Postprint: Research on Agricultural Knowledge Graph Representation Model Based on HARP Framework

Authors: Chen Caiming, Feng Jianzhong, Bai Linyan, Wang Jian, Xie Nengfu, Zou Jun

Date: 2024-04-03T00:00:00+00:00

Abstract

[Purpose/Significance] As the data scale of agricultural knowledge graphs continues to expand, the complexity of nodes and relationships within the graphs increases progressively, presenting novel challenges for their training and representation. Against this backdrop, investigating approaches to preserve knowledge graph structure while concurrently reducing resource consumption and accelerating embedding speed holds substantial research and application significance. [Method/Process] To address this challenge, this study proposes a hierarchical representation model for agricultural knowledge graphs grounded in the HARP framework. Capitalizing on the hierarchical nature of agricultural knowledge graphs, the model adopts an improved relationship-path-based random walk strategy that effectively retains the hierarchical and asymmetric relationship structures of nodes within the graph. [Results/Conclusion] 1) In comparison with the HARP framework, the HRWP model employing LEIDEN exhibits superior preservation of spatial structure and achieves expedited convergence speed; 2) The training time of the fusion model incorporating HRWP is fundamentally less than the sum of the individual training times, with minimal impact on the time complexity of the original algorithm; 3) Traditional algorithms integrated with HRWP demonstrate an average improvement of 2% across metrics, with non-neural network models manifesting particularly pronounced enhancements. In conclusion, the model is capable of accurately representing agricultural knowledge graphs while effectively curtailing training time.

Full Text

Preamble

Authors: Chen Caiming[1], Feng Jianzhong[1]*, Bai Linyan[2,3], Wang Jian[1], Xie Nengfu[1], Zou Jun[1]
Affiliations:
1. Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081
2. Key Laboratory of Digital Earth, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094
3. International Research Center of Big Data for Sustainable Development Goals, Beijing 100094

Abstract:
[Objective/Significance] As agricultural knowledge graphs continue to grow in scale, the complexity of nodes and relationships increases exponentially, posing new challenges for training and representation. In this context, exploring methods to accelerate embedding while preserving graph structure and reducing resource consumption holds significant research and practical importance. [Method/Process] To address this challenge, this study proposes a hierarchical representation model for agricultural knowledge graphs based on the HARP framework. Leveraging the hierarchical characteristics of agricultural knowledge graphs, the model employs an improved random walk strategy based on relational paths to effectively preserve both the hierarchical structure and asymmetric relationship patterns within the graph. [Results/Conclusions] 1) Compared with the HARP framework, the HRWP model using LEIDEN better preserves spatial structure and achieves faster convergence; 2) The fusion model employing HRWP requires training time substantially less than the sum of individual model training times, with minimal impact on the original algorithm' s time complexity; 3) Traditional algorithms combined with HRWP show an average improvement of 2% across metrics, with non-neural network models demonstrating significant enhancement. In summary, the proposed model can accurately represent agricultural knowledge graphs while effectively reducing training time.

Keywords: knowledge graph; random walk; representation learning; HARP framework
Classification: TP391.1

---

1 Introduction

Knowledge graphs have emerged as a critical infrastructure for intelligent knowledge services due to their superior information organization and reasoning capabilities. This triple-based storage structure, typically represented as (h, r, t) where h and t denote head and tail entities respectively, is widely applied across knowledge engineering domains. However, such structures often suffer from data sparsity issues, yielding suboptimal performance in semantic computation and relational reasoning. In practice, knowledge representation learning is employed to map entities and relations into low-dimensional dense vectors, thereby enhancing knowledge acquisition, fusion, and inference performance.

As data organization technologies evolve and knowledge graph scales expand, traditional representation algorithms demand not only extensive training time but also consume massive memory resources. Current rapid learning models for large-scale graphs often overemphasize local features while neglecting long-range global dependencies, resulting in embeddings that fail to capture crucial semantic hierarchies. To enable efficient representation learning for large-scale knowledge graphs while preserving conceptual hierarchical structures, this paper proposes a fast hierarchical random walk learning model based on the Hierarchical Representation Learning for Networks (HARP) framework. The model utilizes hierarchical random walks for initial knowledge graph representation and addresses relational learning challenges by treating relation embedding as separate embeddings for head and tail entities. It employs frustrated random walks to effectively prevent node sinking and capture relationship asymmetry, enabling better integration with other learning models.

