The Application of Graph Embedding Based on Random Walk

Zeyue Zhang

Department of statistics, Rutgers University, New Jersey, United State

zz466@rutgers.edu

Abstract. In the historical process of scientific development, computers have a lofty position, and in recent years, graph embedding algorithms and models are one of the most popular subjects. A large number of similar data structures are indistinguishable by humans, but graph embedding can quickly compare and analyze these data structures. Existing research on random walk-based graph embedding methods is very rich. In order to summarize and classify the status quo of the more mature classical models and compare and integrate them, many different classical models are discussed in this paper. Based on different models, the problems solved, algorithm ideas, strategies, advantages, and disadvantages of the models are discussed in detail, and the application performance of some models is evaluated. DeepWalk model, Node2Vec model, HARP model are three graph embedding models based on the classical random walk model. Calculations for different data can occur by generating different node sequences. The three most important models in attribute random walk models are TriDNR model, GraphRNA model and FEATHER model. The model that only targets the information data in the shallow network is no longer suitable for the rapidly developing network. Attribute random walk models can handle data in deeper networks. At the end of this paper, the full text is summarized and the future prospect of this field is made.

Keywords: Graph embedding, Random walk, DeepWalk model, TriDNR model.

1. Introduction

In recent years, graph form statistics have vied a crucial role in several systems, and graph analysis has aroused widespread interest within the technical information and connected applications of PCs. As a longtime information store, graph evaluation graphs not solely support a lot of convenient storage and more economical access to relative information between interacting entities, however additionally play an important role in today' system administration tasks. Machine domain tasks use graphs as feature statistics to amass and change access to find new models. Social networks, linguistics, biology (protein-protein networks) and recommender systems: data in these and related fields are often simply modeled as graphs showing connections, interactions (e.g. edges) between individual entities (e.g. nodes) capture. From one perspective, the task of graph truth analysis is whether high-dimensional non-Euclidean statistics of graph shapes are often simply encoded as low-dimensional function vectors [1]. Graph-based comprehension discovers an applied mathematics thanks to plug information into a system domain model. Due to the on top of issues, a completely unique approach to the graph domain to programmatically learn representations of structural statistics (graph embedding) has attracted plenty of interest.

The article introduces and explains the development of graph embedding models today in detail from the three parts of Classical random walk model, Attribute walk model, and Method summary. The classical random model is the basis of the graph embedding model, but after the in-depth development of research, a new attribute random walk model has emerged. Many new methods have also emerged based on these different random walk model graph embedding.

2. Classical Random Walk Model

2.1. Classical random walk model

The random processes that support the principles of graph embedding algorithms are involving word vector models in spoken language processing. A completely different stochastic process

technique is employed to derive a sequence of random walk nodes to create a corpus. Therefore, the Skip-Gram model or its variants are wont to generate graphs. Vector Embedding The random walk-based graph embedding model involves affordable performance, particularly within the case of very massive graphs, the random walk technique can preserve the structural options of the graph and generate graphs that may part replicate nodes [2]. Attribute Embedding Vector standard random walk-based graph embedding models be 2 categories: classical random method models and attribute walk models [3].

2.1.1 DeepWalk model

The DeepWalk model is one of the classic random walk models. Its first step is to generate a sequence of network graph vertices, which is done through random walks [4]. Fig. 1 is an example of a sequence of network graph vertices.

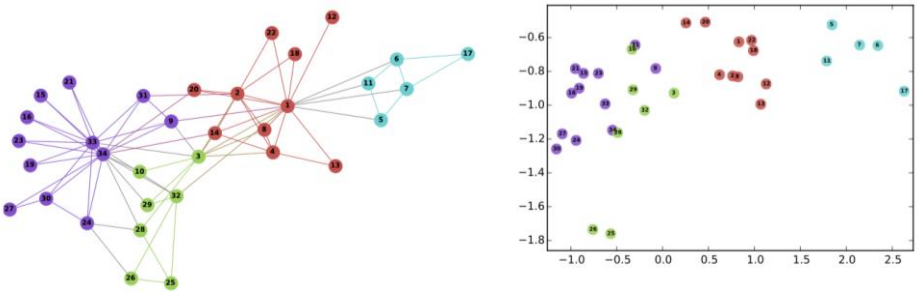

Figure 1. Vertex sequence

But only the vertex sequence can't get the model we need, so the network graph vertex sequence needs to be trained in the skip-gram language model. The vector of each vertex in the network can form a matrix $\Phi \in R^{(|V| \times d)}$. The d in the formula represents the dimension of the embedding space. Mapping nodes in the embedding space is a way to store node information and community structure. In practical applications, the cost of training and computing directly using the model is very high. The shorter paths traversed by the nodes aggregated by random walks can make the training process faster and save costs. Converting the vertex set to a binary tree is one way [4-5]. Embedding DeepWalk for both small graphs and large graphs can be done well. This is based on the condition of unweighted graphs.

