Lifting Heritage Data Integration to the Next Level with Heterogeneous Information Networks

Introduction

Networks can model interconnected real-world data from any domain by connecting node objects by an edge if they are related, allowing additional perspectives on the modeled information. Modeling heterogeneous, information-rich data in a network without type information may lead to information loss. Heterogeneous information networks (HINs) preserve heterogeneous information of the input data by considering different types of objects and relationships, leading to richer semantics and more complex structural information (Yu and Li, 2023). Analyzing these information-rich, diverse graphs enables discovering new knowledge (Sun and Han, 2012).

This work introduces a novel approach utilizing HINs to revolutionize the management and analysis of cultural heritage data, focusing on the Poseidon Community Archive[1] as a practical use case. Numerous methods exist for presenting identical information through HINs, thereby facilitating the requirement for interdisciplinary exchange regarding modeling decisions.

Methodology

HINs are closely related to the concept of knowledge graphs (e.g., Hogan et al., 2021), networks used to model semantic relations. Sun et al. (2022) state that knowledge graphs are special cases of HINs with a richer network schema and the option to represent a hierarchical structure. An HIN is a directed graph consisting of a set of typed nodes, representing objects, and a connecting set of typed edges, representing relationships (Sun et al., 2011). The meta-level structure of the HIN (network schema) contains all possible object types connected by the respective relationship types. It serves as a template for networks belonging to the represented domain.

Network Generation

HINs are often created by informatics experts without specific domain knowledge. However, the different object types need to be discovered from (tabular) data files that do not explicitly represent them. Furthermore, the discovery of relationship types can be challenging, but it is necessary to obtain an information-rich network structure.

The conventional data integration workflow, consisting of schema matching, duplicate detection and data fusion (Bleiholder & Naumann, 2009, p.2), is extended to the stepwise process depicted in Figure 1.

[1] https://github.com/poseidon-framework/community-archive/tree/master

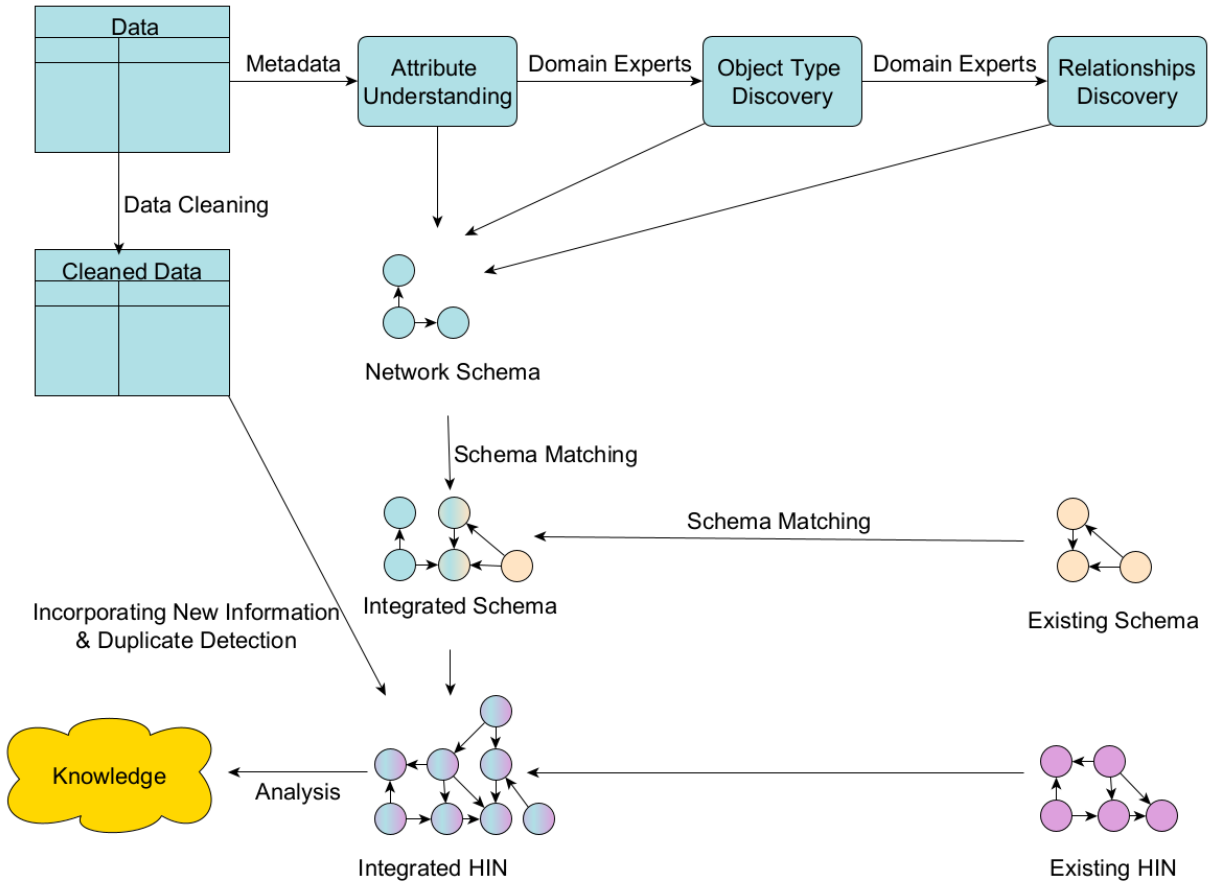

Figure 1: Stepwise HIN Creation Workflow (one Iteration).

