SMTP: Stedelijk Museum Text Mining Project

Jeroen Smeets
Maastricht University
smeetsjeroen@hotmail.com

Prof. Dr. Ir. Johannes C. Scholtes
Maastricht University
j.scholtes@maastrichtuniversity.nl

Dr. Claartje Rasterhoff
CREATE
University of Amsterdam
C.Rasterhoff@uva.nl

Dr. Margriet Schavemaker
Stedelijk Museum Amsterdam
M.Schavemaker@stedelijk.nl

Introduction

This paper addresses how text-mining, machine-learning and information retrieval algorithms from the field of artificial intelligence can be used to analyze Art-Research archives and conduct (art-)historical research. Two aspects are to focus on in order to gain quick insight into the archive: relations between groups of people using community detection, and global content changes over time using topic modeling. To develop and test the validity and relevance of existing tools, close collaboration was established between the AI researchers, museum staff, and researchers in CREATE, a digital humanities project that investigates the development of cultural industries in Amsterdam over the course of the last five centuries.

Data

The research draws on two datasets. The principal dataset is the digitized archive of the Stedelijk Museum Amsterdam, a renowned international museum dedicated to modern and contemporary art and design. The archive of the Stedelijk Museum Amsterdam is a static collection of approximately 160,000 OCR digitized documents from the period 1930-1980. The second dataset is drawn from Delpher (Koninklijke Bibliotheek Nederland, 2015), which provides a collection of digitized newspapers, books and magazines that is available for research. A selection of newspapers was made that is used as an additional dataset for this project.
Methodology

The methodology uses two approaches to obtain a quick and detailed overview of the content of a digitized archive that contains unstructured information.

Relation networks and Community Detection

The first approach focuses on the relations between named entities and aims at finding communities in the relation network. In its most basic form, a relation between two named entities can be said to exist when both occur in the same document. The relation is characterized by its strength, the number of documents in which both named entities occur, and by the sentiment content of the documents. The hypothesis is that relations between individuals with a high sentiment are more interesting than relations with a low sentiment. This is because sentiments around trigger-events are often higher than around common-day events. Community detection algorithms (Fortunato, 2010) are applied to the relation network, shown in Figure 1. Communities such as group exhibitions, art movements or a group of artists closely related to the museum director, could be identified with the help of a museum expert.

Name Extraction

In order to build the relation network, a name extraction method is developed that is able to handle multiple causes of name variations such as OCR induced errors, spelling mistakes, name abbreviations and first and last name combinations. The method makes use of lists of name
variations that are extracted from the document collection, relying on a name database such as RKArtists& (RKD, 2015). A similarity score is calculated between the original name and the possible name variation, based on an n-gram set matching technique described in (Song and Chen, 2007). Using a threshold of 0.9 on the similarity score, the method was tested on 50 randomly chosen names, resulting in an average precision of 81 percent.

**Time based Topic Modeling**

The information content and its evolution over time is analyzed using topic modeling. A time-based collective matrix factorization technique is used, as suggested in (Vaca et al., 2014). It is based on Non-Negative Matrix Factorization (NMF) (Arora et al., 2012), extended by introducing a topic transition matrix that allows to track topics as they emerge, evolve and fade over time. The algorithm was applied to both the archive of the Stedelijk Museum Amsterdam and newspaper articles from the Delpher database. The results are visualized over time in the form of stacked topic rivers (Wei et al., 2010), shown in Figure 2 and Figure 3. Several exhibitions and events could be identified and are annotated on the chart.

![Figure 2: Time based topic modeling for the archive of the Stedelijk Museum Amsterdam](image-url)
Figure 3: Time based topic modeling for Delpher newspaper articles

Conclusion

For the humanities researchers in this project, the main aim was to assess the research potential of computational analysis of digitized art archives in general, and the Stedelijk Museum in particular. Community detection, relying on a robust name extraction method, together with time based topic modeling were applied to the archives. The results demonstrate how AI tools may uncover unexpected relationships between people, artworks and organizations, as well as changing patterns over time. A second possible application can be found in connecting previously isolated content with other relevant digital sources, such as RKDartists& and Delpher. The developed techniques therefore enable unstructured art historical archives to speak up in unprecedented ways – even if the results are not clean at the first try and do not capture all historical nuance. By allowing the archive to ‘open up’, the proposed approaches offer ways to reveal hidden story lines that subvert and augment prevailing historical narratives.

References


