Enhancing Biochemical Extraction with BFS-driven Knowledge Graph Embedding approach.

Bhushan Zope[1,*], Sashikala Mishra[1] and Sanju Tiwari[2]

[1]Symbiosis Institute of Technology, Symbiosis International (Deemed University) (SIU), Lavale, Pune 412115, India
[2]Universidade Autonoma de Tamaulipas, Mexico

Abstract

Knowledge Graph (KG) embedding is a representation of nodes and edges in lower-dimension space. It has many applications, including knowledge graph completion. Extracting the knowledge trapped in thousands of research papers in the biochemical domain is one such application. This work proposes a model that combines the Breadth-first search (BFS) technique and Word2Vec algorithms to generate the node embeddings for each node. Firstly, The knowledge graph is explored using the BFS to construct the various paths. The Word2Vec model is then trained using these paths to obtain the embeddings for the respective nodes. Results have shown that this unsupervised approach produces reasonably good knowledge embeddings. hits@50 results for edge types 'compound name' and 'specie' are 0.83 and 0.81, which are 415% and 184% better than the existing best method, respectively. For other edge types like 'bio-activity' and 'collection-site,' results are reasonably close to the best.

Keywords

Knowledge Graph Embedding Models, Natural Language Processing, Knowledge Representation,

1. Introduction

Biochemical extraction is critical in various scientific domains, including drug discovery, molecular biology, and bioinformatics [1]. It involves identifying and extracting relevant information from vast amounts of biomedical literature. The extracted information is crucial in understanding biological processes, discovering new drugs, and gaining insights into complex molecular interactions. Traditional extraction methods typically rely on manual curation or keyword-based approaches, which are time-consuming, labor-intensive, and prone to errors [2]. Furthermore, they struggle to handle biomedical data's increasing volume and heterogeneity, limiting their effectiveness in extracting comprehensive and structured knowledge.

The primary objective of this research is to enhance the efficiency and accuracy of biochemical extraction by leveraging a BFS-driven Knowledge Graph Embedding approach. By representing biochemical entities and their relationships as a knowledge graph, combined with

BiKE'23: First International Biochemical Knowledge Extraction Challenge, May 28 - Jun 1, 2023, co-located with Extended Semantic Web Conference (ESWC), Hersonissos, Greece

*Corresponding author.
†These authors contributed equally.
bhushan.zope@hotmail.com (B. Zope); sashikala.mishra@sitpune.edu.in (S. Mishra); sanju.tiwari.2007@gmail.com (S. Tiwari)
0000-0003-2636-223X (B. Zope); 0000-0002-5433-4917 (S. Mishra); 0000-0001-7197-0766 (S. Tiwari)

the power of graph embedding algorithms, this approach aims to provide a comprehensive and structured representation of biochemical knowledge. The BFS-driven technique facilitates exploring relationships within the knowledge graph, enabling efficient and effective extraction of biochemical information.

Researchers and professionals working in the fields of bioinformatics and drug discovery can gain from a more thorough and organized representation of biochemical knowledge by implementing this BFS-driven Knowledge Graph Embedding approach. This can speed up data analysis, hypothesis creation, and decision-making processes, leading to faster scientific breakthroughs and progress in the biomedical field.

Overall, this study aims to fill the gap between conventional biochemical extraction techniques and the increasing complexity of biomedical data, providing a promising method for obtaining important information from the vast amount of information already available and enabling researchers to pursue novel insights and discoveries.

2. Related work

The term "embedding" is popular right now. The number of studies on the subject has exploded in recent years, particularly those that deal with word embeddings. Word embeddings are vector representations of a word that maintain the word's meaning and are generally in a Euclidean space. Following the introduction of the Word2Vec model [3], word embeddings have gained enormous popularity. Word2Vec and other language models have been extended to graph structures, as demonstrated by DeepWalk [4]. To anticipate nearby words in a text, Word2Vec trains a neural network. The sentences of the text are composed of the sequence of nodes visited during walks. Then, word embedding models, such Word2Vec, may be used to find the embedding of nodes by treating them as words in the sentences. Although DeepWalk employs a random uniform random walk, each network has unique connection patterns that must be considered when creating node representations. Based on this understanding, node2vec[5] introduced a more complex random walks technique that outperformed DeepWalk and can be more readily modified to various graph connection patterns.

