Personalised
access to
cultural heritage
spaces

Paths

Semantic Enrichment of Cultural Heritage Content in PATHS

Authors:
Mark Stevenson and Arantxa Otegi with Eneko Agirre, Nikos Aletras, Paul Clough, Samuel Fernando and Aitor Soroa.

www.paths-project.eu

Semantic Enrichment of Cultural Heritage Content in PATHS

Mark Stevenson and Arantxa Otegi with Eneko Agirre, Nikos Aletras, Paul Clough, Samuel Fernando and Aitor Soroa.

Introduction

The aim of the PATHS project is to enable exploration and discovery within cultural heritage collections. In order to support this the project developed a range of enrichment techniques which augmented these collections with additional information to enhance the users' browsing experience. One of the demonstration systems developed in PATHS makes use of content from Europeana. This document summarises the semantic enrichment techniques developed in PATHS, with particular reference to their application to the Europeana data. Further details are available in project deliverables and associated publications, details of which are provided below.

Data

The PATHS system makes use of three collections from Europeana: 1) CultureGrid, 2) Cervantes Biblioteca Virtual Miguel De Cervantes and 3) The Biblioteca Nacional de Espana. These collections were chosen since they contain meta-data in the two languages being explored in the PATHS project (English and Spanish) in addition to containing a reasonable number of items.

Motivation for Semantic Enrichment

Analysis of the three data sets revealed that adding additional information would support exploration and discovery in Europeana and would also be helpful for interpretation of the items within them. In particular, two issues with the meta-data were identified:

- **Limited information**. Many of the items have only limited information associated with them, for example a single word title and no description.

- **Incomaptible indexing schemes.** Items in Europeana are indexed using a range of hierarchical structure (e.g. Library of Congress Subject Headings and the Art and Architecture Thesaurus) which are not compatible. In addition, many items are not indexed at all. Consequently there is no single index that covers all items, even for a single language.

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

Semantic Enrichment

A range of Natural Language Processing techniques were developed for cultural heritage data and applied to the Europeana collections used in PATHS. The aim of these techniques was to provide users of cultural heritage collections with additional information that would help them to navigate and interpret the collections.

1) Identification of Key Entities

Descriptions of items often mention key entities, such as people, locations and dates. There were identified automatically in the meta-data of the three collections by applying FreeLing (Padro et al, 2010), an open source library of language processing tools which carries out several stages of linguistic analysis: identification of nouns via part of speech (PoS) tagging. Lemmatization, multiword-unit recognition and recognition of named entities (dates, places, people, organisations etc.)

2) Item Similarity

Information about which items are similar within a collection is useful for navigation, grouping together related content and recommendation of interesting content. The PATHS project developed techniques for determining the similarity between items in cultural heritage collections using Latent Dirichlet Allocation (LDA) to discover latent "topics" within the collection. The approach is described by Aletras et al. (2012).

The similarity between all items in the three Europeana collection was computed in a pairwise fashion and the 25 items with the highest score are retained for each item. This allows uses to quickly navigate from one item to other related ones in the collection, even if they are from a different content provider.

3) Typed Similarity

The approaches to identifying similar items were extended to provide information about the reason pairs of items could be considered similar. This additional knowledge assists users in their understanding about how items in the collection are related together. Various types of similarity were identified: similar author, similar people involved, similar time period, similar location, similar events, similar location and similar description. Similar pairs of items were identified using a range of techniques described in Agirre et. al. (2013). The majority of these were based on comparison of the text in the relevant fields of item's meta-data, for example the <dc:Creator> field was used to identify similar authors. Other types of similarity, e.g. similar people involved and similar location made use of the named entities that has been identified in an earlier stage of enrichment.

4) Background Links

The information associated with each item, which is sometimes very limited, was augmented by providing links to Wikipedia. These were generated using Wikipedia Miner (Milne and Witten, 2008). See Fernando and Stevenson (2012) for further

details about how Wikipedia Miner was adapted for cultural heritage documents. This process generated in-line links in the item's meta-data. In addition, items were also mapped to Wikipedia when appropriate articles could be found, see Agirre et. al. (2012).

5) Hierarchies

The Wikipedia background links added to the item meta-data were used to automatically generate hierarchies that the cover the entire collection. Two approaches are used to generate hierarchies, WikiFreq and WikiTax. WikiFreq uses Wikipedia link frequencies across the Europeana collection to organise the items. The links in the meta-data associated with each item are ordered based on their frequency in the entire collection and that set of links then inserted into the hierarchy. The WikiTax approach uses the Wikipedia Taxonomy (Ponzetto and Strube, 2011), a taxonomy derived from Wikipedia categories. Europeana artefacts are inserted into this taxonomy using the links added by Wikipedia Miner with each artefact being added to the taxonomy for all categories listed in the links. This leads to a taxonomy in which artefacts can occur in multiple locations. The two approaches are combined to create the WikiMerge hierarchy. See Fernando et al. (2012) for further details

Application of Semantic Enrichment

The semantically enriched data from Europeana was used in the PATHS demonstration system. The similar items and background links (2, 3 and 4) for were displayed to the user when they viewed an individual item. The hierarchy (5) was used to provide high-level navigation of the collection and was displayed in a variety of ways, including using a standard thesaurus-type view, a tag cloud and a map. See Agirre et. al. (2013) for further details about how semantic enrichment is used in the demonstrator.

