PATHS: A System for Accessing Cultural Heritage Collections

Eneko Agirre[*], **Nikolaos Aletras**[†], **Paul Clough**[†], **Samuel Fernando**[†], **Paula Goodale**[†], **Mark Hall**[†], **Aitor Soroa**[*] and **Mark Stevenson**[†]

(∗) IXA NLP Group, University of the Basque Country
Manuel Lardizabal, 1, 20.018 Donostia, Basque Country

(†) Department of Computer Science, Sheffield University
211 Portobello, Sheffield S1 4DP, United Kingdom

Abstract

This paper describes a system for navigating large collections of information about cultural heritage which is applied to Europeana, the European Library. Europeana contains over 20 million artefacts with meta-data in a wide range of European languages. The system currently provides access to Europeana content with meta-data in English and Spanish. The paper describes how Natural Language Processing is used to enrich and organise this meta-data to assist navigation through Europeana and shows how this information is used within the system.

1 Introduction

Significant amounts of information about cultural heritage has been digitised in recent years and is now easily available through online portals. However, this vast amount of material can also be overwhelming for many users since they are provided with little or no guidance on how to find and interpret this information. Potentially useful and relevant content is hidden from the users who are typically offered simple keyword-based searching functionality as the entry point into a cultural heritage collection. The situation is very different within traditional mechanisms for viewing cultural heritage (e.g. museums) where artefacts are organised thematically and users guided through the collection.

This paper describes a system that allows users to explore large cultural heritage collections. Navigation is based around the metaphor of pathways (or trails) through the collection, an approach that has been widely explored as an alternative to standard keyword-based search (Furuta et al., 1997; Reich et al., 1999; Shipman et al., 2000; White and Huang, 2010). Pathways are sets of artefacts organised around a theme which form access points to the collection.

Pathways are a useful way to access information about cultural heritage. Users accessing these collections are often unfamiliar with their content, making keyword-based search unsuitable since they are unable to formulate appropriate queries (Wilson et al., 2010). Non-keyword-based search interfaces have been shown to be suitable for exploratory search (Marchionini, 2006). Pathways support this exploration by echoing the organised galleries and guided tours found in museums.

2 Related Work

Heitzman et al. (1997) describe the ILEX system which acts as a guide through the jewellery collection of the National Museum of Scotland. The user explores the collection through a set of web pages which provide descriptions of each artefact that are personalised for each user. The system makes use of information about the artefacts the user has viewed to build up a model of their interests and uses this to customise the descriptions of each artefact and provide recommendations for further artefacts in which they may be interested.

Grieser et al. (2007) also explore providing recommendations based on the artefacts a user has viewed so far. They make use of a range of techniques including language modelling, geospatial modelling and analysis of previous visitors' behaviour to provide recommendations to visitors to the Melbourne Museum.

Grieser et al. (2011) explore methods for determining the similarity between museum artefacts, commenting that this is useful for navigation through these collections and important for personalisation (Bowen and Filippini-Fantoni, 2004; O'Donnell et al., 2001), recommendation (Bohnert et al., 2009; Trant, 2009) and automatic tour generation (Finkelstein et al., 2002; Roes et al., 2009). They also use exhibits from Melbourne

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 151–156,
Sofia, Bulgaria, August 4-9 2013.

Museum and apply a range of approaches to determine the similarity between them, including comparing descriptions and measuring physical distance between them in the museum.

These approaches, like many of the systems that have been developed for online access to cultural heritage (e.g. (Hage et al., 2010)), are based around virtual access to a concrete physical space (i.e. a museum). They often provide tours which are constrained by the physical layout of the museum, such as virtual museum visits. However, these approaches are less suitable for unstructured collections such as databases of cultural heritage artefacts collected from multiple institutions or artefacts not connected with existing physical presentation (e.g. in a museum). The PATHS system is designed for these types of collections and makes use of natural language analysis to support navigation. In particular, similarity between artefacts is computed automatically (see Section 4.1), background information automatically added to artefact descriptions (see Section 4.2) and a hierarchy of artefacts generated (see Section 4.3).

3 Cultural Heritage Data

The PATHS system has been applied to data from Europeana[1]. This is a web-portal to collections of cultural heritage artefacts provided by a wide range of European institutions. Europeana currently provides access to over 20 million artefacts including paintings, films, books, archival records and museum objects. The artefacts are provided by around 1,500 institutions which range from major institutions, including the Rijksmuseum in Amsterdam, the British Library and the Louvre, to smaller organisations such as local museums. It therefore contains an aggregation of digital content from several sources and is not connected with any one physical museum.