2.1.2 Node2Vec model

Node2vec is an algorithmic framework. It utilizes the features of nodes in the network to learn the node-to-feature correspondence in the position space to retain the most nodes [6]. Fig. 2 shows the process of Node2Vec.

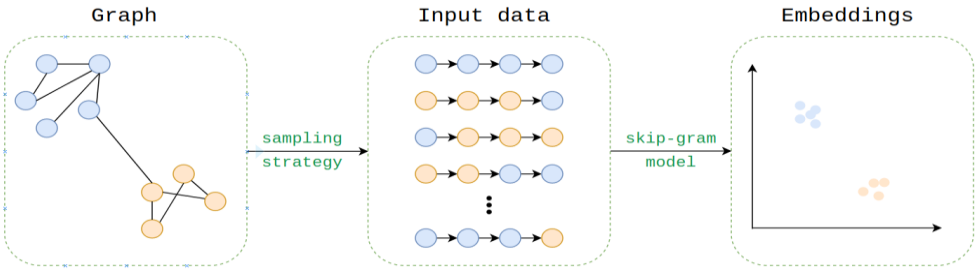

Figure 2. The Process of Node2vec

The Node2Vec model and the DeepWalk model are highly similar. But there is also an essential difference between them. The Node2Vec model adds a biased random walk process, so that the random walk graph embedding model can ensure structure while ensuring homogeneity. This particularity makes the area that the Node2Vec model can explore more extensive [7]. It is also based on this advantage that it can obtain higher quality generated embedding vectors.

2.1.3 HARP model

Both DeepWalk models and Node2Vec explore local neighborhoods, which leads to the same problem: the resulting solution may be a local optimum. The HARP model found this problem and improved the model based on this problem. Optimization of weight initialization is the best way to avoid local optima. Therefore, this model is divided into three parts: coarse-grained graph, graph embedding and rendering enhancement. Retain the information in the original graph and use the particularity of graph coarse-grained to merge all the information recursively, make the original graph smaller and embed the coarse-grained nodes. The embedding of the original graph is continuously optimized by propagating the hierarchy and optimizing the embedding. The basic structure of the graph will ensure the correctness of the data, and the reduction of the size of the graph can allow the HARP model to optimize the random walk strategy to avoid the possibility of obtaining a local optimal solution [8]. However, the reduction of the scale will inevitably lead to the disappearance of some information, so that the accuracy of the embedding vector cannot meet expectations.

2.2. Classical random walk model Summary

In a homogeneous network, random walks have different models, and different models use different random walk strategies and generate completely different node sequences. It not only enriches the diversity of node sequences but also obtains more information about the structure of the graph. In addition to homogeneous networks, heterogeneous networks can also generate node sequences. But what is different from the homogeneous network is the method of generating the sequence of nodes. Random walks in heterogeneous networks are guided according to meta-paths. But these two different ways both make important contributions to the graph embedding of random walks and show good adaptability and ability. As shown in Table 1.

Table 1. Comparison based on the classic random walk model

Model	Year	Strategy	Advantages	Disadvantages	Application
DeepWalk	2014	The node sequence generated by random walk is used to maximize the co-occurrence probability of nodes	Simple, efficient and helps nodes embed	Does not apply to places outside the unauthorized network	Multilabel classification
Node2Vec	2016	A combination of DeepWalk and biased random Walk	Both homogeneity and structure are considered	The scale of the node sequence is naturally generated	Multilabel classification, Prediction Links
HARP	2018	In the case of weight optimization, a compressed graph can be generated	Can skip the local optimal solution and go straight to the optimal solution	Cannot be used on a sparse network	Multilabel classification, Visualization

