Building a Graph Database for Digital Humanities Scientists

Triet Ho Anh Doan[1 [0000−0002−7247−9108]], Péter Király[1 [0000−0002−8749−4597]], and Sven Bingert[1 [0000−0001−9547−1582]]

Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
Burckhardtweg 4, 37077 Göttingen
{triet.doan|peter.kiraly|sven.bingert}@gwdg.de
https://www.gwdg.de

Abstract

Graph database has developed rapidly and plays an important role in research nowadays. It helps scientists in various ways, e.g., finding related works, exploring works in a research area, or gaining knowledge from connections between different nodes. There are already some graph databases for research available on the Internet [1, 3, 4]. However, they do not meet the needs of Digital Humanity (DH) scientists, who mainly work with historical data. The first reason is that they only focus on modern data since it is easier to get them. The second problem comes from the way how documents are connected. In those systems, two documents are connected if they have direct citations, or share a large overlapping of citation lists. But the information of citation is not available in historical texts because they are not just research papers, but also other types of text, like newspapers or advertisements.

For those reasons, we create a graph database specifically for DH scientists. This database is part of MINE [2], a service that facilitates data acquisition and big data analysis. It allows users to search across multiple data sources at the same time and also offers a workspace as a service. At the core of MINE, there are two components: Elasticsearch, a search engine, and Neo4j, a graph database. For each data source, a specific component called Connector is built. A connector is responsible for getting data from the source, extracting metadata and full text, standardizing metadata, assigning Persistent Identifiers (PIDs) to them, and indexing them to Elasticsearch and Neo4j. Due to the fact that data comes in a variety of formats (METS-MODS, TEI header, and custom XML for metadata; PDF, TEI, custom XML, and even images for full text), having the full text extracted greatly assists scientists not only in full-text search but also in having the textual data ready at hand without having to repeat the full-text extraction process on their own. Among the two components, both metadata and full text are indexed to Elasticsearch, while Neo4j only contains metadata. At this point, a graph is created with four node types: Corpus, Document, Author, and Publisher. After this indexing step, data are annotated using Named-Entity Recognition and Topic Modelling approach. For each detected entity and topic, we also do Named-Entity Linking with Wikidata. If the entity is found, its ID (or

IDs, if it appeared as a label of multiple entities) and its English description is retrieved via Wikidata Search API [5]. In addition to the entity, its "ancestors", encoded with the property "subclass of" [6], are also retrieved (e.g., "physical sciences" is a parent of "chemistry" and "physics"). This extracted knowledge greatly enriches our graph. Document nodes in the graph are now not only connected via Corpus, Author or Publisher but also Entity node.

The MINE graph greatly enhances the use of the system. After finding a document through Elasticsearch, users can explore related information by traversing the graph. Users are also able to receive suggestions based on similar authors or topics. In the future, we are going to integrate more type of analyses and include the results in this graph as well. Furthermore, more and more data sources will be analyzed. These sources contain data in various languages, such as German, Middle French, Italian, and so on. Multilingualism is always a challenge in text analysis, but if we were able to overcome it, it would bring a great deal of benefit to the graph and to the end users as well.

References

1. Connected Papers: `https://www.connectedpapers.com/`, [Online; accessed 24-Feb-2023]
2. Doan, T.H.A., Király, P., Bingert, S.: Mine – workspace as a service for text analysis. In: Silvello, G., Corcho, O., Manghi, P., Di Nunzio, G.M., Golub, K., Ferro, N., Poggi, A. (eds.) Linking Theory and Practice of Digital Libraries. pp. 328–334. Springer International Publishing, Cham (2022)
3. Inciteful: `https://inciteful.xyz/`, [Online; accessed 24-Feb-2023]
4. Litmaps: `https://www.litmaps.com/`, [Online; accessed 24-Feb-2023]
5. Wikidata: MediaWiki API help. `https://www.wikidata.org/w/api.php?action=help&modules=wbsearchentities`, [Online; accessed 24-Feb-2023]
6. Wikidata: subclass of. `https://www.wikidata.org/wiki/Property:P279` (February 2023), [Online; accessed 24-Feb-2023]