The contributions of this work are threefold: First, we extend the application scenarios of random walk-based graph embedding algorithms to agricultural knowledge graphs. Second, we optimize existing models' deficiencies in relational learning through hierarchical walks and improved sampling strategies. Third, we validate the performance of the proposed model through quantitative experiments on real-world datasets.

Experimental results demonstrate that our model can be effectively integrated with representation learning methods including TransE and RotatE, providing superior representations for downstream tasks across various application scenarios.

---

2 Background

Node positional representations can reflect graph structural information, enabling applications such as information propagation, project recommendation, and topic search. As data science advances alongside big data and cloud computing technologies, numerous knowledge graph embedding models have been proposed for different scenarios, generally categorized into translation-based models, semantic matching models, and neural network-based models.

Translation models represent entities as vectors and relations as translation rules connecting head and tail entities. TransE models relations as translation vectors between head and tail entities, while subsequent models like TransF and TransA relax the $h + r \approx t$ assumption for greater flexibility. RotatE extends this approach by modeling entities and relations in complex space as rotational transformations, demonstrating effectiveness in handling asymmetric relations and relation composition.

Semantic matching models employ similarity-based scoring functions to measure fact plausibility by matching latent semantics in vector space. RESCAL uses vectors for entities and matrices for relations, capturing interactions within triples through custom scoring functions. DistMult simplifies RESCAL by restricting relation matrices to diagonal matrices, while ComplEx further extends this by introducing complex space and asymmetric scoring functions to better model non-symmetric relationships.

Neural network-based approaches have also been integrated into knowledge graph embedding. R-GCN first applies Graph Convolutional Networks (GCN) to relational data, while SACN introduces weighted convolutional networks. ConvE utilizes 2D convolutional operations, and WGCN employs weighted GCN architectures. In practical applications, Google' s Knowledge Vault project combines latent factor models with random walk models for knowledge evaluation tasks, while the Never-ending Language Learning project uses random walk-based path ranking algorithms for relational reasoning. Random walk models offer parallel computation capabilities better suited for large-scale knowledge graphs, meeting practical requirements for rapid training.

---

3 Feasibility Analysis

Hierarchical embedding effectively preserves knowledge graph hierarchical information, which is crucial for entity relation prediction and knowledge graph completion tasks. By mapping entities and relations to low-dimensional spaces, embedding reduces computational complexity and enhances large-scale graph processing efficiency. This dimensionality reduction increases representational intuitiveness and interpretability.

Agricultural knowledge inherently contains hierarchical concepts such as biological taxonomy, ecosystems, and organizational structures, exhibiting natural hierarchical organization. As illustrated in [Figure 2: see original paper], the constructed agricultural knowledge graph demonstrates clear hierarchical characteristics. Empirically, existing agricultural knowledge graphs exhibit clustering coefficients significantly higher than random graphs (0.357 vs. 0.297, 0.307, and 0.300 for three random graph controls), indicating nodes tend to form tight clusters. The average modularity of agricultural knowledge graphs reaches 0.357, substantially exceeding random graph counterparts (0.014, 0.040, and 0.027), demonstrating more pronounced community structures.

[Figure 3: see original paper] illustrates the relationship between clustering coefficients and node proportions across different datasets. Random graphs show nodes predominantly distributed in low clustering intervals, whereas agricultural knowledge graphs exhibit more uniform distribution across clustering coefficients, indicating significant clustering effects. Node degree, defined as the number of edges directly connected to a node, comprises in-degree (edges pointing to the node) and out-degree (edges originating from the node). The clustering coefficient $C(i)$ calculates the ratio of existing edges to possible edges among node i' s neighbors: $C(i) = \frac{2E(i)}{k(k-1)}$, where $E(i)$ is the number of edges among neighbors and k is node i' s degree. This hierarchical characteristic suggests the layering effect is a global property rather than arising from special nodes.

