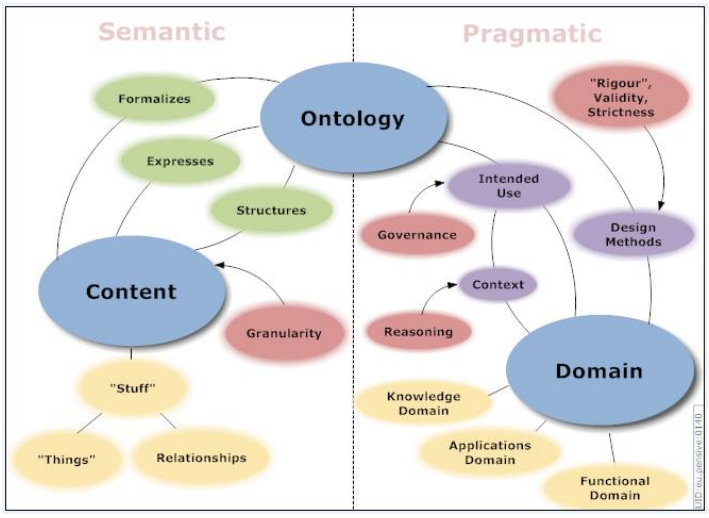
ENTITY IDENTIFICATION



Submitted to:

Dr. Vijay Raghavan

Center for Advanced Computer Studies

University of Louisiana at Lafayette

Advisor:

Shaimaa

elshaimaa.ali@hotmail.com

Prepared by:

Ramesh Sunkara (rxs6616)

ram@louisiana.edu

Abstract

Wikipedia articles consist mostly of free text, but also include structured information embedded in the articles, such as "InfoBox" tables, that contains categorization information, images, geo-coordinates and links to external Web pages. This structured information is extracted and put in a uniform dataset which can be queried. DBpedia is a project which extracts structured content from the information created as part of the Wikipedia project. This structured information is then made available in the form of RDF triples. Entity identification (entity extraction) is a subtask of information extraction that aims to classify concepts or entities into predefined categories such as person, organization, locations…etc.

Table of Contents:

Problem Statement:

Wikipedia articles consist mostly of free text, but also include structured information embedded in the articles, such as “InfoBox" tables, that contains categorization information, images, geo-coordinates and links to external Web pages. This structured information is extracted and put in a uniform dataset which can be queried.

DBpedia is a project which extracts structured content from the information created as part of the Wikipedia project. This structured information is then made available in the form of RDF triples.

Entity identification (entity extraction) is a subtask of information extraction that aims to classify concepts or entities into predefined categories such as person, organization, locations…etc.

As part of this project, I want to classify concepts present in the given RDF files which will be taken from DBPedia dataset. I will be using Apache Jena API to interact with RDF files and to execute SPARQL queries to extract the data present in RDF files.

This project is part of the ongoing research on "*Building Light weight ontologies for the purpose of annotations*”.

Implementation Details:

This would be stand-alone java project which can be used as a library in any other projects where classification of concepts is required.

Tools and Dataset:

Following are the different tools identified to use for the development of this project:

- Eclipse Juno – Java IDE
- Apache Maven – Project Management Tool
- GitHub – Version Control and Cloud Host service for source code
- Apache Log4j – Logging different events
- Apache Jena – Api to interact with RDF files
- SPARQL – Query language to extract data from RDF files

DataSet:

- DBPedia data set will be used. [http://wiki.dbpedia.org/Datasets]

Roadmap:

1. **November 6** – Design, Architecture, Environment and skeleton code base should be available.
2. **November 11**- Understand the data set. Identify all the use cases to be implemented. Implement the utility functions that would be needed in the project.
3. **November 18** – Implement at least 2 main workflows of the project.
4. **November 25** – Initial demo
5. **November 27** – Complete patch works and implement all other reaming use cases.
6. **December 4** – Final demo and draft report.
7. **December 9** – Submit report.

Evaluation:

Check the classified class of models with the given classes.

References:

- http://dbpedia.org/About
- http://jena.apache.org/
- https://github.com/
- http://maven.apache.org/