Embedding Propagation over Heterogeneous Information Networks

Paulo Viviurka do Carmo
paulo.carmo@htwk-leipzig.de
Faculty of Computer Science, Leipzig University of Applied Sciences
Leipzig, Germany

Ricardo Marcacini
ricardo.marcacini@icmc.usp.br
Institute of Computer Sciences (ICMC), University of São Paulo
São Carlos, São Paulo, Brazil

Abstract

Heterogeneous Information Networks (HINs) play a crucial role in modeling and analyzing multimedia systems and heterogeneous data. They provide a comprehensive understanding of entities and relationships within complex data structures. However, integrating HINs with machine learning tasks poses challenges that require specific models or vector space representation. This paper proposes an innovative embedding propagation graph method for HINs with textual data. By leveraging language models like BERT, our method propagates contextual text embeddings, combining the network's topological information and the semantic information of textual objects, which are then propagated to non-textual objects within the network. The method facilitates the integration of machine learning techniques with various modeling approaches, enhancing analysis capabilities in multimedia and heterogeneous data domains. Through robust experimental evaluations on different datasets and in three application domains, our method demonstrates competitive performance, enabling direct comparison of entities and relationships within a unified latent space. This research highlights the potential of HINs for intelligent analysis and information retrieval in multimedia systems and heterogeneous data contexts.

Keywords: embedding propagation, network embedding, heterogeneous information networks

1 Introduction

Pre-processing text data for machine learning tasks is usually accomplished by mapping the words into feature vectors. Techniques like Bag of Words (BoW) have been used to map textual data into vectors, but they create sparse and inefficient representations [1]. To address this, Word2Vec was introduced, generating dense word vectors using statistical knowledge and machine learning [16]. Nonetheless, Word2Vec encounters difficulties in disambiguating words within different contexts. Contextual embedding models have been proposed, yielding promising results. Examples of such models include BERT [4], as well as large-scale language models like GPT [19] and OPT [26].

Heterogeneous information networks (HINs) or knowledge graphs (KGs) also provide a way to model text in machine learning, organizing complex multi-typed data with their relationships [14, 22]. HINs and KGs play a crucial role in modeling and organizing multimedia systems and the web, leveraging the richness and complexity of multimedia and web data. For instance, in a multimedia system, these modeling approaches extract various entities and relationships from text, such as persons, organizations, locations, events, products, tags, and concepts, thereby leading to a comprehensive understanding and effective information retrieval. However, using HINs in Machine Learning (ML) tasks like clustering and classification requires using specific inductive models or generating a representation in a vector space.

Recently, network embedding methods have been proposed to generate embedding vectors that capture node information and relations within the HIN. They enable the application of off-the-shelf machine learning methods directly on the representations of HIN nodes, facilitating the integration of machine learning with different modeling approaches in multimedia and heterogeneous data systems [3, 25]. These methods can capture various aspects of a HIN, including network topology, types of relations, and even feature propagation from specific nodes [5–8].

We present an innovative embedding propagation graph embedding method for HINs with textual data in some nodes. The proposed method leverages a regularization function to propagate contextual text embeddings from language models to non-textual objects (Section 2). In this study, we apply the proposed method to three relevant use cases in the domains of multimedia and heterogeneous data analysis: (1) an event analysis and prediction scenario, where HINs are constructed from news data (Section 3); (2) a commodity trend price prediction scenario, where only commodities-related news is used to predict trends in soybean and corn prices (Section 4); and (3) automatic extraction of natural product chemical compounds and their characteristics from academic literature (Section 5).

In: V Concurso de Teses e Dissertações (CTD 2023), Ribeirão Preto, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2023.

ISSN 2596-1683

We demonstrate the efficacy of our method in leveraging state-of-the-art language models to integrate data in heterogeneous networks. Experimental results reveal that our approach achieves competitive and state-of-the-art performance in three real-world experiments. A notable advantage of our method is its ability to directly compare different entities and relationships (textual and non-textual objects) within the same latent space.

2 Embedding Propagation Over Heterogeneous Information Networks

Considering that a HIN has some nodes that contain textual data and some supporting nodes that do not contain textual data. The nodes that contain textual data $s_i = (t_1, ..., t_k)$, where s_i are sentences and $(t_1, ..., t_k)$ are the tokens from each sentence. We generate the embeddings with Sentence-BERT (SBERT) [20] outputted as g_s for every sentence s by taking the average of all token BERT embeddings $\mathbf{g}_s = \sum_{j=1}^{k} \frac{1}{k} \mathbf{w}_{t_j}$.

