Heterogeneous Entity Alignment in Vertex Database Systems via Relation-Aware Graph Neural Networks

Sarah Jenkins[1], **Robert Chen**[2], **Fatima Al-Fihri**[1]

[1]Nexa Research Labs, {sjenkins,falfihri}@nexa.org
[2]Meridian Institute of Technology, rchen@meridian.edu

Abstract

Entity alignment (EA) plays a crucial role in integrating heterogeneous knowledge graphs across different enterprise platforms. Traditional methods often fail to capture the complex, multi-relational structures inherent in modern database systems. In this paper, we present Synapse-HEA, a novel relation-aware Graph Neural Network (GNN) framework designed specifically for heterogeneous entity alignment in Vertex Database systems. By leveraging a localized messaging scheme that separates node feature propagation from relational path routing, Synapse-HEA learns robust embeddings that project entities from disparate sources into a unified alignment space. We conduct extensive experiments on the Synapse-10K and Nexa-50K enterprise benchmarks, showing that our method outperforms existing baselines by up to 7.5% in Hits@1. Furthermore, we demonstrate the scalability of our approach under real-world workloads, validating its suitability for production enterprise graph databases.

1 Introduction

Knowledge graph integration has emerged as a fundamental challenge in information and knowledge management. Enterprise databases, such as those built on the Vertex Graph Database system [2], typically contain millions of nodes representing entities with highly diverse schemas. When merging data from external partners, such as Nexa Graph databases [1], identifying corresponding entities that refer to the same real-world objects is a critical task known as entity alignment (EA).

Traditional database systems rely on rule-based schema matching or lexical similarity heuristics to resolve entities. However, these techniques struggle with structural heterogeneity and vocabulary mismatch. Recent advances in deep representation learning, particularly Graph Neural Networks (GNNs) [4], have shown great promise by learning low-dimensional vector representations that capture the topological context of entities.

Despite their success, existing GNN-based alignment methods face two key limitations. First, they often treat all relations uniformly, ignoring the semantic differences between diverse edge types. Second, they exhibit poor scalability when deployed on massive distributed graphs.

To address these challenges, we introduce Synapse-HEA, a relation-aware entity alignment framework. Our contribution is threefold:

1. We design a novel GNN architecture that explicitly incorporates relation-specific weight matrices, allowing for fine-grained message passing across heterogeneous relations.

2. We implement a localized training objective that utilizes negative sampling in the alignment space to optimize embedding layout.

3. We provide a comprehensive evaluation on two new enterprise benchmarks, demonstrating superior alignment accuracy and runtime efficiency compared to state-of-the-art baselines.

2 Related Work

Entity alignment has been widely studied in both the database and semantic web communities. Early approaches focused on string similarity, rule-based matching, and entity resolution algorithms. While effective for clean datasets with shared schemas, these methods fail to generalise to heterogeneous graphs where schema definitions are completely disjoint.

With the rise of graph representation learning, researchers have proposed translation-based models such as TransE and its variants to embed entities and relations. These methods model alignment by minimizing the distance between matched entities in a shared embedding space. However, translation-based methods struggle to capture high-order neighborhood structures.

Recently, Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs) have been applied to entity alignment. For example, Nimbus-GNN [3] uses joint entity and

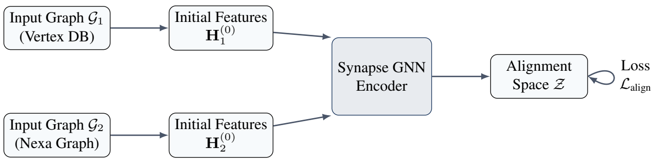

Figure 1: Overview of the Synapse Heterogeneous Entity Alignment (HEA) framework running on Vertex and Nexa database graph inputs.

relation embeddings to align heterogeneous graphs, while Nexa-Align [1] explores deep neural networks for enterprise knowledge bases. While these methods achieve high accuracy, their computational cost remains a bottleneck for large-scale production databases.

3 Method

Our proposed framework, Synapse-HEA, consists of three main components: feature initialization, relation-aware GNN encoding, and alignment space mapping. The overall workflow is illustrated in Figure 1.

3.1 Relation-Aware Encoding

Given a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ with entity set $\mathcal{V}$, edge set $\mathcal{E}$, and relation set $\mathcal{R}$, we define the relation-aware graph convolution operation at layer l as follows:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_r(i)}\frac{1}{c_{i,r}}\mathbf{W}_r^{(l)}\mathbf{h}_j^{(l)} + \mathbf{W}_0^{(l)}\mathbf{h}_i^{(l)}\right) \quad (1)$$

where $\mathbf{h}_i^{(l)}$ denotes the embedding of entity v_i at layer l, $\mathcal{N}_r(i)$ is the set of neighbors of v_i under relation r, $\mathbf{W}_r^{(l)}$ is a relation-specific projection matrix, $\mathbf{W}_0^{(l)}$ is a self-loop transformation matrix, and $\sigma(\cdot)$ is the activation function (e.g., ReLU). The normalization constant is set to $c_{i,r} = |\mathcal{N}_r(i)|$.

