AUTOMATIC ONTOLOGY CONSTRUCTION APPROACH FROM WEB PAGES

By

Naglaa Elsayed Ahmed Mohamed Elmesalmy

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
COMPUTER ENGINEERING

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2019

AUTOMATIC ONTOLOGY CONSTRUCTION APPROACH FROM WEB PAGES

By

Naglaa Elsayed Ahmed Mohamed Elmesalmy

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
COMPUTER ENGINEERING

Under the Supervision of

Prof. Magda B. Fayek

Computer Engineering Department
Faculty of Engineering, Cairo University

Dr. Mayada M.Hadhoud

Computer Engineering Department
Faculty of Engineering, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2019

AUTOMATIC ONTOLOGY CONSTRUCTION APPROACH FROM WEB PAGES

By

Naglaa Elsayed Ahmed Mohamed Elmesalmy

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
COMPUTER ENGINEERING

Approved by the
Examining Committee

_____

Prof. Dr. **Magda B. Fayek**, Thesis Main Advisor

_____

Prof. Dr. **Nevin M. Darwish**, Internal Examiner

- Professor in Computers Engineering Department, Faculty of Engineering, Cairo university

_____

Prof. Dr. **Samia A. Mashali**, External Examiner

- Professor in Computers & Systems Department, Electronics Research Institute (ERI)

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2019

Engineer's Name: Naglaa Elsayed Ahmed Mohamed
Date of Birth: 25/11/1986
Nationality: Egyptain
E-mail: Naglaa.elmesalmy@gmail.com
Phone: 01008879638
Address: 21 Sheikh Al-Kafer st.- El Gamma Dept
Zagazig- El shariqua

Registration Date: 1/10/2011
Awarding Date: …./…./2019
Degree: Master of Science
Department: Computer Engineering

Supervisors:

Prof. Dr. Magda B. Fayek
Dr. Mayada M. Hadhoud

Examiners:

Porf. Dr. Magda B. Fayek (Thesis main advisor)
Prof. Dr. Nevin M. Darwish (Internal examiner)
Prof. Dr. Samia A. Mashali (External examiner)

- Professor in Computers & Systems Department, Electronics Research Institute (ERI)

Title of Thesis:

AUTOMATIC ONTOLOGY CONSTRUCTION APPROACH FROM WEB PAGES

Key Words:
Ontologies ; semantic web ; automatic ontology Construction ; web page ;Wordnet

Summary:

In the thesis, a new approach for automatic ontology construction from web pages is proposed. The proposed approach works as follows: first the triples of the document are extracted, next it utilizes natural language processing techniques to process the extracted triples, and finally it applies ontology design patterns before inserting triples into ontology. Our proposed approach is compared with the previous related work and the results show that we are closed to cover all the information in the documents correctly. Our system is better at reducing repeat rate where it do not insert the same object twice and do not repeat the relation between two concepts. Also our ontology is simpler representation than other work.

Disclaimer

I hereby declare that this thesis is my own original work and that no part of it has been submitted for a degree qualification at any other university or institute.

I further declare that I have appropriately acknowledged all sources used and have cited them in the references section.

Name: Date:

Signature:

Acknowledgments

All praises and thanks are due to Allah, the most gracious, the most merciful, for providing me with the strength, and patience to complete this work.

I am grateful to my supervisors, Prof. Dr. Magda Fayek, and Dr. Mayada for their guidance, advice, and encouragement toward successful completion of this work. They were very helpful indeed, reading and correcting me all the way.

I would like also to express my gratitude to my parents for their endless care and support from the very beginning and my husband for his wonderful support and encouragement.

Finally, I would like to dedicate this work to my husband, and my family.