3. Attribute walk model

3.1. Attribute walk model

The classical random stroll version is loosely used in various footage analysis tasks. exploitation those classical fashions to come up with established node sequences cannot best seize the topological form of graphs, but in addition reduce the exiguity and size of experience illustration. Disaster trouble.

a great deal of statistics show that the particular community consists of rich statistics, currently not merely nodes. Attribute traversal completely} totally fashions arrange to outline such sophisticated statistics into attributes, however characteristic networks are typically heterogeneous. Considering that attributes can complicate node interaction, it makes version creation bigger difficult. so as to remedy this trouble, many students have tried to use this trouble to remedy the difficulty of showing random walks on characteristic networks and also the usage of them to check the illustration of community nodes.

3.1.1 TriDNR model

In data community mining, in general, the affiliation dating among nodes needs to be checked for analysis. historical stochastic process-primarily based totally approaches completely deal with the nodes themselves and forget about node facts, but maximum actual networks include lots of facts. A 3-element deep community instance version is born: The TriDNR version is born, that makes use of facts from the 3 additives of node shape, node content, and node label to conjointly research the most effective node representation. The TriDNR version includes 2 primary steps: first, the era of a random stroll sequence: it takes the community shape as enter and arbitrarily generates a sequence of walks at the node; Second, gaining knowledge of the coupled neural community version [9].

3.1.2 GraphRNA model

However, it is not clear how to design the random walk of attribute networks to extract effective joint information, and the attribute information of nodes makes the network structure more complex. GraphRNA is a new attribute-based network embedding framework, which combines AttriWalk cooperative walking mechanism with GRN cycle graph network to learn node representation in attribute network more effectively [10]. The GraphRNA model can be roughly divided into three parts: (1) Unified walking mechanism: In order to achieve the purpose of scanning nodes with complex attributes, binary graphs are constructed based on node attributes to help generate different node sequences; (2) Graph recursive network (GRN): a deep structure that effectively supports node representation and hides the state sequence generated by data to adapt to the interaction between sensing nodes; (3) Generating node embedding: select a part of the sequence to construct sentences with nodes as starting nodes, and then use pooling method to generate node embedding vector.

3.1.3 FEATHER model

In a real network, translating community options may be terribly complicated. A community contains several attributes with special distributions, unit nodes, and community characteristics. The FEATHER version was born to resolve this problem. A feature motion theme is planned that flexibly defines the feature motion at the vertices of the graph and explains the distribution of vertex choice at multiple scales. This can be a collection of utterly rational rules that outline the potential weights of feature features thanks to the interchangeability of random swimming [11]. The FEATHER version can perform feature inferences of varied feature maps at high speed in linear time to make a geometrical vector position map of nodes. The FEATHER version has powerful applied math knowledge destruction capabilities, provides ancient vector position explanations for constant graph, and permits for a fast and sturdy switch from graph to graph.

3.2. Attribute walk model Summary

The large quantity of data management of the node itself makes the model special. This quality model is supported, not just for nodes, however conjointly for common topological properties. exploit neighborhood attribute inertia at scale are useful for several totally different applications. The work of uncounted scientists has also discovered the powerful vitality of ensemble models. getting the feature information of the network within the shallow network is presently the foremost wide used thanks to engraft the attribute network model. However, this approach ends up in the disappearance of deep feature information in nonlinear networks. the fact that the data will lead to the ultimate result's an area best answer instead of an optimal solution [12-13]. during this case it becomes

significantly vital to engraft the deep attribute network into the model. customized stochastic process models will capture deeper structural and attribute data and create the ensuing solutions additional reliable and persuasive. So in table 2, there is a side-by-side comparison of the three different models mentioned in this section.

Table 2. Comparison of attribute based walk models

Model	Year	Strategy	Advantages	Disadvantages	Application
TriDNR	2016	Maximize the co-occurrence probability of nodes	The rich attributes of nodes can be integrated	Lack of basis for weight adjustment	Node classification, Visualization
GraphRNA	2019	Attribute collaboration and sample collection can take place within the network and nodes	Increase diversity and flexibility	There is a loss of information along the way	Node classification
FEATHER	2020	The attributes of vertex features can be flexibly defined according to the characteristic functions on the vertices of graphs	Feature functions on large attribute graphs can be derived quickly based on linear time	---	Node classification, Graph classification and transfer learning

4. Conclusion

Artificial intelligence is one of the hottest research projects in the world, and graph embedding is one of the research hotspots of artificial intelligence. In order to study the current situation of graph embedding model, this paper studies the existing classical models, analyzes the characteristics and application scenarios of these models, and finally makes an in-depth comparison between the different models of chance classical random walk and attribute walk to find the advantages that can be further strengthened and the disadvantages that need to be improved.

