Journal of Electronic Research and Application, 2024, Volume 8, Issue 5
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Research on Heterogeneous Information Network Link Prediction Based on Representation Learning

Yan Zhao[1], Weifeng Rao[1], Zihui Hu[1], Qi Zheng[2]*

[1]School of Information Engineering, Jingdezhen Ceramic University, Jingdezhen 333403, China
[2]School of Mechanical and Electronic Engineering, Jingdezhen Ceramic University, Jingdezhen 333403, China

***Corresponding author:** Qi Zheng, 15172480512@163.com

Abstract: A heterogeneous information network, which is composed of various types of nodes and edges, has a complex structure and rich information content, and is widely used in social networks, academic networks, e-commerce, and other fields. Link prediction, as a key task to reveal the unobserved relationships in the network, is of great significance in heterogeneous information networks. This paper reviews the application of presentation-based learning methods in link prediction of heterogeneous information networks. This paper introduces the basic concepts of heterogeneous information networks, and the theoretical basis of representation learning, and discusses the specific application of the deep learning model in node embedding learning and link prediction in detail. The effectiveness and superiority of these methods on multiple real data sets are demonstrated by experimental verification.

Keywords: Heterogeneous information network; Link prediction; Presentation learning; Deep learning; Node embedding

Online publication: October 9, 2024

1. Introduction

With the rapid development of the Internet and artificial intelligence, data and information have grown exponentially, and people, data, information, and computers have formed a complex network system [1]. How to dig out valuable information from massive data and information has always attracted many, and the research of network science has gradually risen [2]. A heterogeneous information network is a complex network structure containing multiple nodes and multiple types of edges, which can well simulate the diversity and complexity of the real world [3–4]. In a network, nodes can represent different entities or concepts, while edges represent various relationships between them [5–6]. Traditional information networks are divided into homogenous networks and heterogeneous networks according to the types of nodes and edges. Only one type of nodes and edges in a network is a homogeneous network, while heterogeneous networks are information networks composed of various types of nodes and connected edges [7–8]. Different types of nodes or edges contain different attributes, as shown in **Figure 1**. For example, users, goods, and comments in social networks, as well as their interaction behaviors, constitute a typical heterogeneous information network.

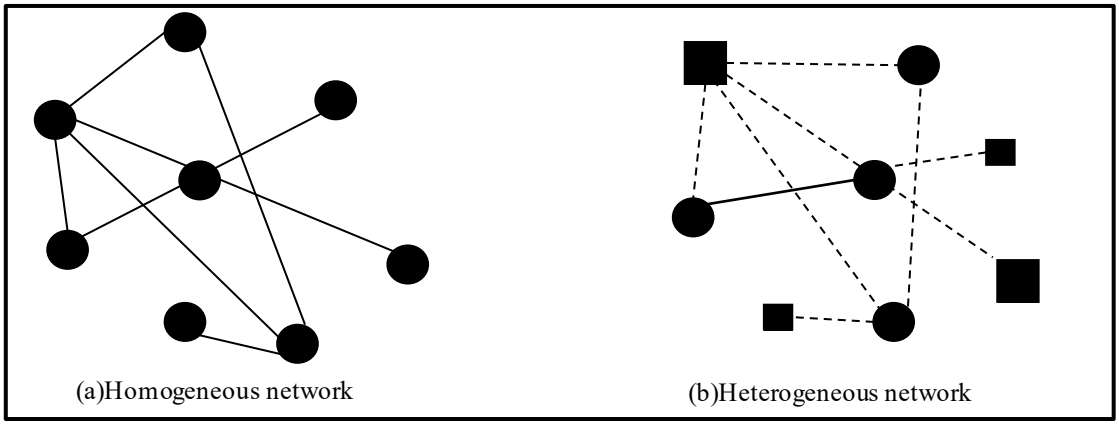

Figure 1. Homogenous network vs. heterogeneous network

Link prediction aims to predict the presence or absence of links between unobserved pairs of nodes based on known network structure information [9]. In heterogeneous information networks, traditional prediction methods are difficult to deal with effectively due to the diversity of nodes and edges [10]. In recent years, with the development of deep learning techniques, presentation learning has become an effective tool to deal with this problem. Presentation learning can learn low-dimensional vector representations of nodes, thereby preserving structural and semantic information between nodes in vector space, providing a new solution for link prediction.

