Ain Shams University

Faculty of Computer & Information Sciences

Computer Science Department

Developing a Framework For Building an Arabic Ontology

A THESIS

Presented to Computer Science Department, Faculty of Computer and Information Sciences, Ain shams University.

Submitted in partial fulfillment of the requirements for the degree of doctor of philosophy in computer science.

Faculty of Computer and Information Sciences

Computer Science Department

Ain Shams University

BY

Dalia Sayed Mohamed Hassan Fadl

Assistant Lecturer Department of Computer Science, Akhbar El-Yom Academy.

M.Sc. of Computer and Information, Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University

Supervisors

Prof. Mostafa Mahmoud Aref

Professor of computer science,

Faculty of Computer and Information Sciences

Ain Shams University

Dr. Safia Abbas

Associate Professor in Computer Science

Faculty of Computer and Information Sciences

Ain Shams University

2018

To My father, My Mother, and My family.

Abstract

In Computer Science, ontology is a shared and common understanding of some domain that can be communicated across people and application systems or enabling knowledge sharing. It is a specification of a conceptualization. The rise of linguistic ontologies is a consequence of two simultaneous circumstances. Data organizing and description to encourage its usage by users later. Language is the best approach to vehicle data and information. So the requirement for semantic information is vital in all research fields. Web information always depends on the language which is written in; the accessibility of data identified with the language that would be much ideal as per the client would be an expanding need for today. Ontology is similar to a structure, dictionary or glossary, however, with more exceptional detail and structure. There is a robust requirement for Arabic language support since the ontology in English cannot be converted into Arabic. Distinctive languages have contained the particular semantic environment and the cultural context, which has brought on the need to build up the other ontology for various languages. Arabic ontology is an important natural language processing field it helps to enrich the Arabic language resources.

The aim of the framework is to automate the process of ontology generation, generating ontologies from pre-existing XML documents. It is an innovative framework, annotated as NAAO (Novel Automatic Arabic Ontology), which automates the ontology generation process from XML documents. The novelty of NAAO resides in generating the Arabic ontology, in the form of XML graph schema (XSG), from semi-structured data (XML documents associated with graph schema). The definition of this automation process was through four main steps necessary to achieve our goal. These steps conclude the main tasks of the automation process for building ontologies from XML documents. This thesis represents a framework that generates an Arabic Ontology from a semi-structured data (XML documents associated with graph schema), in which, XML schema is created and used in the graph schema development (XSG). The thesis provides two case studies, insectivore's case and mammal's case study where the developed Arabic ontology is applied. The results consist of 143 words, 10 concepts, 10 elements and 20 relationships. The generated ontology is evaluated using data-driven evaluation methods. 65% of the source XML documents have been included in the insectivore's case study.

Finally, the thesis provides the implementation of the framework for generating Arabic ontology containing the animal kingdom automatically. The ontology is divided into two parts to be more representative. The first part is the vertebrate's ontologies which provide 1576 concepts, 3836 element, and 2689 relations. Moreover, the second is the invertebrate's ontology which contains 320 concepts, 603 elements, and 783 relations. This result can be refined more than one time to reach satisfying results. The generated Arabic ontology is going to be evaluated using data-driven ontology measures cosine similarity measures and tree path mining. Finally, a comparison of the generated framework and three other system is provided.

Acknowledgments

Thanks to **Allah** before and after. And heartily thankful to my supervisors; **Professor Mostafa Aref** for his encouragement guidance and support from the initial to the final level, enabled me to develop and understand Arabic Ontology Generation. He was not just a supervisor but a father in all the situations. Moreover, I would like to thank my supervisor; **Doctor Safia Abbas**. I would like to thank my family for all their love and encouragement; My parents who raised me with a love of science and supported me in all my pursuits; My brother Mohamed and my sister Dina. And, this thesis would not have been possible without my loving husband Tarek Khattab, who helped, supported, encouraged and believed in me all the time at all the situations. Thank you, is a very small word to express my feeling. Thanks to my son and my daughter .I am thankful to my friends and my colleges how supported me.

Publications

Parts of this thesis have been published as original papers in the following references:

- Dalia Fadl, Safia Abbas, and Mostafa Aref, Automatic Arabic Ontology Construction framework: Insectivore's case study, AISI 2017: Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017 pp 458-466, Cairo, Egypt.

- Dalia Fadl, Safia Abbas, and Mostafa Aref, Approach for Automatic Arabic Ontology Generation, International Journal of An Intelligent Computing and Information Sciences (IJCICIS) Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt, 2017.