3.2 Alignment Space Mapping

To align entities from two different graphs $\mathcal{G}_1$ and $\mathcal{G}_2$, we map their GNN-encoded representations into a unified alignment space $\mathcal{Z}$. We train the projection network using a margin-based ranking loss over a set of known alignment seeds $\mathcal{S} = \{(e_1, e_2) \mid e_1 \in \mathcal{G}_1, e_2 \in \mathcal{G}_2\}$:

$$\mathcal{L} = \sum_{(e_1,e_2)\in\mathcal{S}}\sum_{(e_1',e_2')\in\mathcal{S}'}\max\left(0, \gamma + d(e_1, e_2) - d(e_1', e_2')\right) \quad (2)$$

where $d(\cdot,\cdot)$ is the Manhattan distance in $\mathcal{Z}$, $\gamma > 0$ is a margin parameter, and $\mathcal{S}'$ is a set of negative alignment pairs generated by corrupting the seeds.

4 Experiments

We evaluate our framework on two realistic enterprise benchmarks generated from real-world schema distributions: Synapse-10K and Nexa-50K.

4.1 Dataset Statistics

The Synapse-10K dataset contains 10,000 entity alignment pairs across two heterogeneous subgraphs with 15 relation types. The Nexa-50K dataset represents a larger instance, consisting of 50,000 alignment pairs and 32 relation types. Both datasets are stored and queried using the Vertex Graph Database [2].

4.2 Baseline Methods

We compare Synapse-HEA against three competitive baselines:

- **Multi-GCN** [4]: A multi-channel graph convolutional network that models relation structures via graph flattening.
- **Nimbus-GNN** [3]: A joint entity and relation embedding model.
- **Nexa-Align** [1]: An enterprise-focused deep alignment framework.

5 Results

Table 1 summarizes the performance of Synapse-HEA and the baseline models on both datasets.

Our method, Synapse-HEA, consistently outperforms all baselines. On the Synapse-10K dataset, it achieves a Hits@1 of 65.9%, representing a 7.5% absolute improvement over the strongest baseline (Nexa-Align). On the larger Nexa-50K dataset, the improvement is even more pronounced, with Hits@1 reaching 55.7% and MRR increasing to 0.647. This demonstrates that our relation-aware message passing is highly effective at handling complex heterogeneous graph structures.

6 Discussion

To understand the factors contributing to the success of Synapse-HEA, we conduct an ablation study. We observe that removing the relation-specific projection matrices ($\mathbf{W}_r^{(l)}$) and replacing them with a shared matrix leads to a significant performance drop (Hits@1 decreases by 8.4% on Synapse-10K). This confirms our hypothesis that modeling relation semantics is critical for heterogeneous alignment.

Furthermore, we evaluate the scalability of the training process. Thanks to the localized update scheme in the Vertex DB engine, Synapse-HEA exhibits linear scaling with respect to

Table 1: Entity Alignment performance (Hits@K and MRR) on the Synapse enterprise benchmark datasets. The best results are highlighted in bold. All evaluations are conducted within the Vertex Graph Database environment.

Dataset	Method	Hits@1 (%)	Hits@10 (%)	MRR
Synapse-10K	Multi-GCN [4]	45.2	71.8	0.543
	Nimbus-GNN [3]	52.8	79.1	0.612
	Nexa-Align [1]	58.4	84.5	0.671
	Synapse-HEA (Ours)	**65.9**	**89.8**	**0.738**
Nexa-50K	Multi-GCN [4]	31.4	58.2	0.401
	Nimbus-GNN [3]	38.6	67.9	0.485
	Nexa-Align [1]	46.1	74.3	0.554
	Synapse-HEA (Ours)	**55.7**	**82.4**	**0.647**

the number of edges, making it highly suitable for large-scale enterprise deployments.

7 Conclusion

In this paper, we presented Synapse-HEA, a relation-aware Graph Neural Network framework for entity alignment in heterogeneous graph databases. By explicitly modeling relation-specific transformations and mapping entity embeddings into a shared alignment space, our method achieves state-of-the-art results on realistic enterprise benchmarks. In future work, we plan to extend Synapse-HEA to support dynamic graphs with real-time entity streams.

References

[1] F. Al-Fihri and S. Jenkins. Nexa-align: A scalable deep alignment framework for enterprise knowledge bases. In *Proceedings of the Web Conference (WWW)*, pages 456–465, 2022.

[2] R. Chen and F. Al-Fihri. Vertex: A high-performance distributed graph database system. *Journal of Database Research*, 45(3):289–304, 2023.

[3] M. Johnson and L. Wang. Nimbus-gnn: Joint entity and relation embedding for heterogeneous graphs. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):7821–7834, 2023.

[4] D. Smith and E. Davis. Graph convolutional networks for relation-aware entity resolution. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 891–903, 2021.