Unsupervised Creation of Semantic Graphs to Navigate Intangible Cultural Heritage Using Transformers

Maria Teresa Artese[0000-0001-8453-0807], Isabella Gagliardi[0000-0002-3706-0993]

National Research Council, Institute of Applied Mathematics and Information Technologies "Enrico Magenes" (CNR-IMATI), 12, Alfonso Corti Str., 20133, Milan, Italy
artese@mi.imati.cnr.it, gagliardi@mi.imati.cnr.it

Abstract. The success of cultural heritage archives depends on their ease of use and navigation. A simple and intuitive user interface can improve accessibility and inclusiveness. This paper presents a new approach to create semantic graphs from archive contents. The resulting graphical representation allows users to explore and navigate the data in new and intuitive ways.

Keywords: Semantic Graph, QueryLab Intangible Cultural Heritage Portal, Clustering, Data Visualization, Bert-like Transformers.

1 Introduction

In today's digital age, cultural heritage institutions are increasingly creating digital archives of their collections to make them more accessible to the public. This trend has been made even more evident and necessary because of the pandemic. However, the success of these digital archives often depends on their ease of use and navigation. Users must be able to find what they are looking for quickly and easily, without getting lost in a maze of menus and options.

A simple and intuitive user interface can make a difference in the way users interact with digital archives. A logical hierarchy of information and an intuitive design can make navigating through an archive a flowing and enjoyable experience. In addition, simplicity can also increase inclusiveness and accessibility. People with different levels of digital literacy and cognitive abilities should be able to use an archive with ease. Designing with simplicity in mind can also ensure that the archive is accessible to people with disabilities, such as those who are visually impaired or have reduced mobility.

In this paper, we present a new approach to visualizing information in an Intangible Cultural Heritage (ICH) archive and navigating it in a simple and intuitive way. The paper presents a prototype for the unsupervised creation of semantic graphs that relies on the pre-trained language models, using them both to cluster the data and to create the similarity matrices. A semantic graph is a graphical representation of words and their relationships to each other. It can be used to show semantic similarity between words or to explore connections between related concepts.

Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings. Vol. 13, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2023. ISSN: 1314-4006, eISSN: 2535-0366

The main features that make this approach both effective in retrieving information and useful and engaging for users are:

- once some hyperparameters and transformers models are optimized, the pipeline is completely unsupervised;
- in the case of very large archives, semantic graphs are layered as the clustering levels can increase, allowing users to explore a few dozen items at a time, letting them get involved in deep exploration;
- the use of language pre-trained models allows this approach to be used even for a few hundred items;
- by using pre-trained models for multilingual texts, archives with documents in English, French, Italian, ... can also be handled efficiently and effectively.

The approach proposed in the article uses several state-of-the-art tools, such as the UMAP dimensional reduction tool or transformers such as BERT, integrating them synergistically to achieve a user-friendly tool that can be adapted to many contexts and is valuable in providing an overview of the contents of archives.

The pipeline has been tested using data from the QueryLab portal, a framework for the management of tangible and intangible cultural heritage (Artese & Gagliardi, 2022).

The article is structured as follows: after a brief survey of related works on graphs used for cultural heritage, the pipeline is described with some technical notes. Following this, the experiment is presented with some results of both clustering and semantic graphs. Conclusions and future works follow.

2 Related Works

In this paper, we present an approach to unsupervised semantic graph creation. They are meant, in this paper, as knowledge graphs constructed on the basis of the semantic similarity of graph nodes.

Many articles and special issues in the literature are devoted to knowledge graphs. For example, the journal 'Heritage' in 2021 published a special issue on "Knowledge Graphs for Cultural Heritage".

Much research is related to the use of linked open data and ontologies for Knowledge Graph (KG) creation. Ryen et al. (Ryen, Soylu, & Roman, 2022) focus on research related to knowledge graph creation and publication within the Semantic Web domain.

Another example is Arco (Carriero, et al., 2019) which allows the construction of knowledge graphs based on Linked Open Data (LOD).

