Automating Metadata Extraction: Genre Classification

Yunhyong Kim[1,2] and **Seamus Ross**[1,2,3]

[1] Digital Curation Centre (DCC)
[2] Humanities Advanced Technology Information Institute (HATII),
University of Glasgow, Glasgow, UK
[3] Oxford Internet Institute (2005/6), University of Oxford
{y.kim, s.ross}@hatii.arts.gla.ac.uk

Abstract

A problem that frequently arises in the management and integration of scientific data is the lack of context and semantics that would link data encoded in disparate ways. To bridge the discrepancy, it often helps to mine scientific texts to aid the understanding of the database. Mining relevant text can be significantly aided by the availability of descriptive and semantic metadata. The Digital Curation Centre (DCC) has undertaken research to automate the extraction of metadata from documents in PDF([22]). Documents may include scientific journal papers, lab notes or even emails. We suggest genre classification as a first step toward automating metadata extraction. The classification method will be built on looking at the documents from five directions; as an object of specific visual format, a layout of strings with characteristic grammar, an object with stylo-metric signatures, an object with meaning and purpose, and an object linked to previously classified objects and external sources. Some results of experiments in relation to the first two directions are described here; they are meant to be indicative of the promise underlying this multi-faceted approach.

1. Background and Objective

Text mining has received attention in recent years as a means of providing semantics to scientific data. For instance, Bio-Mita ([4]) employs text mining to find associations between terms in biological data. Descriptive, administrative, and technical metadata play a key role in the management of digital collections ([25], [15]). As the DELOS/NSF ([8], [9], [10]) and PREMIS working groups ([23]) noted, when done manually, metadata are expensive to create and maintain. The manual collection of metadata can not keep pace with the number of digital objects that need to be documented. Automatic extraction of metadata would be an invaluable step in the automation of appraisal, selection, and ingest of digital material. ERPANET's Packaged Object Ingest Project ([12]) illustrated that only a limited number of automatic extraction tools for metadata are available and these are mostly geared to extracting technical metadata (e.g. DROID ([20]) and Metadata Extraction Tool ([21])). Although there are efforts to provide tools (e.g. MetadataExtractor from University of Waterloo, Dublin Core Initiative ([11], [7]), Automatic Metadata Generation at the Catholic University of Leuven([1])) for extracting limited descriptive metadata (e.g. title, author and keywords) these often rely on structured documents (e.g. HTML and XML) and their precision and usefulness is constrained. Also, we lack an automated extraction tool for high-level semantic metadata (such as content summary) appropriate for use by digital repositories; most work involving the automatic extraction of genres, subject classification and content summary lie scattered around in information extraction and language processing communities(e.g. [17], [24], [26], [27]). Our research is motivated by an effort to address this problem by integrating the methods available in the area of language processing to create a prototype tool for automatically extracting metadata at different semantic levels.

The initial prototype is intended to extract Genre, Author, Title, Date, Identifier, Pagination, Size, Language, Keywords, Composition (e.g. existence and proportion of images, text and links) and Content Summary. Here we discuss genre classification of documents represented in PDF ([22]) as a first step. The ambiguous nature of the term genre is noted by core studies on genre such as Biber

([3]) and Kessler et al. ([17]). We follow Kessler who refers to genre as “any widely recognised class of texts defined by some common communicative purpose or other functional traits, provided the function is connected to some formal cues or commonalities and that the class is extensible”. For instance, a scientific research article is a theoretical argument or communication of results relating to a scientific subject usually published in a journal and often starting with a title, followed by author, abstract, and body of text, finally ending with a bibliography. One important aspect of genre classification is that it is distinct from subject classification which can coincide over many genres (e.g., a mathematical paper on number theory versus a news article on the proof of Fermat's Last Theorem).

By beginning with genre classification it is possible to limit the scope of document forms from which to extract other metadata. By reducing the metadata search space metadata such as author, keywords, identification numbers or references can be predicted to appear in a specific style and region within a single genre. Independent work exists on extraction of keywords, subject and summarisation within specific genre which can be combined with genre classification for metadata extraction across domains (e.g. [2], [13], [14], [26]). Resources available for extracting further metadata varies by genre; for instance, research articles unlike newspaper articles come with a list of citations closely related to the original article leading to better subject classification. Genre classification will facilitate automating the identification, selection, and acquisition of materials in keeping with local collecting policies.

We have opted to consider 60 genres and have discussed this elsewhere [initially in 18]. This list does not represent a complete spectrum of possible genres or necessarily an optimal genre classification; it provides a base line from which to assess what is possible. The classification is extensible. We have also focused our attention on information from genres represented in PDF files. Limiting this research to one file type allowed us to bound the problem space further. We selected PDF because it is widely used, is portable, benefits from a variety of processing tools, is flexible enough to support the inclusion of different types of objects (e.g. images, links), and is used to present a diversity of genre.

