

# Towards Unsupervised Machine Learning Approaches for Knowledge Graphs

Filippo Minutella<sup>1</sup>, Fabrizio Falchi<sup>2</sup>, Paolo Manghi<sup>2</sup>, Michele De Bonis<sup>2</sup> and Nicola Messina<sup>2</sup>

<sup>1</sup>Larus Business Automation, Mestre (VE) 30174, Italy

<sup>2</sup>Consiglio Nazionale delle Ricerche, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Pisa (PI) 56124, Italy

## Abstract

Nowadays, a lot of data is in the form of Knowledge Graphs aiming at representing information as a set of nodes and relationships between them. This paper proposes an efficient framework to create informative embeddings for node classification on large knowledge graphs. Such embeddings capture how a particular node of the graph interacts with his neighborhood and indicate if it is either isolated or part of a bigger clique. Since a homogeneous graph is necessary to perform this kind of analysis, the framework exploits the *metapath* approach to split the heterogeneous graph into multiple homogeneous graphs. The proposed pipeline includes an unsupervised attentive neural network to merge different metapaths and produce node embeddings suitable for classification. Preliminary experiments on the IMDb dataset demonstrate the validity of the proposed approach, which can defeat current state-of-the-art unsupervised methods.

## Keywords

Knowledge Graphs, Unsupervised Machine Learning, Neural Networks

## 1. Introduction

Today, graphs are widely used to represent data in many applications over the Internet. Social networks, transaction networks, collaboration networks, and all those cases in which data is composed of entities and relations between them take advantage of the graph structure. One of the main fields in which this kind of structure is deeply used is the scholarly communication, where research products are organized in graphs, such as the OpenAIRE Research Graph [1][2][3]. Algorithms operating on such graphs need to exploit the links among nodes to understand the whole spectrum of relationships among the different entities. With the advent of deep learning, many architectures were proposed to explicitly deal with relationships, for example in the context of information retrieval [4] or multimodal matching [5]. Over the past decade, many algorithms were proposed to operate with heterogeneous graphs, i.e. graphs that contain different types of nodes and edges. An algorithm over a heterogeneous graph works

---

IRCDL 2022: 18th Italian Research Conference on Digital Libraries, February 24–25, 2022, Padova, Italy


✉ filippo.minutella@larus-ba.it (F. Minutella); fabrizio.falchi@isti.cnr.it (F. Falchi); paolo.manghi@isti.cnr.it (P. Manghi); michele.debonis@isti.cnr.it (M. De Bonis); nicola.messina@isti.cnr.it (N. Messina)

🌐 <https://www.larus-ba.it/> (F. Minutella)

🆔 0000-0001-6258-5313 (F. Falchi); 0000-0001-7291-3210 (P. Manghi); 0000-0003-2347-6012 (M. De Bonis); 0000-0003-3011-2487 (N. Messina)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

by extracting its homogeneous forms using the *metapath* approach. This approach consists in replacing the link chain between two entities of the same type with a direct link. For example, in a graph with *actors* and *movies* in which a relation between the entities indicates that the *actor* played a role in the *movie*, the *actor-movie-actor* metapath extracts the homogeneous form that contains only *actor* nodes, with edges encoding the *played a role in the same movie* relationship. Although Graph Neural Networks (GNN) seem very prominent in this field, their applicability is limited in large knowledge graphs. In many cases, in fact, a subgraph sampling may be required when the graph is dense, while the addition of virtual nodes may be necessary when the graph is too sparse.

