briefing paper

digital preservation europe

Automating semantic metadata extraction

Since May 2005 the Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, has undertaken an initiative to automate the extraction of semantic metadata from digital objects. This is motivated by an earlier study into automating or semi-automating the ingest and preservation processes (e.g. [2]). The construction of sufficient metadata, describing the content, bibliographic information, provenance, and technical and administrative requirements of an object, is a crucial element in the management and sustenance of digital repositories, libraries and archives ([8], [9]). The manual collection of such metadata is a labour intensive process, and, the exponential rate at which digital objects are produced will eventually make it impossible to rely on manual methods. The objective of this initiative is to take steps to understand the extent to which metadata creation can be automated, before the urgency arises.

Scope

Document Format (PDF): By selecting a specific format we can reduce the problem space to a manageable size. More specifically, a tool which can handle PDF, a widely adopted format throughout digital repositories, libraries and archives, as well as commercial sectors and individuals, is expected to be of immediate use to a broad range of communities.

Initial efforts are focused on text documents: Natural Language Processing (NLP) methods have proven effective in retrieval, extraction and classification of documents and terms within documents. This presents NLP and other machine learning techniques on text documents as an obvious candidate for applying the first stages of metadata extraction. The development of extraction tools for text is also expected to have consequences in other objects, as many extraction processes for other media (e.g. image or audio-visual material) depend on mining associated text.

Progress

At HATII, Automated Genre Classification has been identified as a fundamental step in realising Automated Semantic Metadata Extraction. Genre is a structural and functional classification of a document which reflects one or more of the following:

- the intention of the creator (e.g. to inform, to argue, to instruct),
- the interpretation of the user community (e.g. as a collection of facts, as an expression of opinion, as a piece of research),
- the prescription of a process (e.g. article for journal publication, job description for recruitment, minutes of a meeting), and,
- the type of data structure (e.g. table, graph, chart, list).

Why genre classification?

1. Identifying the genre will limit the structural scope of document forms from which to extract other metadata:
    - The search space for further metadata will be reduced; within a single genre, metadata such as author, keywords, identification numbers or references can be expected to appear in a similar style and region.

2. Identifying the genre will create an over-arching tool which will bind genre-specific work:
    - Independent work ([1], [3], [4], [10], [11]) exists for extraction of metadata within a specific genre which can be combined with a general genre classifier for metadata extraction over many domains.
    - Resources available for extracting further metadata are different for each genre; for instance, research articles, unlike newspaper articles, come with a list of reference articles closely related to the original document leading to better subject classification.

3. Scoping new genres not apparent in the context of conventional libraries is necessary to keep management practices of digital materials up to date.

4. Different institutional collecting policies might focus on digital materials in disparate genres. Genre classification will support automating the identification, selection, and acquisition of materials in keeping with local collecting guidelines.

The classification tool

The genre classification tool is intended to evaluate documents statistically based on its visual layout (e.g. the amount of white space in the document), stylistic elements (e.g. the frequency and use of definite articles), language model (significant terms which characterise the document), semantic patterns (e.g. the use of subjective noun phrases), and available external resources (e.g. the source URL), to determine its genre class. The results of this research effort have been published in several papers ([5], [6], [7]).

Metadata extraction workflow

The study in this project will be tied into a general architecture of automated metadata extraction and ingest process using the following workflow:

- receive digital object,
- determine genre or structural class of object,
- scope best metadata extraction tool for the identified genre, or best option based on structure,
- if no such tool exists, request the creation of tool,
- once tool has been selected or created, extract necessary metadata,
- ingest object and metadata into repository or archive.

Conclusions

The automation of ingest, preservation, and selection processes within a digital repository is no longer a convenience but a necessity. The storage, sharing, and cross-institutional use of information has now become an active reality, and the manual control of such processes is slow, ineffective and inefficient. The challenges of adapting to the information highway have to be met with innovative automated processes of extraction, authentication, and evaluation. Automated semantic metadata extraction is still at its infancy. It is essential that the effort is continued to push forward and refine extraction tools, and integrate these with other processes across the library, archival and related communities.

References

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

[2] ERPANET: Packaged Object Ingest Project. http://www.erpanet.org/events/2003/rome/presentations/ross_rusbridge_pres.pdf

[3] Han, H., Giles, L., Manavoglu, E., Zha, H., Zhang, Z., Fox, E. A. (2000) Automatic Document Metadata Extraction using Support Vector Machines. Proceedings 3rd ACM/IEEECS Conference on Digital libraries, 37-48.

[4] Giurida, G., Shek, E. Yang, J. (2000) Knowledge-based Metadata Extraction from PostScript File. Proceedings 5th ACM International Conference on Digital Libraries, 77-84.

[5] Kim, Y. and Ross, S. (2007) Detecting family resemblance: Automated genre classification. CODATA Data Science Journal, Volume 6, S172-S183, ISSN:1683-1470.

[6] Kim, Y. and Ross, S. (2006) Genre classification in automated ingest and appraisal metadata. In J. Gonzalo, editor, Proceedings European Conference on advanced technology and research in Digital Libraries (ECDL), volume 4172 of Lecture Notes in Computer Science, pages 63–74. Springer.

[7] Kim Y. and Ross, S. (2006) "The Naming of Cats": Automated genre classification. To appear International Journal of Digital Curation, preprint available at http://eprints.erpanet.org/123

[8] PREMIS (PREservation Metadata: Implementation Strategy) Working Group: http://www.oclc.org/research/projects/pmwg/

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

[10] Sebastiani F.: 'Machine Learning in Automated Text Categorization', ACM Computing Surveys, Volume 34 1-47.

[11] Thoma,G. (2001) Automating the production of bibliographic records. R&D report of the Communications Engineering Branch, Lister Hill National Center for Biomedical Communications, National Library of Medicine.

Yunhyong Kim, DCC-EPSRC Research Fellow
HATII University of Glasgow