This paper systematically introduces the research progress of link prediction in heterogeneous information networks based on representation learning. Firstly, the basic concepts of heterogeneous information networks and the theoretical basis of representation learning are summarized. Secondly, how to apply the deep learning model to node embedding learning and link prediction tasks is discussed in detail. Finally, the effectiveness and applicability of these methods in different scenarios are verified by experimental analysis.

2. Heterogeneous information network modeling

Heterogeneous information networks are typically modeled as a multigraph where different types of nodes and edges are represented as different entities and relationships, respectively. For example, in an academic network, there may be types of nodes for authors, papers, and conferences, as well as many types of edges for author-author partnerships, author-paper writing relationships, and many more. To effectively analyze and utilize complex relationships in heterogeneous information networks, they are represented by modeling methods. The main modeling methods include the following.

2.1. Graph representation

The heterogeneous information network is represented as a multiple graph, each type of node and edge has its own unique identification. **Figure 2** shows the information network constructed by document data, and **Figure 2(b)** illustrates the network pattern describing the document heterogeneous network objects and the types of relationships between them. **Figure 2(a)** is an example of the network of **Figure 2(b)**, in which there are 3 types of objects: paper (P), author (A), and conference (C). Links connect different types of objects, and the type of link is defined by the relationship between the two object types. For example, a link between an author and a paper indicates a relationship between writing or being written, while a link between a conference and a paper indicates

