Capstone Project: Understanding 'Things' using Semantic Graph Classification

Rahul Parundekar

Machine Learning Nanodegree, Udacity
rparundekar@gmail.com
https://github.com/rparundekar

Abstract. The abstract should summarize the contents of the paper and should contain at least 70 and at most 150 words. It should be written using the *abstract* environment.

Keywords: Ontology, Semantic Web, Graph Kernels, Graph Classification, Deep Learning

1 Introduction

The world around us contains different types of things (e.g. people, places, objects, ideas, etc.). Predominantly, these things are defined by their attributes like shape, color, etc. These things are also defined by the "roles that they play in their relationships with other things. For example, Washington D.C. is a place and U.S.A is a country. But they have a relationship of Washington D.C. being the capital of USA, which adds extra meaning to Washington D.C. This same role is played by Paris for France.

The field of Knowledge Representation and Reasoning within Artificial Intelligence deals with representing these things, types, attributes and relationships using symbols and enabling the agent to reason about them. As it happens, a convergence has come about in this field of knowledge representation from the Databases domain - Graph databases can use graphs with nodes and edges to represent data much similar to the knowledge graphs.

Many domains use semantic grahps to represent their information because nodes, properties and edges of graphs are very well suited to describe the attributes and relationships of things in the domain. For example:

- Spoken systems - the output of Natural Language Processing is a parse tree.
- Social networks are graphs.
- High level semantic information in images are graphs of arrangements of things.
- The arrangement of objects on the road for autonomous driving is a graph.
- A users browsing pattern of products, usage graphs, etc. is a graph (e.g. browing products, plan of actions, etc.).

In the classic sense, Machine Learning focuses on specific kinds of understanding - classification, clustering, regression, etc. The algorightims in these deal with feature vectors (e.g. the features used for classification, etc.) and are aimed at essentially discriminating between different types of input to produce some output. To make decisions based on the state of the world an A.I. Agent can read from the world using sensors etc., can easily perform a classification task once it learns the relation between the data to its output decision. For the most part, the feature vectors used in such cases as input encode the attributes of the things, BUT not necessarily the relationships between things. And while the designer of the inputs and outputs of the algorithms may manually craft features to represent some of these relationships, the Agent has no automatic way of comprehending and using these relationships.

Can we use machine learning to make Agents better understand Things, including their attributes AND their relationships? If we are able to inspect the attributes and relationships of the things together and infer their roles, find its types, etc. our agent can act on those. If an Agent is able to classify things by understanding its semantic relationships, we could in the future generalize it to an Agent that can act on the meaning of the things.

1.1 Dataset

An Ontology is a formalism used to represent semantic data about *things*. It specifies the facts about the *things'* attributes and their relationships that the Agent believes to be true. DBpedia[1] is one such Ontology whose goal is to create a knowledge repository of general knowledge about the world such that Agents can have a grounding of the popular concepts and entities and their relationships. DBpedia datasets contains structured information extracted out of Wikipedia (e.g. page links, categories, infoboxes, etc.)[1].

The semantic data in DBpedia can be represented as a graph of nodes and edges. In this case, the nodes are *things* (i.e. entities) and the edges are links/relationships between the *things*. Each *thing* has both a **type** and a **category** associated with it. Our goal is to try and estimate these values from the data about the *thing*. We model this as a classification problem. An agent looking at graphical data can then use this classifier to identify the **type** or **category** of any *thing* that it sees for the first time and then perform possible action for that *thing*.

We use DBpedia as an exemplary dataset as a starting point to study Semantic Graph Classification. This method could then be used in a variety of domains like the ones mentioned earlier. For example:

- Spoken systems - Have a dialog with the user based on the Natural Language Processing parse tree.
- Social networks - recommend interest groups to similar users.
- Semantic image understanding - situation awareness.
- Autonomous driving - perform the next driving action.
- Product browsing pattern - recommend ads.

[1] http://wiki.dbpedia.org/

1.2 Problem Statement

Given the semantic data about things (i.e. their attributes and relationships), can we identify their types (e.g. hierarchy of classes) or categories (e.g. roles it plays) interest?

For example, if you look at examples of categories in DBpedia, Achilies has been put into the categories - demigods, people of trojan war, characters in Illead, etc. What makes him part of those categories? Can we learn the definitions of these based on the attributes and relationships of Achilies?

2 Preprocessing

References

1. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web Journal 6(2), 167–195 (2015), http://jens-lehmann.org/files/2014/swj_dbpedia.pdf