The PATHS system makes use of three collections from Europeana. The first of these contains artefacts from content providers in the United Kingdom which has meta-data in English. The artefacts in the remaining two collections come from institutions in Spain and have meta-data in Spanish.

CultureGrid Culture Grid[2] is a digital content provider service from the Collection Trust[3]. It contains information about over one million artefacts from 40 different UK content providers such as national and regional museums and libraries.

Cervantes Biblioteca Virtual Miguel De Cervantes[4] contains digitalised Spanish text in various formats. In total, the online library contains about 75,000 works from a range of periods in Spanish history.

Hispana The Biblioteca Nacional de España[5] contains information about a diverse set of content including text and drawings. The material is collected from different providers in Spain including museums and libraries.

Europeana stores metadata for each artefact in an XML-based format which includes information such as its title, the digital format, the collection, the year of creation and also a short description of each artefact. However, this meta-data is created by the content providers and varies significantly across artefacts. Many of the artefacts have only limited information associated with them, for example a single word title. In addition, the content providers that contribute to Europeana use different hierarchical structures to organise their collections (e.g. Library of Congress Subject Headings[6] and the Art and Architecture Thesaurus[7]), or do not organise their content into any structure. Consequently the various hierarchies that are used in Europeana only cover some of the artefacts and are not compatible with each other.

3.1 Filtering Data

Analysis of the artefacts in these three collections revealed that many have short and uninformative titles or lack a description. This forms a challenge to language processing techniques since the artefact's meta-data does not contain enough information to model it accurately.

The collections were filtered by removing any artefacts that have no description and have either fewer than four words in their title or have a title that is repeated more than 100 times in the collection. Table 1 shows the number of artefacts in each of the Europeana collections before and

[1] http://www.europeana.eu
[2] http://www.culturegrid.org.uk
[3] http://www.collectionstrust.org.uk
[4] http://www.cervantesvirtual.com
[5] http://www.bne.es
[6] http://authorities.loc.gov/
[7] http://www.getty.edu/research/tools/vocabularies/aat/

after this filter has been applied. Applying the heuristic leads to the removal of around 31% of the artefacts, although the number varies significantly across the collections with 61% of the artefacts in CultureGrid being removed and only 1% of those in Hispana.

Collection	Lang.	Total	Filtered
CultureGrid	Eng.	1,207,781	466,958
Hispana	Sp.	1,235,133	1,219,731
Cervantes	Sp.	19,278	14,983
		2,462,192	1,701,672

Table 1: Number of artefacts in Europeana collections before and after filtering

4 Data Processing

A range of pre-preprocessing steps were carried out on these collections to provide additional information to support navigation in the PATHS system.

4.1 Artefact Similarity

We begin by computing the similarity between the various artefacts in the Europeana collections. This information is useful for navigation and recommendation but is not available in the Europeana collections since they are drawn from a diverse range of sources.

Similarity is computed using an approach described by Aletras et al. (2012). in which the topics generated from each artefact's metadata using a topic model are compared. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a widely used type of topic model in which documents can be viewed as probability distributions over topics, θ. The similarity between a pair of documents can be estimated by comparing their topic distributions. This is achieved by viewing each distribution as a vector of probabilities and then computing the cosine of the angle between them:

$$sim(a,b) = \frac{\vec{\theta_a}.\vec{\theta_b}}{|\vec{\theta_a}| \times |\vec{\theta_b}|} \quad (1)$$

where $\vec{\theta_a}$ is the vector created from the probability distribution generated by LDA for document a.

This approach is evaluated using a set of 295 pairs of artefacts for which human judgements of similarity were obtained using crowdsourcing (Aletras et al., 2012). Pearson correlation between the similarity scores and human judgements was 0.53.

The similarity between all the artefacts in the collection is computed in a pairwise fashion. The 25 artefacts with the highest score are retained for each artefact.

4.2 Background Links

The metadata associated with Europeana artefacts is often very limited. Consequently links to relevant articles in Wikipedia were added to each the meta-data of each artefact using Wikipedia Miner (Milne and Witten, 2008) to provide background information. In addition to the link, Wikipedia Miner returns a confidence value between 0 and 1 for each link based on the context of the item.

