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

November 1, 2015

1 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. To gain quick insight into the archive, two aspects are focused on: relations between groups of people using community detection, and global content changes over time using topic modeling. For such archives pre-tagged ground-truth collections are generally not available, and the archives are often too large, geographically distributed, and not always available in digital formats to build such a ground-truth at reasonable costs. 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.

2 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 contains documents from the period 1930-1980. The corpus is a static collection of approximately 160.000 text documents that were digitized using OCR. The second dataset is drawn from Delpher, developed by Koninklijke Bibliotheek Nederland [1]. Delpher 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. Only articles from 1930-1980 that resulted from the query "Stedelijk Museum" AND "Amsterdam" were used, forming a set of 18.290 articles.

3 Methodology

The following methodology uses two approaches to obtain a quick and detailed overview of the content of a digitized archive that contains unstructured information. The first one focuses on the relations between named entities and aims at finding communities in the relation network. The second approach uses time based topic-modeling to get an overview of content changes over time. Finally, a name extraction method is presented that is able to handle multiple causes of name variations.

3.1 Relation networks and Community Detection

In its most basic form, a relation between two named entities can be said to exist when they occur together in the same document. The strength of a relation can be characterized by the number of documents in which both named entities occur. When all the co-occurrences are found, a relation network can be constructed.

In addition, sentiment analysis can be done to further characterize a relation. A sentiment score is assigned to each document, indicating the sentiment content of the document. No distinction is made between positive and negative sentiment polarity. 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. A lexicon based approach is used with lists of language specific sentiment words. The sentiment score of a document is then given by the sigmoid of the count of the sentiment words in the document, normalized by the number of words in the document.

Finally, community detection algorithms can be applied to the relation network. These types of algorithms aim at finding clusters of groups of entities that have dense connections between members of the clusters and sparse connections with members of other clusters [2]. The relation weight measure that is used to calculate the communities, is taken as the product of the strength of the relation, i.e. the number of documents where both entities occur in, and the average sentiment score of the documents of a relation. It was found that combining these two measures, resulted in more meaningful communities.

3.2 Time based Topic Modeling

In the next approach, topic modeling algorithms are applied to analyze the information content and their evolution over time. Topic modeling tries to discover the underlying thematic structure in a collection of documents. Non-Negative Matrix Factorization (NMF) is being used as a tool for topic modeling [3]. NMF is an unsupervised method where a matrix is approximated by two low rank non-negative matrices. The extracted semantic feature vectors have only non-negative values and are sparse so they are easily interpretable. Furthermore, NMF is shown to generate more consistent results over multiple runs [4], compared to other tools used for topic modeling such as LDA [5].

The approach suggested in [6] uses a time-based collective matrix factorization based on NMF and is used in this project. It extends NMF by introducing a topic transition matrix that allows to track topics as they emerge, evolve and fade over time.

3.3 Name Extraction

The following method was used to extract named entities from a collection of documents in order to build the relation network. It handles different causes of name variations such as OCR induced errors commonly found in digitized document collections, spelling mistakes, name abbreviations and first and last name combinations.

The method makes use of lists of name variations. Starting from a set of names extracted from a name database, such as RKDArtists& [7], the document collection is searched for possible name variations. These variations are found by searching for the last name using a fuzzy search. The similarity between the group of tokens around the found last name, and the original name is then calculated as a similarity score. The similarity score calculation is based on the idea described in [8], which uses a n-gram set matching technique. The lists of name variations can then be evaluated manually or a threshold on the similarity score can be used to identify name variations that correspond to the original name. The method using a threshold of 0.9 on the similarity score was tested on 50 randomly chosen names. The average precision was found to be 81 percent.

4 Results

A relation network was constructed for the document collection of the archive of the Stedelijk Museum Amsterdam. Only artists with the qualification 'graphic artist' in the RKDArtists& database were used. The methods were implemented using available open source software libraries such as the Apache Lucene text search engine library [9] and the Gephi platform [10]. The standard community detection feature in Gephi was used, which is based on the Louvain method [11]. The result is shown in figure 1. The color of the relation between the nodes indicates the average sentiment score of the relation, starting from blue (neutral) to red (high sentiment content). 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.

