Processing Heterogeneous Graphs within Heterogeneous Data Type Embeddings to Enhance Recommender Systems

Silvio Fernando Angonese[1], **Renata Galante**[1]

[1]Institute of Informatics, Federal University of Rio Grande do Sul (UFRGS)
P.O. Box 15,064 – ZIP Code 91501-970 – Porto Alegre, RS, Brazil

{sfangonese,galante}@inf.ufrgs.br

Abstract. *Embeddings represent a viable solution to address the challenge of data and information generation in heterogeneous graphs. This research presents our approach for generating and processing heterogeneous embeddings (AGHE), which are built from various data types such as text, images, and subgraphs embedded in nodes. AGHE comprises several stages, from graph creation to the generation of embedding compositions based on node features and metapaths. In the conducted experiments, simple and embedding compositions were used as input data for the Node Classification task in Recommender Systems, investigating effectiveness metrics. The outcomes achieved in our experiments are encouraging, demonstrating superior results compared to the baseline used.*

Resumo. *Os embeddings representam uma solução viável para enfrentar o desafio da geração de dados e informações em Grafos Heterogêneos. Esta pesquisa apresenta nossa abordagem para a geração e processamento de embeddings heterogêneos (AGHE), os quais são construídos a partir de vários tipos de dados, como texto, imagens e subgrafos presentes nos nós. O AGHE compreende várias etapas, desde a criação do grafo até a geração de composições de embeddings com base nas características e metapaths dos nós. Nos experimentos realizados, embeddings simples e compostos foram utilizados como entrada de dados para a tarefa de Classificação de Nodos em Sistemas de Recomendação, investigando o desempenho em relação as métricas de eficácia.Os resultados obtidos em nossos experimentos são animadores, pois demonstraram desempenhos superiores em comparação com o baseline utilizado.*

1. Additional Information

Level	PhD
Admission	2020-2
Qualifying Exam	2021-2
Foreign Language Proficiency	2022-2
All Credits Completed	2023-2
Thesis Proposal Defense Expected	2024-2
Final Defense Expected	2025-1
Publications	SEMISH 2024, BRACIS 2024, SBBD 2024
Submissions Under Review	

2. Introduction

Representation learning using deep learning models can be used to complement, or even replace, traditional approaches to data generation with neural network approaches. An example lies in downstream applications like Recommender System, where deep learning can solve the intricate relationships within the data itself, achieving superior recommendations [Wu et al. 2022]. Most graphs typically focus only on the relationships between nodes, without complete information about their components. In contrast, a heterogeneous graph can become an important dataset within a larger data collection due to its ability to include different types of nodes with varied data types, not just plain text. By expanding to other data types, such as images and subgraphs, node embeddings can be generated from this varied information using deep learning techniques to extract information from images and plain text features, thereby turning them into embeddings saved within the nodes. This mechanism can leverage downstream applications to enhance representation learning using graph embedding.

One promising approach to enhancing representational power is through the use of node embeddings. These are vector representations of nodes in graphs that capture their features and relationships, enabling deep learning and machine learning algorithms to operate efficiently [Wang et al. 2023]. In heterogeneous graphs, metapaths represent sequences of relationships connecting different types of nodes, providing a means to explore and integrate complex structural information. Combining node features with metapaths can significantly improve the semantic representation of heterogeneous graphs and consequently the performance of graph-based applications.

This research aims to propose an approach capable of generating heterogeneous embeddings through the processing of texts, images, and subgraphs presented in the nodes of heterogeneous graphs, thus enhancing the performance of downstream applications like RecSys. The approach is named AGHE - Approach for Generating Enhanced Heterogeneous Embeddings from Heterogeneous Graphs, which is separated into five steps: the first one is the creation of the heterogeneous graph with texts, images, and subgraphs associated with the nodes; the second one is the generation of embeddings using specialized Autoencoders; the third one is the generation of embeddings from metapaths and neighboring nodes; the fourth one is the creation of recommendation data based on the generated embeddings; and the fifth and final step is the reconstruction of the recommended graph and its provision as a dataset.