---

4 Model Introduction

This section details the improved algorithm within the network hierarchical representation learning paradigm, explaining how to leverage multi-dimensional structural relationships in knowledge graphs for rapid learning while preserving semantic structure and relational integrity.

4.1 Network Hierarchical Representation Learning Model

Hierarchical representation learning comprises three steps: First, **Graph Coarsening** reduces the original graph G to a series of progressively smaller graphs $G_0, G_1, ..., G_L$. Second, **Representation Learning** embeds the smallest graph G_L where $|V|$ and $|E|$ are minimal, enabling rapid learning of quality representations. Third, **Representation Refinement** iteratively obtains embeddings for each level G_i by using embeddings from G_{i+1} as initial vectors for nodes in G_i, then continuing with the embedding function.

4.2 Improved Random Walk Algorithm

Random walk is a process that starts from a given node and transfers to neighboring nodes with certain probabilities until reaching a termination point. Simple random walk defines transition probabilities as $P_{ij} = \frac{\omega_{ij}}{D_i}$, where D_i is node i' s degree. However, this symmetric structure cannot represent asymmetric relationships in knowledge graphs, causing embedding biases for important nodes.

Consider the example in [Figure 4: see original paper] (partial crop planting graph). Simple random walks may terminate prematurely at nodes without other neighbors, artificially inflating hit probabilities. To address this, we improve the sampling model' s transition matrix B_{ij}. The improved algorithm ensures that higher-degree nodes are more likely to be visited by defining transition probabilities asymmetrically: the probability of moving from i to j is $\frac{\omega_{ij}}{D_i}$, while the acceptance probability at j is $\frac{\omega_{ji}}{D_j}$. This blocking mechanism allows the starting point to be rejected with probability $1 - \sum_{r \in R_v} P(v, r, x)$, where R_v is the set of relations associated with node v.

4.3 LEIDEN Clustering Algorithm

LEIDEN is a non-overlapping community clustering algorithm based on multi-level modularity optimization, building upon the LOUVAIN algorithm. A good partition exhibits high internal similarity within communities and low similarity between communities. LEIDEN initializes each node as a separate community, then attempts to assign node i to neighboring communities, calculating modularity gains ΔQ. The algorithm selects the node with maximum gain to join adjacent communities, refines partitions, updates ring weights based on internal node weights, and iterates until no further improvements occur, as illustrated in [Figure 5: see original paper].

The modularity Q is defined as: $Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \gamma\frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$, where c_i represents communities, k_c is the sum of degrees within communities, m is the total number of edges, and γ is a resolution parameter. LEIDEN' s advantages include: customizable community definitions, applicability to directed weighted graphs, guaranteed community connectivity, and generation of a community partition tree through iteration.

4.4 Hierarchical Embedding Based on Improved Random Walk

To effectively learn latent knowledge while addressing issues of high-dimensional information loss and asymmetric representation during sampling, we define the objective function $f(G) : V \to \mathbb{R}^{|V|\times d}$ where $d \ll |V|$. The model employs a hybrid coarsening algorithm combining Edge Collapsing and Star Collapsing to hierarchically reduce large graphs.

Edge Collapsing merges any two edges connected to the same node into a single edge, with merging order having minimal impact on results. **Star Collapsing** preserves second-order similarity by merging nodes around central nodes with similar structural characteristics. The model first performs star collapsing, then edge collapsing, until reaching a sufficiently small graph, as shown in [Figure 6: see original paper].

Using LEIDEN' s intermediate variable community tree (Dendrogram), we implement graph sampling through hierarchical representation learning. For each partition level, the node with highest degree within a community is designated as the center node, progressively merging until node count falls below a threshold.

The **Hierarchical Random Walk with Path (HRWP)** model processes the blocking mechanism by converting rejection probabilities into self-loops with transition probability $1-\sum_{r\in R_v} P(v, r, x)$. Except for top-level nodes, all nodes transfer upward with this probability, while top-level nodes remain as self-loops, as illustrated in [Figure 7: see original paper].

For relation path biased sampling, given a relation sequence $P = T_0 \to \ldots \to T_l$ where $T_i = \text{range}(r_i) = \text{domain}(r_{i+1})$, the walk probability along path r is defined to ensure semantic proximity between target and current nodes while capturing asymmetric relationships. This approach prevents neglecting sporadic relations and ensures flexible representation.

