Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/52/20241560

Research on the application of graph neural networks in text classification

Chuangyu Xu

School of Cybersecurity, Zhengzhou University, Zhengzhou, 450000, China

15378756819@163.com

Abstract. With the rapid development of the Internet and social media, a large amount of text data is constantly generated, and text classification has become an important task. Text classification is an important branch in the field of natural language processing and has attracted the attention of many researchers in recent years. Due to the ease of editing text data, most texts in the network are manually constructed and uploaded by users. Therefore, the standardization of online texts and the classification of texts with different granularity are of great significance in the field of information retrieval. This article introduces the basic concepts of text classification and three text classification methods: knowledge engineering (rule) based text classification methods, traditional machine learning-based text classification methods, and deep neural network-based text classification methods. This article also analyzes the advantages and disadvantages of the three methods. Subsequently, the basic concepts and common models of graph neural networks were explored. An application case of graph neural networks in text classification is provided. This article provides effective suggestions for text classification applications in the field of natural language processing.

Keywords: Text classification, Figure Attention Network, Hypergraph neural network, Natural language processing.

1. Introduction

With the rapid development of the Internet and social media, a large amount of text data is constantly generated, and text classification has become an important task. Traditional text classification methods are usually based on word bag models or TF-IDF weight calculations, using machine learning algorithms to train classifiers for classification. However, these methods have some limitations, such as being unable to capture semantic and contextual information in the text and being difficult to handle long texts. In recent years, graph neural networks, as a new type of neural network architecture, have received widespread attention due to their ability to capture relationship information and global structural information between nodes. In the field of text classification, graph neural networks have strong representation modeling capabilities, which can capture global word co-occurrence and high-order neighbor information, and have node classification capabilities. Therefore, introducing graph neural networks into the field of text classification is expected to break through the limitations of traditional text classification methods and effectively improve the performance of text classification. This article studies the different frameworks and applications of exploratory graph neural networks in text

DOI: 10.54254/2755-2721/52/20241560

classification through a literature review, summarizes some existing work, and hopes to provide some ideas for the development and improvement of future text classification models[1].

2. Overview of Graph Neural Networks

Graph neural network is a deep learning technique used to process graph-structured data. It views nodes and edges as learnable features and models the graph using neural networks to extract feature representations of nodes. Common graph neural networks include Graph Convolutional Network (GCN), Graph Convolutional Network (GraphSAGE), Graph Attention Network (GAT), and so on.

GCN is a deep learning model that processes graph data through convolutional operations. It can perform complex analysis and prediction on nodes, edges, and the entire graph. In GCN, each node can interact with its neighboring nodes to update its representation. This interaction is achieved by weighted averaging the features of nodes with those of adjacent nodes. GCN can effectively handle various types of graph structures such as heterogeneous graphs and directed graphs.

GraphSAGE is a model that learns feature representations from neighboring nodes through sampling methods. It can be seen as a lightweight GCN, characterized by the use of random sampling to reduce computational complexity and memory consumption. In GraphSAGE, each node interacts with its neighboring nodes and updates its representation using a shared aggregation function. This aggregation function can be average, maximum, minimum, etc. GraphSAGE can effectively handle large and heterogeneous graphs.

GAT is a graph neural network that introduces an attention mechanism. In GAT, each node can interact with its neighboring nodes, but the weight of this interaction is determined by a shared weight matrix. This weight matrix can be regarded as a shared attention matrix, which can adjust the intensity of interaction based on the characteristics of neighboring nodes. This attention mechanism can make GAT more flexible and effective in handling heterogeneous and directed graphs[2-3].

3. Introduction to Text Classification

Text classification is a fundamental task in the field of natural language processing. Its core task is to analyze and extract text features and divide the text into corresponding categories based on the analysis results. In the past few decades, text classification has been widely studied and applied in many practical tasks, such as news classification, sentiment analysis, question-answering systems, dialogue systems, etc. Because text data is easy to edit, conservative to store, and widely dispersed, text data accounts for a large promotion of a large number of network information, and most of the text is manually constructed and shared on the network, so there must be noise data Theoretical, it is of great significance to standardize the network text and classify the text to different degrees in the field of information retrieval Text processing kits such as JIEBA [1] make data cleaning simple and easy to operate, which greatly promotes the development and landing of text classification First, text can be divided into different coalesce-grained categories depending on the application field, such as news text, medical text, geographic text, etc. Second, according to the different application scenarios, text can also be divided into a variety of fine-grained [4-5].