The challenge with representation learning is the variety of node and link types, which makes it difficult to use traditional network embedding approaches. The metapath2vec model [6] uses a heterogeneous skip-gram model to conduct node embeddings after establishing meta-path-based random walks to create a node's heterogeneous neighborhood. In numerous heterogeneous network mining tasks, the metapath2vec model can beat state-of-the-art embedding models and identify the structural and semantic linkages between different network objects.

The majority of embedding techniques used today focus on network topology. Still, EPHEN [7] uses a language model-based embedding propagation method that uses both textual information about events and the complex relationships between that event & a low-dimensional vector space. This results in the possibility of gradual and adaptive embedding updation.

Our research builds on these earlier investigations and uses a BFS-driven Knowledge Graph Embedding technique to improve biochemical extraction. The proposed method aims to capture biological entities' structural and semantic context by combining the benefits of semantic embedding and graph traversal techniques. This comprehensive strategy has the potential

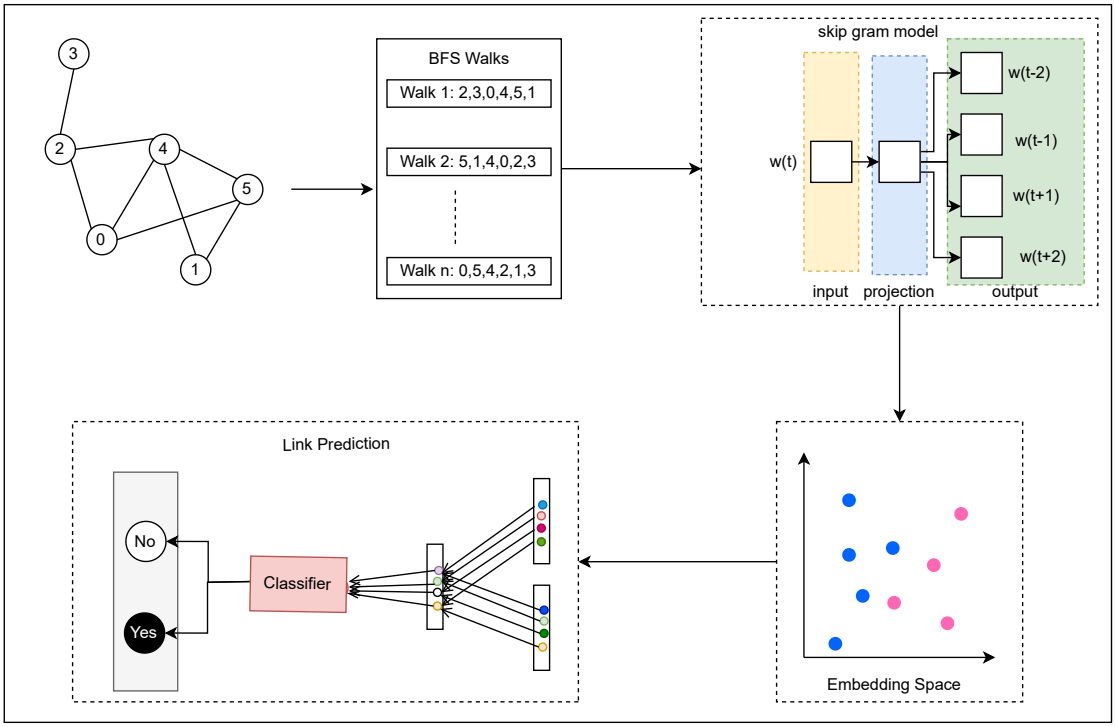


Figure 1: Methodology

to enhance the extraction process's precision, effectiveness, and interpretability, allowing researchers to draw important conclusions from vast biological information repositories.

