RDF Graph Summarization using Relational Graph Convolutional Networks

PhD Course Assignment

Emil Riis Hansen[1], Kashif Rabbani[1], and Haridimos Kondylakis[2]

[1] Aalborg University, Denmark
emilrh , kashifrabbani @cs.aau.dk
[2] Institute of Computer Science, Foundation of Research & Technology (FORTH)
kondylak@ics.forth.gr

Abstract. The proliferation of various unstructured, irregular, and incomplete Big data sources on the Web has led to the need to explore, query, and understand such data sources. RDF Summarization facilitates these tasks by extracting the precise and essential information from these heterogeneous Big data sources. We studied various state-of-the-art approaches in PhD course "RDF Graph Summarization: Principles, Techniques, and Applications". We aim to investigate the usage of Relational Graph Convolutional Networks (RGCN) to create summaries of the Knowledge Graphs. *We propose an approach to apply RGCN to the RDF graph to vectorize the instance data iteratively to produce an informed vectorized network of RDF graph categorized as quotient graph to understand the structural nature of the input RDF graph.* The implementation was done in Python using the DGL library. We conclude with an open-ended discussion on our approach. We discuss that manual inspection could provide insights for an evaluation of our end goal, if fully implemented and evaluated.

Keywords: RDF Graph Summarization · RGCN and RDF Graphs

1 Introduction

The requirement for enterprise data to provide value to the business has never been stronger in the last two decades than now. The rapid increase in complex enterprise data has made it difficult for the data stewards, researchers, scientists, and practitioners to quickly explore, query, and understand such data sources. Graph data provide tremendous utility to organizations dealing with multiple heterogeneous connected data sources. The major challenge with graphs data is difficult to understand or explore them along with their exponential growth due to their irregular or unknown structure.

Graph Summarization is applied to cope with this challenge for the purpose of understanding complex large graph data sources and help enterprises in decision making at a large-scale. Summarization is to extract meaningful information from graphs. There exist many approaches of graphs summarization in the literature, where the most recent surveys are [1, 2]. The authors of [2] highlight several challenges in this area, e.g., identifying the quality of the RDF summaries, comparing the summaries produced by different summarization algorithms, and providing some guarantees over the combined summaries, *updates* of the RDF summaries. We propose an RDF graph summarization approach using Relational Graph Convolution Network (RGCN). This approach also addresses the challenge of updating the RDF summaries when the input RDF graph changes without generating the complete summary again.

2 Our Approach

Relational Graph Convolution Network (RGCN) has successfully been used in multiple graph applications for node classification, link prediction, and whole graph prediction [3]. RGCN extends the classical graph convolution network (GCN) by learning a weight matrix for each relation, in contrast to the GCN, which does not consider relation types. RGCN utilizes the message passing framework consisting of an aggregation step and an update step to update entity vector representations. The aggregation step can be seen from Equation 1.

$$m_{\mathcal{N}(i,l)} = \sum_{r\in\mathcal{R}} \sum_{j\in\mathcal{N}_i^r} \frac{1}{c_{i,r}} W_r^{(l)} \cdot h_j^{(l)} \tag{1}$$

Here $\mathcal{R}$ is the set of relations, $c_{i,r}$ is a normalization constant such as $|\mathcal{N}_i^r|$ and $W_r^{(l)}$ is a weight matrix for relation r in layer l of the network. By keeping a weight matrix for each relation, entity hidden states will be updated (according to Equation 2) according to the types relations they participate in.

$$h_i^{(l+1)} = \sigma(m_{\mathcal{N}(i,l)} + W_0^{(l)} \cdot h_i^{(l)}) \tag{2}$$

Here σ is a non-linear element-wise activation function such as the rectified linear unit or the sigmoid function. To the best of our knowledge, our approach is the first quotient technique that fully utilizes deep graph representation learning for summarization. Furthermore, utilizing GCN-based summarization trivially makes the model inductive, thus overcoming a known open problem related to RDF summarization [2]. The pipeline is illustrated in Figure 1. The pipeline takes as input an RDF graph. Every entity and relation matrix is then initialized using *Xavier initialization.* Entity weights are smoothed using multiple forward passes of the RGCN graph algorithm. The intuition behind the smoothing is that relation matrices are shared between all entities, thereby guiding entities with similar neighborhood structures towards the same position in the n-dimensional space. After n iterations of the RGCN algorithm, we perform clustering on the resulting n-dimensional space. Cluster statistics can now be computed to summarize the different structures inherent in the original RDF graph.