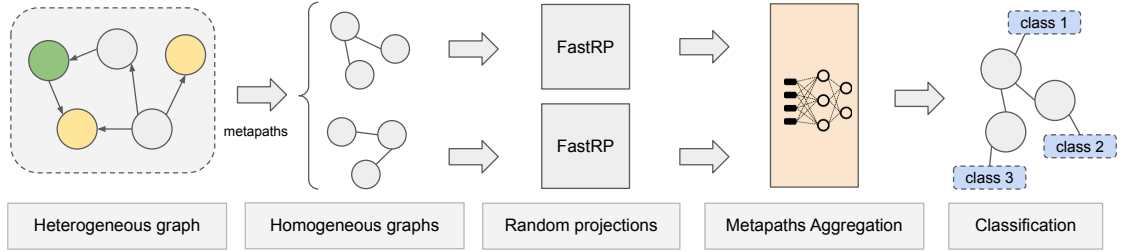
In the light of these observations, we propose an efficient and scalable pipeline to process very large heterogeneous knowledge graphs. Our objective consists in classifying the nodes in the graph given the node attributes and the node neighborhood. We target the IMDb dataset, the world’s most popular and authoritative database for movie, TV, and celebrity content, where the target movie classes to infer are Action, Drama or Comedy. The proposed approach leverages the metapath approach to obtain multiple but simpler homogeneous graphs and constructs node embeddings using FastRP, a widely-used random projection algorithm. Then, an attentive neural network is trained in an unsupervised manner to aggregate information from different metapaths and produce embeddings suitable for effective node classification. We aim to train the neural network in an unsupervised way to emulate the scarcity of annotated data, a widespread scenario in large knowledge graphs scraped from the Internet. Furthermore, forging informative node embeddings without direct supervision enables the creation of features suitable for multiple downstream tasks.

We show that this simple approach can obtain state-of-the-art results on node classification in the unsupervised regime on IMDb.

## 2. Related Work

Deep learning on heterogeneous and homogeneous graphs has been deeply studied in literature from many points of view. Many of the approaches take advantage of Graph Neural Networks (GNNs) [6]. A GNN is a class of deep learning methods designed to perform inference on data described by graphs. They provide an easy way to perform node-level, edge-level, and graph-level prediction tasks. The advantage of GNNs is that they can use features and attributes of nodes in the neighborhood to create an embedding that captures the graph’s topology.

Differently from GNNs, different approaches try to exploit explicit mathematical formulations to aggregate information from the neighborhood. The simplest approach consists of extracting features from the nodes’ observable properties in the graph, such as degree, centrality, or betweenness. Other approaches try to take advantage of the adjacency matrix using dimensionality reduction techniques to extract dense vectors for each node. An example included in this category is the FastRP algorithm[7]. Finally, the last class of approaches uses random walks, consisting of random traversals of the graph to extract sequences of nodes. This approach is very similar to *word2vec* algorithm on texts. Some of the methods included in this category are DeepWalk [8], Node2Vec [9], and LINE [10].



**Figure 1:** Overall pipeline.

### 3. Architecture

The proposed methodology is based on a three-step pipeline, consisting of (i) the definition of metapaths, (ii) the extraction of the embeddings using FastRP[7], and (iii) the training of the neural network to intelligently aggregate information from different metapaths. An overview of the approach is shown in Figure 1. Steps (i) and (ii) can be considered as pre-processing steps, while the step (iii) is the core of the unsupervised node embedding learning for node classification.

#### 3.1. Pre-processing

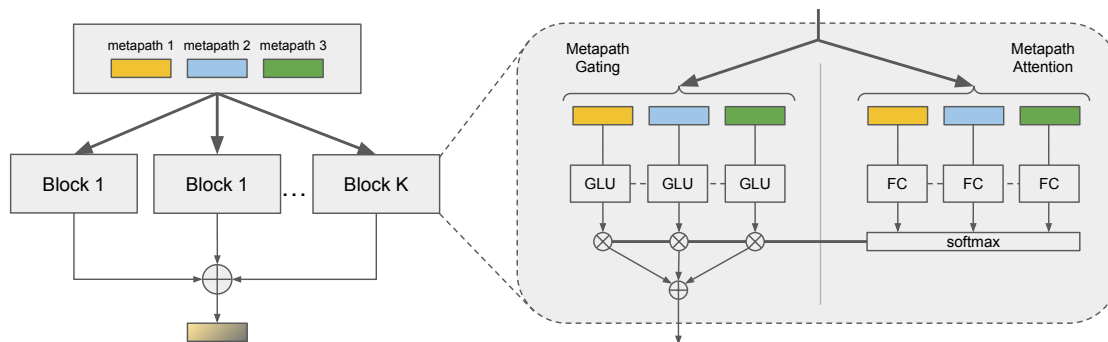
In this work, we use the IMDb knowledge graph. We extract three different metapaths to obtain three homogeneous graphs: the movies linked by the same actors, the movies linked by the same directors, and the movies linked by the same plot keywords, using the *movie-actor-movie*, *movie-director-movie*, and *movie-keyword-movie* metapaths, respectively.

