Personalised
access to
cultural heritage
spaces

Paths

Recommendations for the automatic enrichment of DL content using open source software

Authors:
Eneko Agirre and Arantxa Otegi

www.paths-project.eu

Recommendations for the automatic enrichment of DL content using open source software

1 Introduction

PATHS uses text processing software to enable the following functionality in the prototypes (see deliverables D2.1 and D2.2 http://paths-project.eu/eng/Resources):

- intra-collection links
- background links
- ontology extension

The aim of PATHS is to investigate the use of those functionalities to better serve exploration by users. Most of the software is in-house and relatively complex. The release of the software as open source was not in the DOW, and the exploitations plan goes in the software-as-a-service direction, where the processing is done in servers prepared by the partners. For instance, in a closely related Best Practice Network (LoCloud http://www.locloud.eu/), servers for automatic content enrichment of digital library items are to be produced.

Alternatively, projects like OpeNER (http://www.opener-project.org/) are devoted to the release of open source tools which are similar to the ones used for text processing in PATHS. The release of all PATHS software as open source out-of-the-box packages would be a major undertaking, on a par to those European projects.

The goal of this document is to present an overall set of recommendations for the automatic enrichment of Digital Libraries content using open source software. We think this would be useful for third-parties who would like to offer similar services. Note that this is not a step-by-step guide for reimplementation, but an overall view of the required software and programming effort involved.

This document is structured according to each of the enrichment tasks described in PATHS deliverables D2.1 and D2.2.

PATHS is funded by the European Commission FP7 programme under Digital Libraries and Digital Preservation

2 Producing intra-collections links

PATHS produced both generic similarity links and more specific typed-similarity links.

2.1 Similarity links

The target items would need to be parsed by the following Natural Language Processing (NLP) tools: Pos tagging and lemmatization. Several open source products exist, including the two used in PATHS:

- **CoreNLP** (http://nlp.stanford.edu/software/corenlp.shtml)
  and
- **Freeling** (http://nlp.lsi.upc.edu/freeling/).

Regarding multilinguality PATHS worked on Spanish and English texts. Freeling covers both, and Stanford works out-of-the-box for English. Recent projects like OpenNER also offer a suite of open source NLP tools including, in addition to English and Spanish, four other European languages.

In addition, we used in-house scripts to process the internal representation of the items, extract the textual pieces, and produce the enriched representations.

The actual production of the similarity links requires additional software, which in this case has been developed in-house. No out-of-the box alternatives exist. The interested parties would need to replicate the software described in (Aletras et al. 2012).

As an alternative, the PATHS server demonstrating similarity links described in (Agirre et al. 2013b) uses the functionality provided by a search engine (Solr http://lucene.apache.org/solr/) to provide similar items. Note that this alternative does not require additional NLP tools, as it uses their own stop words and stemming algorithm.

2.2 Typed-similarity links

The target items would need to be parsed by the following NLP tools: Pos tagging and lemmatization (see previous section).

In addition, we used in-house scripts to process the internal representation of the items, extract the textual pieces, and produce the enriched representations.

The actual production of the typed-similarity links requires additional software, including open source machine learning software (Weka http://www.cs.waikato.ac.nz/ml/weka/) and the in-house scripts to extract features from items, train the machine learning models on the publicly available typed-similarity datasets produced by PATHS (http://ixa2.si.ehu.es/sts/), and use the machine learning models on the target items. No out-of-the-box alternatives exist. The interested parties would need to replicate the software as described in (Agirre et al. 2013a).

3 Producing background links

In order to produce background links we used Wikipedia Miner, an open source software available at http://wikipedia-miner.cms.waikato.ac.nz. We used wikipedia miner out-of-the-box. In addition, we used in-house scripts to process the internal representation of the items, extract the textual pieces, and produce the enriched representations.

4 Ontology extension

The target items would need to be parsed by the following NLP tools: Pos tagging and lemmatisation (see previous section).

In addition, we used in-house scripts to process the internal representation of the items, extract the textual pieces, and produce the enriched representations.

The actual production of the vocabulary requires additional software. We first extract the background links (see previous section), and then find the most relevant Wikipedia articles per item. This is done globally, analysing the statistics of the whole collection. Those articles are used to categorize the items according to a Wikipedia-based category system, which is trimmed-down to only cover the categories which are relevant to the collection at hand. No out-of-the-box alternatives exist. The interested parties would need to replicate the software as described in (Fernando et al. 2012).

References

Agirre E., Aletras N., Gonzalez-Agirre A., Rigau G., Stevenson M. (2013a). *UBC UOS-TYPED: Regression for typed-similarity.* The Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)

Agirre E., Barrena A., Fernandez K., Miranda E., Otegi A., Soroa A. (2013b). *PATHSenrich: a Web Service Prototype for Automatic Cultural Heritage Item Enrichment* in Research and Advanced Technology for Digital Libraries, International Conference on Theory and Practice of Digital Libraries, TPDL 2013, Valletta, Malta, September 22-26, 2013. Lecture Notes in Computer Science, vol. 8092, pp 462-465.

Aletras N., Stevenson M., Clough P. (2012). *Computing Similarity between Items* in a Digital Library of Cultural Heritage. In ACM JOCCH.

Fernando S., Hall M., Agirre E., Soroa A., Clough P., Stevenson M. (2012). *Comparing taxonomies for organising collections of documents*. In Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012), pages 879-894, Mumbai, India, December 2012.