Intuitive Semantic Graph Tool for Enhanced Archive Exploration

Isabella Gagliardi
gagliardi@mi.imati.cnr.it
IMATI MI
CNR
Milan, Italy

Maria Teresa Artese
artese@mi.imati.cnr.it
IMATI MI
CNR
Milan, Italy

ABSTRACT

The paper introduces a new method for visualizing and navigating information in a cultural heritage archive in a simple and intuitive way. The proposed approach employs pre-trained language models to cluster data and create semantic graphs. The creation of multilayer maps enables deep exploration of archives with large datasets, while the ability to handle multilingual datasets makes it suitable for archives with documents in various languages. These features combine to provide a user-friendly tool that can be adapted to different contexts and provides an overview of archive contents, to allow even non expert users to successfully query the archive.

CCS CONCEPTS

• **Information systems → Digital libraries and archives**; • **Applied computing → Arts and humanities**.

KEYWORDS

Bert, pre-trained language models, transformers, clustering, data visualization, archives, non-expert users

ACM Reference Format:
Isabella Gagliardi and Maria Teresa Artese. 2023. Intuitive Semantic Graph Tool for Enhanced Archive Exploration. In *34th ACM Conference on Hypertext and Social Media (HT '23), September 4–8, 2023, Rome, Italy.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3603163.3609069

1 INTRODUCTION

In the digital age, cultural institutions are increasingly creating digital archives of their collections to improve public accessibility. The ongoing pandemic has further emphasized the need for such archives. However, the success of digital archives depends on their usability and ease of navigation. Users must be able to find easily and quickly what they need without getting lost in a maze of options and menus.

The paper is focused on semantic graphs, meant as knowledge graphs constructed based on the semantic similarity of graph nodes. Much research is related to the use of linked open data and ontologies for KG creation. In [10] authors focus on research related to knowledge graph creation and publication within the Semantic Web domain. Arco [4] allows the construction of knowledge graphs based on LOD. To the best of our knowledge, this is one of the first experiments to create semantic graphs using pre-trained transformer-based language models.

The paper presents a pipeline for the unsupervised creation of multilayer semantic graphs using pre-trained language models to cluster data and create semantic graphs, which are graphical representations of words and their relationships. These graphs can be thought of as hypertext where each node is linked to related items, allowing for easy navigation and exploration of archive contents. The approach integrates several state-of-the-art tools and models, such as pre-trained language models based on transformer architecture, UMAP for dimension reduction, and HDBSCAN for clustering, to create an intuitive and adaptable tool suitable for a variety of contexts. The defining characteristics of this approach are its unsupervised nature and its ability to create multi-layer graphs. It is also capable of handling multilingual datasets, as demonstrated by its successful testing on a portal for managing intangible cultural heritage data in multiple languages [2, 3].

2 THE APPROACH

In the poster, we will present an innovative way of presenting the contents of an archive, as a multilevel graph. Each node of the top-level graph represents one or more sets of elements, grouped through clustering algorithms. The lowest level graphs are 1 to 1 with the documents in the archive. If necessary, the clustering step is iterated, creating intermediate graphs, until the number of nodes in each graph is below a certain threshold. Clustering is performed on vectorized elements, using pre-trained language models based on neural networks and later fine-tuned. In case the level of graphs is higher than two, it is necessary to create one vector for all its underlying elements. Several ways of combining have been investigated, described in more details in the poster. The pipeline depicted in Figure 1 includes several steps, starting with a standard preprocessing stage that also extracts candidate words [8, 9] to describe the essential content of the item, as an additional way of processing the data:

Clustering: BERT and other transformers models [5, 11] have been used (and tuned) to transform text data into high-dimensional vectors that capture semantic meaning. We then apply UMAP [7] to the vectors to obtain a lower-dimensional space that is the input to HDBSCAN [6] for clustering similar texts. This approach has been shown to be very effective in this context. Several tests have been performed to evaluate the best parameter values of UMAP and HDBSCAN and pre-trained language models, e.g., BERT, RoBERTa and ALBERT [1]. Clustering is iterated until the number of elements is below a certain threshold, identified as 30, with the essential contribution of target users and domain experts.

HT '23, September 4–8, 2023, Rome, Italy

ACM ISBN 979-8-4007-0232-7/23/09.
https://doi.org/10.1145/3603163.3609069

Semantic graphs: The first step is the creation of the similarity matrices. To create these matrices, we used pre-trained transformers: each element, corresponding to a centroid or a document, is represented by a single vector, obtained by preprocessing the input text and inserting it into the transform model. To represent a document, [CLS] or [AVG] tokens are used, while for the centroid, in this experiment, the averaged [CLS] tokens from the lower-level documents were used. These tokens are compared pairwise using a distance metric like cosine similarity to create the similarity matrix. A qualitative evaluation has been performed at this prototype stage, focused on two aspects: 1) whether the clustering and similarity matrices could identify significant elements and 2) whether users found browsing the archive via graph intuitive, useful and helpful. An initial qualitative assessment yielded a positive response for both aspects.