In the experiments which follow we worked with a data set of 4000 documents collected via the internet using a randomised PDF-grabber. Currently 570 are labelled with one of the 60 genres and manual labelling of the remainder is in progress. A significant amount of disagreement is apparent in labelling genre even between human labellers; we intend to cross check the labelled data later with the help of other labellers. However, the assumption is that an experiment on data labelled by a single labeller, as long as the criteria for the labelling process are consistent, is sufficient to show that a machine can be trained to label data according a preferred schema, thereby warranting further refinement complying with well-designed classification standards.

2. Classifiers

For the experiments described here we implemented of two classifiers. First, an *Image classifier***,** which depends on features extracted from the PDF document when handled as an image. It converts the first page of the PDF file to a low resolution image expressed as pixel values. This is then sectioned into ten regions for an assessment of the number of non-white pixels. Each region is rated as level 0, 1, 2, 3 with the larger number indicating a higher density of non-white space. The result is statistically modelled using the Maximum Entropy principle with MaxEnt developed by Zhang ([28]). Second we implemented a *Language model classifier***,** which depends on an N-gram model on the level of words. N-gram models look at the possibility of word w(N) coming after a string of words W(1), W(2), ..., w(N-1). A popular model is the case when N=3. This has been modelled by the BOW toolkit ([19]) using the default Naïve Bayes model without a stoplist.

3. Experiment Design

An assumption in the two experiments described here is that PDF documents are one of four categories: Business Report, Minutes, Product/Application Description, Scientific Research Article. This is, of course, a false assumption and limiting the scope in this way changes the meaning of the resulting statistics considerably. However, our contention is that high level performance on a limited data set combined with a suitable means of accurately narrowing down the candidates to be labelled would achieve the end objective.

For the first experiment we took the 70 documents in our labelled data belonging to the above four genres, randomly selected a third of them as training data (27 documents) and the remaining documents as test data (43), trained both classifiers on the selected training data, and examined result. In the second experiment we used the same training and test data. We allocated the genres to two groups each containing two genres: Group I included business reports and minutes while Group II encompassed scientific research articles and product descriptions. Then we trained the image classifier to differentiate between the two groups and used this to label the test data as documents of Group I or II. Concurrently we trained two language model classifiers: Classifier I which distinguishes business reports from minutes and Classifier II which labels documents as scientific research articles or product descriptions. Subsequently we took the test documents labelled Group I and labelled them with Classifier I and those labelled Group II and labelled them with Classifier II. We examined the result.

4. Results and Conclusions

The precision and recall for the two experiments are presented in Tables 1 and 2. Although the performance of the language model classifier shown in Table 1 is already high, this, to a great extent, reflects on the four categories chosen. In fact, when the classifier was extended to include 40 genres, it performed only at an accuracy of about 10%. When a different set was employed which included Periodicals, Thesis, Minutes and Instruction/ Guideline, the language model performs at an accuracy of 60.34% and the image classifier on Group I (Periodicals) and Group II (Thesis, Minutes, Instruction/Guideline) performs at 91.37%. Note also that, since Thesis, Minutes and Instruction/Guidelines can be intuitively predicted to have distinct linguistic characteristics, the language classifier's performance on each group is also predicted to perform at a high level of accuracy (results pending). It is clear from the two examples that such a high performance can not be expected for any collection of genres. Judging from the result of the classifiers, the current situation seems to be a case of four categories which are similar under the image classifier but which differ linguistically. A secondary reason for involving images in the classification and information extraction process arises because some PDF files are textually inaccessible due to password protection, and even when text is extracted, text processing tools are quite strict in their requirements for input data. In this respect, images are much more stable and manageable. Combining a soft decision image classifier with the language model both increases the overall accuracy and results in a notable increase in recall for most of the four genres (see Table 2).

Table 1. Result of Language Model Only

Overall accuracy: 77%

Genres	Prec (%)	Rec (%)
Business Report	83	50
Sci. Res. Article	88	80
Minutes	64	100
Product Desc	90	90

Table 2. Result of Image & Language Model

Overall accuracy: 87.5%

Genres	Prec (%)	Rec (%)
Business Report	83	50
Sci. Res. Article	75	90
Minutes	71	100
Product Desc	90	100

The results of the experiments indicate that the combined approach is a promising candidate for further experimentation with a wider range of genres. The experiments show that, although there is a lot of confusion visually and linguistically over all 60 genres, subgroups of the genres exhibit statistically well-behaved characteristics. This encourages the search for groups which are similar or different visually or linguistically. Further experiments are planned to enable us to refine this hypothesis.

Further improvement can be envisioned, including integrating more classifiers. An *extended image classifier* could examine pages other than the just first page (as done here), or examine the image of several pages in combination: different pages may have different levels of impact on genre classification, while processing several pages in combination may provide more information.. A *language model classifier on the level of POS and phrases would use* a N-gram language model built on the Part-of-speech tags (for instance, tags denoting words as a verb, noun or preposition) of the underlying text of the document and also on partial chunks resulting from detection of phrases (e.g. noun, verb or prepositional phrases). A *stylometric classifier* taking its cue from positioning of text and image blocks, font styles, font size, length of the document, average sentence lengths and word lengths. A *semantic classifier* would combine extraction of keywords, subjective or objective noun phrases (e.g. using [24]). Finally a c*lassifier based on available external information* such features as name of the journal, subject or address of the webpage and anchor texts can be gathered for statistical analysis or rule-based classification

5. Putting it into Context

The genre extractor provides the basis for constructing an efficient tool. Extension of the tool to extract author, title, date, identifier, keywords, language, summarizations, and other compositional properties can be targeted based upon genre and will, thereby, improve the precision of these other extractors. When the genre classifier is refined for PDF documents, extending it to cover other document format types (e.g. Open Office, Word, LATEX) will be straightforward. Our aim is eventually to pass the prototype to colleagues in the Digital Curation Centre ([6]) who will integrate it with other extraction tools and eventually an automated ingest model.