Many research projects are related to KG. An example is CHLOD, a project that focuses on creating a linked data infrastructure for cultural heritage institutions in the Netherlands. The project involves the development of ontologies, data models, and data mapping tools.

Semantic Web for Cultural Heritage-SW4CH is a project and series of workshops (SWODCH, 2022) that aim to use Semantic Web technologies to provide access to cultural heritage data. The project involves the development of ontologies, vocabularies, and tools for publishing and querying cultural heritage data.

To the best of our knowledge, this is one of the first experiments to create semantic graphs using transformers.

3 Our Approach

The goal of this work is to define a pipeline that enables the unsupervised creation of a semantic graph for content navigation from a textual dataset. The innovative elements involved in the pipeline include a series of interconnected processes and techniques that enable the automated generation of meaningful relationships and connections between different pieces of data. These processes typically involve several stages, including dimension reductions, data clustering, choice of pre-trained models and ways to define a single vector per item, and graph construction.

3.1 Innovative Elements of the Pipeline

UMAP (Uniform Manifold Approximation and Projection) (McInnes, Healy, & Melville, 2018) is a dimensionality reduction technique used in machine learning and data analysis, in competition with other popular dimension reduction methods such as PCA and t-SNE. UMAP uses manifold learning for mapping high-dimensional data to a lower-dimensional space while preserving the local structure of the data. UMAP uses a combination of topological and geometric techniques, handling non-linear relationships in the data, with the ability to handle high-dimensional data well. The algorithm requires two main parameters, the number of dimensions and the distance metric used.

HDBSCAN (McInnes, Healy, & Astels, 2017) is a clustering algorithm that groups similar data points together, even if the clusters have different shapes, densities, and sizes. It does this by identifying areas of high density within the data set and grouping points together based on their proximity to these high-density areas.

Unlike other clustering algorithms, HDBSCAN does not require the user to specify the number of clusters or the size of the neighbourhood to be searched. It automatically detects the number of clusters and the shape of clusters. Additionally, HDBSCAN can identify points that do not belong to any cluster as noise. This makes it useful for datasets with a high degree of density variation. The algorithm has a couple of parameters that can be adjusted depending on the dataset, such as the minimum cluster size and the minimum samples to form a dense region.

An advantage of using UMAP and HDBSCAN over other clustering algorithms is that they can handle nonlinear relationships between data points. This is especially important in natural language processing, where the relationships between words and sentences can be very complex and nonlinear.

Transformers and pre-trained language models: Transformers are a type of neural network that has greatly improved natural language processing (Wolf, et al., 2020).

Unlike traditional methods, which relied on hand-created features and statistical models, transformers use self-attention mechanisms to focus on different parts of the input and capture long-term dependencies. BERT is one of the most popular pre-training language models and one of the earliest (Devlin et al., 2019). There are several pre-trained language models, such as:

- GPT (Generative Pre-Training Transformer) is a set of pre-trained language models developed by OpenAI, which uses a transformer architecture similar to that of BERT;
- RoBERTa is a pre-trained language model developed by Facebook AI, similar to BERT but trained using a larger and more diverse corpus of data.
- ALBERT (A Lite BERT) is a smaller and more efficient version of BERT, developed by Google.

These models use a transformer-based neural network architecture and are trained on large amounts of unlabeled text data to learn a deep understanding of natural language. One of the main features of BERT (and its competitors trained by Microsoft, Facebook, OpenAI or HuggingFace) is the ability to handle bidirectional language modeling, that is, to analyze and understand the context of a word by looking both forward and backward in the sentence. This provides a better understanding of the relationships between words and the nuances of language. In this paper, we test different pre-trained models as described below.

3.2 The Pipeline

Dataset preparation includes all those operations that are performed on the data, in a manner preliminary to their actual use on the algorithms. This stage may require removing stopwords, replacing accents, and reducing to normal form, etc. In this case, preprocessing may also include the automatic identification of tags or keywords. The results of this task lead to the identification of items of interest to be used to test algorithms and pipelines.