3 RESULTS AND CONCLUSIONS

The goal of this experiment has been to create a way to visualize the entire contents of an archive through graphs of increasing detail. The poster will report the results of the experiments testing different values of the hyperparameters for clustering and pre-trained models on both monolingual Italian /English /French/ German and mixed datasets. In the presented experiments, textual metadata from an intangible cultural heritage inventory was utilized. This metadata includes titles, keywords, and words extracted from the descriptions. Typically, the data used is in either Italian or English. However, if data in the desired language is not available, the existing metadata is utilized. Since low-level graphs are pointers to archive documents, datasets created from different metadata were separately tested. Qualitative results have demonstrated the effectiveness of the pipeline in generating meaningful semantic graphs and exploring different visualization techniques that can be used to communicate the relationships between data to experts and users. We collected feedback from experts in (intangible) heritage and web users, which indicated that the simplicity and usability of the graph visualization were highly appreciated. They have highlighted that graphs with a number of nodes greater than 30 make understanding and navigating the graph difficult. We also discovered that low-level clusters included unrelated elements or that some related elements were scattered across multiple clusters when there were either too few or too many clusters. These issues need to be addressed in order to enhance the visualization. Future activities will focus on quantitative evaluation of the results, tools to suggest to users how to traverse the graphs and integration of fish-eye views to overcome the problem of overly dense graphs. Additional materials can be accessed at https://arm.mi.imati.cnr.it/papers/ht2023.

REFERENCES

[1] [n. d.]. Hugging face models. Retrieved 2023-03-29 from https://huggingface.co/models

[2] Maria Teresa Artese and Isabella Gagliardi. 2020. Language independent searching tools for cultural heritage on the QueryLab platform. In *Euro-Mediterranean Conference*. Springer, 657–665.

[3] Maria Teresa Artese and Isabella Gagliardi. 2022. Integrating, Indexing and Querying the Tangible and Intangible Cultural Heritage Available Online: The QueryLab Portal. *Information* 13, 5 (2022), 260.

[4] Valentina Anita Carriero, Aldo Gangemi, Maria Letizia Mancinelli, Ludovica Marinucci, Andrea Giovanni Nuzzolese, Valentina Presutti, and Chiara Veninata. 2019. ArCo: The Italian cultural heritage knowledge graph. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18*. Springer, 36–52.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* 2, 11 (2017), 205.

[7] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[8] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*. 404–411.

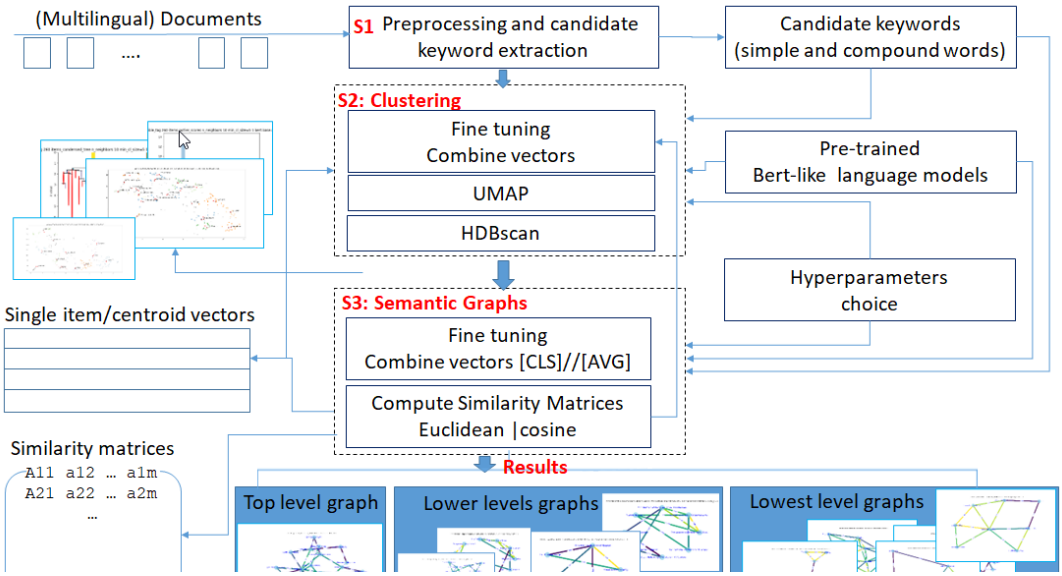

Figure 1: The pipeline for the unsupervised creation of multilayer semantic graphs.

[9] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory* (2010), 1–20.

[10] Vetle Ryen, Ahmet Soylu, and Dumitru Roman. 2022. Building semantic knowledge graphs from (semi-) structured data: a review. *Future Internet* 14, 5 (2022), 129.

[11] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 38–45.