Semantic-Based Dataset Discovery in Data Lakes

Nassima KAID[1,2] and Zoubida KEDAD[1,2]

[1] Laboratoire DAVID UVSQ
[2] Université de Versailles Saint Quentin

Abstract. Data lakes have become essential components of modern data ecosystems, providing a flexible repository for storing raw data in various formats. However, the lack of a global schema and reliance on incomplete metadata pose significant challenges in discovering relevant data within these lakes. This paper addresses the limitations of current approaches to dataset discovery by proposing a semantic-based method. We introduce a two-phase approach: an offline phase that converts tables into structured semantic graphs and an online phase that leverages these graphs for efficient table search, focusing on union and join operations. Preliminary results demonstrate the potential of our approach, and our early findings suggest that incorporating semantics significantly improves the discovery process in data lakes.

Keywords: Dataset Discovery · Data Lakes · Semantic Search.

1 Motivation

Data lakes have emerged as an essential element in modern data ecosystems. Unlike traditional data warehouses, where data is extracted, transformed, cleaned, aggregated, and then made available to users, a data lake serves as a repository for raw data, which can be structured (tables), semi-structured (XML, JSON), or unstructured (text). In a context where businesses are confronted with large volumes of data, data lakes have proven to be an indispensable solution for leveraging data in various ways. Common use cases for data lake solutions include machine learning, visualizations and dashboards, enabling businesses to extract value regardless of the format of data source.

A key challenge within data lakes is discovering relevant data. Unlike data warehouses, data lakes lack a global schema, and the reliance on metadata is often insufficient due to the raw nature of the stored data. For instance, consider a data scientist seeking to enhance a model that analyzes the performance of competing firms. The scientist might have a table containing basic information and needs to identify relevant tables in the data lake for two potential scenarios: either enriching the model with additional features by finding tables that can be joined, or verifying the model's generalizability by locating similar tables for union operations. Formally, the objective is to identify tables within a data lake that can augment a query table, either through union (adding rows) or join (adding columns).

Various approaches in the literature address this challenge. For instance, some methods [1] focus on finding unionable tables by encoding them into embeddings and measuring table similarity. However, these embedding-based approaches often fail to capture semantic meaning, relying primarily on table values without revealing underlying context. Alternatively, other approaches [2] leverage knowledge bases, although these are often limited by their coverage. When it comes to identifying joinable tables, many existing methods [3][4] treat the task as an overlap set similarity search problem, relying primarily on table values. However, this approach may be inadequate in scenarios where data quality is compromised.

The search for these relevant tables involves uncovering meaningfull semantic relationships between tables. Our objective is to address the previous limitations and to enhance the discovery of relevant tables in data lakes by integrating semantics into the process. We seek to answer the following key questions: How can semantics be incorporated effectively? Where should it be used in the search process? And how can leveraging semantics improve the identification of tables for union or join operations? Our proposed approach explores these questions and offers insights to advance the data lake discovery techniques.

2 Our Proposal

Our approach aims to facilitate table discovery in a data lake using semantic representations. It is divided into two main phases: **Offline** and **Online**.

In the **Offline phase**, each table in the data lake is converted into a structured semantic representation in the form of a graph. For each *Table* T_i in the data lake, we generate a semantic graph G_i. In this graph, each *Node* v_{ij} represents a column in *Table* T_i and is annotated with a semantic type s_{ij} (such as "person" or "date of birth"). Each *Edge* $e_{ij,ik}$ between the nodes v_{ij} and v_{ik} represents a semantic relationship r_{ijk} between these columns. Formally, the semantic graph for *Table* T_i is defined as:

$$G_i = (V_i, E_i, S_i, R_i)$$

where $V_i = \{v_{i1}, v_{i2}, \ldots, v_{in}\}$ is the set of nodes (columns in the table), $E_i = \{e_{ij,ik}\}$ is the set of edges between nodes, $S_i = \{s_{i1}, s_{i2}, \ldots, s_{in}\}$ is the set of semantic types, and $R_i = \{r_{ijk}\}$ is the set of semantic relationships between columns.

To generate these semantic representations, we use a dedicated **Annotation Module**. This module combines **Language Models**, which involve fine-tuning an LM (preferably Transformer based) to annotate column types and relationships, with **Knowledge Graphs (KG)**, which are used to enrich the annotation process. The resulting annotations are indexed using an inverted index and the semantic graphs generated are then stored for use during the online phase.