If the number of elements is too large for the creation of a single semantic graph, a preliminary clustering step is required. In this work, clustering is performed using UMAP and HDBSCAN on the vectors obtained by fine-tuned transformers. In order to evaluate which model is optimal for the specific purpose, several pre-trained models were tested and then subsequently fine-tuned on the items of interest. Additionally, to obtain better results, specific to each dataset, this task includes choosing the hyperparameters of UMAP and HDBSCAN. The result of this task is the clustering of the items: both the centroids and the set of items in each cluster.

Once the items were clustered (if necessary), again individual items were vectorized using BERT-like pre-trained transformers, fine-tuned and then averaged (or something similar) to obtain a single vector per item. A similarity matrix is obtained by applying semantic similarity to these vectors. Based on this similarity matrix, a semantic graph is produced by connecting the k most similar items. In the experiment reported here, the values of k range from 1 to 4. The similarity between the elements is calculated using the cosine of the obtained vectors. The result of this task is the semantic graph

created in an unsupervised manner. If the elements have been clustered, n+1 graphs are produced, corresponding to the n clusters plus the centroid graph.

Since the purpose of the project is to create tools for simplified navigation in archives, the final task involves a preliminary evaluation of semantic graphs by experts in the field and/or Web users to assess their real usability.

Table 1: Steps of the proposed approach

Task 1: Dataset Preparation
- Preprocessing (possibly strip stopwords, accents, …)
- Process data to extract items to be used
- Output: items of interest
Task 2: Items clustering
- Choice of transformers and pre-trained models
- Fine tuning of pre-trained Bert-like models to obtain the vectors
- Choice of hyperparameters for UMAP and HDBSCAN
- Output: centroids of clustered items, and elements of each cluster
Task 3 Semantic graph creation
- Choice of transformers and pre-trained models, both on raw data and on clustered items and fine tuning
- Creation of similarity matrix using [AVG] or [CLS] tokens
- Output: Semantic graphs con k most similar items, with k=1…4
- Preliminary evaluation of the results with domain experts and web users

4 The Experimentation

4.1 Dataset: QueryLab Platform

The defined pipeline has been tested on data from QueryLab, a portal specifically designed to manage intangible cultural heritage data. The portal handles two types of data: those data stored locally, and data queried on the fly from remote repositories via REST API web services. In this experiment, we used locally stored data and various visualization techniques were tested on this data. By experimenting with different visualizations on locally stored data, we can gain insights into the patterns and relationships of the data and develop effective ways to communicate this information to experts and users.

The experimental datasets used in this study can be classified into the following types:

Tags: Expert-defined tags associated with the records in the archive. Two different datasets were used in this study: one from the Ethnography and Social History archive (Artese & Gagliardi, 2017), where tags were defined by expert ethnographers, and the other from UNESCO for managing intangible cultural heritage assets (Unesco ICH, 2023). The former tags were originally defined in Italian and then translated into Eng-

lish, French, and German. The terms used in the latter dataset can be simple or compound and provide valuable insights into the nature and characteristics of the cultural assets.

Title: The name of the intangible cultural asset, which may be a simple name or a name with a short description.

Description: A description of an intangible cultural asset, which could be a dance, ritual, or knowledge, among other things. These descriptions are created to safeguard the asset and preserve its knowledge for future generations. The length of the descriptions varies from a few words to several paragraphs.

Rake/Textrank Keywords: Additionally, using the Rake (Rose, Engel, Cramer, & Cowley, 2010) or Textrank (Mihalcea & Tarau, 2004, July) algorithm, simple or compound words were in an automatic and unsupervised manner extracted from these descriptions, forming another dataset of valuable terms for analysis and modeling purposes.

The main problems with this dataset are related to the fact that we dealt with real data related to intangible cultural heritage, which is created by communities with the purpose of preserving and transmitting their intangible cultural heritage, which en-compasses traditions, sayings, dialect expressions, local object names, masks, and other elements. However, due to the unique and specialized nature of this cultural heritage, the terms and concepts associated with it are not commonly found in pretrained models.

