RDF Graph Analysis for Fact Verification with Machine Learning

Neha Pokharel

Paderborn University

1 Introduction

The project focuses on the problem of fact verification, specifically, assessing the veracity of facts represented as Resource Description Framework (RDF) triples. The facts are embedded into a graph, and the truthfulness of the RDF facts is predicted using machine learning techniques.

2 Methodology

2.1 Data Preprocessing

Data is loaded from a N-Triples(.nt) file format, which contains RDF triples [1]. Each RDF statement is represented as a subject-predicate-object triple, and each fact is associated with a unique ID. These facts are parsed and stored in a dictionary for subsequent processing.

In addition, the RDF data is converted into a NetworkX graph where nodes represent subjects and objects of the facts and edges are predicates linking them. The graph structure provides a useful medium for visualizing relationships between different entities and is a basis for generating node embeddings.

2.2 Feature Extraction

The project employs Node2Vec [2], a graph embedding technique that learns continuous feature representations for the graph nodes. Node2Vec operates in a two-step process. First, it uses a random walk to generate paths through the graph. Then, it utilizes the Word2Vec algorithm to create vector embeddings for the nodes encountered in these paths. This method maintains the network topology while producing high-quality embeddings, preserving local and global node similarities. The feature vector for a fact is calculated as the mean vector of its subject and object node embeddings.

2.3 Model Training and Evaluation

With the feature vectors and their corresponding truth values, the next step is to train a machine-learning model. For this task, a Random Forest Classifier was selected due to its robustness against overfitting and ability to handle high-dimensional data.

The dataset was split into a training set (80%) and a validation set (20%). The ROC AUC score was used to evaluate the model's performance. The Receiver Operating Characteristic Area Under the Curve (ROC AUC) is a performance measurement for classification problems at various threshold settings.

2.4 Fact Verification

For fact verification, each fact is converted into a feature vector using node embeddings. The trained model then predicts the truth value of each fact.

3 Results

After implementing the above process, the model achieved a ROC AUC score around 0.83. This result indicates that the model has a good discriminative ability to distinguish between true and false facts.

References

1. "Rdf 1.1 turtle." `https://www.w3.org/TR/turtle/`, 2014.
2. A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864, 2016.