Table of Contents

List of Tables

List of Figures

Nomenclature

AJAX	Asynchronous Javascript And Xml
API	Application Programming Interface
ASCII	American Standard Code For Information Interchange
ATCIS	Army Tactical Command Information System
CERN	The European Organization For Nuclear Research
CIDOC	International Committee For Documentation Of The International Council Of Museums
CRM	Conceptual Reference Model
D-FDED	Number Of Duplicates In FRED Ontology
D-ONTO	Number Of Duplicates In Our Ontology
DRS	Discourse Representation Structures
DTiMS	Defense Transaction Interface Module Systems
GWT	Google Web Toolkit
HTML	Hyper Text Markup Language
HTTP	Hypertext Transfer Protocol
IC-FRED	Number Of Incorrect Results In FRED Ontology
IC-ONTO	Number Of Incorrect Results In Our Ontology
IDF	Inverse Document Frequency
Info- onto	Number Of Information which Covered By Our Ontology
Info-FRED	Number Of Information which Covered By FRED Ontology
Info-No	Number Of Information
MOBM	Mixed Ontology Building Methodology
NER	Name Entity Recognizer
NLP	Natural Language Processing
NotCON-ONTO	Non-Connected Class In Our Ontology
NotCON-FRED	Non-Connected Class In FRED
OWL	Web Ontology Language
POM	Probabilistic Ontology Model
PVC	Professional Virtual Communities
RDF	Resource Description Framework
RSS	Really Simple Syndication
SQL	Structured Query Language
SVD	Singular Value Decomposition
TF	Term Frequency
TNO	Text Number
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
WWW	World Wide Web
XML	Extensible Markup Language

Abstract

Over the past decades, the Semantic Web got a great attention as the next generation Web. The Semantic Web adds semantics to the Web pages to become meaningful, machine-processable, and understandable. Ontology is a main component of the Semantic Web. In various fields, Ontology has been studied by many researchers.

Ontology plays an important role in exchanging knowledge between different organizations and within various Systems. It provides the description of data and the relations between them.

This thesis presents an approach for automatically constructing ontology from web pages thus transforming data from web into ontology. The proposed approach applies to web pages in any domain. It is a step towards the utilization of the huge amount of data published on the web to build ontology.

The proposed approach first extracts a set of triples from Web page, next it utilizes natural language processing techniques to process the extracted triples, and finally it applies ontology design patterns before inserting these triples into ontology.

We defined a number of performance metrics for automatic ontology generation. These metrics were used to compare our proposed system with other systems.

Compared to other related work and results, our proposed approach, show that we are closer to cover all the information in the documents correctly. Our system is good at reducing repeat rate where it does not insert the same object twice and does not repeat the relation between two concepts. Also our ontology is a simpler representation than other automatic systems. Total precision of our system ontology is 0.84 and our system recall is 0.79.

Chapter 1 : Introduction

The Semantic Web is the next generation of the current Web in which smart devices can understand the meaning of web contents because it contains Semantic annotations which provide additional information in the form of markup about various concepts (such as places, organizations, people, things etc.). It exchanges principles of the WWW (World Wide Web) from shared documents to shared data through a common framework which permits data to be shared and reused across applications, enterprise, and community boundaries. Ontology is the backbone of Semantic Web. The definition of Ontology is "a formal conceptualization of a particular domain that can be shared by a group of people (in and between organizations)"[1].

In semantic web, Ontology provides a sound semantic basis for the definition of meaning. It is typically used to present the natural language for communication between machines and humans [1]. It is a collection of concepts and the relationships between the concepts within a domain. It organizes the concepts into categories of concepts, each with its attributes, and describes relationships between concepts. When data is marked up using Ontology, softbots can well understand the semantics and therefore more intelligently locate and integrate data for a wide variety of tasks.

1.1. Motivation

Beside the significance of semantic networks, there are two more reasons that have encouraged me to propose this research:

1- Ontology availability is the most challenge of semantic web. Most proposed methods of ontology construction were for a specific domain. Also, ontology should be constructed by experts.

2- Amount of available knowledge on the current web is huge, important and in various fields. Therefore, there is a need to prepare this knowledge for usage in the semantic web.

These reasons led me think to develop a totally automatic approach for transforming the huge amount of knowledge in the web to ontology that can be used in the semantic web.

1.2. The Study Problem

The Wide World Web is a vast and rapidly growing source of information in all fields of science, biology, business, medicine, military, etc. The success of the Semantic Web depends strongly on transforming this information into a huge ontology, which provides a controlled vocabulary or conceptualization in these fields. Ontology is the core of the Semantic Web. In fact, Ontology is used to provide semantics and present a comprehensible, common foundation for resources on the Semantic Web. Also, ontology can present a common vocabulary and a rule for publishing data. Moreover, ontology can provide data with the semantic description. In this case, the WWW will be transformed from being machine-readable to machine-understandable.