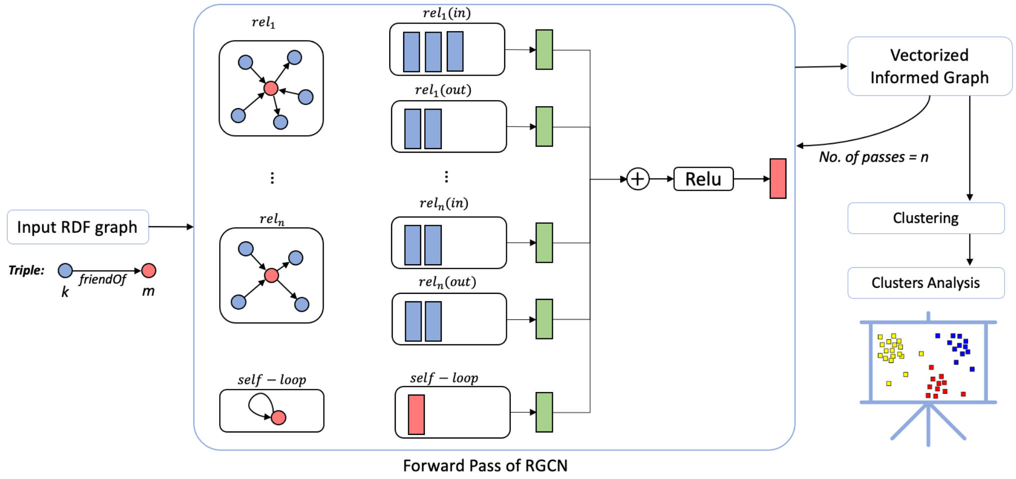

Fig. 1. Illustration of our RGCN approach.

We choose the AIFB dataset which describes the AIFB[3] research institute in terms of staff, research group, and publications. The dataset describes the inter-relationships between persons, projects, publications, and research topics etc. It contains 4 classes and 178 members belonging to the research group. It contains 8,285 nodes, 91 relations, and 66,371 number of edges.

The initial implementation of the pipeline has been made available online from a GitHub repository [4]. The implementation includes automatic download of the AIFB dataset, random initialization of node features, and a model for performing multiple forward passes using RGCN.

3 Conclusion and Discussion

In this course assignment, we proposed an approach to summarize RDF graphs using Relational Graph Convolution Network (RGCN). RGCN utilizes a message-passing framework that captures the multi-relational structure of graphs. This has been shown to provide state-of-the-art results in node classification, link prediction, and whole classification tasks.

We extend the utilization of RGCN for the use-case of RDF summarization. Taking an RDF graph as input, our model first initializes an n-dimensional hidden state for graph entities and shared relation-specific weight matrices. Multiple forward passes of RGCN are then performed to smoothen the hidden state representations with the prospect of similarly structured entities having similar hidden states. Clustering techniques can now be used to combine nodes from the

[3] Institute for Applied Informatics and Formal Description Methods at the Karlsruhe Institute of Technology.

[4] https://github.com/IKnowLogic/RDF-Summarization-Miniproject

initial graph based on the distance between entity hidden states. Statistics can now be calculated for each cluster in order to investigate the structures of the initial graph. We were not able to implement clustering and hence not able to further investigate our proposed method due to limited time constraints.

In future work, we propose to finalize the implementation and investigate the proposed method by manual inspection of cluster statistics. We believe these statistics will be useful in the summarization of RDF graph entities with similar structures. Furthermore, we propose that the utilization of specialized loss functions for training RGCN using Stochastic Gradient Descent (SGD) could significantly improve the quality of generated summaries for various purposes depending on the loss function. This could be used to guide the process of creating summaries based on user-defined goals.

References

1. Š. Čebirić, F. Goasdoué, H. Kondylakis, D. Kotzinos, I. Manolescu, G. Troullinou, and M. Zneika. Summarizing semantic graphs: a survey. *The VLDB Journal*, 28(3):295–327, 2019.
2. H. Kondylakis, D. Kotzinos, and I. Manolescu. Rdf graph summarization: principles, techniques and applications. In *EDBT*, pages 433–436, 2019.
3. J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.