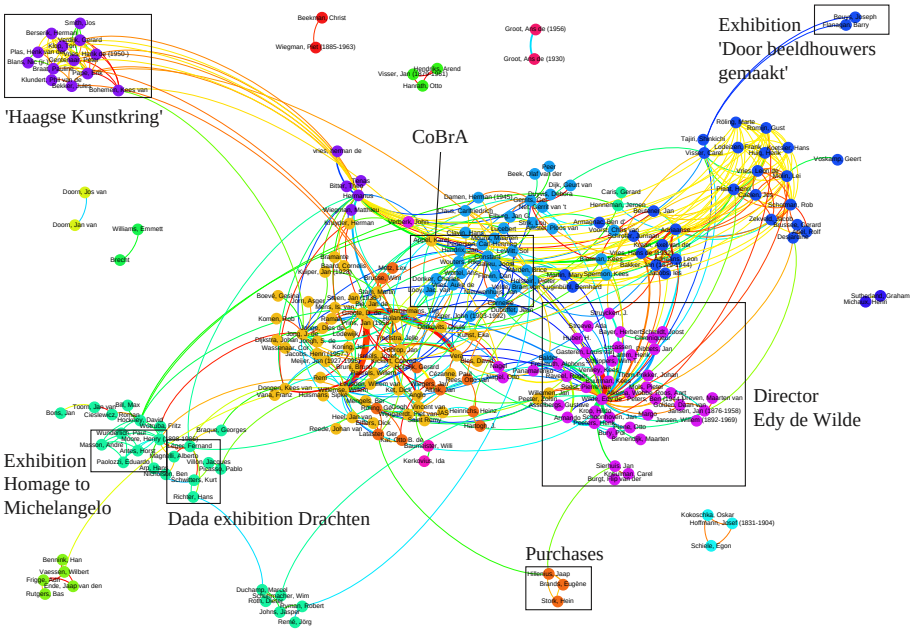


Figure 1: Found communities for ”graphic artists” in the archive of the Stedelijk Museum

The time based topic modeling algorithm suggested in [6] was implemented in MATLAB and Java. The algorithm was applied to both the archive of the Stedelijk Museum Amsterdam and newspaper articles from the Delpher database [1]. The results are visualized over time in the form of *stacked topic rivers* [12], shown in figure 2. Several exhibitions and events could be identified and are annotated on the chart.

5 Conclusion

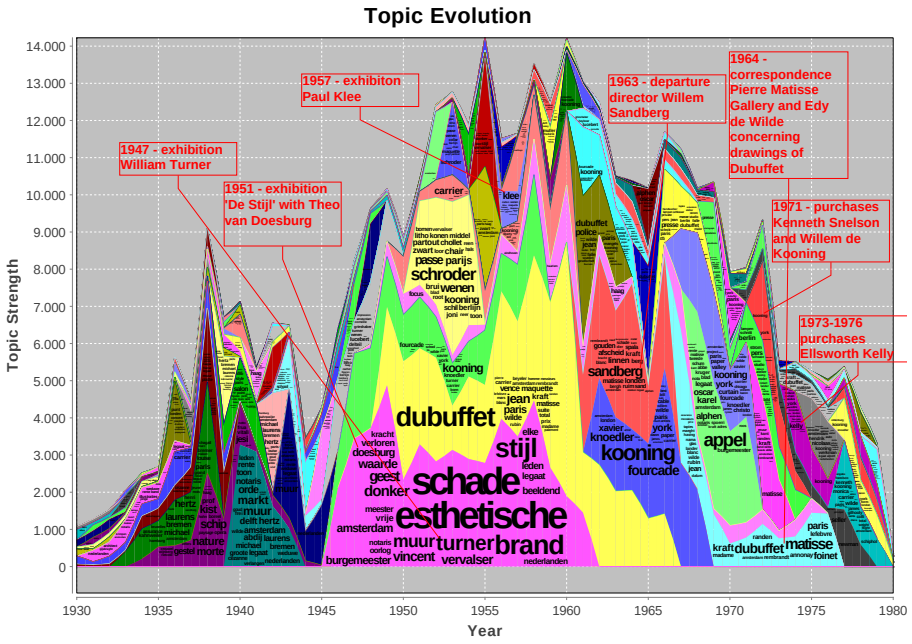
This paper discusses two approaches to gain insight into a digitized archive. Relation networks of persons with community detection are considered, relying on a robust name extraction method. Furthermore, the evolution of content over time can be explored using time based topic modeling.

