Enabling the Discovery of Digital Cultural Heritage Objects through Wikipedia

Conference or Workshop Item

How to cite:

For guidance on citations see FAQs.

© 2012 Association for Computational Linguistics

https://creativecommons.org/licenses/by/

Version: Version of Record

Link(s) to article on publisher’s website:
https://www.aclweb.org/anthology/W12-1013

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online’s data policy on reuse of materials please consult the policies page.
Enabling the Discovery of Digital Cultural Heritage Objects through Wikipedia

Mark M Hall
Paul D Clough
Information School
Sheffield University
Sheffield, UK
m.mhall@shef.ac.uk
p.d.clough@shef.ac.uk

Oier Lopez de Lacalle1,2
IKERBASQUE
Basque Foundation for Science
Bilbao, Spain
p.d.clough@shef.ac.uk
oier.lopezdelacalle@gmail.es

Aitor Soroa
IXA NLP Group
University of the Basque Country
Donostia, Spain
a.soroa@ehu.es

Eneko Agirre
University of Edinburgh
Edinburgh, UK
e.agirre@ehu.es

Abstract

Over the past years large digital cultural heritage collections have become increasingly available. While these provide adequate search functionality for the expert user, this may not offer the best support for non-expert or novice users. In this paper we propose a novel mechanism for introducing new users to the items in a collection by allowing them to browse Wikipedia articles, which are augmented with items from the cultural heritage collection. Using Europeana as a case-study we demonstrate the effectiveness of our approach for encouraging users to spend longer exploring items in Europeana compared with the existing search provision.

1 Introduction

Large amounts of digital cultural heritage (CH) information have become available over the past years, especially with the rise of large-scale aggregators such as Europeana1, the European aggregator for museums, archives, libraries, and galleries. These large collections present two challenges to the new user. The first is discovering the collection in the first place. The second is then discovering what items are present in the collection. In current systems support for item discovery is mainly through the standard search paradigm (Sutcliffe and Ennis, 1998), which is well suited for CH professionals who are highly familiar with the collections, subject areas, and have specific search goals. However, for new users who do not have a good understanding of what is in the collections, what search keywords to use, and have vague search goals, this method of access is unsatisfactory as this quote from (Borgman, 2009) exemplifies:

“So what use are the digital libraries, if all they do is put digitally unusable information on the web?”

Alternative item discovery methodologies are required to introduce new users to digital CH collections (Geser, 2004; Steemson, 2004). Exploratory search models (Marchionini, 2006; Pirolli, 2009) that enable switching between collection overviews (Hornb[Please insert into preamble]k and Hertzum, 2011) and detailed exploration within the collection are frequently suggested as more appropriate.

We propose a novel mechanism that enables users to discover an unknown, aggregated collection by browsing a second, known collection. Our method lets the user browse through Wikipedia and automatically augments the page(s) the user is viewing with items drawn from the CH collection, in our case Europeana. The items are chosen to match the page’s content and enable the user to acquire an overview of what information is available for a given topic. The goal is to introduce new users to the digital collection, so that they can then successfully use the existing search systems.

2 Background

Controlled vocabularies are often seen as a promising discovery methodology (Baca, 2003). However, in the case of aggregated collections such as Europeana, items from different providers are frequently aligned to different vocabularies, requiring an integration of the two vocabularies in

1http://www.europeana.eu
order to present a unified structure. (Isaac et al., 2007) describe the use of automated methods for aligning vocabularies, however this is not always successfully possible. A proposed alternative is to synthesise a new vocabulary to cover all aggregated data, however (Chaudhry and Jiun, 2005) highlight the complexities involved in then linking the individual items to the new vocabulary.

To overcome this automatic clustering and visualisations based directly on the meta-data have been proposed, such as 2d semantic maps (Andrews et al., 2001), automatically generated tree structures (Chen et al., 2002), multi-dimensional scaling (Fortuna et al., 2005; Newton et al., 2009), self-organising maps (Lin, 1992), and dynamic taxonomies (Papadakos et al., 2009). However none of these have achieved sufficient success to find widespread use as exploration interfaces.

Faceted search systems (van Ossenbruggen et al., 2007; Schmitz and Black, 2008) have arisen as a flexible alternative for surfacing what metadata is available in a collection. Unlike the methods listed above, faceted search does not require complex pre-processing and the values to display for a facet can be calculated on the fly. However, aggregated collections frequently have large numbers of potential facets and values for these facets, making it hard to surface a sufficiently large fraction to support resource discovery.

Time-lines such as those proposed by (Luo et al., 2012) do not suffer from these issues, but are only of limited value if the user’s interest cannot be focused through time. A user interested in examples of pottery across the ages or restricted to a certain geographic area is not supported by a time-line-based interface.

The alternative we propose is to use a second collection that the user is familiar with and that acts as a proxy to the unfamiliar collection. (Villa et al., 2010) describe a similar approach where Flickr is used as the proxy collection, enabling users to search an image collection that has no textual meta-data.