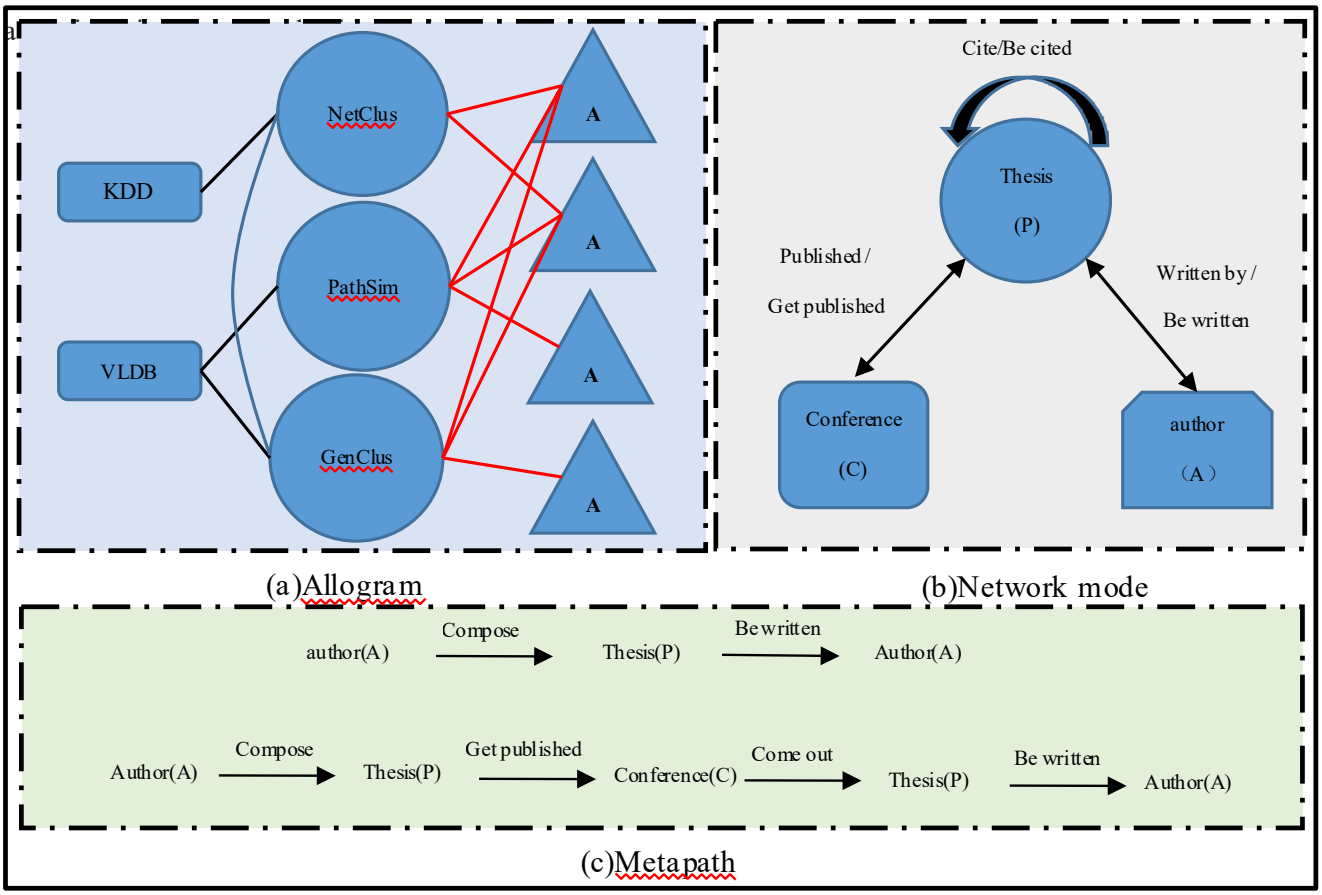

Figure 2. Heterogeneous network of literature data

2.2. Node attributes and relationship types

Nodes in heterogeneous information networks usually have rich attribute information, such as text content, labels, etc., while edges have different relationship types and weights. The weighted heterogeneous network is constructed by synthesizing the temporal and semantic features of text nodes. The temporal features reflect the dynamics of the research direction, and the semantic features use the abstracts of literature nodes to construct the literature node relationships, thus mining more information. Through the representation learning of the heterogeneous network nodes, the academic information recommendation including scholars and literature is finally completed.

The figure above shows a simplified example of a weighted heterogeneous network literature model network with different types of nodes and different types of edges (solid lines for one relationship and dashed lines for another).

