Call for Master Thesis

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

Be precise. Making the most out of little information! Predicting appropriate Infobox Types for not so popular Wikipedia Pages

Are you interested in making a big impact with your thesis? Work with us on an innovative approach for predicting Wikipedia Infobox Types for less known Wikipedia articles.

Wikipedia has emerged as the largest multilingual, web-based general reference work on the Internet. A huge amount of human resources has been invested in the creation and updating of Wikipedia articles which are ideally complemented by so-called **infobox**[1] templates defining the type of the underlying article. Structured data in Wikipedia is represented in the form of an infobox containing property value pairs summarizing the information content of the article. It has been observed that the Wikipedia infobox type information is often incomplete and inconsistent due to various reasons. However, the Wikipedia infobox type information plays a fundamental role for the RDF type information of Wikipedia based Knowledge Graphs such as DBpedia[2].

In this thesis, your focus will be to **predict infobox Types for the Wikipedia articles** which are less known and hence are provided with very little information. Currently, in Wikipedia, these articles do not have an infobox type or are assigned to a generalized type. As a result in DBpedia, these entities get assigned to the rdf:type `owl:Thing`. Therefore, this infobox type prediction approach will eventually help in various question answering applications and NLP related problems such as Named Entity Recognition etc..

The aim of this thesis is to develop a semi-supervised classification based approach to predict the infobox types. The students will extract different features (such as words, Table of Contents (TOC), Named Entity mentions in the articles, categories etc.) from the Wikipedia articles to predict the type information. Possible approaches will have to use different types of Machine Learning based approaches to do the feature engineering followed by semi-supervised classification approaches to predict the infobox type information.

This thesis will be supervised by **Prof. Dr. Harald Sack, Information Service Engineering at Institute AIFB, KIT, in collaboration with FIZ Karlsruhe**.

[1] https://en.wikipedia.org/wiki/Help:Infobox
[2] http://wiki.dbpedia.org/

DBpedia

Wikipedia
The Free Encyclopedia

Which prerequisites should you have?

- Good programming skills in Python or Java
- Interest in Deep Learning technologies
- Interest in Machine Learning approaches
- Interest in Semantic Web technologies

Contact person:
Russa Biswas
russa.biswas@kit.edu
russa.biswas@fiz-karlsruhe.de

aifb
Institute of
Applied Informatics and
Formal Description Methods
http://www.aifb.kit.edu/

Karlsruhe Institute of Technology