The accuracy of the links added by Wikipedia Miner were evaluated using the meta-data associated with 21 randomly selected artefacts. Three annotators analysed the links added and found that a confidence value of 0.5 represented a good balance between accuracy and coverage. See Fernando and Stevenson (2012) for further details.

4.3 Hierarchies

The range of hierarchies used by the various collections that comprise the Europeana collection make navigation difficult (see Section 3). Consequently, the Wikipedia links added to the artefact meta-data were used to automatically generate hierarchies that the cover the entire collection. These hierarchies are used by the PATHS system to assist browsing and exploration.

Two approaches are used to generate hierarchies of Europeana artefacts (WikiFreq and WikiTax). These are combined to generate the WikiMerge hierarchy which is used in the PATHS system.

WikiFreq uses link frequencies across the entire collection to organise the artefacts. The first stage in the hierarchy generation process is to compute the frequency with which each linked Wikipedia article appears in the collection. The links in each artefact are these analysed to construct a hierarchy consisting of Wikipedia articles. The links in the meta-data associated with each artefact are ordered based on their frequency in the entire collection and that set of links then inserted into the hierarchy. For example, if the set of ordered links for an artefact is $a_1, a_2, a_3 \cdots a_n$ then the artefact is then inserted into the hierarchy under the branch $a_1 \rightarrow a_2 \rightarrow a_3 \cdots \rightarrow a_n$, with a_1 at the top level in the tree and the artefact appearing under the node a_n. If this branch does not already exist in the tree then it is created.

The hierarchy is pruned to removing nodes with fewer than 20 artefacts in them. In addition, if a node has more than 20 child nodes, only the 20 most frequent are used.

WikiTax 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.

Each approach was used to generate hierarchies from the Europeana collections. The resulting hierarchies were evaluated via online surveys, see Fernando et al. (2012) for further details. It was found that WikiFreq performed well at placing items into the correct location in the taxonomy and grouping together similar items under the same node. However, the overall structure of WikiTax was judged to be more coherent and comprehensible.

WikiMerge combines combines WikiFreq and WikiTax. WikiFreq is used to link each artefact to Wikipedia articles $a_1 \ldots a_n$, but only the link to the most specific article, a_n, is retained. The a_n articles are linked to their parent WikiTax topics based on the Wikipedia categories the articles belong to. The resulting hierarchy is pruned removing all WikiTax topics that do not have a WikiFreq child or have only one child topic. Finally top-level topics in the combined hierarchy are then linked to their respective Wikipedia root node.

The resulting WikiMerge hierarchy has WikiFreq topics as its leaves and WikiTax topics as its interior and root nodes. Experiments showed that this approach was successful in combining the strengths of the two methods (Fernando et al., 2012).

5 The PATHS System

The PATHS system provides access to the Europeana collections described in Section 3 by making use of the additional information generated using the approaches described in Section 4. The interface of the PATHS system has three main areas:

Paths enables users to navigate via pathways (see Section 5.1).

Search supports discovery of both collection artefacts and pathways through keyword search (see Section 5.2).

Explore enables users to explore the collections using a variety of types of overview (see Section 5.3).

5.1 Paths Area

This area provides users with access to Europeana through pathways or trails. These are manually generated sets of artefacts organised into a tree structure which are designed to showcase the content available to the user in an organised way. These can be created by users and can be published for others to follow. An example pathway on the topic "railways" is shown in Figure 1. A short description of the pathway's content is shown towards the top of the figure and a graphical overview of its contents at the bottom.

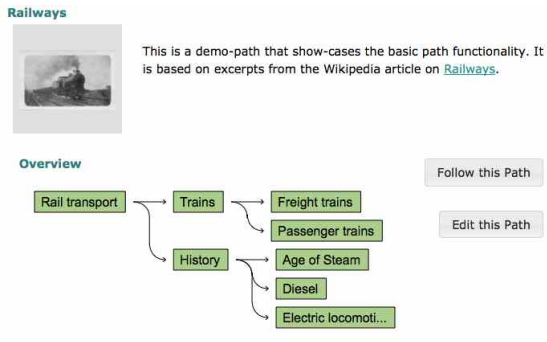

Figure 1: Example pathway on the topic "railways"

Figure 2 shows as example artefact as displayed in the system. The example artefact is a portrait of Catherine the Great. The left side of the figure shows information extracted directly from the Europeana meta-data for this artefact. The title and textual description are shown towards the top left together with a thumbnail image of the artefact. Other information from the meta-data is shown beneath the "About this item" heading. The right side of the figure shows additional information

