briefing paper

digital preservation europe

Extracção automatizada de metadados semânticos

Desde Maio de 2005 que o Humanities Advanced Technology and Information Institute (HATII), da Universidade de Glasgow, desenvolve um projecto de extracção automática de metadados semânticos de objectos digitais. O projecto baseia-se num estudo precedente sobre a automatização ou a semi-automatização dos procedimentos de entrada e de preservação (e.g. [2]). A constituição dos metadados necessários para descrever o conteúdo, as informações bibliográficas, a proveniência, as necessidades técnicas e administrativas de um objecto, é um elemento crucial de gestão e de subsistência dos repositórios digitais, bibliotecas e arquivos ([8], [9]). A recolha manual de tais metadados é um processo intensivo, e perante a rapidez exponencial com que tais objectos digitais são produzidos, é impossível confiar apenas nos métodos manuais. O objectivo deste projecto é avançar por etapas para perceber até que ponto a criação de tais metadados pode ser feita de forma automatizada, antes que a urgência apareça.

Âmbito

Formato de documento (PDF): a partir de um formato específico é possível limitar a amplitude do problema a um tamanho manuseável. Mais especificamente, uma ferramenta que apreenda o PDF - um formato largamente adoptado nos repositórios digitais, bibliotecas e arquivos, assim como pelo sector comercial e por particulares - é presumivelmente de utilização imediata por parte de um largo espectro de comunidades.

Os esforços iniciais concentram-se nos documentos textuais: os métodos de processamento em linguagem natural (NLP) revelaram-se eficazes no que concerne à pesquisa, à extracção e à classificação de documentos e seus termos. Este facto torna a NLP, e outras técnicas automatizadas de leitura de documentos textuais, um candidato evidente para aplicar os primeiros passos de extracção de metadados. Espera-se ainda que o desenvolvimento de ferramentas de extracção de texto tenha consequências em outros objectos, na medida que muitos procedimentos de extracção para outros media (e.g. imagem ou material audiovisual) dependem da exploração dos textos que lhe estão associados.

Desenvolvimento

No HATII, a classificação automática por tipologia tem sido identificada como um passo fundamental para realizar a extracção automática de metadados. A tipologia é uma classificação estrutural e funcional de um documento que é o reflexo de um dos seguintes aspectos:

- A intenção de um criador (informar, argumentar, instruir),
- A interpretação dos utilizadores (um conjunto de factos, a expressão de uma opinião, um artigo de investigação),
- A descrição de um procedimento (e.g. artigo para publicação numa revista, curriculum vitae, acta de uma reunião),
- O tipo de estrutura dos dados (tabela, gráfico, mapa, lista)

Porquê a classificação por tipologia?

1. Identificar a tipologia limitará o âmbito estrutural dos tipos de documentos a partir dos quais extrair outros metadados:
   - O espaço de pesquisa para os restantes metadados será reduzido; para uma mesma tipologia, metadados como autor, palavras-chave, números de identificação ou referências devem surgir em estilos e localizações idênticas;
2. Identificar a tipologia criará uma ferramenta dedicada que homogeneizará o trabalho específico dessa tipologia:
   - Existem trabalhos independentes ([1], [3], [4], [10], [11]) para a extracção de metadados de uma tipologia específica, que podem ser incorporados num classificador geral de tipologias para a extracção de metadados em vários domínios;
   - Os recursos disponíveis para a extracção de outros metadados são diferentes para cada tipologia específica; por exemplo, os artigos de investigação, ao contrário dos artigos de jornais, comportam uma lista de referências de artigos relacionados com o documento original e permitem, dessa forma, uma melhor classificação.

3. Introduzir novas tipologias que não fazem parte do contexto das bibliotecas tradicionais promove a manutenção de práticas de gestão de materiais digitais actualizadas;
4. Diferentes políticas institucionais de recolha podem ser aplicadas a diferentes tipos de materiais digitais. A classificação por tipologia suportará a automatização da identificação, da selecção e da aquisição dos materiais de acordo com as directrizes locais de recolha.

A ferramenta de classificação?
A ferramenta de classificação das tipologias destina-se a avaliar os documentos estatisticamente baseando-se no seu aspecto visual (e.g. quantidade de espaço em branco no documento), nos elementos de estilo (a frequência e a utilização de determinados artigos), nos modelos de linguagem (os termos significativos que caracterizam o documento), nas formas semânticas (a utilização de frases com nomes subjectivos) e nos recursos externos (e.g. o endereço URL), com o objectivo de determinar a sua tipologia. Os resultados destes trabalhos de pesquisas foram publicados em vários artigos ([5], [6], [7]).

O processo de extracção dos metadados

Lo O estudo neste projecto integrará a arquitectura geral de extracção e ingestão automática de metadados de acordo com o seguinte procedimento:
1. Receber o objecto digital,
2. Determinar a tipologia ou a classe estrutural do objecto,
3. Avaliar a melhor ferramenta de extracção de metadados para a tipologia identificada, ou a melhor opção baseada na sua estrutura,
4. Uma vez escolhida ou criada a ferramenta, extrair os metadados necessários,
5. Adicionar o objecto e respectivos metadados no repositório ou arquivo.

Conclusões

Os processos automatizados de ingestão, preservação e selecção num repositório digital não são tanto uma facilidade mas antes uma necessidade. O armazenamento, a disponibilização e a utilização inter-institucional da informação tornaram-se uma realidade activa; o controle manual de tais processos é moroso, ineficaz e insuficiente. Os desafios de adaptação à auto-estrada da informação devem ser atingidos através de processos automatizados e inovadores de extracção, autenticação e avaliação. É essencial que os esforços continuem no sentido de desenvolver e aperfeiçoar ferramentas de extracção, tendo em vista a sua integração com outros procedimentos em bibliotecas, arquivos e outras comunidades relacionadas.

Further information and resources:

[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: ImplementationStrategy) 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.
Formatted: Bullets and Numbering

Author:**Di Yunhyong Kim**, *DCC-EPSRC Research Fellow, HATII University of Glasgow*
Translation: **Cristina Faria**, fariacristi@gmail.com