3. Proposed BFS-driven Knowledge Graph Embedding Approach

3.1. Overview of the Approach

Word2Vec is a popular algorithm for creating word embeddings in natural language processing. It represents the words in the form of numerical vectors. It looks at the surrounding words and finds out the context of the word. This way, it learns the relationships and meanings of the words.

General idea of the proposed method, as shown in Figure 1, is to utilize the Word2Vec approach for node embedding generations. However, Word2Vec relies on the sentences to find the word embeddings. Hence, in the knowledge graph context, the sequence of nodes appearing in a particular path can be treated as a sentence. Multiple such paths can then be given to Word2Vec for node embedding generation.

3.2. Dataset

The dataset given for the BiKE challenge [8] is used for experimentation. The dataset [9] was generated by extracting information from peer-reviewed scientific articles. These articles served

as the primary source of information for natural product extraction. It focuses on five NuBBE properties: Compound Name, Bioactivity, Species from Extraction, Collection Site, and Isolation Type. Figure 2 shows each property type's number of distinct values. It consists of four different split rations viz. 20/80, 40/60, 60/40, and 80/20 percent for testing and training respectively. For each percentage, ten randomly split knowledge graphs were given.

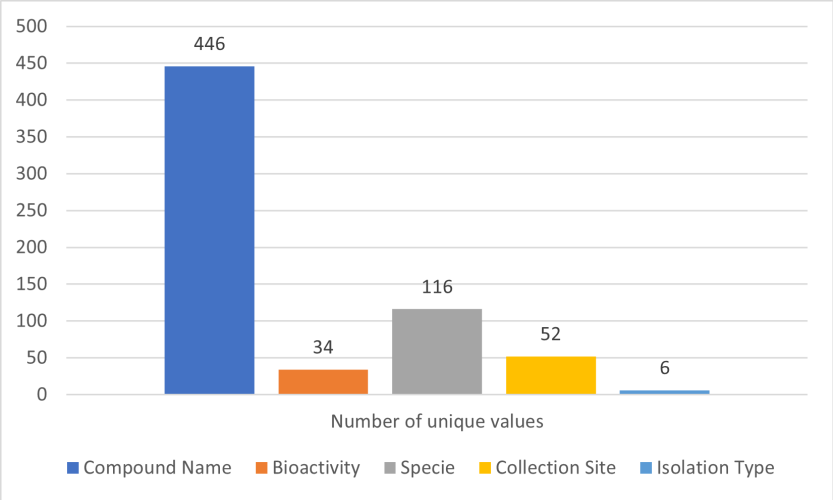


Figure 2: Diversity in values

3.3. BFS Exploration for Graph Embedding

Breadth First Search (BFS) is a simple algorithm to explore and navigate through a graph or a network. Here we have traversed the knowledge graph in an all-source-BFS manner. BFS explores all the nodes at the same level from the starting point before moving on to the nodes from the level one step further away. To generate the node sequence, the proposed method uses the BFS approach.

BFS ensures that it explores the graph layer by layer. It guarantees that nodes at a shallower level (closer to the starting point) are visited before moving on to nodes at deeper levels. This leads to structural awareness, which is the main advantage of BFS. As shown in Figure 3, neighborhood nodes are closer in the sequence, contributing more to the node's context in the knowledge graph.

