yes
relation.references	Yes	Yes
subject	Sometimes	Yes
identifier.*	Yes	No
description	Sometimes	No

With the exception of subject and description textual values for all the fields considered here are generally amenable to a document code analysis for the purposes of information extraction. In the case of subject, unless subject keywords are embedded in document metadata or pre-defined sections of the document text, noun phrases must be identified in the document text. For the description field, the proposed system extracts abstracts, tables of contents, or other direct quotations of self-descriptive document content which are formatted predictably but only present in some types of textual documents such as books and theses; interpretation of the content of a description field is beyond the scope of this study.

Having been extracted from the source document, references corresponding to the specified dc fields will be disambiguated against a Knowledge Base (KB). Herein lies the novelty of the approach: in previous work, the metadata assignment task consists simply in information extraction; meanwhile, name disambiguation is applied in isolation to names of a pre-specified type (notably anthroponyms naming authors or toponyms naming places). In contrast, the proposed approach extracts a wide variety of metadata for subsequent disambiguation. The KB will be pre-populated with a set of instances gleaned from a suitably thorough review of the current contents of the TAMU institutional repository[7], a repository that contains on the order of 40000 documents by 20000 authors indexed by 40000 subject keywords, many of which are synonyms. These instances will represent documents, topics, and authors, and a instance may play more than one role in different contexts. Instances in the KB will serve as labels for a series of statistical classifiers to be applied to the extracted references. Below, we graphically depict the steps involved in the extraction and disambiguation of metadata field values.



First, a naive Bayes classifier, specifically a probabilistic latent semantic analysis (PLSA) classifier, will determine likely topics for documents based the initial feature set. This approach requires a pre-determined number of topics, which for our purposes will be instantiated in the KB. We hypothesize that the selected labels, corresponding to instances in a KB rather than ambiguous words as in [7] will afford a more predictive feature set for subsequent statistical classification than would the venerable term vector. This manner of representation compares with the approach of Niepert, et al. [3] in that the topics of documents are signified by document metadata field values and encoded as entities in a KB.

Next, a random forest classifier like that used by of Treeratpituk [9] for author name disambiguation will disambiguate references for dc.creator, dc.title, dc.relation.references, and dc.subject.

Supervised learning techniques like random forests can provide disambiguation when referents are represented in the KB but not when new, unrecognized reference/referent pairs are encountered. Therefore, when supervised learning techniques fail to offer a convincing classification of a reference, its referent may be assumed to be absent from the KB. Different classifiers afford different measurements of the surety of a classification. For example, random forests measure distances between cases, enabling detection of outliers. A different threshold for certainty about classifications will need to be calibrated for each type of reference. For example, titles can be inferred to refer to identical entities only if their containing document-context feature sets are extremely close or co-located in the feature space, whereas author names would face a less stringent criterion to infer identity of their referents.

An unsupervised classifier that has been shown success-

[7] http://repository.tamu.edu

ful at unsupervised disambiguation of author names is the hierarchical agglomerative clustering (HAC) technique described by Song et al. [7, p 37]. We propose to adapt this algorithm for the disambiguation of other sorts of references/names captured in the DC fields. The HAC algorithm will, without pre-defined categories, categorize ambiguous references and propose referents to instantiate in the KB. Finally, the document curator must confirm the veracity of these entities. Interfaces for evaluation and acceptance of suggested new instances have been described by Niepert et al. [3, p 296];

Suggested KB instances that are referents of `dc.subject` values present an interesting challenge to the knowledge-based approach. As described by Niepert et al. in the context of their *Idea* sub-ontology, the concepts that are the topics of our documents are decomposable into an acyclic hierarchical taxonomy of sub-concepts representing part-of relations, or a *mereonymy*.

Implementation of most of the required software components is straightforward, but two components stand out as particularly complex and at present ill-defined: the initial information extraction component and the KB component. These components involve extensive coding projects that will require a high fidelity correspondence with local repository practice in order to be successful. The information extraction component's development will be informed by a document code analysis of existing repository content and application of rule-learning algorithms where practical. Considerable time must be allocated for populating the KB, and dozens of items per day must be cataloged to complete the task in a matter of months.

4. CURRENT STATUS

Texas A&M University Libraries has enabled the development of a prototype *geoparser* to extract and disambiguate place-names in electronic theses and dissertations (ETDs), and this work is to be presented as a poster at JCDL 2012. This software involves some of the first steps in the proposed work, and has begun to encourage institutional buy-in for the approach.

The geoparser begins with a rudimentary document code analysis component that partitons the ETD into prefatory material (title pages, tables of contents, etc.), main body text, references, and suffixed material (vitas, appendices, etc.). Only the body text is subsequently processed, as location names elsewhere in the document are not typically relevant to the subject matter.