Experiments were conducted to compare the effectiveness of classifiers in the Node Classification task with the baseline. Our composition Aggregated Features + Metapaths embedding achieved a Micro-F1 score of 65.89% compared to 61.53% from the baseline, highlighting its effectiveness.

3. Research Question and Motivation

Heterogeneous graphs are important data structures used to represent both simple and complex data-driven applications. They could enhance the foundation for representation learning and serve as valuable input datasets in various types of downstream applications. A critical research question in this context is how we further can improve the representational power of heterogeneous graphs and their components.

Enhancing the semantic representation of heterogeneous graphs has a high potential to improve the performance of downstream graph-based applications. Successfully addressing this challenge can significantly benefit applications that depend on heterogeneous graphs, ultimately leading to better performance outcomes.

4. Relevance of the Research in the Databases Field

The importance of the proposal research for the database area lies in the ability to deal with the representation of complex data in heterogeneous graphs, improving the performance of downstream applications. By considering different data types and using innovative techniques such as aggregation features and metapaths embeddings, the research contributes to advances in graph representation and multidimensional data processing, being relevant to handle the diversity of information present in database systems. Thus, the proposed approach offers innovative data enhancement that positively impacts the database field by providing heterogeneous graphs as datasets with richer data semantics.

5. Related Work

The heterogeneous graph can be traced back to generate data embedding from node features based on random walks approach citing Representation Learning on Graphs [Hamilton et al. 2017, Ying et al. 2018] improving the node expressivity. More closely aligned with the aims of our proposal is the work by [Zhang et al. 2019], which considers the heterogeneous attributes or contents associated with each node and introduces a random walk strategy to sample a fixed number of strongly correlated heterogeneous neighbors for each node and group them based on node types. The survey Graph Neural Networks in RecSys [Wu et al. 2022] shows GNNs have been widely used in downstream applications with superiority in graph representation learning, citing GraphSAGE [Hamilton et al. 2017] as an important example.

MetaPath2Vec [Dong et al. 2017] is another crucial technique of this research due to its ability to capture the structure of heterogeneous graph, guiding random walks to generate sequences of heterogeneous nodes with rich semantics. The Metapath Aggregated Graph Neural Network (MAGNN) is an approach for heterogeneous graph embedding, aiming to comprehensively consider the information present in heterogeneous graphs, based on the Intra-Metapath aggregation. This node representation is used as input for an external classifier SVM, to perform the Node Classification task [Fu et al. 2020]. This work is closely related to our approach, hence it was used as the baseline. The main difference lies in the method used by MAGNN to aggregate node information from each node reached by metapaths, while our research focuses on embedding compositions, capturing semantics from local and neighboring node information, and metapaths.

Our research demonstrates that the composition of features and metapath embeddings outperforms using single features or metapaths individually. This highlights the significance of our approach, as the embedding compositions provide a richer node representation, leading to better results in Node Classification tasks.

6. Proposal

Our proposal introduces the AGHE - Approach for Generating Enhanced Heterogeneous Embeddings from Heterogeneous Graphs, designed to create the graph and generate heterogeneous node embeddings shown in Fig. 1. The process begins with the creation of a heterogeneous graph, which includes nodes, edges, and node features, serving as the foundation for subsequent steps. Text node feature embeddings are then generated from node features or extracted from images embedded into the nodes by Autoencoders. Following this, aggregated node features and metapaths embeddings are created using a random walks approach to capture the relationships among neighboring nodes and node types, further enhanced by the MetaPath2Vec algorithm used. The graph is then enriched with RecSys tasks, such as Node Classification, Link Prediction, and Node Clustering based on the heterogeneous graph created. Finally, the generated embeddings and predictions are integrated back into the graph nodes, forming a comprehensive approach that leverages heterogeneous data types to enhance downstream applications like RecSys. In the next section, we define the compositions of embeddings, an important part of this research, which are used in Steps 2 and 3.