However, BFS paths suffer from very important problems. The nodes adjacent to the sequence may not be immediate neighbors of each other. For example, as shown in Figure 3, nodes 2, 1, and 3 are adjacent in the given BFS sequence but are a few hops away. If only nodes 2 and 3 are considered while finding the embedding for node 1, then it won't be appropriate. To mitigate this problem, a large window size of 10 is used during embedding generation. The large window size enables the model to handle long-range dependencies, resulting in a meaningful node representation. Additionally, Five walks are constructed from each node by visiting four neighbors in each BFS iteration for four iterations. This allows more opportunities for the node

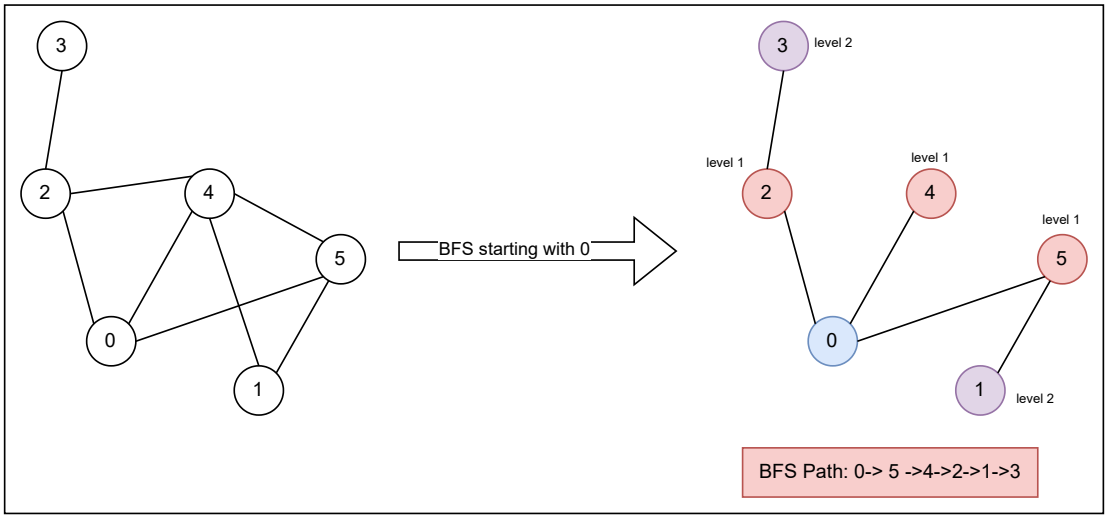


Figure 3: BFS explores the graph level-wise leading to structural awareness

to get surrounded by relevant, adjacent nodes. Thus these constructed walks capture the context of the neighborhood. These walks are then used for training the Word2Vec model, giving the embeddings for each node.

4. Results and Discussions

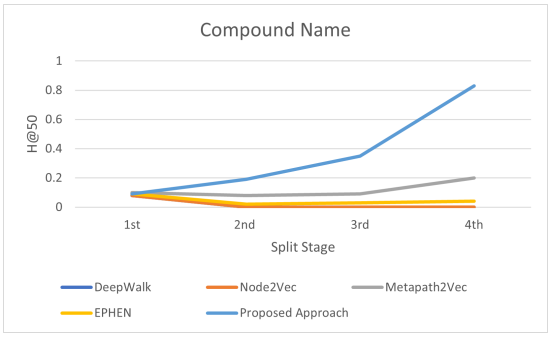
The evaluation was performed using the official BiKE challenge benchmark NatUKE [9], and the results[1] are listed in Table 1. Results for the proposed method are compared to DeepWalk, Node2Vec, Metapath2Vec, and EPHEN, which are taken from [9].
We have used the property prediction task and hits@k performance metric on the dataset for experimentation. Not all properties have the same unique values. Therefore, a different value for k in hits@k is used for different property predictions.

Evidently, the proposed method gave excellent results for 'Compound Name' and 'Specie' properties. For the 'Compound Name' property, hits@50 for 1st evaluation stage is 0.9, slightly less than the results for Metapath2vec. However, results improved progressively in the subsequent evaluation stages, with 415% better results than the previous. Similarly, the Results for 'Specie' are the best among all the other four models for all the evaluation stages.