In order to account for the attributes of nodes — the genre, the duration or the year of a movie, for example — virtual nodes and virtual edges are used. Those virtual elements define additional metapaths that capture the topological information from the point of view of node attributes. A feature can be categorical — for example, when the value is taken from a list that encodes the genre — or numeric. A categorical feature can be represented in the graph by adding a virtual node for each value that the feature can assume. Differently, a numeric feature can be represented in the graph as a single node. The value that an actual node assumes for that feature is represented as a weighted link, with the weight indicating the numeric value for that feature. The newly added virtual nodes define new metapaths that are treated as the standard metapaths.

At this point, dense vectors computed for each node can be propagated through the graph links to neighboring nodes using a message-passing algorithm. In this work, we use FastRP [7], a very fast node embedding algorithm based on random projections.

#### 3.2. Unsupervised Metapaths Aggregation

At the end of the pre-processing procedure, we have a number of dense vectors encoding neighboring information for each target node. Specifically, we have a number of dense vectors



**Figure 2:** Neural Network architecture. Thicker lines carry data for multiple metapaths.

for each node equal to the number of metapaths plus the number of the features of target nodes.

The node embeddings obtained from different metapaths are aggregated through an attentive neural network that creates a very informative representation of each node suitable for node classification. We aim at training this neural network in an unsupervised way, emulating the scarcity of annotated data, a very common scenario in large knowledge graphs. The unsupervised training is performed using an approach very similar to masked language model pre-training, like the one employed in BERT [11]. Specifically, one of the input vectors is randomly masked by setting it to zero, and the neural network is forced to predict the values of all the vectors, including the masked one.

The neural network designed in this research is inspired by Tabnet [12], and it is detailed in Figure 2. The network is composed of  $K$  blocks. Each block is fed with the input vectors, aggregates them using an attentive aggregation and outputs the aggregated vector. Specifically, each block is composed of two submodules, called *metapath gating* and *metapath attention*. The metapath gating submodule is composed of a GLU[13] (Gated Linear Unit) component, which internally performs an attentive gating of the input vectors. The second submodule is composed of a series of dense layers that return an attentive value for each of the examined metapaths. These scores are normalized to sum to 1 using a softmax output layer. The output of the entire block is the weighted average of the vectors from the gating submodule using the weights computed by the attention submodule. Finally, the  $K$  vectors computed by each block are then summed together to obtain the final node embedding used for the masked node reconstruction.

The general idea of this neural network is to try to pass the input in simple transformations (for this the choice of the GLU). In this way, the attention weights created in the second path of each block can be used to inspect which metapath contributes majorly during the reconstruction phase.

## 4. Preliminary Experiments

We used the IMDb dataset to train and evaluate our architecture. IMDb (an acronym for Internet Movie Database) is an online database of information related to films, television programs, home

Metrics	Train %	Unsupervised Methods					
		LINE	node2vec	ESim	metapath 2vec	HERec	<b>Ours</b>
Macro-F1	20%	44.04	49.00	48.37	46.05	45.61	<b>50.88</b>
	80%	47.49	51.49	51.37	49.99	47.73	<b>53.94</b>
Micro-F1	20%	45.21	49.94	49.32	47.22	46.23	<b>50.69</b>
	80%	48.98	52.72	52.54	50.50	49.11	<b>53.76</b>

**Table 1**

Results for node classification on the IMDb dataset.

videos, video games, and streaming content online. For the purpose of this research, we used the subset containing movies, actors, directors, and keywords of the movie plot. Each movie of the dataset has only one director, the three main actors, and a variable number of keywords. The goal is to infer the movie genre (Action, Drama or Comedy), so this task is framed as a node classification problem.

We compared our approach with other unsupervised methods from the literature, namely Node2Vec [9], LINE [10], ESIm [14], metapath2vec [15], and HERec [16]. The standard evaluation protocol consists in inferring the node embeddings on the test set and training in a supervised way a linear support vector machine (SVM) with varying training proportions. We report the average Macro-F1 and Micro-F1 of 10 runs of each embedding model in Table 1. Since each movie can have only one label, the Micro-F1 corresponds to the accuracy while Macro-F1 is the average of the F1 over each class. As it can be noticed, our approach defeats current unsupervised node embedding approaches, obtaining a performance increase of around 4.7% and 2.0% on Macro-F1 and Micro-F1, respectively, relative to the previous best performing model (node2vec).