The PATHS demonstrator is available at http://explorer.paths-project.eu/

Semantic Enrichment Using Open Source Software

Most of the enrichment techniques applied in the PATHS project were developed in-house and are relatively complex. An overall set of recommendations for the automatic enrichment of cultural heritage collections using open source software is presented in a project report available from the PATHS project web site: Agirre and Otegi (forthcoming).

Representation of Semantic Enrichment using Europeana Data Model

The semantically enriched cultural heritage data in PATHS is encoded using the ESEPaths format, which is derived from Europeana Semantic Elements (ESE) and adds the enrichment information described above. ESE is the metadata scheme used

to describe cultural heritage objects in Europeana (see http://europeana.eu/schemas/ese/). However, Europeana is moving from ESE to a new data representation, called Europeana Data Model (EDM) (Doerr et al., 2010). A project report describing how to represent the semantically enriched data can be representing using the EDM schema will be made available on the project website: (Sora et. al, forthcoming).

Web service for semantic enrichment

All the semantic enrichment has been done offline in PATHS. However, the project provides a web service prototype which allows independent content providers to enrich their cultural heritage items online. Specifically, the service enriches the items with two types of information: links to similar items within the PATHS collection and links to Wikipedia articles which are related to it. The Web service is described by Agirre et. al. (2013).

The web service can be accessed via http://ixa2.si.ehu.es/paths_wp2/paths_wp2.pl

Project Deliverables

Further information about the enrichment is available in the following two project deliverables which are available from the PATHS project web site: http://www.paths-project.eu/eng/Resources

- D2.1 Processing and representation of Content for the first prototype by Eneko Agirre and Oier Lopez de Lacalle with Aitor Soroa, Mark Stevenson, Samuel Fernando, Nikos Aletras, Antonis Kukurikos, Kate Fernie

- D2.2 Processing and representation of Content for the second prototype by Eneko Agirre, Arantxa Otegi and Aitor Soroa with, Nikos Aletras, Constantinos Chandrinos, Samuel Fernando, Aitor Gonzalez-Agirre

References

E. Agirre, A. Barrena, O. Lopez de Lacalla, A. Soroa, M. Stevenson and S. Fernando. 2012. Matching Cultural Heritage items to Wikipedia. In Proceedings of the 8th International Conference on Language Resources and Evaluation, pages 1729--1735, Istanbul, Turkey

E. Agirre, N. Aletras, P. Clough, S. Fernando, P. Goodale, M. Hall, A. Soroa and M. Stevenson. 2013. PATHS: A System for Accessing Cultural Heritage Collections. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 151--156, Sofia, Bulgaria

E. Agirre, N. Aletras, A. Gonzalez-Agirre, G. Rigau and M. Stevenson. 2013. UBCUOS-TYPED: Regression for typed-similarity. In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, pages 132--137, Atlanta, Georgia, USA

E. Agirre, A. Barrena, K. Fernandez, E. Miranda, A. Otegi and A. Soroa. 2013. PATHSenrich: A Web Service Prototype for Automatic Cultural Heritage Item Enrichment. In *TPDL 2013, Lecture Notes in Computer Science volume 8092*, pages 462--465.

E. Agirre and A. Otegi (forthcoming) Recommendations for the automatic enrichment of DL content using open source software. PATHS Project Report

N. Aletras, M. Stevenson, and P. Clough. 2012. Computing similarity between items in a digital library of cultural heritage. *Journal of Computing and Cultural Heritage*, 5(4):no. 16.

M. Doerr, S. Gradmann, S. Hennicke, A. Isaac, C. Meghini and H. van de Sompel. 2010. The europeana data model (EDM). In *World Library and Information Congress: 76th IFLA general conference and assembly*, pages 10--15.

S. Fernando and M. Stevenson. 2012. Adapting Wikification to Cultural Heritage. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 101–106, Avignon, France.

S. Fernando, M. Hall, E. Agirre, A. Soroa, P. Clough, and M. Stevenson. 2012. Comparing taxonomies for organising collections of documents. In *Proc. of COLING 2012*, pages 879– 894, Mumbai, India.

D. Milne and I. Witten. 2008. Learning to Link with Wikipedia. In *Proc. of CIKM 2008*, Napa Valley, California.

L. Padro, M. Collado, S.Reese, M. Lloberes, and I. Castellon. 2010. FreeLing 2.1: Five Years of Open-Source Language Processing Tools. In Proceedings of 7th Language Resources and Evaluation Conference (LREC).

S. Ponzetto and M. Strube. 2011. Taxonomy induction based on a collaboratively built knowledge repository. *Artificial Intelligence*, 175(9-10):1737– 1756.

A. Soroa, E. Agirre and A. Otegi (forthcoming) RoadMap from EEPaths to EDMPaths: a user case. PATHS Project Report.