Lab 2 - Graph Databases

Big Data Management for Data Science (2024)

Natalia Beltrán, Clarice Mottet

Under the supervision of Besim Bilalli

May 25, 2024

May 25, 2024

A Modeling, Loading, Evolving

A.1 Modeling

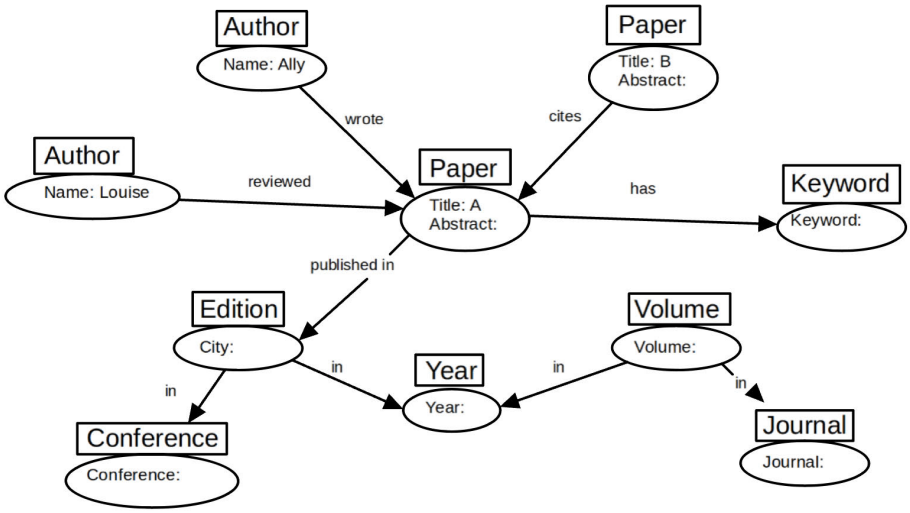

Figure 1: Graph Storage Model

With the goal of designing a graph to model the relationships and attributes of research papers, we determined that each paper should be represented as its own node. Next we decided what attributes of a paper to include directly in its node, and which to represent as separate nodes. The deciding factor for much of the graph was the publication year which we initially considered as an attribute within the paper node but chose to represent as a separate node. This approach better connects the year to specific journal volumes or conference editions, reflecting the framework of how papers are published. To distinguish between the non-overlapping characteristics of conferences and journals, aside from both publishing papers, each is modeled as distinct entities with their own nodes. Similarly, we represent authors with individual nodes to streamline connections to their papers and reviews. Keywords are also node-based, optimizing the search process for papers with specific keywords instead of traversing through multiple papers. Citation data is stored separately from a paper node to give ease of access to the retrieval of related citation information. Distinctly paper attributes like the title and abstract remain within the paper node, focusing on the direct aspects of the research itself.

A.2 Instantiating/Loading

The Cypher expressions for creating and instantiating the data for modelling can be found on the attached text file called 'Big Data - Instantiating NEO4J'. With the challenge of accessing research papers, we opted to instead develop an automated process utilizing Faker to create our own data to input into our graph database with the characteristics listed in Figure 1.

A.3 Evolving the graph

The proposed modifications as detailed in the assignment details were understood to be the following:

- Reviews should include the actual review and whether the reviewer accepted the results of the paper.
- Papers are only verified/published if at least three reviewers have accepted the results of a paper.
- Authors are affiliated with institutions or universities.

To incorporate the inclusion of actual reviews and the acceptance status of a review (represented as a zero or one), we remove all previous review connections and create new ones where the edge relationship between an author and paper includes this information. Next we add a variable to the papers published in volume and papers published in edition relationships that is a one if the paper has three or more reviews with an acceptance status of one and zero otherwise. Lastly, we create new nodes that store university information and add connections from author to university indicating an author is affiliated with a university.

B Querying

In order to save space in the report we have opted to provide all the queries for this section in the attached document "Big Data - B. Query NEO4J".

C Graph algorithms

For this section we looked at the Similarity Algorithm - Node similarity in particular focusing on identifying research papers that have similar citation patterns. For this section we utilized the GDS graph project syntax to create a graph projection that includes the 'paper' nodes and our $'p2p_cites'$ relationship. The cypher script for the projection can be found on the attached text file 'Big Data - C. Graph Algorithm NEO4J'.

For this section we are obtaining similarity scores between pairs of research papers. The results reveal how closely related different papers are based on their citation patterns. We had to create synthetic data for this project, as we had issues downloading real data from the web, because of this our results are very inconclusive. We see values of roughly 0.27 as the highest similarity between papers and go as low as 0.1. If the data was genuine then a high similarity would mean that the two papers indicate that they cite many of the same sources, showing that they cover similar topics. While lower score show very different topics overall. For our synthetic data is makes sense to have low/moderate low similarity as the data is random.