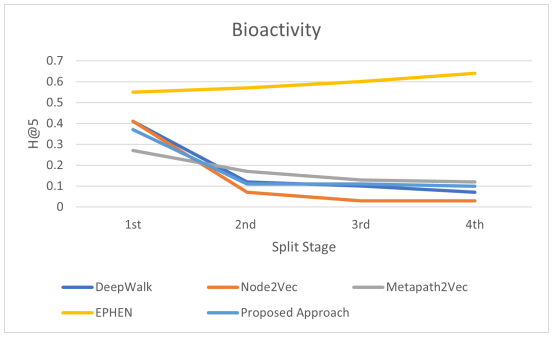
Furthermore, the results for the other two properties, i.e., 'Bioactivity' and 'Collection Site,' are also motivating. However, results for EPHEN are distinguishably apart from all other methods. The results for the proposed method are similar to the remaining methods.

On the other hand, results for the 'isolation type' property are not encouraging and very similar to the Node2vec method, which is very similar to the proposed method. There are only six unique values for the 'Isolation Type' property compared to 446 for the 'Compound Name' property. Since there are few unique values, one value may appear with different types of

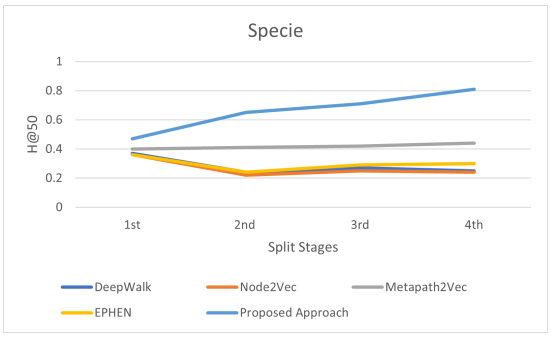
[1]Code and result files are kept at the GitHub repository: https://github.com/bhushan-zope/BiKE.



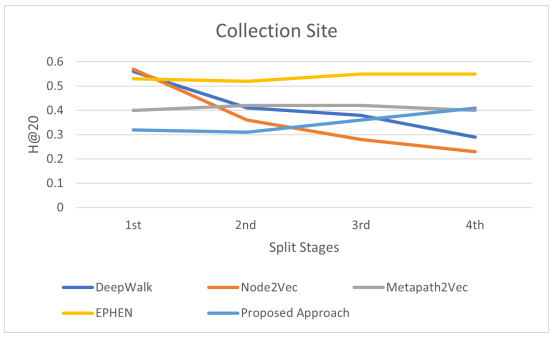
(a) hits@50 result for 'Compound Name' property



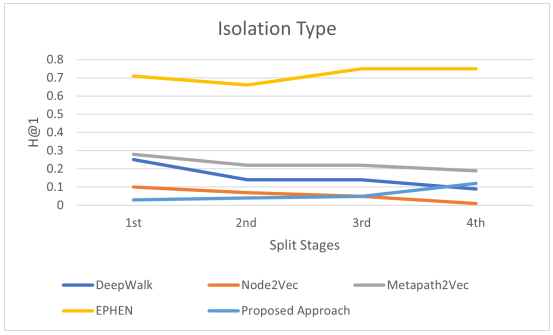
(b) hits@5 result for 'Bioactivity property



(c) hits@50 result for 'Specie' property



(d) hits@20 result for 'Collection Site' property



(e) hits@1 result for 'Isolation Type' property

Figure 4: Results comparison for compound name (C) with hits@50, bioactivity (B) with hits@5, specie (S) with hits@50, collection site (L) with hits@20, and isolation type (T) with hits@1 performance metric.

nodes in the path. Making it very difficult to discriminate the context. Thus, limited diversity in distinct values contributes to relatively poor outcomes. Moreover, because of the diversity in distinct values, each value appears with a specific set of nodes in the path, resulting in a precise understanding of context.