4.2 Clustering

In this paper, we use transformers together with UMAP and HDBSCAN to cluster data. BERT and other Bert-like models have been used (and fine-tuned) to transform text data into high-dimensional vectors that capture semantic meaning. We then apply UMAP to the vectors to obtain a lower dimension space which constitutes the input of HDBSCAN to group similar text. This approach has proven to be very effective in this context. Several tests were done to evaluate the best values of the hyperparameters for UMAP and HDBSCAN and the pre-trained models in relation to different datasets.

The only parameter we tested for UMAP is n_neighbor. This parameter controls how UMAP balances local versus global data structure. The default in the python implementation is 15. Our tests evaluated the following values: 20,15,10,5.

For HDBSCAN different values were tested for min_cluster_size and min_samples. Min_cluster_size is intuitively set to the smallest cluster size that you want to consider a cluster. It should be considered together with min_samples, which somehow provide a measure of how conservative the clustering is. The values tested for min_cluster_size (HDBSCAN) were 15,10,5, the values for min_samples (HDBSCAN) 15,10,5,1. The pre-trained transformers models tested were.

- BERT Base: This is the original pre-trained BERT model released by Google. It has 12 transformer layers and is trained on a large corpus of text data from Wikipedia and the Book Corpus dataset.

- BERT Large: This is a larger version of the BERT model with 24 transformer layers. It is also trained on a large corpus of text data and has been shown to perform better than BERT Base on some tasks.
- MiniLM-L6-v2: This is a smaller version of the BERT model developed by Microsoft. It has only 6 transformer layers and is trained on a subset of the data used to train BERT Base.
- Bert-base-Wikipedia-sections-mean-tokens: This is a pre-trained BERT model released by the Hugging Face team. It is trained on a large corpus of text data from Wikipedia and uses a mean pooling strategy to create a fixed-length representation of the input text.

The problem of terms not present in the pre-trained model is overcome by the use of Bert-like transformers and their tokenizers. This solution has had much better results than the use of single word level tokenizers or models with Word2vec or GloVe.

Results for MiniLM-L6-v2 on titles are reported here, limiting to the first 500 records. Table 1 reports the significant tested values of n_neighbors, min_cluster_size and min_samples allowing for a number of clusters ranging from 11 to 45. Obviously as the three parameters decrease, the number of clusters increases. The following values n_neighbors=20, min_cluster_size=5 and min_samples =5 with a number of clusters equal to 17 were used in the graphs in Paragraph 4.3.

Table 1. Number of clusters obtained as the values of n_neighbors, min_cluster_size and min_samples vary.

N neighbors	min_cluster_size	min_samples	Number cluster
20	15	5	11
20	15	1	12
20	10	5	14
20	10	1	15
20	5	5	17
20	5	1	30
15	10	1	18
15	5	1	34
10	10	5	13
10	10	1	20
10	5	5	18
10	5	1	33
5	10	1	21
5	5	5	26
5	5	1	45

4.3 Semantic Graph Creation

The first step in creating semantic graphs is to generate similarity matrices. In this experiment, similarity matrices were created between words or phrases using BERT-like transformers. The similarity matrix requires only one vector for each element. In the

case of word embeddings such as Word2Vec or GloVe, the vectors are usually averaged, possibly weighted according to the frequency or importance of the word. When using a BERT-like model, in addition to average [AVG] token, the [CLS] tokens are used, representing the entire sentences. The input text is preprocessed by tokenizing and encoding it into numerical vectors, which are then fed into the transformer model.

The model generates contextualized embeddings for each token in the input text. To construct a similarity matrix, the [CLS] or [AVG] tokens are compared pairwise using a distance metric like cosine similarity. The resulting scores represent the similarity relationships between the [CLS] or [AVG] tokens of the input text and can be used to create the matrix. The graphs shown in the figures were obtained using [CLS] tokens to generate a single vector for each title.