For example, the review text of a movie can be divided into multiple levels to determine the user's love and preference for the movie, and the user's search log can be divided into multiple labels to determine the user's real search intent, etc. These fine-grained classification tasks are also one of the differences in the field of natural language processing In this study, geographic text, and user retrieval logs are taken as the main research objects The research goal of the form is to identify geographic texts from a large number of other categories of texts, which is essentially a text binary classification task The research goal of the latter is to assign different labels to the search requests entered by users in search engines to identify the real search intent of users, which is essentially a text multi-classification task[6].

DOI: 10.54254/2755-2721/52/20241560

4. Text classification methods

At present, text classification methods are mainly divided into three categories: knowledge engineering-based text classification methods, traditional machine learning-based text classification methods, and deep neural network-based text classification methods[6].

4.1. Text Classification Method Based on Knowledge Engineering (Rules)

This method mainly classifies text manually. This method relies on human experts' understanding of the text content and their mastery of domain knowledge. The key steps of this method include defining classification standards, manually annotating samples, model training and testing, and classification implementation.

Advantages: This method can accurately classify text based on specific needs and backgrounds, and does not require a large amount of training data.

Disadvantages: Text classification methods based on knowledge engineering require manual participation, resulting in high costs. The accuracy and reliability of classification depend on the professional knowledge and experience of domain experts and data annotators. With the continuous accumulation of network text resources, the number of rules written by experts increases, and rules designed by different experts are prone to conflicts, resulting in higher maintenance costs and a significant decrease in text matching speed. In addition, this method is difficult to handle large-scale, complex, or dynamic text data[7].

4.2. Text Classification Method Based on Traditional Machine Learning

The traditional machine learning-based text classification method mainly uses machine learning algorithms to classify texts. This method requires first extracting features from the text, then using these features to train the model, and finally using the model for classification. Common traditional machine learning algorithms include naive Bayes, logistic regression, support vector machines (SVM), and decision trees.

Advantages: Text classification methods based on traditional machine learning can utilize a large amount of unlabeled data for pre-training, thereby improving the accuracy of classification to a certain extent. In addition, compared to knowledge engineering-based methods, this method has lower costs and is easier to achieve automation.

Disadvantage: This method requires manual feature setting or the use of specific feature extraction methods to extract features from text, which may affect the accuracy and reliability of classification. In addition, traditional machine learning-based text classification methods may not achieve ideal results for large-scale, complex, or dynamic text data.

4.3. Text Classification Method Based on Deep Neural Networks

The text classification method based on deep neural networks has emerged in recent years. This method utilizes deep neural networks (DNNs) for automatic feature extraction and classification of text. Common deep neural networks include recurrent neural networks (RNNs), short-term memory networks (LSTMs), and transformers.

Advantages: Text classification methods based on deep neural networks can automatically extract features from text using neural networks, avoiding the need for manually setting features or using specific feature extraction methods, thereby improving the accuracy and reliability of classification to a certain extent. In addition, text classification methods based on deep neural networks can better handle large-scale, complex, or dynamic text data.

Disadvantage: This method requires a large amount of training data, and the training process of the model usually requires a large number of computational resources and time. In addition, for certain tasks or domains, specific training or adjustments may be required for the model.

DOI: 10.54254/2755-2721/52/20241560

5. Application Cases of Graph Neural Networks in Text Classification

In text classification, text can be viewed as a graph structure. Each word or phrase can be considered as a node in the graph, and the semantic relationship between words can be considered as an edge. By converting text into a graph structure and modeling the graph using a graph neural network, the feature representation of the text can be effectively extracted for classification.