As shown in Figure 2, bioactivity, collection site, and isolation type has very limited diversity. Results for these properties are exceptionally well for the EPHEN method. Whereas results for compound name and specie properties, which have more unique values, are outstanding for the proposed method. It follows that the suggested approach is better suited to properties with

more distinct values, whereas EPHEN is better suited to properties with less distinct values.

Table 1
Results comparison for compound name (C) with hits@50, bioactivity (B) with hits@5, specie (S) with hits@50, collection site (L) with hits@20, and isolation type (T) with hits@1 performance metric. The figures in bold are the best results.

Property	DeepWalk			
	1st	**2nd**	**3rd**	**4th**
C	0.08	0.00	0.00	0.00
B	0.41	0.12	0.10	0.07
S	0.37	0.24	0.27	0.25
L	0.56	0.41	0.38	0.29
T	0.25	0.14	0.14	0.09

Property	Node2Vec			
	1st	**2nd**	**3rd**	**4th**
C	0.08	0.00	0.00	0.00
B	0.41	0.07	0.03	0.03
S	0.36	0.22	0.25	0.24
L	**0.57**	0.36	0.28	0.23
T	0.10	0.07	0.05	0.01

Property	Metapath2Vec			
	1st	**2nd**	**3rd**	**4th**
C	**0.10**	0.08	0.09	0.20
B	0.27	0.17	0.13	0.12
S	0.40	0.41	0.42	0.44
L	0.40	0.42	0.42	0.40
T	0.28	0.22	0.19	0.19

Property	EPHEN			
	1st	**2nd**	**3rd**	**4th**
C	0.09	0.02	0.03	0.04
B	**0.55**	**0.57**	**0.60**	**0.64**
S	0.36	0.24	0.29	0.30
L	0.53	**0.52**	**0.55**	**0.55**
T	**0.71**	**0.66**	**0.75**	**0.75**

Property	Proposed Approach			
	1st	**2nd**	**3rd**	**4th**
C	0.09	**0.19**	**0.35**	**0.83**
B	0.37	0.11	0.11	0.10
S	**0.47**	**0.65**	**0.75**	**0.81**
L	0.32	0.31	0.36	0.41
T	0.03	0.04	0.05	0.12

5. Conclusion

This research paper presented an approach for enhancing biochemical knowledge extraction through a BFS-driven Knowledge Graph Embedding method. Our BFS-driven Knowledge Graph Embedding approach offered several advantages. The knowledge graph is traversed using a Breadth-First Search algorithm to capture context and relationships between biochemical entities. The results of our experiments showcased the potential of our BFS-driven Knowledge Graph Embedding approach.

References

[1] Q.-C. Bui, P. M. Sloot, A robust approach to extract biomedical events from literature, Bioinformatics 28 (2012) 2654–2661. doi:10.1093/bioinformatics/bts487.

[2] L. Tari, S. Anwar, S. Liang, J. Hakenberg, C. Baral, Synthesis of pharmacokinetic pathways through knowledge acquisition and automated reasoning, in: Biocomputing 2010, World Scientific, 2010, pp. 465–476.
[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, ArXiv abs/1310.4546 (2013).
[4] B. Perozzi, R. Al-Rfou, S. S. Skiena, Deepwalk: online learning of social representations, Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (2014).
[5] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016).
[6] Y. Dong, N. Chawla, A. Swami, metapath2vec: Scalable representation learning for heterogeneous networks, Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2017).
[7] P. V. D. Carmo, R. M. Marcacini, Embedding propagation over heterogeneous event networks for link prediction, 2021 IEEE International Conference on Big Data (Big Data) (2021) 4812–4821.
[8] 2023. URL: https://aksw.org/bike/.
[9] P. V. do Carmo, E. Marx, R. Marcacini, M. Valli, J. V. S. e Silva, A. Pilon, NatUKE: A Benchmark for Natural Product Knowledge Extraction from Academic Literature, in: 17th IEEE International Conference on Semantic Computing, IEEE, 2023.