The similarity metric was then used to construct the semantic graphs, taking, for each item, the k most similar items with k ranging from 1 to 4. With k=1 sometimes disconnected graphs are formed, in which some nodes are connected to each other but not to others. With k>=2 graphs are created that are fully connected, with k =4 practically everything connects with everything, making the graphs difficult to read.

Figure 1 shows the semantic graph of the entire dataset, represented by 17 nodes. In case of clustering, the name of each node is taken from the most similar element in the corresponding cluster. Figure 2 represents the semantic graph for the Cegni carnival with k=2.

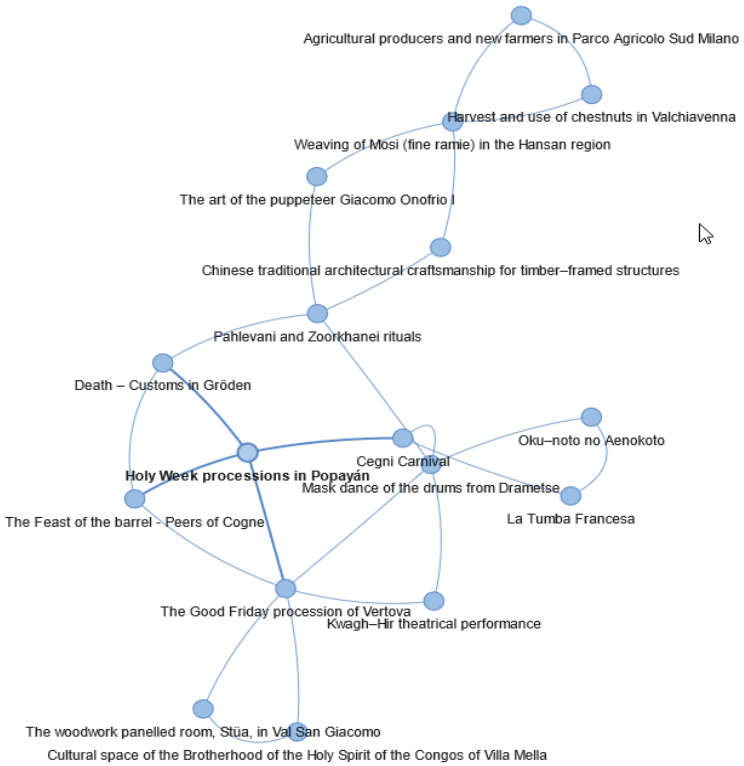

Fig. 1. Semantic graph of the whole dataset represented by 17 nodes.

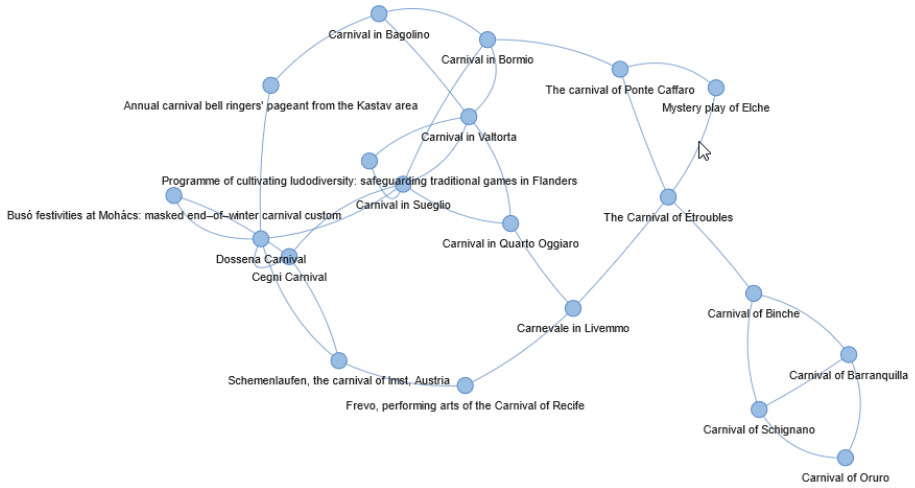

Fig. 2. Semantic graph of the Cegni Carnival, with k=2.