In the **Online phase**, we use the generated semantic graphs to perform table searches for union and join operations. This phase comprises two distinct processes: Unionable table search and Joinable table search.

For the **Union Table Search**, we aim to identify tables in the data lake that can be unioned with a query table T_q. Two tables are considered unionable if they share the same semantic column types and similar relationships between these columns. The process involves generating a semantic representation of the query table T_q to create a semantic graph G_q, searching for candidate tables whose graphs G_i are similar to G_q using the inverted index to accelerate the search, and calculating the union score $U(T_q, T_i)$ between the query table T_q and a candidate table T_i by exploring several graph similarity measures.

For the **Join Table Search**, our goal is to find tables in the data lake that are most likely joinable with the query table T_q. Our approach, called "semantic join search", differs from existing methods that focus only on exact matches, it also explores all columns in the query table rather than limiting to a single predefined join key. The process involves searching for candidate columns in the data lake tables with matching semantic annotations for each column C_q in the query table T_q, and calculating the join score $J(C_q, C_i)$ between a query column C_q and a candidate column C_i based on value overlap.

With this framework in place, we have begun implementing our approach. We are developing the annotation module, fine-tuning the RoBERTa model for column type and relationship annotation. We have started the fine-tuning process, and we are working on the improvement of the accuracy of our result. We are also exploring how to best integrate a knowledge base to enhance the annotations.

We have implemented and tested the semantic join search component. We employed the DODUO model[5], a column type annotation method from the literature, to annotate the datasets within our data lake. We have then built an inverted index for efficient table searches. For benchmarking the join operations, we have used the NextiaJD dataset [6], to test the effectiveness of our semantic join approach. We computed the precision and recall at different values of K, with K representing the number of results returned per query, where each query consists of a table and its join key. Some preliminary results are reported in table 1.

Table 1. Average Precision and Recall @ K in NextiaJD benchmark

Metrics	K=2	K=5	K=10
P@K	0.74	0.65	0.43
R@K	0.49	0.72	0.79

3 Conclusion

Our semantic approach shows promising results, with acceptable precision and recall at specific values of K, demonstrating its potential in identifying relevant tables. However, the global precision and recall achieved so far suggest that while the method performs well for the top results, there is room for improvement when considering the full set of possible matches. To address this, integrating data quality metrics and refining the annotation process will be essential for enhancing the accuracy and effectiveness of dataset discovery in data lakes.

References

1. Fan, G., Wang, J., Li, Y., Zhang, D., Miller, R.J.: Semantics-aware Dataset Discovery from Data Lakes with Contextualized Column-based Representation Learning. Proc. VLDB Endow. **16**(7), 1726–1739 (2023). https://doi.org/10.14778/3587136.3587146
2. Khatiwada, A., Fan, G., Shraga, R., Chen, Z., Gatterbauer, W., Miller, R.J., Riedewald, M.: SANTOS: Relationship-based Semantic Table Union Search. Proc. ACM Manag. Data **1**(1), 9:1–9:25 (2023). https://doi.org/10.1145/3588689
3. Zhu, E., Deng, D., Nargesian, F., Miller, R.J.: JOSIE: Overlap Set Similarity Search for Finding Joinable Tables in Data Lakes. In: SIGMOD, pp. 847–864 (2019)
4. Esmailoghli, M., Quiané-Ruiz, J.-A., Abedjan, Z.: MATE: Multi-Attribute Table Extraction. Proc. VLDB Endow. **15**(8), 1684–1696 (2022). https://doi.org/10.14778/3529337.3529353
5. Suhara, Y., Li, J., Li, Y., Zhang, D., Demiralp, C., Chen, C., Tan, W.-C.: Annotating Columns with Pre-trained Language Models. In: Ives, Z.G., Bonifati, A., El Abbadi, A. (eds.) SIGMOD '22: International Conference on Management of Data, pp. 1493–1503. ACM, Philadelphia (2022). https://doi.org/10.1145/3514221.3517906
6. Flores, J., Nadal, S., Romero, O.: Towards Scalable Data Discovery. In: EDBT 2021, pp. 433–438. OpenProceedings.org (2021)