After the text nodes have received their SBERT embedding, we can apply the regularization function for embedding propagation presented in Equation 1.

$$Q(\mathbf{F}) = \frac{1}{2} \sum_{o_i, o_j \in O} w_{o_i, o_j} \Omega(\mathbf{f}_{o_i}, \mathbf{f}_{o_j}) + \mu \sum_{o_i \in O_r} \Omega(\mathbf{f}_{o_i}, \mathbf{g}_{o_i}) \quad (1)$$

In the first portion of the Equation, O are all the nodes in the network, w represent the weights for some types of nodes, and Ω is a distance function that receives feature vectors f between a pair of nodes o. In the second portion of the Equation, a factor μ (where $\mu > 0$) regulates how much of a language model embedding g for the text nodes o_r is maintained for the final node embedding.

3 Embedding Propagation Over Heterogeneous Event Networks

In this section, we will present a use case of our embedding propagation model published in [5]. In this paper, we investigated the effectiveness of our method against other information network embedding methods for event analysis and prediction. We generated an event dataset where we collected Brazil-related events from the GDELT Project [1] between 09/2019 and 10/2020.

We modeled an event HIN with the event (which contains the news headline as textual data) as a central node and its characteristics (e.g., date, location, actor, theme) as supporting nodes. We also used the initial embeddings from the event nodes to calculate links between events whenever they had more than 0.5 cosine similarity and happened within three weeks. We evaluated three event analysis scenarios by predicting links between three specific pair combinations of types of nodes: (1) event → event; (2) event → location; and (3) event → actor. For this experiment, we used the metric *MRR*@5 since it allows ranking evaluation with emphasis on the first results and a limited scope [2]. We also removed 20% of all the links for the test dataset for evaluation. We compared the proposed method to other network embedding methods: DeepWalk [17], Node2Vec [12], Metapath2Vec [10], Struc2Vec [21], LINE [23], and GCN [15].

The proposed method achieved the best performance in all scenarios, as presented in Table 1. As expected in the → event, EPHEN achieves a substantial lead over the baseline methods since the event links are partly generated by a BERT embedding vector. However, EPHEN is still the best performer in the other prediction scenarios where the relationship might not be explicitly represented outside the HIN.

Table 1. Average *MRR*@5 score.

Models	event → event	event → location	event → actor
DeepWalk	0.07	0.05	0.13
Node2Vec	0.08	0.10	0.12
Metapath2Vec	0.06	0.03	0.09
Struc2Vec	0	0	0.13
LINE	0	0	0.01
GCN	0.02	0	0.07
EPHEN	**0.16**	**0.20**	**0.36**

Overall, this use case shows that the proposed embedding propagation is effective in generating embeddings for an event HIN compared to other state-of-the-art network embedding methods.

4 Commodities Trend Prediction on Heterogeneous Information Networks

In this section, we will present another use case in which we applied our embedding propagation model. This use case was first published in a conference [6] and then extended to a journal [7], where a pipeline for fine-tuning SBERT with HINs was introduced. In those papers, we described how we adapted our embedding propagation method, and other network embedding baselines, for commodities trend prediction.

Firstly, we must describe the dataset used for the experiments. We used a collection of news from soybean and corn related to the Brazilian market and production extracted from the website Soybean & Corn Advissor[2] and the historical prices from the "Centro de Estudos Avançados em Economia Aplicada" (CEPEA)[3]. The HIN we used contained the news as a central node, surrounded by 5 of the six components from 5W1H (what, where, who, why, and how), the date of the publication, and a trend node generated by comparing the current weeks' prices with the last weeks' price in a sliding window. We extracted the 5W1H with Hamborg et al.'s [13] method and excluded the "when" component since it was

[1]Available at: https://www.gdeltproject.org/

[2]Available at: http://soybeansandcorn.com

[3]Available at: https://www.cepea.org.br

too inconsistent for our news dataset. The trend nodes can be *big_up*, *up*, *down*, or *big_down*.