Due to ontology's significance, many efforts have been exerted to develop ontology. The manual and semi-automatic constructional approaches of large ontology form web pages will not be feasible because of the effort, time and costs required. This gives place to the initiation of proposing approaches for constructing automatic ontology generation.

In this thesis, a new approach is proposed for automatic ontology construction that extracts the concepts and their relationship from web page. Then, it stores the concepts with their relationship on the ontology after making sure that they do not exist in ontology with the same relationship. Also, it uses the ontology design patterns in constructing the ontology.

1.3. The Thesis Organization

The remainder of the thesis is organized as follows. In chapter 2, an introduction about the basic concepts of semantic web and its layers is presented. In chapter 3, a literature survey about different ontology construction methods is presented. In chapter 4, the proposed approach for automatic ontology construction from web pages is explained in details. In chapter 5, the results of the comparison between the proposed approach and the previous work are pointed out and discussed. Finally, conclusion and future work are given in chapter 6.

Chapter 2 : Background

2.1. Introduction

The World Wide Web (WWW) is a collection of web pages and other web resources which are linked together through the internet. It was invented in 1989 by the English scientist Tim Berners-Lee who is the director of the World Wide Web Consortium (W3C). He was an employee at the European Organization for Nuclear Research (CERN) in Switzerland and presented a proposal for CERN communication system. Thus, the idea was implemented throughout the world. At the beginning of 1990, Belgian with Tim proposed the hyperlink to navigate across web pages. Tim also released the first web browser computer program in 1990, and consequently WWW began to appear to the world [2, 3, 4].

The World Wide Web plays a real role in the evolution of the Information Age and is the main tool for billions of people to interact on the Internet. There are billions of web pages, published and indexed as shown in figure 2.1[5]. Web pages contain text that annotated and formatted with Hypertext Markup Language (HTML). In addition to formatted text, web pages can display images, audio, video, etc.

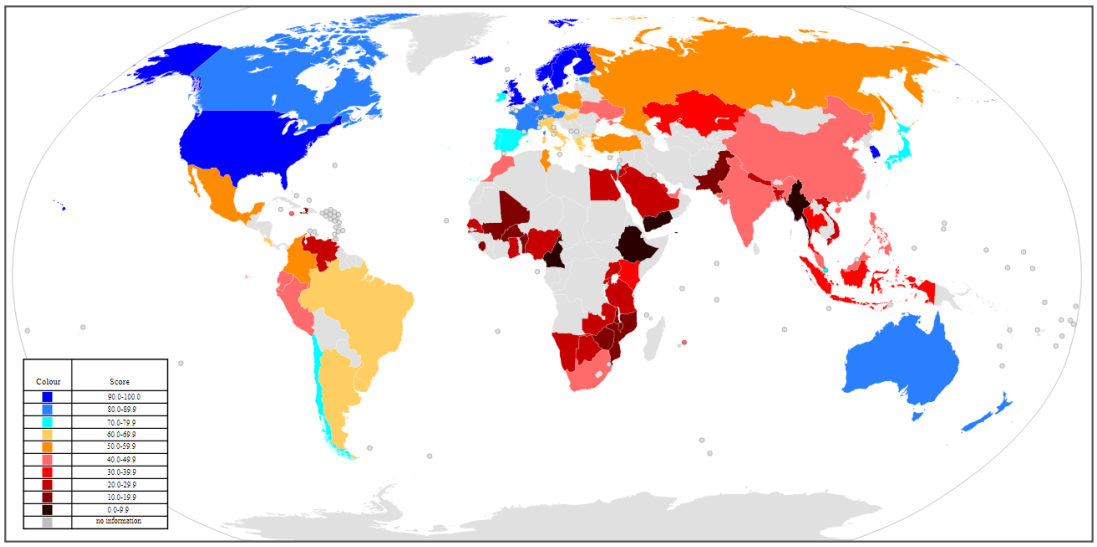

Figure 2.1: A global map of the web index for countries

2.2. Web Evolution And Semantic Web

Since the early 1990s, Web has been evolving in response to the never ending needs of users. Table 2-1 shows that the number of Internet users is 51.7% of the World population, and the growth rate is 976.4% in the period between 2000 to 2017[6]. This means that the Web should be more and more intelligent and sophisticated as the users' expectation is increasing.