- Dalia Fadl, Safia Abbas, and Mostafa Aref, Automatic Arabic Ontology Generation for the Animal Kingdom, accepted, Journal Of Theoretical And Applied Information Technology(JATIT) (E-ISSN 1817-3195 / ISSN 1992-8645),2017

- Dalia Fadl, Safia Abas, Mostafa Aref, "Arabic Ontology Using Different Ontology Learning Techniques**,"** Proceedings of the Conference of Egypt Society of Language Engineering ESOLEC' 2014 Pages 68-67, Faculty of Engineering Ain Shams University, Cairo, Egypt.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1	CR	Context Resemblance
2	IDF	Inverse Document Frequency
3	NAAO	Novel Automatic Arabic Ontology
4	OWL	Ontology Web Language
5	OOT	OWL Ontology Term
6	TF	Term Frequency
7	TO	Term Occurrence
8	TPM	Tree Path Mining
9	XML	Extensible Marked Up Language

Chapter 1

Introduction

1.1. Overview

The term "Ontology" has been introduced with the data sciences and research fields amid the 1990's by a few Artificial Intelligence (AI) research groups. AI analysts received the expression "Ontology" basically to depict what they thought would be (from the outlook of computational aspects) a legitimate representation of the world in a program code. Ontologies are of essential enthusiasm for a broad range of fields, to a great extent because of what they guarantee: a mutual and normal comprehension of some area that can be the reason for correspondence ground over the crevices amongst individuals and PCs. They (Ontology approaches) take into account sharing and reuse of knowledge bodies in computational shape. The same numbers of traditional activities are changing their way in the realm of today because of the accessibility of data brought by the World-Wide-Web (WWW), Ontologies are probably going to turn increasingly when the learning is organized in a machine-readable way, and the abstract concepts it contains are shared.

Of the many definitions which have aroused for Ontology the following "*Ontology is a formal, explicit specification of a shared conceptualization.*" A "conceptualization" is a theoretical model of a phenomenon, made by ID of the related concepts of the phenomenon. The concepts, the relations amongst them and the imperatives on their usege are unequivocally characterized. "Formal" implies that Ontology is machine-readable and rejects the use of natural languages. For instance, in restorative spaces, the concepts are illnesses and manifestations, the relations between them are causal, and an imperative is that an infection cannot bring itself.

Ontology is a "shared conceptualization" states that Ontologies expect to speak to consensual information proposed for the use of the group. In a perfect world, the Ontology catches knowledge autonomously of its usege and in a way that can be shared all around, however for all intents and purposes unique errands and utilizations call for various portrayals of the knowledge in Ontology. Ontology is some of the time mistook for taxonomy, which is an arrangement of the information in a domain.

The distinction between them is in two vital contexts:

1. Ontology has a richer internal structure as it includes relations and constraints between the concepts.
2. Ontology claims to represent a certain consensus about the knowledge in the domain.

This consensus is among the intended users of the knowledge, e.g. doctors using a hospital Ontology regarding a certain disease, artists relating to historical art and so on. Ontologies are divided into types in accord with the degree of generality of the principles they contain. Using the distinctive languages in the investigation of Ontology can be a challenge to the many endeavors of the Web designs to cater the huge number of users on the World Wide Web. Web data is typically language dependent. The accessibility of data identified with the language that would be much ideal as indicated by the client would be an expanding need of today.

The developing enthusiasm for ontologies for some common language applications in the current years has prompted to the production of ontologies. These ontologies are for various purposes and with various elements frameworks. Additionally, the current work in Artificial Intelligence is investigating the usege of formal ontologies.

Its use is a way of specifying content-specific agreements for the sharing and reuse of knowledge among software entities. There are various studies conducted in the Arabic language in Semantic Web. The propose of this studies is to improve the Arabic information retrieval on the web [1]. The ontology development life cycle had many questions around it in the last few years. Are there common designing criteria or not? Although Arabic is the language of hundred millions of people over the world, little has been done regarding computerized linguistic resources, tools or applications.

There are six parts in the life cycle in the development of ontology: Creation, Population, Validation, Deployment, Maintenance, and Evolution [3]. Manual ontology building is a time-consuming activity that requires many efforts for knowledge domain acquisition and knowledge domain modeling. To overcome these problems; many methods have been developed, including systems and tools that automatically or semi-automatically, using text mining and machine learning techniques, allows generating ontologies. The research fields which study this issues is