Secondly, we must describe the fine-tuning scenario for SBERT. Our proposed pipeline for fine-tuning takes advantage of the Siamese neural networks SBERT uses to generate a single output for a sentence [20]. In Equation 2, we present the setup we proposed for fine-tuning, where for every pair of nodes that contain textual data (n_i,n_j), we adjust the back-propagation error function to consider the difference between the Shared Nearest Neighbor (SNN) from the nodes in the HIN and the cosine similarity between the aggregated pair of BERT embeddings (g_{n_i},g_{n_j}).

$$MSE = \frac{1}{k}\sum_{i=1}^{k}(cos(\mathbf{g}_{n_i}, \mathbf{g}_{n_j}) - SNN(n_i, n_j))^2 \qquad (2)$$

For evaluation we collect the F1, precision, and recall scores [18] from the proposed embedding propagation method for trend prediction (TPHIN) with its fine-tuned version and the baselines: DeepWalk [17], Node2Vec [12], Metapath2Vec [10], Struc2Vec [21], LINE [23], and GCN [15]. To obtain the final prediction, we used a Long-Short Term Memory (LSTM) [24] for multiclass classification, while GCN was used in its semi-supervised inductive form.

In Table 2, we present the results for a sliding window of 24 weeks for Corn and Soybeans. From these experiments, we can observe that TPHIN and FT-TPHIN are competitive compared to the baselines for this task. We can also observe that the pre-trained achieved the best $F1$ and recall for Corn, while the FT-TPHIN was the best in all three metrics for Soybean by a significant margin.

Table 2. Average F1, precision (pre), and recall (rcl) score.

	Corn			Soybean		
Models	*F1*	**pre**	**rcl**	*F1*	**pre**	**rcl**
DeepWalk	0.24	0.24	0.26	0.23	0.23	0.26
Node2Vec	0.27	0.30	0.27	0.23	0.23	0.25
Metapath2Vec	0.22	**0.32**	0.27	0.20	0.23	0.25
Struc2Vec	0.17	0.22	0.22	0.20	0.23	0.23
LINE	0.24	0.25	0.24	0.25	0.26	0.26
GCN	0.27	**0.32**	**0.30**	0.21	0.22	0.24
TPHIN	**0.29**	0.30	**0.30**	0.21	0.22	0.24
FT-TPHIN	0.21	0.22	0.21	**0.33**	**0.34**	**0.34**

Overall, these results show us that both the embedding propagation method and the fine-tuning pipeline can achieve competitive results with the state-of-the-art in the task of commodity trend prediction. However, these experiments also show that the pre-trained and fine-tuned versions should be used in conjunction as they obtain better results in different properties for the evaluated dataset.

5 NatUKE: Natural Product Knowledge Extraction from Academic Literature

In this section, we present the last use case we applied our embedding propagation method on automatic knowledge extraction from academic literature, and the work was published in [8]. Natural products are chemical compounds generated by living organisms, contributing to as much as 67% Most newly discovered natural products are published in academic literature (i.e., papers, patents). In this work, we focus on the automatic extraction from research papers, which are unstructured text data and might not follow a single template.

In order to circumvent these challenges, we propose to apply HIN modeling and our embedding propagation to extract knowledge about natural products. More specifically, we aim to extract the following characteristics, that are described and stored in NuBBE$_{KG}$[4] for it is Knowledge Graph (KG) version: (I) compound name (rdfs:label), (II) bioactivity (nubbe:biologicalActivity), (III) species from where natural products were extracted (nubbe:collectionSpecie), (IV) collection site of these species (nubbe:collectionSite), and (V) isolation type (nubbe:collectionType). Therefore, we modeled a HIN for processing in Python with a paper's DOI as the central node that contains textual data and the characteristics we wish to extract as supporting nodes. We added other molecule attributes and BERTopic [11] topics for extra completion of the HIN.

For evaluation, we collect the *hits@k* scores [9] in different values of k as described per this rule: the scores are calculated with k values from 1 to 50 in multiples of 5, and the k is determined whenever any score surpasses 0.50. We compare the proposed embedding propagation method (EPHEN) with the baselines: DeepWalk [17], Node2Vec [12], and Metapath2Vec [10]. The final list of nodes that contain the correct characteristics of a natural product mentioned by the paper is given by cosine similarity of the node embeddings from the DOI and the desired characteristic type. We evaluated with 40% of the links to characteristics removed.

In Table 3, we present the results from this experiment. We can see that EPHEN achieves the best results for three of the five extraction tasks, while Metapaht2Vec achieves the best results for the other two. The two tasks where Metapath2Vec achieved the best results are the most challenging since there are more nodes from these types. However, EPHEN had the highest scores overall, particularly at the isolation type extraction, where it performs 56% better on average than Metapath2Vec, the second-best performer.