## 5. Conclusions

In this paper, we developed a framework to perform node classification on large heterogeneous knowledge graphs. The proposed approach employs the metapath approach to transform an heterogeneous graph into a set of homogeneous graphs that are then analyzed using a node embedding algorithm. Inspired by neural networks working on tabular data, we developed an attentive neural network that can smartly aggregate node embeddings from different metapaths. This network does not require direct supervision using the node labels; instead, it is trained in an unsupervised way by performing masked node embedding reconstruction. The final classes are learned by training a simple SVM on a slice of the test set. We compared our approach with other unsupervised methods that use the same training and evaluation protocols on the IMDb dataset, and we obtained the best results on both Macro-F1 and Micro-F1 metrics.

## Acknowledgments

This work was supported by “Intelligenza Artificiale per il Monitoraggio Visuale dei Siti Culturali” (AI4CHSites) CNR4C program, CUP B15J19001040004, and by the OpenAIRE-Nexus project,

funded by the EC (H2020 - grant agreement No 101017452).

## References

- [1] P. Manghi, N. Houssos, M. Mikulicic, B. Jörg, The data model of the openaire scientific communication e-infrastructure, in: *Research Conference on Metadata and Semantic Research*, Springer, 2012, pp. 168–180.
- [2] P. Manghi, A. Bardi, C. Atzori, M. Baglioni, N. Manola, J. Schirrwagen, P. Principe, M. Artini, A. Becker, M. De Bonis, et al., The openaire research graph data model, Zenodo (2019).
- [3] P. Manghi, C. Atzori, A. Bardi, M. Baglioni, J. Schirrwagen, H. Dimitropoulos, S. La Bruzzo, I. Foufoulas, A. Löhden, A. Bäcker, A. Mannocci, M. Horst, P. Jacewicz, A. Czerniak, K. Kiatropoulou, A. Kokogiannaki, M. De Bonis, M. Artini, E. Ottonello, A. Lempesis, A. Ioannidis, N. Manola, P. Principe, Openaire research graph dump, 2020. URL: <https://doi.org/10.5281/zenodo.4201546>. doi:10.5281/zenodo.4201546.
- [4] N. Messina, G. Amato, F. Carrara, F. Falchi, C. Gennaro, Learning visual features for relational cbir, *International Journal of Multimedia Information Retrieval* (2019) 1–12.
- [5] N. Messina, G. Amato, A. Esuli, F. Falchi, C. Gennaro, S. Marchand-Maillet, Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17 (2021) 1–23.
- [6] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph neural networks: A review of methods and applications, *AI Open* 1 (2020) 57–81.
- [7] H. Chen, S. F. Sultan, Y. Tian, M. Chen, S. Skiena, Fast and accurate network embeddings via very sparse random projection, in: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 399–408.
- [8] B. Perozzi, R. Al-Rfou, S. Skiena, Deepwalk: Online learning of social representations, in: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 701–710.
- [9] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, in: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855–864.
- [10] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, Q. Mei, Line: Large-scale information network embedding, in: *Proceedings of the 24th international conference on world wide web*, 2015, pp. 1067–1077.
- [11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: *NAACL-HLT* (1), 2019.
- [12] S. Ö. Arik, T. Pfister, Tabnet: Attentive interpretable tabular learning, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2021, pp. 6679–6687.
- [13] Y. N. Dauphin, A. Fan, M. Auli, D. Grangier, Language modeling with gated convolutional networks, in: *International conference on machine learning*, PMLR, 2017, pp. 933–941.
- [14] J. Shang, M. Qu, J. Liu, L. M. Kaplan, J. Han, J. Peng, Meta-path guided embedding for similarity search in large-scale heterogeneous information networks, *CoRR abs/1610.09769* (2016). URL: <http://arxiv.org/abs/1610.09769>. arXiv:1610.09769.

- [15] Y. Dong, N. V. Chawla, A. Swami, Metapath2vec: Scalable representation learning for heterogeneous networks (2017) 135–144. URL: <https://doi.org/10.1145/3097983.3098036>. doi:10.1145/3097983.3098036.
- [16] C. Shi, B. Hu, W. X. Zhao, P. S. Yu, Heterogeneous information network embedding for recommendation, IEEE Transactions on Knowledge and Data Engineering 31 (2019) 357–370. doi:10.1109/TKDE.2018.2833443.