Project provides a strong numerous to walk-based random graph embedding for representing graph data which will be used for an expansion of tasks, which could be broadly divided into: network reconstruction, node classification, graph clustering, outlier detection, link prediction, and visualization. the foremost necessary distinction between graph embedding based on random walk and type ways that is that the technique of resolution graph form within the graph. it' appointed to low-dimensional vector regions and endlessly optimized to verify that the embedded region vector shows the distinctive type of the chart as accurately as possible. Since downstream systems capture the opportunistic input characteristics, i.e. the graph truth preprocessing step is skipped once the graph is embedded, the graph preprocessing is usually done forthwith as a results of the system itself detects a risk. enhancements in graph embedding techniques allow random walks to be used to represent homes on a graph, but not node positions or similarities. this technique is useful once viewing simple chart elements or once the chart is simply too large. once transactions as a full don't work, this flashy strategy can't be helped by asking the community to network from time to time, but iteratively accommodating very little changes inside the chart.

With the occasion of large statistics technology, graph facts or graph systems showing superior homes just like large, high-dimensional, sparse, dynamic, and heterogeneous showcase most important demanding situations for facts evaluation and extraction. Graphs. The version plays nicely on many obligations but conjointly faces a few demanding situations inclusive of making normalized an identical the same a regular an even theoretical framework to enhance the overall performance of these approaches to deliver a reference widespread for any research. Embedding models for large scale networks and complex networks are still to be improved. The model based on random wandering is one of the better ways to react to the data information of large-scale networks. However, the cost of generating embedding vectors in this way is very high. The amount of information that needs to

be collected is increasing as the Web becomes more important in people's daily lives. It is a huge problem to cope with the mega-scale graph embedding vectors in the future. Secondly, there are complex networks, which may be homogeneous or heterogeneous. There is also an urgent need for the way to handle different types of networks. These are very promising research directions.

References

[1] Cai, Hongyun, Vincent W. Zheng, and Kevin Chen-Chuan Chang. A comprehensive survey of graph embedding: Problems, techniques, and applications. IEEE Transactions on Knowledge and Data Engineering 2018, 30.9: 1616-1637.

[2] Goyal, Palash, and Emilio Ferrara. Graph embedding techniques, applications, and performance: A survey. Knowledge-Based Systems 2018, 151: 78-94.

[3] Xu, Mengjia. Understanding graph embedding methods and their applications. SIAM Review, 2021, 63.4: 825-853.

[4] Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014: 1-11.

[5] Tu, Cunchao, et al. Max-margin deepwalk: Discriminative learning of network representation. IJCAI. Vol. 2016: 20-26.

[6] Grover, Aditya, and Jure Leskovec. node2vec: Scalable feature learning for networks. Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016, 12: 353-362.

[7] Hu, Fang, et al. Community detection in complex networks using Node2vec with spectral clustering. Physica A: Statistical Mechanics and its Applications 2020, 545: 123633.

[8] Wang, Yashen, and Huanhuan Zhang. Harp: a novel hierarchical attention model for relation prediction. ACM Transactions on Knowledge Discovery from Data (TKDD) 2021, 15.2: 1-22.

[9] Liao, Lizi, et al. Attributed social network embedding. IEEE Transactions on Knowledge and Data Engineering 2018, 30.12: 2257-2270.

[10] Huang, Xiao, et al. Graph recurrent networks with attributed random walks. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 83-91.

[11] Streit, Lisa, and Wolfgang Heidrich. A Biologically-Parameterized Feather Model. Computer Graphics Forum. Oxford, UK: Blackwell Publishing, Inc, 2002, 21(3): 46-55.

[12] Yan, Shuicheng, et al. Graph embedding: A general framework for dimensionality reduction. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). 2005, 2:742-751.

[13] Robles-Kelly, Antonio, and Edwin R. Hancock. A Riemannian approach to graph embedding. Pattern Recognition 2007 40.3: 1042-1056.