The complete HRWP algorithm is presented in [Figure 8: see original paper], with the overall model schematic shown in [Figure 9: see original paper]. For a triple (h, r, t), the head entity embedding emb_h is represented as $h + r_h$, where r_h is the relation embedding result. This flexible representation allows seamless integration with existing algorithms.

5 Experiments

Experiments were conducted on a system with Intel(R) Core(TM) i7-11800H processor and NVIDIA GeForce RTX 3070 GPU. We evaluated performance using FB15K-237 and WN18RR datasets, with statistics provided in . The model code was implemented in Python using the PyTorch framework.

5.1 Comparative Analysis of Clustering Results

To visually demonstrate the structural preservation of HRWP' s coarsening algorithm, we visualized multi-level representations for different datasets in two dimensions ([Figure 10: see original paper]). Results show FB15K-237 exhibits stronger central aggregation than WN18RR. LEIDEN better preserves spatial structure and serves as a superior foundation for subsequent layer layouts compared to the force-directed algorithm used in original HARP.

[Figure 11: see original paper] illustrates the effectiveness of the hybrid coarsening method, showing that edge fusion converges faster than node fusion, consistent with the model' s combined edge-star collapsing characteristics. [Figure 12: see original paper] depicts modularity changes across hierarchy levels, revealing LEIDEN' s rapid convergence to maximum modularity as a greedy algorithm, though with stability limitations.

5.2 Comparative Analysis of Training Time

[Figure 13: see original paper] compares convergence times between HRWP and baseline models. HRWP alone requires significantly less time than baseline models on both datasets. Fusion models employing HRWP framework train faster than the sum of individual model times, demonstrating HRWP' s effectiveness as a pre-training method that reduces training time without compromising convergence.

To analyze sampling algorithm impact on training speed, we conducted experiments on agricultural graphs with node counts ranging from 100 to 100,000 while controlling average degree. [Figure 14: see original paper] shows HRWP' s runtime trend matches the frustrated random walk algorithm, with graph coarsening and prolongation overhead being negligible. For large-scale graphs, HRWP minimally impacts original algorithm time complexity.

5.3 Comparative Analysis of Representation Effectiveness

We compared HRWP against other algorithms on FB15k-237 and WN18RR datasets using Mean Reciprocal Rank (MRR) and Hits@k metrics. MRR, defined as the arithmetic mean of reciprocal ranks across test triples, measures overall performance, while Hits@k represents the ratio of triples ranked in top k positions.

Results in show that on FB15K-237, TransE-HRWP generally matches or outperforms original TransE across all metrics. When compared with neural methods, HRWP-integrated models achieve an average 2% improvement. Non-neural models show particularly significant enhancement, with TransE-HRWP approaching RotatE performance in handling asymmetric relations.

To validate performance on agricultural knowledge graphs, we conducted relation prediction experiments (). Under equivalent convergence conditions, RotatE and RotatE-HRWP-RW require over 60 minutes training time, while RotatE-HRWP-H and RotatE-HRWP show reduced training time. This confirms the hierarchical representation framework' s effectiveness in accelerating training while maintaining accuracy.

Cross-validation experiments () demonstrate that HRWP-H (hierarchical embedding only) maintains performance comparable to original models, while HRWP-RW (random walk only) improves performance by approximately 2%. The full RotatE-HRWP model combines both benefits, confirming that hierarchical embedding accelerates training while improved random walk enhances representation quality.

References

[1] ZHAO Z X, WANG Y T, TIAN J T, et al. A novel algorithm for community discovery in social networks based on label propagation[J]. Journal of computer research and development, 2011, 48(S3): 8-15.

[2] SUN T, DING P, HUANG Y W, et al. Review on the application and development strategies of text mining in agriculture knowledge services[J]. Journal of library and information science in agriculture, 2021, 33(1): 4-16.

[3] CAO S J, WU Y B, WEI J Z, et al. History, schools and trend in knowledge map: Investigation and visualization based on SSCI and CSSCI[J]. Journal of library science in China, 2015, 41(5): 16-34.