Overall, this use case poses a unique advantage of an embedding propagation method as it combines text data and knowledge extraction to HINs and KGs. We also show the high adaptability of the method in some knowledge extraction tasks.

[4]Available at: https://nubbe-kg.aksw.org/ontology/index.html

Table 3. Average *Hits@k* score.

Property	*k*	*DeepWalk*	*Node2Vec*	*Metapath2Vec*	**EPHEN**
Compound name	50	0.00	0.00	**0.09**	0.03
Bioactivity	5	0.10	0.03	0.13	**0.60**
Specie	50	0.27	0.25	**0.42**	0.29
Location	20	0.38	0.28	0.42	**0.55**
Isolation type	1	0.14	0.05	0.19	**0.75**

6 Conclusion

We have introduced a novel approach that combines language models and network embedding techniques to integrate textual data into HINs. Our method propagates language model embeddings and employs a regularization function to generate a unified latent space that facilitates analyzing and comparing entities and relationships within the network. Our results have been published in conferences [5, 6, 8] and journal [7], highlighting the potential of our approach for intelligent analysis and information retrieval in multimedia systems and heterogeneous data contexts.

Future work extends our proposal for evolving networks or dynamic data sources, enabling real-time analysis and prediction. Additionally, we aim to integrate our approach with large language models. While our experiments already utilized real-world datasets, we also intend to explore large-scale networks and optimize their computational efficiency.

Acknowledgments

We acknowledge CAPES for the financial support under process number 88887.513429/2020-00.

References

[1] Charu C Aggarwal. 2018. *Machine learning for text.* Springer.

[2] Nick Craswell. 2009. *Mean Reciprocal Rank.* Springer US, Boston, MA, 1703–1703. https://doi.org/10.1007/978-0-387-39940-9_488

[3] Peng Cui, Xiao Wang, Jian Pei, and Wenwu Zhu. 2018. A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering* 31, 5 (2018), 833–852.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Paulo do Carmo and Ricardo Marcacini. 2021. Embedding propagation over heterogeneous event networks for link prediction. In *2021 IEEE International Conference on Big Data (Big Data)*. 4812–4821.

[6] P do Carmo, IJ Reis Filho, and R Marcacini. 2021. Commodities trend link prediction on heterogeneous information networks. In *Anais do IX Symposium on Knowledge Discovery, Mining and Learning*. SBC, 81–88.

[7] P. do Carmo, I. J. Reis Filho, and R. Marcacini. 2023. TRENCHANT: TRENd PrediCtion on Heterogeneous informAtion NeTworks. *Journal of Information and Data Management* 13, 6 (Jan. 2023). https://doi.org/10.5753/jidm.2022.2546

[8] Paulo Viviurka Do Carmo, Edgard Marx, Ricardo Marcacini, Marilia Valli, João Victor Silva e Silva, and Alan Pilon. 2023. NatUKE: A Benchmark for Natural Product Knowledge Extraction from Academic Literature. In *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*. IEEE, 199–203.

[9] AmpliGraph Docs. 2019. Hits at n score.

[10] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 135–144.

[11] Maarten Grootendorst. 2020. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics. https://doi.org/10.5281/zenodo.4381785

[12] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 855–864.

[13] Felix Hamborg, Soeren Lachnit, Moritz Schubotz, Thomas Hepp, and Bela Gipp. 2018. Giveme5W: main event retrieval from news articles by extraction of the five journalistic w questions. In *International Conference on Information*. Springer, 356–366.

[14] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge graphs. *ACM Computing Surveys (CSUR)* 54, 4 (2021), 1–37.

[15] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*.

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[17] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 701–710.

[18] David Powers. 2008. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Mach. Learn. Technol.* 2 (01 2008).

[19] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. *Technical report, OpenAI* (2018).

[20] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3982–3992.

[21] Leonardo FR Ribeiro, Pedro HP Saverese, and Daniel R Figueiredo. 2017. struc2vec: Learning node representations from structural identity. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 385–394.

[22] Rafael G Rossi, Alneu A Lopes, and Solange O Rezende. 2014. A parameter-free label propagation algorithm using bipartite heterogeneous networks for text classification. In *Proceedings of the 29th annual acm symposium on applied computing*. 79–84.

[23] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*. 1067–1077.

[24] Ronald J Williams and Jing Peng. 1990. An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural computation* 2, 4 (1990), 490–501.

[25] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[26] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068* (2022).