For the humanities researchers in this project, the main aim was to asses the research potential of computational analysis of digitized art archives in general, and the Stedelijk Museum in particular. Two types of preliminary research questions were developed to do so. The first type had to do with identifying patterns of change and continuity, across time and place. These include for instance tracing the position of the Stedelijk Museum as an intermediary in Dutch design industries, or the development of the Stedelijk Museum as an increasingly international player. The second type of question is less concerned with general historical patterns, and more with specific art-historical research questions, regarding for instance (networks of) particular artists, artworks or exhibitions. But before we could start asking such questions to digitized art-historical archives, the quality and accessibility of the texts needed to be established. Secondly, specific methods needed to be explored and adapted in order to clean, identify, retrieve, extract, and structure the texts. The first results presented in this paper demonstrate that even though they may not be clean at the first try or capture all historical nuance, they do help archives to open up and show unexpected relationships and patterns, to answer specific questions, and to get connected with other relevant sources, such RKDartists& and Delpher. The community detection in relation with sentiment mining, the topic modeling and name extraction method developed in this project therefore provide a solid basis for the next step in assessing the research potential of art-historical archives: developing in-depth case studies, again in close collaboration with art-historians and historians, allowing the archive to speak up in unprecedented ways, offering access to hidden story lines that subvert and augment prevailing historical narratives.

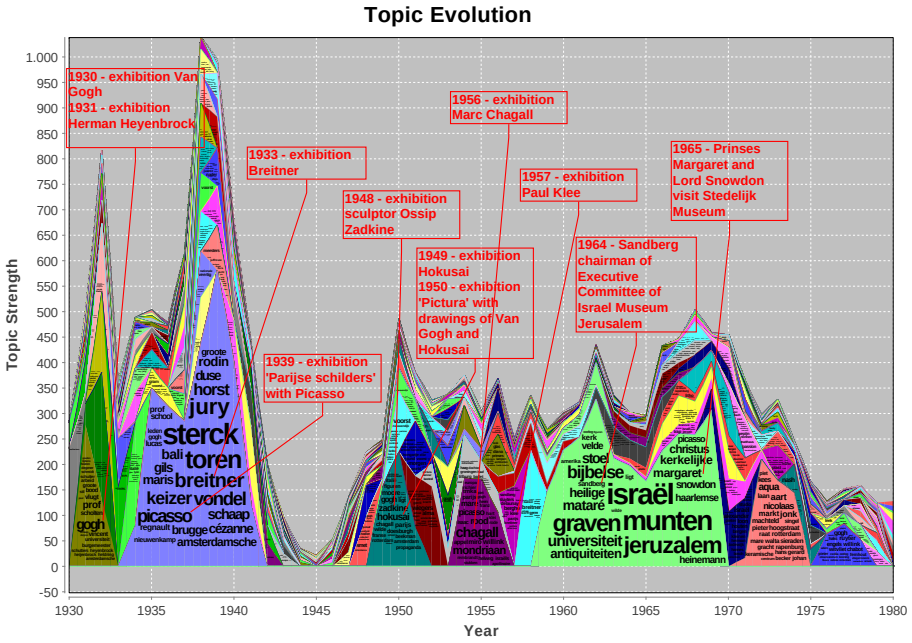
References

[1] Koninklijke Bibliotheek Nederland. Delpher - Boeken Kranten Tijdschriften, 2015. `http://www.delpher.nl/`.

[2] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.



(a) Archive Stedelijk Museum Amsterdam



(b) Delpher newspaper articles

Figure 2: Time-based Topic Modeling

[3] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 1–10. IEEE, 2012.

[4] Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Heejung Park. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):1992–2001, 2013.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[6] Carmen K Vaca, Amin Mantrach, Alejandro Jaimes, and Marco Saerens. A time-based collective factorization for topic discovery and monitoring in news. In *Proceedings of the 23rd international conference on World wide web*, pages 527–538. ACM, 2014.

[7] RKD. explore.rkd.nl - RKDartists&, April 2015. `https://rkd.nl/en/info/rkdartists`.

[8] Shaoxu Song and Lei Chen. Similarity joins of text with incomplete information formats. In *Advances in Databases: Concepts, Systems and Applications*, pages 313–324. Springer, 2007.

[9] The Apache Software Foundation. Apache Lucene http://lucene.apache.org/, October 2015. `http://lucene.apache.org/`.

[10] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009.

[11] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[12] Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 153–162. ACM, 2010.