In our proposed approach items from the unfamiliar collection are surfaced via their thumbnail images and similar approaches for automatically retrieving images for text have been tried by (Zhu et al., 2007; Borman et al., 2005). (Zhu et al., 2007) report success rates that approach the quality of manually selected images, however their approach requires complex pre-processing, which

3 Discovering Europeana through Wikipedia

As stated above our method lets users browse Wikipedia and at the same time exposes them to items taken from Europeana, enabling them to discover items that exist in Europeana.

The Wikipedia article is augmented with Europeana items at two levels. The article as a whole is augmented with up to 20 items that in a pre-processing step have been linked to the article and at the same time each paragraph in the article is augmented with one item relating to that paragraph.

Our system (Wikiana, figure 1) sits between the user and the data-providers (Wikipedia, Europeana, and the pre-computed article augmentation links). When the user requests an article from Wikiana, the system fetches the matching article from Wikipedia and in a first step strips everything except the article’s main content. It then queries the augmentation database for Europeana items that have been linked to the article and selects the top 20 items from the results, as detailed below. It then processes each paragraph and uses
Figure 2: Screenshot of the augmented article “Mediterranean Sea” with the pre-processed article-level augmentation at the top and the first two paragraphs augmented with items as returned by the Europeana API.

The system makes heavy use of caching to speed up the process and also to reduce the amount of load on the backend systems.

3.1 Article augmentation

To create the article-level augmentations we first create a Wikipedia “dictionary”, which maps strings to Wikipedia articles. The mapping is created by extracting all anchor texts from the inter-article hyperlinks and mapping these to the articles they link to. For instance, the string “roman coin” is used as an anchor in a link to the Wikipedia article Roman currency. As a result, we create a many-to-many mapping between Wikipedia articles and Europeana items. The Wikiana application displays at most 20 images per article, thus the Europeana items need to be ranked. The goal is to rank interesting items higher, with “interestingness” defined as how unusual the items are in the collection. This metric is an adaption of the standard inverse-document-frequency formula used widely in Information Retrieval and is adapted to identify items that have meta-data field-values that are infrequent in the collection. As in original IDF we diminish the weight of values that occur very frequently in the collection, the non-interesting items, and increases the weight of values that occur rarely, the interesting items. More formally the interestingness $\alpha_i$ of an item $i$ is calculated as follows:

$$\alpha_i = \frac{\#\{\text{title}_i\}}{\mu_{\text{title}}} \log \frac{N_{\text{title}}}{c(\text{title}_i) + 1} + \frac{\#\{\text{desc}_i\}}{\mu_{\text{desc}}} \log \frac{N_{\text{desc}}}{c(\text{desc}_i) + 1} + \frac{\#\{\text{subj}_i\}}{\mu_{\text{subj}}} \log \frac{N_{\text{subj}}}{c(\text{subj}_i) + 1}$$

where $\#\{\text{field}_i\}$ is the length in words of the field of the given item $i$, $\mu_{\text{field}}$ is the average length in words of the field in the collection, $N_{\text{field}}$ is the total number of items containing that field in the collection.
The Roman Empire (Latin: Imperium Romanum) was the post-Republican period of the ancient Roman civilization, characterised by an autocratic form of government and large territorial holdings in Europe and around the Mediterranean.

“Latin language” OR “Roman Republic” OR “Ancient Rome” or “Autocracy”

Figure 4: Example paragraph with the Wikipedia hyperlinks in bold. Below the search keywords extracted from the hyperlinks and the resulting thumbnail image.

entire collection, and $c(field_i)$ is the frequency of the value in that field.

Items are ranked by descending $\alpha_i$ and the for the top 20 items, the thumbnails for the items are added to the top of the augmented page.

3.2 Paragraph augmentation

The items found in the article augmentation tend to be very focused on the article itself, thus to provide the user with a wider overview of available items, each paragraph is also augmented. This augmentation is done dynamically when an article is requested. As stated above the augmentation iterates over all paragraphs in the article and for each article determines its core keywords. As in the article augmentation the Wikipedia hyperlinks are used to define the core keywords, as the inclusion of the link in the paragraph indicates that this is a concept that the author felt was relevant enough to link to. For each paragraph the Wikipedia hyperlinks are extracted, the underscores replaced by spaces and these are then used as the query keywords. The keywords are combined using “OR” and enclosed in speech-marks to ensure only exact phrase matches are returned and then submitted to Europeana’s OpenSearch API (fig. 4). From the result set an item is randomly selected and the paragraph is augmented with the link to the item, the item’s thumbnail image and its title. If there are no hyperlinks in a paragraph or the search returns no results, then no augmentation is performed for that paragraph.

4 Evaluation

The initial evaluation focuses on the paragraph augmentation, as the quality of that heavily depends on the results provided by Europeana’s API and on a log-analysis looking at how users com-

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Familiar</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Appropriate</td>
<td>9</td>
<td>27</td>
</tr>
<tr>
<td>Supports</td>
<td>4</td>
<td>32</td>
</tr>
<tr>
<td>Visually interesting</td>
<td>13</td>
<td>23</td>
</tr>
<tr>
<td>Find out more</td>
<td>3</td>
<td>33</td>
</tr>
</tbody>
</table>

Table 1: Evaluation experiment results reduced from the 5-point Likert-like scale to a yes/no level.