Wu Xinmata designed a dual sliding window composition method that integrates the global co-occurrence information of words into the text graph structure. Next, a graph neural network is used to learn the low-dimensional embedding representation of words. By aggregating neighbor node information and edge weights, local and global information is fully integrated into the low-dimensional word representation. Finally, using the attention mechanism to learn the importance of word nodes in the graph, and combining maximum pooling and average pooling to transform word embedding into graph embedding, the graph embedding representation is fully learned from the overall and local perspectives, respectively, to achieve text classification tasks[8].

Teng Sijie proposed the HVHMC model, which takes into account the horizontal and vertical dependencies between label structures, enabling the model to better capture the associations between categories and documents, thereby improving the accuracy of text classification [9].

Zhao Ke proposed a graph neural network algorithm model that integrates BERT. The algorithm idea is to use a pre-trained BERT model to obtain the feature vectors containing contextual semantic order information for each word in the corpus document and use them as the initial embedding features of the document graph. Thus, the DGNN model is used to achieve document graph classification[10].

6. Conclusion

This article introduces the basic concepts of text classification and three text classification methods and analyzes the advantages and disadvantages of these three methods. Discussed the basic concepts and commonly used models of graph neural networks, and provided application examples of graph neural networks in text classification. Although graph neural networks have achieved certain results in text classification, there are still some issues that need further research. Firstly, selecting appropriate model parameters is one of the key factors affecting model performance. Future research can explore more effective parameter selection methods to improve the performance of the model. Secondly, how to expand the application scope of graph neural networks is also one of the future research directions. At present, graph neural networks are mainly applied in short-text classification and long-text classification, and their applications in other fields such as sentiment analysis and information retrieval can be explored in the future. Finally, how to improve the generalization ability of the model is also one of the future research directions. At present, the training and testing datasets of most graph neural networks are separated. In the future, we can explore how to combine the training and testing datasets to improve the model's generalization ability.

References

[1] Wang Bing. Tabbed text classification algorithm based on neural network diagram study [D]. Jilin University, 2023. The DOI: 10.27162 /, dc ski. Gjlin. 2023.005218.

[2] Wang Jin, Chen Chongyuan, Deng Xin et al. Multi-state graph neural network text classification algorithm [J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition),2023,35(02):193-201. (in Chinese)

[3] Zheng Cheng, Xiao Shuang. Combined with the rules of grammar and text classification method of the neural network [J/OL]. Small microcomputer system: 1-12 [2023-11-02]. HTTP: / / http://kns.cnki.net/kcms/detail/21.1106.tp.20230918.1827.034.html.

[4] Wang Zhixue. Figure neural network application in text categorization research [J]. Computer knowledge and technology, 2023 (24): 79-81. The DOI: 10.14004 / j.carol carroll nki CKT. 2023.1268.

DOI: 10.54254/2755-2721/52/20241560

[5] Wang Jin, Chen Chongyuan, Deng Xin, et al. Multi-state graph neural network text classification algorithm [J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition),2023,35(02):193-201. (in Chinese)
[6] Zheng Cheng, Xiao Shuang. Combined with the rules of grammar and text classification method of the neural network [J/OL]. Small microcomputer system: 1-12 [2023-11-02]. HTTP: / / http://kns.cnki.net/kcms/detail/21.1106.tp.20230918.1827.034.html.
[7] Wang Jiawei, Yu Xiao. Deep learning based on the text classification research review [J/OL]. Electronic science and technology: 1-6 [2023-11-02]. https://doi.org/10.16180/j.cnki.issn1007-7820.2024.01.012.
[8] Wu Xinyu. Text classification algorithm based on neural network diagram study [D]. Southwest University, 2023. The DOI: 10.27684 /, dc nki. GXNDX. 2022.000713.
[9] TengSiJie. Hierarchical text categorization based on map neural network [D]. University of science and technology of China, 2023. The DOI: 10.27517 /, dc nki. Gzkju. 2022.001275.
[10] Zhao Ke. Based on the sequence diagram text classification algorithm of neural network research [D]. Jilin University, 2023. The DOI: 10.27162 /, dc nki. Gjlin. 2022.003193.