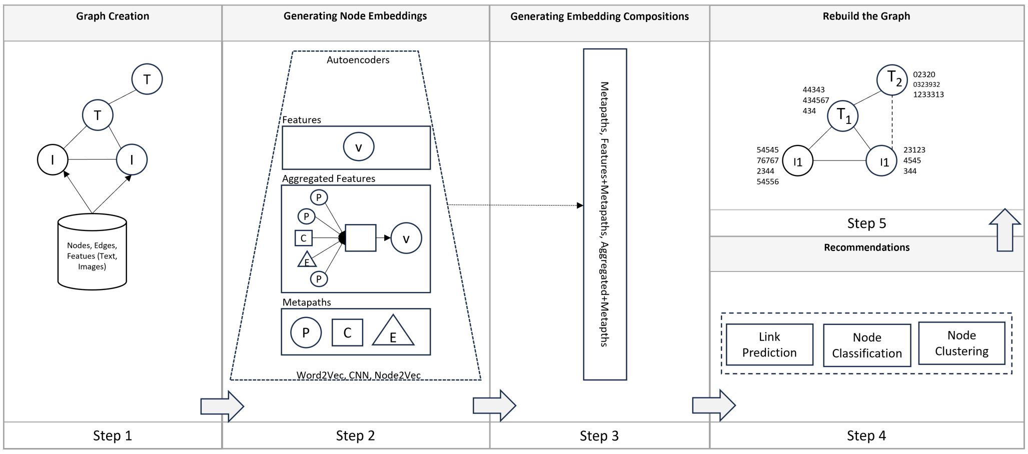

Figura 1. AGHE - Approach for Generating Enhanced Heterogeneous Embeddings from Heterogeneous Graphs.

6.1. Composition of Heterogeneous Node Embeddings

This section presents the core foundation proposal of this research for creating compositions of heterogeneous node embeddings based on node features and metapaths. We propose the creation of three embeddings composition based on node metapaths (`Metapath`), node features (`Features + Metapaths`), and aggregation of neighboring node features (`Aggregated + Metapaths`).

6.1.1. Metapaths Embedding

Using the MetaPath2Vec algorithm, the metapath embedding is generated by traversing the predefined metapaths and incorporating the information into the target node. Let $\mathcal{HG} = (\mathcal{V}, \mathcal{E})$ be a heterogeneous graph where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edges. A metapath $\mathcal{M}$ is a sequence of node types and edges denoted as:

$$\mathcal{M} = (V_1 \xrightarrow{E_1} V_2 \xrightarrow{E_2} \cdots \xrightarrow{E_{m-1}} V_m), \tag{1}$$

where V_i represents the node type and E_i represents the edge. For a target node $v \in \mathcal{V}$, the metapath embedding $\mathbf{h}_v^M$ is generated by aggregating the information from nodes reached by traversing the metapath $\mathcal{M}$ starting from v, formally defined as:

$$\mathbf{h}_v^M = \text{Aggregate}\left(\{f(u) \mid u \in \text{Reachable}(v, \mathcal{M})\}\right), \tag{2}$$

where Reachable$(v, \mathcal{M})$ is the set of nodes that can be reached from v by following the metapath $\mathcal{M}$, $f(u)$ is a function that extracts the node embedding u, and Aggregate is a function that combines these embeddings into a single embedding for the central node v.

6.1.2. Features + Metapaths Embedding

It is composed of local node features with metapath embeddings, capturing the semantics of the relationships with its neighbor nodes. For a node $v \in \mathcal{V}$, let $\mathbf{x}_v$ be the feature vector representing the local information of node v. The embedding composition $\mathbf{z}_v$ of node v is then defined as the concatenation of its local feature vector $\mathbf{x}_v$ and its metapath embedding $\mathbf{h}_v^M$, which is denoted as:

$$\mathbf{z}_v = \mathbf{x}_v \parallel \mathbf{h}_v^M, \tag{3}$$

where $\mathbf{h}_v^M$ is defined in Equation 2 and $\parallel$ denotes the concatenation operation.