Acknowledgements

The Digital Curation Centre (DCC) is supported by a grant from the Joint Information Systems Committee (JISC) and the e-Science Core Programme of the Engineering and Physical Sciences Research Council (EPSRC). The EPSRC supports (GR/T07374/01) the DCC's research programme.

References

[1] Automatic Metadata Generation, http://www.cs. kuleuven.ac.be/~hmdb/amg/documentation.php

[2] Bekkerman R, McCallum A, and Huang G, 'Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora', *CIIR Technical Report*, IR-418 (2004).

[3] Biber D, *Dimensions of Register Variation: a Cross-Linguistic Comparison*, Cambridge (1995).

[4]Bio-Mita, http://personalpages.manchester.ac.uk/ staff/G.Nenadic/BioMITA.htm

[5]Boese E S, 'Stereotyping the web: genre classification of web documents', Master's thesis, Colorado State University (2005).

[6] Digital Curation Centre, http://www.dcc.ac.uk

[7] DC-dot, Dublin Core metadata editor, http:// www.ukoln.ac.uk/metadata/dcdot/

[8] DELOS, http://www.delos.info/

[9] NSF, http:// www.dli2.nsf.gov/intl.html

[10] DELOS/NSF Working Groups, 'Reference Models for Digital Libraries: Actors and Roles' (2003), http://www.dli2.nsf.gov/internationalprojects/working_group_reports/actors_final_report.html

[11] Dublin Core Initiative, http://dublincore.org/tools/#automaticextraction

[12] ERPANET, Packaged Object Ingest Project, http://www.erpanet.org/events/2003/rome/presentations/ ross rusbridge pres.pdf

[13] Giufirida G, Shek E, and Yang J, 'Knowledgebased Metadata Extraction from PostScript File', *Proc. 5th ACM Intl. conf. Digital Libraries* (2000) 77-84.

[14] Han H, Giles L, Manavoglu E, Zha H, Zhang Z and Fox E A, 'Automatic Document Metadata Extraction using Support Vector Machines', *Proc. 3rd ACM/IEEE-CS conf. Digital Libraries* (2000) 37-48.

[15] NSF-DELOS Working Group on Digital Archiving: 'Invest to Save', Report DELOS and NSF Workgroup on Digital Preservation and Archiving (2003) http://eprints.erpanet.org/94/01/NSF_Delos_WG_Pres_final.pdf

[16] Karlgren J and Cutting D, 'Recognizing Text Genres with Simple Metric using Discriminant Analysis', *Proc. 15th conf. Comp. Ling*., Vol 2 (1994) 1071-1075

[17] Kessler B, Nunberg G, Schuetze H, 'Automatic Detection of Text Genre', *Proc. 35th Ann. Meeting ACL* (1997) 32-38.

[18] Kim Y and Ross S, Genre Classification in Automated Ingest and Appraisal Metadata, J. Gonzalo et al. (Eds.): *ECDL 2006*, LNCS 4172, 63–74, 2006.

[19] McCallum A, Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering, http://www.cs.cmu.edu/ mccallum/ bow/ (1998)

[20] National Archives, DROID (Digital Object Identification), http://www.nationalarchives.gov.uk/ aboutapps/pronom/droid.htm

[21] National Library of New Zealand, Metadata Extraction Tool, http://www.natlib.govt.nz/en/ whatsnew/4initiatives.html#extraction

[22] Adobe Acrobat PDF specification, http://partners.adobe.com/public/developer/pdf/index_reference.html

[23] PREMIS Working Group, http://www.oclc.org/research/projects/pmwg/

[24] Riloff E, Wiebe J, and Wilson T, `Learning Subjective Nouns using Extraction Pattern Bootstrapping', *Proc. 7th CoNLL*, (2003) 25-32

[25] Ross S and Hedstrom M, 'Preservation Research and Sustainable Digital Libraries', *Int Journal of Digital Libraries* (Springer) (2005) DOI: 10.1007/s00799-004-0099-3.

[26] Sebastiani F, 'Machine Learning in Automated Text Categorization', *ACM Computing Surveys*, Vol. 34 (2002) 1-47.

[27] Witte R, Krestel R, and Bergler S, 'ERSS 2005:Coreference-based Summarization Reloaded', *DUC 2005 Document Understanding Workshop*.

[28] Zhang L, Maximum Entropy Toolkit for Python and C++, LGPL license, http:// homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html