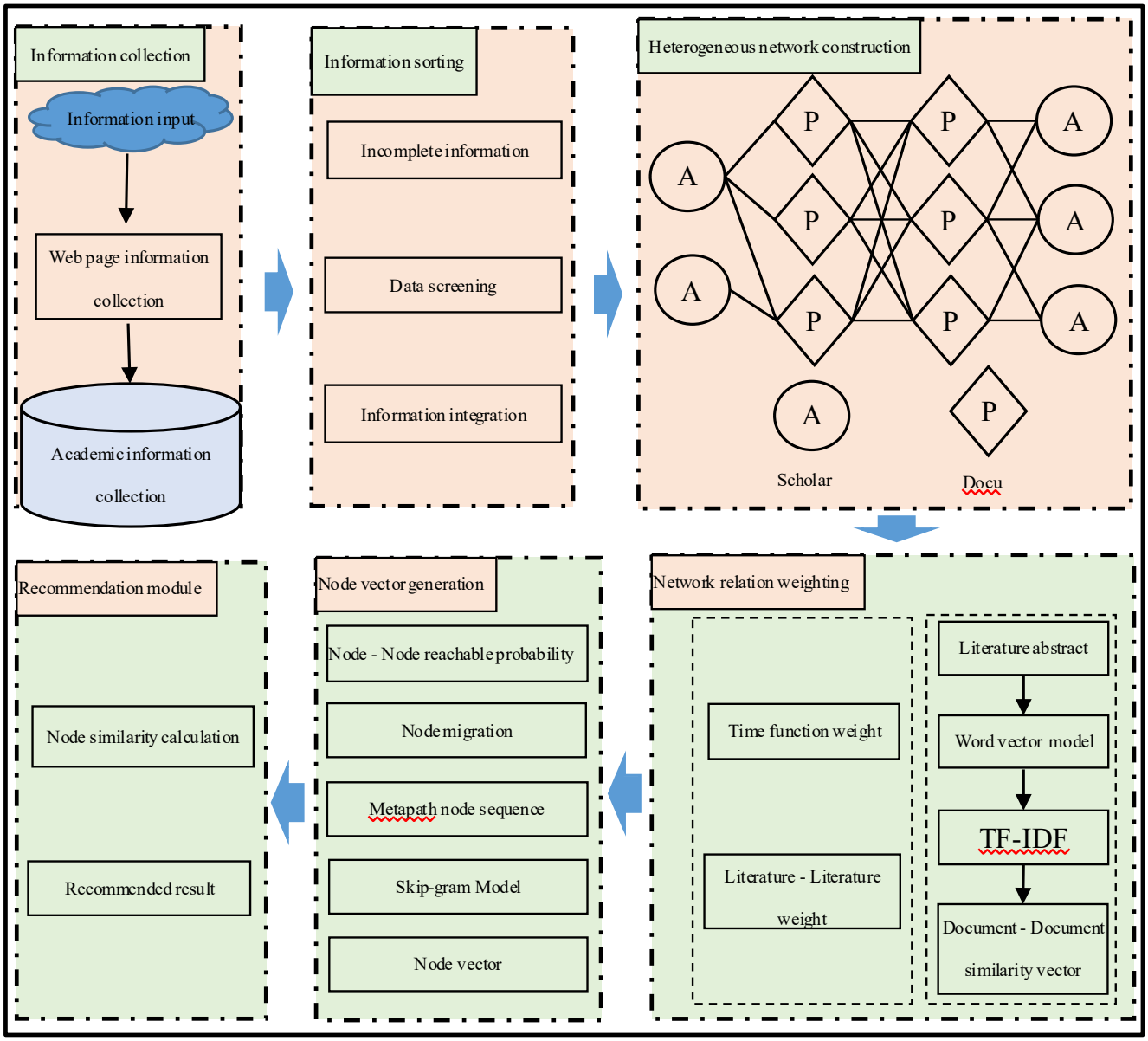

Figure 3. Weighted heterogeneous network literature recommendation

3. Representation learning methods

Presentation learning is a technique widely used in complex network analysis in recent years. Its core goal is to learn low-dimensional vector representations of nodes to capture structural and semantic information between nodes. In heterogeneous information networks, representation learning can not only effectively capture the complex relationships between nodes, but also integrate the content information of nodes (such as node attributes, text information, etc.) to provide a more comprehensive solution for link prediction.

(1) Deep random Walk (DeepWalk): Learning the context information of nodes by simulating the random walk between nodes, and converting it into a vector representation of the nodes.

(2) Graph Convolutional Network (GCN): A deep learning model suitable for processing graph-structured data, learning the embedded representation of nodes through information propagation on the graph through convolution operations.

(3) Heterogeneous Graph Attention Network (HAN): A network of attention mechanisms specifically designed to process heterogeneous information networks, which can effectively capture the association of different types of nodes and edges.

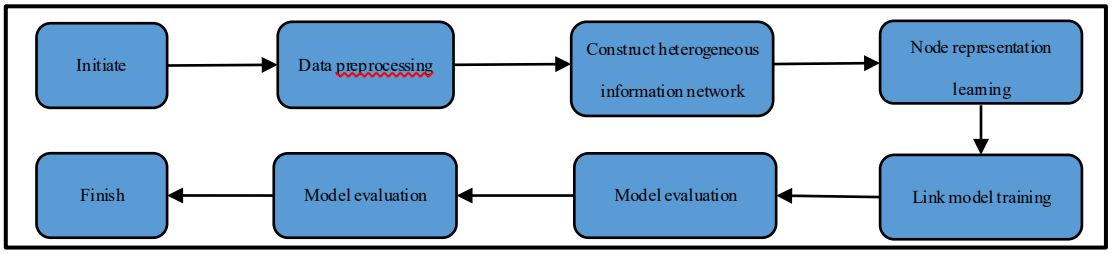

Figure 4. Flow chart of representation-based learning heterogeneous information link prediction

The flowchart shows the main steps and flow of the presentation learning-based heterogeneous information network link prediction method.