The evaluation of this approach can be assessed at different levels of granularity and from different perspectives. At this prototype stage, we were mainly interested in two aspects: 1) whether the clustering and similarity matrix was able to extract the significant elements, and 2) whether users found browsing the archive via graph interesting and useful. For both aspects, an initial qualitative assessment provided a positive response. We gathered feedback from (intangible) heritage experts and web users. Their feedback showed that the graph visualisation's simplicity and usability were highly valued. However, we also found that low-level clusters contained elements that were not closely related, or that some related elements were spread across multiple clusters, in the case of too few or too many clusters. These problems need to be addressed to improve the visualisation.

5 Conclusions and Future Works

In this paper, we presented a prototype for creating semantic graphs in an unsupervised manner. The ultimate goal is to integrate a new way of searching and browsing data into cultural heritage archives, using semantic graphs as a layered map with different granularity. Users are presented with the content of the archive in a powerful visual representation, enabling them to explore and navigate the data in engaging and intuitive ways. The nodes in the graph represent words, concepts, or items, and the edges represent the relationships between them.

The steps of the pipeline require, after an initial pre-processing step, to create a similarity matrix using BERT transformer tokens, to navigate the data by moving from one topic to another, having a complete, high-level view of the content of the archive. If the data are big, and do not allow the creation of a single graph (because it is too dense), one (or more) preliminary clustering steps are performed. The pipeline was tested using data from QueryLab, a growing portal of cultural heritage, tangible and intangible.

Users, experts in the field, and Web users gave an initial, qualitative, positive assessment of the prototype, judging this overall view of the archive and each node positively. More quantitative evaluation of the whole pipeline is being studied.

Future developments will consider providing tools for traversing the graph, to go from one point to another one, chosen by the users. Moreover, fish-eye views could overcome the problem of overly dense graphs. Experimentations on other (multilingual) datasets are planned to evaluate the effectiveness of the prototype.

References

Artese, M. T., & Gagliardi, I. (2017). Inventorying intangible cultural heritage on the web: a life-cycle approach. *International Journal of Intangible Heritage*, *12*, 112-138.

Artese, M. T., & Gagliardi, I. (2022). Integrating, Indexing and Querying the Tangible and Intangible Cultural Heritage Available Online: The QueryLab Portal. *Information, 13*(5), 260.

Carriero, V. A., Gangemi, A., Mancinelli, M. L., Marinucci, L., Nuzzolese, A. G., Presutti, V., & Veninata, C. (2019). ArCo: The Italian Cultural Heritage Knowledge Graph. In *The* Semantic *Web–ISWC 2019: 18th International Semantic Web Conference, October 26-30, 2019. Proceedings, Part II 18*0 (pp. 36-52). Springer International Publishing.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186).

McInnes, L., Healy, J., & Astels, S. (2017). Hdbscan: Hierarchical density based clustering. *J. Open Source Softw, 2*(11), 205.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. https://doi.org/10.48550/arXiv.1802.03426

Mihalcea, R., & Tarau, P. (2004, July). Textrank: Bringing order into text. In *Proceedings of the 2004* conference *on empirical methods in natural language processing* (pp. 404-411).

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1-20.

Ryen, V., Soylu, A., & Roman, D. (2022). Building semantic knowledge graphs from (semi-) structured data: a review. *Future Internet, 14*(5), 129.

SWODCH. (2022). Retrieved May 2023, from Semantic web and ontology Design for Cultural Heritage: https://swodch2022.inf.unibz.it/

Unesco ICH. (2023, May). Retrieved from Browse the Lists of Intangible Cultural Heritage and the Register of good safeguarding practices: https://ich.unesco.org/en/lists

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., . . . Xu, C. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38-45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Received: March 15, 2023
Reviewed: April 06, 2023
Finally Accepted: May 18, 2023