Initially, tabular data is cleaned to ensure accuracy and consistency without giving any details of the challenges occurring. Metadata is (created and) utilized to understand the attributes, which are mapped to object and relationship types with the assistance of domain experts. This leads to the formation of a new network schema, which is matched against any existing schema to obtain a joint schema containing all available information. Finally, the integrated HIN is created using the old HIN and the new pre-processed data. Elements of cross-domain data fusion as described by Zheng (2015) are incorporated in this process by allowing the combination of knowledge from different domains to create an information-rich, diverse HIN.

Network Analysis

There are various ways to gain knowledge from HINs. Sun et al. (2009) proposed RankClus**,** a ranking-based clustering algorithm, to detect communities of similar objects in HINs. Clustering discovers communities of objects in the network. Additionally, PathSim (Sun et al., 2011) captures network-based similarity between objects in HINs. These techniques will be adapted and applied to identify similar sites, finds, or samples based on their connectivity along semantic paths in the network. These network paths define similarity between objects by telling a story using the network semantics defined by node and edge types.

Data

Poseidon (Schmid et al., 2024) is a framework designed to work with ancient DNA data in a FAIR and open way. It aims to organize genotype data, along with relevant metadata and archaeological context data, in a structured yet flexible format. The Max Planck Institute for Evolutionary Anthropology and researchers from around the world maintain public data archives. The Poseidon Community Archive (PCA) is one of these archives containing publication-wise genotype data. The PCA primarily collects author-submitted data, including the exact genotype data analyzed and additional spatio-temporal context information provided by domain experts. The PCA is a significant resource in the field of cultural heritage, emphasizing the reproducibility of computational research by storing the genotype data used in specific academic papers. An HIN is created from the PCA data and context information.

Results & Conclusion

The workflow illustrated in Figure 1 underscores the necessity of close interdisciplinary collaboration to accurately represent cultural heritage data using HINs. Implicit assumptions and tacit knowledge are not clear to all involved parties, and even well-documented metadata is not enough to create an HIN solely based on it. Exchange with domain experts is necessary to successfully represent meaningful information.

Furthermore, ensuring data quality is required to create a useful HIN since the same information has to be modeled the exact same way to refer the same object in the network. These shared objects are essential for combining different data sets and performing successful network-based analysis. Data cleaning during the network creation is also helpful for domain experts and data owners, as unidentified data quality issues may be revealed, e.g. automatically incrementing spatial coordinates for multiple occurrences of an archaeological site caused by the replication of a row in Microsoft Excel.

Discussion

HINs enable new possibilities for the management and analysis of heterogeneous data sets in the cultural heritage domain, leading to potential new insights. The creation of such a network requires a close cooperation of domain experts and technical experts to ensure that both parties share the same assumptions.

HINs also enable cross-domain fusion and enrichment of the modeled knowledge by processing multiple data sets or even combining existing HINs. This may lead to new analysis results due to the enriched data available for analysis. However, effort is required to create accurate and consistent networks. Enrichment of the Poseidon HIN as well as initial network analysis results, such as similarity or clustering results, will be presented at the conference.

References

Bleiholder, J., and Naumann, F. (2009). '*Data Fusion*', ACM Computing Surveys, 41, 1, Article 1, 41 pages. https://doi.org/10.1145/1456650.1456651.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutiérrez, C., Kirrane, S., Neumaier, S., Polleres, A., Rashid, S.M., Rula, A., Schmelzeisen, D., Sequeda, J., Staab, S., and Zimmermann, A. (2021). '*Knowledge Graphs*', Synthesis Lectures on Data, Semantics, and Knowledge, Morgan & Claypool Publishers. https://doi.org/10.1145/3447772.

Schmid, C., Ghalichi, A., Lamnidis, T. C., Mudiyanselage, D. B. A., Haak, W., and Schiffels, S. (2024). *'Poseidon – A framework for archaeogenetic human genotype data management*', bioRxiv. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2024.04.12.589180.

Sun, Y., and Han, J. (2012). '*Mining Heterogeneous Information Networks: Principles and Methodologies'*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00433ED1V01Y201207DMK005.

Sun, Y., Han, J., Yan, X., Yu, P., and Wu, T. (2011). '*PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks'*, PVLDB, 4, pp. 992–1003. https://doi.org/10.14778/3402707.3402736.

Sun, Y., Han, J., Yan, X., Yu, P. S., and Wu, T. (2022). *'Heterogeneous information networks: the past, the present, and the future'*. Proceedings of the VLDB Endowment, 15(12), 3807–3811. https://doi.org/10.14778/3554821.3554901.

Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., and Wu, T. (2009). '*RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis*', In Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology (EDBT '09), ACM, New York, NY, USA, pp. 565–576. https://doi.org/10.1145/1516360.1516426.

Yu, J., and Li, X. (2023). *'Heterogeneous graph contrastive learning with meta-path contexts and weighted negative samples'*. In Proceedings of the 2023 SIAM International Conference on Data Mining (SDM), pp. 37-45.

Zheng, Y. (2015). '*Methodologies for Cross-Domain Data Fusion: An Overview*', IEEE Transactions on Big Data, 1(1), pp. 16–34. https://doi.org/10.1109/TBDATA.2015.2465959.