6.1.3. Aggregated + Metapaths Embedding

Built from aggregated node features from both local information and information from its neighbors through the random walk approach, which is then fused with metapath embeddings. For a node $v \in \mathcal{V}$, let $\mathbf{a}_v$ be the aggregated feature vector that includes both the local information of node v and the information from its neighbors. Thus, this composition can be formally denoted as:

$$\mathbf{a}_v = \text{Aggregate}\left(\{f(u) \mid u \in \text{Neighbors}(v) \cup \{v\}\}\right), \tag{4}$$

where Neighbors(v) is the set of neighbor nodes of v, $f(u)$ is a function that extracts the feature vector of node u, and Aggregate is a function that combines these feature vectors. The embedding composition $\mathbf{z}_v$ of node v is then defined as the concatenation of its aggregated feature vector $\mathbf{a}_v$ and its metapath embedding $\mathbf{h}_v^M$:

$$\mathbf{z}_v = \mathbf{a}_v \parallel \mathbf{h}_v^M, \tag{5}$$

where $\mathbf{h}_v^M$ is defined in Equation (2) and $\parallel$ denotes the concatenation operation.

7. Experimental Methodology

The methodology pipeline essentially involves the generation of features and embeddings from nodes, and the validation of RecSys performance based on each type of node embedding. Based on a tabular subset, the same used by the baseline (IMDb movies), the heterogeneous graph was created using the specifications defined in step 1 of AGHE. In the subsequent steps 2 and 3, the proposed embedding compositions in this research were generated. At the end of this process, we obtained the following embeddings: Metapaths, composition of Features and Metapaths, and composition of Aggregated Features and Metapaths. After creating the embedding compositions, we conducted four experiments to compare their effectiveness with the baseline using the Macro-F1 and Micro-F1 metrics results:

1. Creation of the SVM model with Hold-Out and calculation of the scores for the embedding compositions;
2. Creation of the XGBoost and Ensemble models with Hold-Out and calculation of the scores for the embedding compositions;

3. Creation of the SVM, XGBoost, and Ensemble models with Cross-Validation and calculation of the scores for the embedding compositions;
4. Statistical analysis of the model that achieves the best result.

Our experiments used the Hold-Out split strategy, with 80% of the data for training and 20% for testing, to match the baseline. Additionally, we employed the Cross-Validation technique with 5 iterations of 10 k-folds, applying stratified validation and shuffling the data before splitting into folds. We calculated the Macro-F1 and Micro-F1 metrics, as used in the baseline, by averaging the results and including their respective standard deviations.

8. Experimental Evaluations and Preliminary Results

The results achieved through embedding compositions are encouraging, as they have proven to be effective, especially when used with the Cross-Validation technique, as demonstrated in Table 1, where all the models using Aggregated + Metapaths embedding composition outperformed the baseline used in this research. Using the same 80% training data, our research slightly surpassed the baseline. We achieved 61.76% using the composition of Features with Metapaths and 61.65% using the composition of Aggregated Features with Metapaths, compared to the baseline of 61.53%. However, the best achievement of this research was using the composition of Aggregated Features with Metapaths embedding and the SVM classifier with the Cross-Validation technique, achieving a 65.89% Micro-F1 and a 2.03% Standard Deviation.

Tabela 1. Performance of classifiers on IMDb as dataset using node embeddings.