[4] XU Y W, ZHANG H J, CHENG K, et al. Comprehensive survey on knowledge graph embedding[J]. Computer engineering and applications, 2022, 58(9): 30-50.

[5] ZHOU C, LIU Y Q, LIU X F, et al. Scalable graph embedding for asymmetric proximity[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 31(1): 456-545.

[6] LI E Z, LE Z Y. Frustrated random walks: A faster algorithm to evaluate node distances on connected and undirected graphs[J]. Physical review E, 2020, 102: 052135.

[7] QIAO L, JIANG L, HAN M, et al. Hierarchical random walk inference in knowledge graphs[C]//The 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. Pisa, Italy: Association for Computing Machinery, 2016.

[8] SUN G F, WU L, LIU Q, et al. Recommendations based on collaborative filtering by exploiting sequential behaviors[J]. Journal of software, 2013, 24(11): 2721-2733.

[9] BORDES A, USUNIER N, GARCIA-DURÁN A, et al. Translating embeddings for modeling multi-relational data[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. New York: ACM, 2013: 2787-2795.

[10] FENG J, ZHOU M T, HAO Y, et al. Knowledge graph embedding by flexible translation[J]. arXiv: 1505.05253, 2015.

[11] XIAO H, HUANG M L, HAO Y, et al. TransA: An adaptive approach for knowledge graph embedding[J]. arXiv: 1509.05490, 2015.

[12] SUN Z Q, DENG Z H, NIE J Y, et al. RotatE: Knowledge graph embedding by relational rotation in complex space[J]. arXiv: 1902.10197, 2019.

[13] NICKEL M, TRESP V, KRIEGEL H P. A three-way model for collective learning on multi-relational data[C]//Proceedings of the 28th International Conference on Machine Learning. New York: ACM, 2011: 809-816.

[14] YANG B, YIH W T, HE X D, et al. Embedding entities and relations for learning and inference in knowledge bases[J]. arXiv preprint arXiv:1412.6575, 2014.

[15] SCHLICHTKRULL M, KIPF T N, BLOEM P, et al. Modeling relational data with graph convolutional networks[C]//European semantic web conference. Cham: Springer, 2018: 593-607.

[16] SHANG C, TANG Y, HUANG J, et al. End-to-end structure-aware convolutional networks for knowledge base completion[J]. Proceedings of the AAAI conference on artificial intelligence, 2019, 33: 3060-3067.

[17] DETTMERS T, MINERVINI P, STENETORP P, et al. Convolutional 2D knowledge graph embeddings[C]//Proceedings of the AAAI conference on artificial intelligence, 2018, 32(1): 241-249.

[18] VOORHEES E. The TREC-8 question answering track report[C]. Gaithersburg: Proceedings of TREC-8, 2000.

[19] HERLOCKER J L, KONSTAN J A, TERVEEN L G, et al. Evaluating collaborative filtering recommender systems[J]. ACM transactions on information systems, 2004, 22(1): 5-53.

[20] CHEN H C, PEROZZI B, HU Y F, et al. HARP: Hierarchical representation learning for networks[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1): 336-341.

[21] TRAAG V A, WALTMAN L, VAN ECK N J. From Louvain to Leiden: Guaranteeing well-connected communities[J]. Scientific reports, 2019, 9: 5233.

[22] TOUTANOVA K, CHEN D, PANTEL P, et al. Representing text for joint embedding of text and knowledge bases[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 1499-1509.

[23] XU Z L, SHENG Y P, HE L R, et al. Review on knowledge graph techniques[J]. Journal of university of electronic science and technology of China, 2016, 45(4): 589-606.

[24] LI X F, XIAO S, LIU X Y, et al. A brief analysis of the classification retrieval of Chinese national standards for agriculture[J]. China standardization, 2020(7): 67-71, 75.

[25] DONG X, GABRILOVICH E, HEITZ G, et al. Knowledge vault: A web-scale approach to probabilistic knowledge fusion[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM, 2014.

[26] MITCHELL T, FREDKIN E. Never-ending language learning[C]//2014 IEEE International Conference on Big Data (Big Data). Piscataway, New Jersey: IEEE, 2015: 1.

Note: Figure translations are in progress. See original paper for figures.