DS 294 - Data Analysis and Visualization
Analysis and Visualization of Knowledge Graphs

Shriram R.
M Tech (CDS)
06-02-01-10-51-18-1-15763

January 25, 2019

Abstract

Knowledge graphs provide highly structured information that can be efficiently processed by computers. It is considered as a prime contributor to develop intelligent systems. Consequently, knowledge graphs are already powering multiple Big Data applications in a variety of commercial and scientific domains [1]. In this project, several data analysis tasks will be performed over a knowledge graph and the result will be visualized in an interactive manner.

The primary dataset used for this project will be *YAGO* [2] which is a huge semantic knowledge base derived from different sources like Wikipedia, WordNet and GeoNames. The project can be potentially extended to other open knowledge graphs like *DBPedia* and *NELL*.

Project deliverable will consist of an interactive website having the following capabilities: Visualization of the data in the form of graph which allows complete search of vertex/edges and traversal from one vertex to another interactively. It will also be possible to perform analytic queries over the data and view the result in the form of graph.

One example for the analytic query is that a text document can be uploaded to the site and the contents will be matched with the knowledge graph resulting in finding key entities and relationships among them. The data can also be viewed as it evolved over time since the dataset also has temporal dimension in many of the facts and entities.

The following analytical tasks can also be performed on the graph using the website: *Link prediction* which finds the probability of forming an edge between any two vertices; *Link-based clustering* where vertices are grouped by the similarity of their links; *autocorrelation* which is finding the tendency of vertices to be linked to other vertices of similar characteristics. These tasks will be performed using existing algorithms in literature [1].

The website and data will be hosted in *Microsoft Azure/Google Cloud* platform. *D3.js* library will be used for visualizing the graph. Since the dataset is huge, *MapReduce* jobs using *Apache Spark* or a relational database like *Postgres* or *Neo4j* will be used to process analytic queries.

Though there are existing visualization programs available for *YAGO*, they are limited in their search and analysis functionalities. This project aims to provide an open accessible and interactive platform for visualizing and analyzing such knowledge graphs.

References

1. Nickel, M., Murphy, K., Tresp, V., Gabrilovich, E.: A Review of Relational Machine Learning for Knowledge Graphs. Proceedings of the IEEE 104(1), 11-33 (Jan 2016)

2. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A Large Ontology from Wikipedia and WordNet. Web Semantics: Science, Services and Agents on the World Wide Web 6(3), 203-217 (Sep 2008)