The geoparser uses OpenNLP software to partition the document into sentences, tokenize the sentences, and detect place names among the tokens. Identified place names are issued as queries via the GeoNames[8] webservice API to obtain a list of candidate locations with which to disambiguate the names. The candidates are subjected to a variety of heuristics, including context-based ones that examine the name occurrences in the text for other nearby place names and geographic feature types.

5. NEXT STEPS

The geoparser has not yet undergone an evaluation, but we have prepared manual annotations of about 100 ETDs for this purpose.

The heuristics employed by the geoparser for disambiguation of place names rely on specific properties of geographic locations, such as their proximities and containment relations, and take advantage of the easy availability of location information in public gazetteers. As a next step, we plan to implement a supervised random-forest classifier as proposed above. The manually annotated ETDs for the evaluation can be used as initial training data. This classifier will be extensible to disambiguation tasks for other types of names (such as author names as demonstrated by Treeratpituk et al.).

The special nature of geographic locations provides a great starting point for exploring the potential for domain knowledge to facilitate the metadata assignment task. The GeoNames gazetteer includes extensive information about the relationships between and characteristics of named locations. One of the currently implemented place name disambiguation heuristics examines names occurrences to determine if they occur in the context of a geographic feature-type name, such as "city" or "lake", and favors candidate locations of the mentioned type. The GeoNames list of feature types contains about 600 such types[9]. The current implementation of this heuristic employs a simple mapping from strings to GeoNames feature types; however, additional information could enhance the heuristic. There is considerable conceptual overlap between geographic feature types - to use an example from the GeoNames listing, the feature type name "road" (having unique code `RD`) may be used by some writers to describe a feature listed by GeoNames as a "street" (having unique code `ST`), and vice versa. We plan to encode a similarity matrix of geographic feature-type names to refine this heuristic so that locations of types similar to those mentioned can also be favored as disambiguation candidates.

6. CONTRIBUTIONS

Successful completion of the proposed work entails several contributions. To begin with, the proposed system exhibits an unprecedented integration of institutional repository metadata, document encodings, and domain knowledge. The proposed system extends and combines numerous technologies established in the field of digital libraries: document code analysis, intelligent metadata suggestion, and topic-based name disambiguation. In particular, name disambiguation techniques applied in previous work to author names are extended to apply to other names. The system integrates these technologies by using explicit knowledge about document content and context, which confers two advantages. First, the metadata suggestions direct curators toward canonical labels associated with instances in the knowledge base, thereby curbing the proliferation of synonymous subject keywords and author names. Second, explicit knowledge of topics provides more solid semantic grounding of the topics inferred in the first stage of topic-based name disambiguation, which we expect to provide better feature sets for subsequent classifications. Applied in combination with background domain knowledge, these technologies may push the envelope of precision and recall of metadata values for repository content. If an evaluation confirms the experimental hypothesis, then the system will demonstrate how appropriate interfaces and contexts can empower users to interact with formally encoded knowledge without get-

[8] www.geonames.org

[9] http://www.geonames.org/export/codes.html

ting mired in that formality. Finally, the proposed work promises to enable an institutional repository to grow self-knowledge organically with its collections. In addition to improving metadata accuracy and thereby enhancing existing interfaces to the repository, this self-knowledge could provide the framework for new information discovery interfaces to the repository for humans and machines, a project we leave for future work.

7. REFERENCES

[1] L. F. H. Edvardsen, I. T. Sølvberg, T. Aalberg, and H. Trætteberg. Automatically generating high quality metadata by analyzing the document code of common file types. In *JCDL 2009*, pages 29–38, 2009.

[2] A. R. Marko, J. Bollen, and H. V. de Sompel. A practical ontology for the large-scale modeling of scholarly artifacts and their usage. In *JCDL 2007*, pages 278–287, 2007.

[3] M. Niepert, C. Buckner, and C. Allen. A dynamic ontology for a dynamic reference work. In *JCDL 2007*, pages 288–297, 2007.

[4] G. W. Paynter. Developing practical automatic metadata assignment and evaluation tools for internet resources. In *JCDL 2005*, pages 291 – 300, Denver, CO, USA, June 2005.

[5] A. Ratnaparkhi. A simple introduction to maximum entropy models for natural language processing. Technical Report 97–08, Institute for Research in Cognitive Science, University of Pennsylvania 3401 Walnut Street, Suite 400A Philadelphia, PA 19104-6228, May 1997.

[6] B. Smith. *Formal Ontology in Information Systems*, chapter The Basic Tools of Formal Ontology, pages 19–28. IOS Press, 1998.

[7] Y. Song, J. Huang, I. G. Councill, J. Li, and C. L. Giles. Efficient topic-based unsupervised name disambiguation. In *JCDL 2007*, pages 342–351, 2007.

[8] E. Tonkin and H. L. Muller. Semi automated metadata extraction for preprints archives. In *JCDL 2008*, pages 157–166, 2008.

[9] P. Treeratpituk and C. L. Giles. Disambiguating authors in academic publications using random forests. In *JCDL 2009*, 2009.