4. Link prediction model

After obtaining the embedded representations of the nodes, we need to input these representations into the link prediction model to predict new potential links. Common prediction models include logistic regression, support vector machines (SVMS), neural network models, etc. These models can effectively capture complex interaction patterns between nodes based on their embedded representation and are used to predict whether links exist between unobserved pairs of nodes.

(1) Accuracy: The ratio of the number of correctly predicted links to the total number of predicted links.

(2) Precision: The proportion of links that exist among the pairs of nodes predicted to be links.

(3) Recall: The proportion of pairs of nodes that are linked and correctly predicted to be linked.

(4) F1 score: The harmonic average of accuracy and recall as a comprehensive measure of the predictive performance of the model.

Table 1. Performance evaluation table of heterogeneous information network model

Data set	Methods	Accuracy rate	Precision rate	Recall rate	F1 score
Academic network	DeepWalk + LR	0.85	0.82	0.88	0.85
Social networks	HAN + SVM	0.91	0.89	0.93	0.91

The table shows the experimental results of link prediction using different methods on different datasets, evaluating the performance of each model on accuracy, precision, recall, and F1 score.

5. Experiment and result analysis

By conducting experiments on multiple real data sets, we verify the effectiveness of the representation learning-based method for predicting heterogeneous information network links. The experimental results show that the deep learning model is used for node embedding learning and link prediction.

6. Conclusion

Research on heterogeneous networks has expanded to e-commerce, security, medicine, and other fields, significantly improving the performance of related mining tasks. With the development of artificial intelligence, heterogeneous networks will be used more frequently to handle complex interactions and information processing, and more potential application scenarios await exploration. For example, in software engineering, there are complex relationships between requirement documents, problem reports, and test samples. In biological engineering, there are gene sequences, coding structures, etc. Therefore, applying heterogeneous network analysis to these specific scenarios and fully leveraging its role is a key direction for future development.

Funding

Science and Technology Research Project of Jiangxi Provincial Department of Education (Project No. GJJ211348, GJJ211347 and GJJ2201056)

Disclosure statement

The authors declare no conflict of interest.

References

[1] Wu H, 2022, Research on Heterogeneous Network Link Prediction Based on Graph Representation Deep Learning, thesis, Inner Mongolia University of Science and Technology.

[2] Zhao Y, Wu H, 2023, Based on Diagram Depth Study of Heterogeneous Network Link Prediction Research. Small Microcomputer System, 44(02): 422–428.

[3] Zhao Y, Zhao S, Ma Q, 2021, Heterogeneous Information Network Link Prediction Method Based on Graph Nuclear. Computer Application Research, 38(10): 6. https://doi.org/10.19734/j.iSSN.1001-3695.2021.01.0056

[4] Jiang Z, Li J, 2022, Recommendation Model Based on Heterogeneous Information Network and Multi-Task Learning. Journal of Beijing University of Technology, 48(12): 1289–1297.

[5] Jiao P, Pan T, Jin D, et al., 2023, A Review of Role-Oriented Network Representation Learning. Journal of Computers, 46(2): 274–303.

[6] Jiang T, Qin B, Liu T, 2018, Said Learning Based Open Domain Knowledge Reasoning. Journal of Chinese information, 32(3): 8.

[7] Wang S, Cao J, 2019, Edge in the Heterogeneous Network Community Structure Discovery Algorithm. Computer Engineering, 45(6): 6. https://doi.org/10.19678/j.iSSN.1000-3428.0050734

[8] Hu B, 2019, Research and Implementation of Recommendation Algorithm Based on Representation Learning in Heterogeneous Information Networks, thesis, Beijing University of Posts and Telecommunications.

[9] Li F, Wang J, Chen H, 2023, Link Prediction Method Based on Graph Attention and Feature Fusion. Journal of Sichuan University: Natural Science Edition, 60(5): 96–105.

[10] Ishikawa, Wang R, Wang X, 2022, Review on Heterogeneous Information Network Analysis and Application. Journal of Software, 33(2): 598–621.