For the paragraph augmentation evaluation 18 wikipedia articles were selected from six topics (Place, Person, Event, Time period, Concept, and Work of Art). From each article the first paragraph and a random paragraph were selected for augmentation, resulting in a total set of 36 augmented paragraphs. In the experiment interface the participants were shown the text paragraph, the augmented thumbnail image, and five questions (“How familiar are you with the topic?”, “How appropriate is the image?”, “How well does the image support the core ideas of the paragraph?”, “How visually interesting is the image?”, and “How likely are you to click on the image to find out more?”). Each question used a five-point Likert-like scale for the answers, with 1 as the lowest score and 5 the highest. Neither the topic nor the paragraph selection have a statistically significant influence on the results. To simplify the analysis the results have been reduced to a yes/no level, where an image is classified as “yes” for that question if more than half the participants rated the image 3 or higher on that question (table 1).

Considering the simplicity of the augmentation approach and the fact that the search API is not under our control, the results are promising. 9 out of 36 (25%) of the items were classified as appropriate. The non-appropriate images are currently being analysed to determine whether there are shared characteristics in the query structure or item meta-data that could be used to improve the query or filter out non-appropriate result items.

The difficulty with automatically adding items taken from Europeana is also highlighted by the fact that only 13 of the 36 (36%) items were classified as interesting. While no correlation could be found between the two interest and appro-
Table 2: Summary statistics for the number of items viewed in per session for users coming from our system (Wikiana) and for all Europeana users.

<table>
<thead>
<tr>
<th>Category</th>
<th>Sessions</th>
<th>1st q.</th>
<th>Med</th>
<th>3rd q.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikiana</td>
<td>88</td>
<td>6</td>
<td>11</td>
<td>15.25</td>
</tr>
<tr>
<td>All users</td>
<td>577642</td>
<td>3</td>
<td>8</td>
<td>17</td>
</tr>
</tbody>
</table>

appropriate results, only one of the 23 uninteresting items was judged appropriate, while 8 out of 9 of the appropriate items were also judged to be interesting. We are now looking at whether the item meta-data might allow filtering uninteresting items, as they seem unlikely to be appropriate.

Additionally the approach taken by (Zhu et al., 2007), where multiple images are shown per paragraph, is also being investigated, as this might reduce the impact of non-appropriate items.

4.2 Log analysis

Although the paragraph augmentation results are not as good as we had hoped, a log analysis shows that the system can achieve its goal of introducing new users to an unknown CH collection (Europeana). The system has been available online for three months, although not widely advertised, and we have collected Europeana’s web-logs for the same period. Using the referer information in the logs we can distinguish users that came to Europeana through our system from all other Europeana users. Based on this classification the number of items viewed per session were calculated (table 2). To prevent the evaluation experiment influencing the log analysis only logs acquired before the experiment date were used.

Table 2 clearly shows that users coming through our system exhibit different browsing patterns. The first quartile is higher, indicating that Wikiana users do not leave Europeana as quickly, which is further supported by the fact that 30% of the general users leave Europeana after viewing three items or less, while for Wikiana users it is only 19%. At the same time the third quartile is lower, showing that Wikiana users are less likely to have long sessions on Europeana. The difference in the session length distributions has also been validated using a Kolmogorov-Smirnov test ($p = 0.00287, D = 0.1929$).

From this data we draw the hypothesis that Wikiana is at least in part successfully attracting users to Europeana that would normally not visit or not stay and that it successfully helps users overcome that first hurdle that causes almost one third of all Europeana users to leave after viewing three or less items.

5 Conclusion and Future Work

Recent digitisation efforts have led to large digital cultural heritage (CH) collections and while search facilities provide access to users familiar with the collections there is a lack of methods for introducing new users to these collections. In this paper we propose a novel method for discovering items in an unfamiliar collection by browsing Wikipedia. As the user browses Wikipedia articles, these are augmented with a number of thumbnail images of items taken from the unknown collection that are appropriate to the article’s content. This enables the new user to become familiar with what is available in the collection without having to immediately interact with the collection’s search interface.

An early evaluation of the very straightforward augmentation process revealed that further work is required to improve the appropriateness of the items used to augment the Wikipedia articles. At the same time a log analysis of Europeana browsing sessions showed that users introduced to Europeana through our system were less likely to leave after viewing less than three items, providing clear indication that the methodology proposed in this paper is successful in introducing new users to a large, aggregated CH collection.

Future work will focus on improving the quality of the augmentation results by including more collections into the article-level augmentation and by introducing an “interestingness” ranking into the paragraph augmentation. We will also look at evaluating the system in a task-based setting and with existing, comparable systems.

Acknowledgements

The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 270082. We acknowledge the contribution of all project partners involved in PATHS (see: http://www.paths-project.eu).
References