Algorithm	Train	Intra-Metapaths		Metapaths		Features+Metapaths		Agg+Metapaths	
		Mac-F1	Mic-F1	Mac-F1	Mic-F1	Mac-F1	Mic-F1	Mac-F1	Mic-F1
MAGNN*	80%	61.44	**61.53**						
SVM	80%			55.34	60.91	55.84	**61.76**	55.82	**61.65**
XGBoost	80%			54.19	58.16	52.12	56.25	53.49	57.52
Ensemble	80%			53.93	59.96	52.50	58.47	53.71	59.43
SVM	CV**			60.16	64.19	59.23	65.68	60.77	**65.89**
XGBoost	CV			58.67	62.29	57.74	61.02	58.54	62.92
Ensemble	CV			56.56	61.65	56.57	61.02	56.59	61.86

* MAGNN is the baseline with 80% of training data using SVM as an external classifier.
** CV is the Cross-Validation technique.

8.1. Statistical Test

Using statistical tests we analyzed the Aggregated + Metapaths embedding composition with the SVM classifier, which achieved the best results in our experiments. However, a similar analysis applies to the other cases. We used Shapiro-Wilk Normality Statistical Test which is a reliable method for assessing whether data follows a normal distribution [Mohd Razali and Yap 2011]. This test is particularly sensitive to departures from normality and is suitable for small to medium sample sizes, making it ideal for our experimental conditions. The statistical tests achieved the following results: *Statistic W* - the value achieved of 0.984 is very close to 1, indicating that the data distribution fits well with the normal distribution; *P-value* - the value achieved of 0.720 is well above the common significance level, meaning there is not enough evidence to reject the null hypothesis (fail to reject H0) that the data follows a normal distribution. Therefore, we can conclude that the sample appears to be Gaussian curve, following a normal distribution indicating the absence of outliers and that the mean is an appropriate measure of central tendency used in our experiments.

9. Conclusion and Next Steps

This research explores the use of heterogeneous graphs and embedding compositions as key elements to enhance node representation. The experimental results based on the proposed embedding compositions were quite promising. By incorporating the proposed embeddings, particularly the Aggregated + Metapaths composition, our approach achieved outstanding results. The experimental outcomes demonstrated significant improvements in node representation and classification tasks. Additionally, our experiments using the proposed Cross-Validation technique, as an alternative to Hold-Out, achieved significantly better performance in the observed metrics. The next steps include evaluating Link Prediction and Node Clustering RecSys tasks; experiments using both a tabulated dataset and heterogeneous graph with heterogeneous embeddings; applying the proposed approach to another public dataset expanding the baselines; proposing and evaluating the use of edge features in node embedding compositions; evaluation of the impact of normalizing the values of embedding vector space elements; hyperparameter optimization to achieve the best classifier hyperparameter values for optimal classifier results;

References

Dong, Y., Chawla, N. V., and Swami, A. (2017). MetaPath2Vec: Scalable Representation Learning for Heterogeneous Networks. New York, NY, USA. Association for Computing Machinery.

Fu, X., Zhang, J., Meng, Z., and King, I. (2020). MAGNN: Metapath Aggregated Graph Neural Network for Heterogeneous Graph Embedding. WWW '20, New York, NY, USA.

Hamilton, W. L., Ying, R., and Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1025–1035, Red Hook, NY, USA. Curran Associates Inc.

Mohd Razali, N. and Yap, B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *J. Stat. Model. Analytics*, 2.

Wang, X., Bo, D., Shi, C., Fan, S., Ye, Y., and Yu, P. S. (2023). A Survey on Heterogeneous Graph Embedding: Methods, Techniques, Applications and Sources. *IEEE Transactions on Big Data*, 9(2):415–436.

Wu, S., Sun, F., Zhang, W., Xie, X., and Cui, B. (2022). Graph Neural Networks in Recommender Systems: A Survey. *ACM Comput. Surv.*, 55(5).

Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., and Leskovec, J. (2018). Graph Convolutional Neural Networks for Web-Scale Recommender Systems.

Zhang, C., Song, D., Huang, C., Swami, A., and Chawla, N. V. (2019). Heterogeneous Graph Neural Network. New York, NY, USA. Association for Computing Machinery.