Figure 2: Example artefact displayed in system interface. Related artefacts and background links are displayed on right hand side

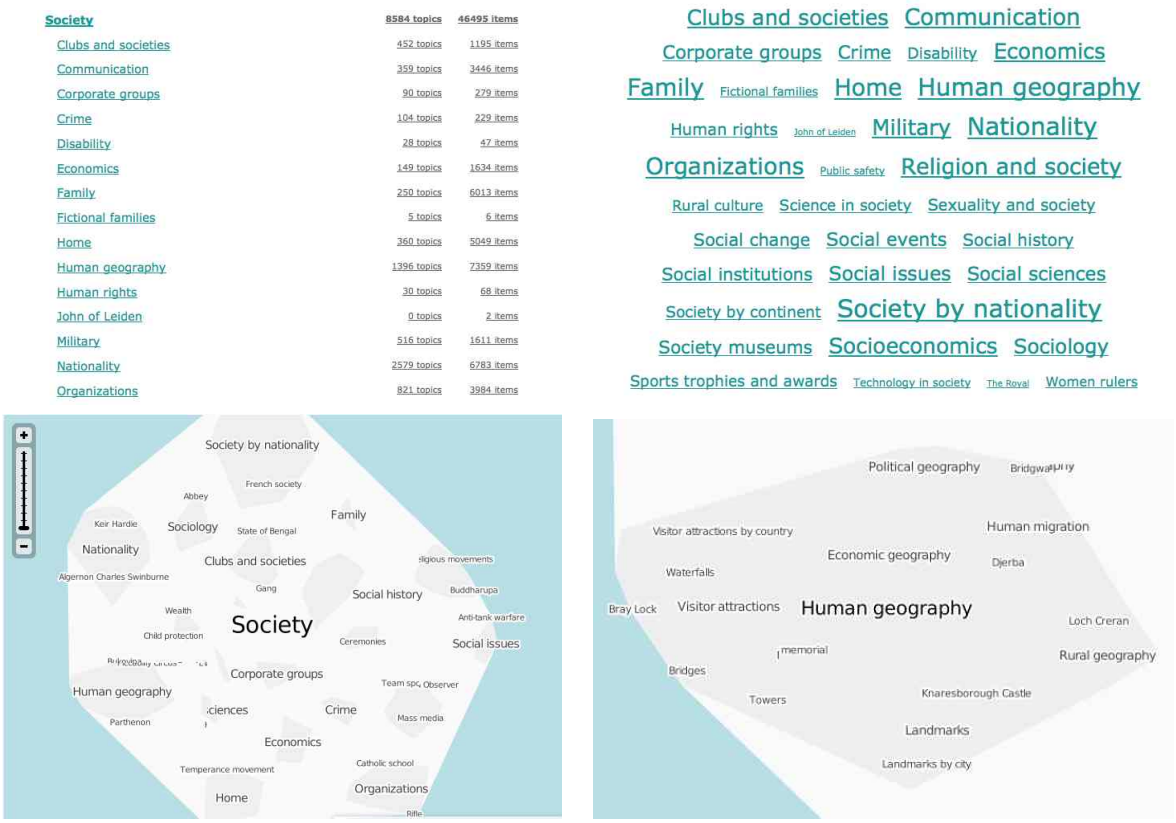

Figure 3: Example visualisations of hierarchy: thesaurus view (top left), tag cloud (top right), map views (bottom)

about the artefact generated using the approaches described in Sections 4.1 and 4.2. Related artefacts are shown to the user one at a time, clicking on the thumbnail image leads to the equivalent page for the related artefact. Below this are links to the Wikipedia articles that are identified in the text of the article's title and description.

5.2 Search Area

This area allows users to search for artefacts and pathways using standard keyword search implemented using Lucene (McCandless et al., 2010).

5.3 Explore Area

The system provides a variety of ways to view the hierarchies generated using the approach described in Section 4.3. Figure 3 shows how these are displayed for a section of the hierarchy with the label "Society". The simplest view (shown in the top left of Figure 3) is a thesaurus type view in which levels of the hierarchy are represented by indentation. The system also allows levels of the hierarchy to be viewed as a tag cloud (top right of Figure 3). The final representation of the hierarchy is as a map, shown in the bottom of Figure 3. In this visualisation categories in the hierarchy are represented as "islands" on the map. Zooming in on the map provides more detail about that area of the hierarchy.

