Semantically Linking/Extracting and Processing the University Network Data in Turkey (Draft)

Uğur Çekmez*
İstanbul Bilgi University
February 3, 2010

Abstract

This study demonstrates a methodology to make the data structured. The project handles linked data model to examine interlinking open data on the web and extracting the unstructured data from the web pages to make them reusable. Essentially, since the term linked data refers to the smart structured data which can be understood by the agents on the servers, firstly it will be examined to be able to understand what is it and how is it published. After that the issues about using the structured data will be discussed. In this respect the term ontology is touched upon to tell how is the structured data arranged to a formal representation, then the university ontology will be referred to deepen the explanation. According to the progresses on the project, the aim will be moving to process the structured data in which type it is expected to get university information such as extracting universities’ location, faculty, student information and determining the relation in between each university in the respect of specific criteria. This study is interested in the universities in Turkey and the explanations in the work supplies suggestions about how to increase relations in between universities in Turkey.

*ucekmez@cs.bilgi.edu.tr
1 Introduction

The web is a system of interlinked hypertext documents accessed via the Internet \(^1\) and it provides an interface to display the pages that have addresses with WWW. The idea behind the web is that if there is information, then everybody should be able to publish their accumulation and retrieve the web documents over the Internet.

In the past, the web sites was made of just static pages. Internet users could only see the pages, thus there were no interaction possibility. The increasing need lead to the foundation of a new model that is compatible with the previous one and supplies an interaction, thus the web pages turned their form from the simple place to store information to dynamic place where people gather the data. This development called Web 2.0 and brought many innovations.

Even if the Web 2.0 provides an advanced interaction between user and the server side, today’s internet users are looking for more than just information or interaction. The requisite is making the machines can process the data instead of people. The new formations are expected to transform the way people use the World Wide Web and make the machines communicate with each other by exchanging information. This is the term Semantic Web supplies the need of making the web more collaborative. This means that interlinking data from various sources and preparing a good infrastructure to make the data always and everywhere available to be able to develop new services such as setting up the services on top of each other. The considerable factor and the focusing point is the structure of the data. (Daconta, Obrst, and Smith 2009)

Many solutions to make the web as a sharing environment are proposed. In this article, the study demonstrates the linked data model to examine interlinking open data on the web and extracting the unstructured data from the web pages to make them reusable. Essentially, since the term linked data refers to the smart structured data which can be understood by the agents on the servers, firstly it will be examined to be able to understand what is it and how is it published. After that the issues about using the structured data will be discussed such as the way of determining will-be structured data. In this respect the term ontology is touched upon to tell how is the structured data arranged to a formal representation and after that, the university ontology will be referred to deepen the explanation. According to the progresses on the project, the aim will be moving to process the data in which type it is expected to get university information such as extracting universities’ location, faculty, student information and determining the relation in between each university in respect of specific criteria. This study is interested in the universities in Turkey and the explanations in the work supplies suggestions about how to increase relations in between universities in Turkey. In conclusion, in the light of the results of the research, the outputs will be evaluated.

2 Interlinking Open Data on the Web

The most important prerequisite of the Semantic Web is to interlink the data that are located on different sources but referred to a single target. This means that any object (or anything) has different relationships and properties can be defined on different sources. Because all of them still refer to these objects, if the characteristics of the relationships and the features are fit to the predetermined patterns, then it would be much simpler to build a knowledge base which is machine-processable.(Palmer 2001)

There are various methods to associate the data to each other such as RDF (Resource Description Framework). RDF is basically used to make the data modelled in various syntax styles. What it means, it is called a general flexible method to separate the knowledge into small parts, which are called as triples, with some basic rules.\(^2\) The triples generally come

\(^{1}\)http://en.wikipedia.org/wiki/The_web

\(^{2}\)http://www.rdfabout.com/intro/#Introducing RDF
about three main factors which are about relating an object to another one and the structures come from the term ‘type’. This means that every object can be defined as a type of another one. When a triple is prepared, firstly the subject part is filled and the second area is a predicate to determine what the subject is/does/has. The last part of a triple is also an object which the subject is related (according to its predicate). The large scale interlinking and automated mechanism can be developed by using these RDF statements (triples). The RDF yields many basic rules and unlimited relating points such as associating something to an object somehow. (Bizer, Heath, Ayers, and Raimond)

In the respect of topic interlinking data on the web, there is a need to structure and interlink university information/data to make them serve the aim of this project. The required data can be in different format and may include various information related universities such as academic, technical, planning (curriculum, courses), materials in the university and information about people in academia. The considerable point in this data is making it categorized by obeying some general patterns, thus preparing an ontology is required for this aim.

2.1 Ontology

In Computer Science, a term ontology refers a formal representation of a set of knowledge which have relations within its domain of interest. Essentially, it can be said as a concept based tree contains hierarchical knowledge. As relevant to the topic of the study, ontologies are considered one of the pillars of the Semantic Web, although they do not have a universally accepted definition.

When the ontology about an object is defined and developed, the information on that object is converted to the format which is created by obeying predetermined basic concepts. The considerable factor when design an ontology is representing the information in a standard format and also machine-readable. In this manner, the common concepts can be used to make a dictionary to be able to create a model of linked data.

2.2 Linked Data Model

The term linked data refers to an execution recipe for integrating data into the web as part of improvement of the current web to a web of interlinked data. In short, the linked data is about encapsulating the structured data and make them reusable and in such a way that data becomes machine-readable according to the requirements. (community 2009)

While creating linked data models by generating ontologies, there is a need to represent the concepts with the objects and the relations in between them. In practice, an ontology (or linked data model) consists of three stages: (1) identification of the concepts or asset class, (2) the arrangement of the class hierarchies, (3) identification of the relations in between the features of the assets.

2.3 University Ontology

There exist a few studies related extracting university data and some ontologies have been created about staffs and administration of university. For instance, in Computer Science Department at University of Maryland, a draft about university ontology is prepared. This study demonstrates elements for describing universities and the activities that occur at them and the project includes some concepts such as departments, faculties, students, courses, researches and publications.

A web crawler will be implemented to search and extract the data by making it structured and also the hyper-links between the university pages will be
determined to be able to prepare a graph shows relation.

3 Conclusion

This study suggests a way to take advantage of using semantic web technologies such as getting information from built knowledge base. However, related works shows that there is no completed project about university ontology and creating a dataset for universities and the aim of this project is starting this process with universities in Turkey. This working can be used to categorize universities in the respect of specific criteria. However, since the project will be done by getting data from different sources such as projects in the Linking Open Data cloud, the target and usage of the system can be moved to that cloud. Since the Semantic Web proposal is a solution for many need, then this study aims to become another small solution in the table. (Auer, Bizer, Idehen, and Community 2009)

3.1 Future Work

In the light of the results, the aim of this study will be demonstrating a methodology to crawl and extract hyper-link data in between universities in Turkey. The retrieved data forms the basis to address following research questions followed by a set of policy suggestions: (1) to what extent universities in Turkey are connected to each other, (2) what overall topological properties they are exhibiting, (3) which universities are at the center, which other universities remain at the periphery or stay isolated. As of policy suggestions, it further proposes a linked data model to better interconnect university web sites to each other, aiming to ease knowledge diffusion in academia at a national level.

3.2 Acknowledgement

- Bülent Özel, MSc, Lecturer, Istanbul Bilgi University, Department of Computer Science
- Çağıl Uluşahin, Teacher Assistant, Istanbul Bilgi University, Department of Computer Science

References