6 Summary and Future Developments

This paper describes a system for navigating Europeana, an aggregation of collections of cultural heritage artefacts. NLP analysis is used to organise the collection and provide additional information. The results of this analysis are provided to the user through an online interface which provides access to English and Spanish content in Europeana.

Planned future development of this system includes providing recommendations and more personalised access. Similarity information (Section 4.1) can be used to provide information from which the recommendations can be made. Personalised access will make use of information about individual users (e.g. from their browsing behaviour or information they provide about their preferences) to generate individual views of Europeana.

Online Demo

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

Acknowledgments

The research leading to these results was carried out as part of the PATHS project (`http://paths-project.eu`) funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 270082

References

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.

D. Blei, A. Ng, and M. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

F. Bohnert, D. Schmidt, and I. Zuckerman. 2009. Spatial Process for Recommender Systems. In *Proc. of IJCAI 2009*, pages 2022–2027, Pasadena, CA.

J. Bowen and S. Filippini-Fantoni. 2004. Personalization and the Web from a Museum Perspective. In *Proc. of Museums and the Web 2004*, pages 63–78.

Samuel Fernando and Mark 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.

Samuel Fernando, Mark Hall, Eneko Agirre, Aitor Soroa, Paul Clough, and Mark Stevenson. 2012. Comparing taxonomies for organising collections of documents. In *Proc. of COLING 2012*, pages 879–894, Mumbai, India.

L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. 2002. Placing Search in Context: The Concept Revisited. *ACM Trans. on Information Systems*, 20(1):116–131.

R. Furuta, F.. Shipman, C. Marshall, D. Brenner, and H. Hsieh. 1997. Hypertext paths and the World-Wide Web: experiences with Walden's Paths. In *Proc. of the Eighth ACM conference on Hypertext*, pages 167–176, New York, NY.

K. Grieser, T. Baldwin, and S. Bird. 2007. Dynamic Path Prediction and Recommendation in a Museum Environment. In *Proc. of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pages 49–56, Prague, Czech Republic.

K. Grieser, T. Baldwin, F. Bohnert, and L. Sonenberg. 2011. Using Ontological and Document Similarity to Estimate Museum Exhibit Relatedness. *Journal of Computing and Cultural Heritage*, 3(3):1–20.

W.R. van Hage, N. Stash, Y. Wang, and L.M. Aroyo. 2010. Finding your way through the Rijksmuseum with an adaptive mobile museum guide. In *Proc. of ESWC 2010*, pages 46–59.

J. Heitzman, C. Mellish, and J. Oberlander. 1997. Dynamic Generation of Museum Web Pages: The Intelligent Labelling Explorer. *Archives and Museum Informatics*, 11(2):117–125.

G. Marchionini. 2006. Exploratory Search: from Finding to Understanding. *Comm. ACM*, 49(1):41–46.

M. McCandless, E. Hatcher, and O. Gospodnetic. 2010. *Lucene in Action*. Manning Publications.

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

M. O'Donnell, C. Mellish, J. Oberlander, and A. Knott. 2001. ILEX: An architecture for a dynamic hypertext generation system. *Natural Language Engineering*, 7:225–250.

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

S. Reich, L. Carr, D. De Roure, and W. Hall. 1999. Where have you been from here? Trails in hypertext systems. *ACM Computing Surveys*, 31.

I. Roes, N. Stash, Y. Wang, and L. Aroyo. 2009. A personalized walk through the museum: the CHIP interactive tour guide. In *Proc. of the 27th International Conference on Human Factors in Computing Systems*, pages 3317–3322, Boston, MA.

F. Shipman, R. Furuta, D. Brenner, C. Chung, and H. Hsieh. 2000. Guided paths through web-based collections: Design, experiences, and adaptations. *Journal of the American Society for Information Science*, 51(3):260–272.

J. Trant. 2009. Tagging, folksonomies and art museums: Early experiments and ongoing research. *Journal of Digital Information*, 10(1).

R. White and J. Huang. 2010. Assessing the scenic route: measuring the value of search trails in web logs. In *Proc. of SIGIR 2010*, pages 587–594.

M. Wilson, Kulesm B., M. Schraefel, and B. Schneiderman. 2010. From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends in Web Science*, 2(1):1–97.