WEB SERVICES FOR
ONTOLOGY BASED INFORMATION EXTRACTION

L. C. T. Silva

138234L

Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

March 2015

WEB SERVICES FOR ONTOLOGY BASED INFORMATION EXTRACTION

L.C.T.SILVA

138234L

This dissertation was submitted to the Department of Computer Science and Engineering of the University of Moratuwa in partial fulfillment of the requirements for the Degree of MSc in Computer Science specializing in Software Architecture

Department of Computer Science & Engineering
University of Moratuwa, Sri Lanka

March 2015

DECLARATION

The work included in this report was done by me, and only by me, and the work has not been submitted for any other academic qualification at any institution.

....

Chamendri Silva Date

I certify that the declaration above by the candidate is true to the best of my knowledge and that this report is acceptable for evaluation for the CS6999 MSc Research Project.

....

Daya Chinthana Wimalasuriya (PhD) Date

ABSTRACT

The amount of data contained in a textual format has increased rapidly in the recent past. Such data includes web sites, documents of business organizations, etc., and contain lots of information. Information Retrieval (IR) is a field that allows identifying relevant document for a given query out of all these available documents. Information Extraction is taking another step in this direction. Instead of returning the set of documents that contains the relevant information, IE recognizes and returns the information among the natural text in these documents.

Ontology is defined as the "*formal, explicit specification of a shared conceptualization*". It contains classes, properties, individuals and values to represent data in a certain domain. Most of the time in Ontology-Based Information Extraction, an IE technique is used to discover individuals for classes and values for properties to build ontology for a given domain. However, sometimes these classes and properties also identified as part of the IE technique rather than using a template with the pre-identified classes and properties in the Ontology.

A traditional Ontology Based Information Extraction system contains two main operations, ontology construction and ontology population. In the component-based approach defined in the "Ontology-Based Components for Information Extraction (OBCIE)", the operation of constructing ontology is not changed. However, the operation to populate the ontology is refined in to a pipeline of three separate components: pre-processors, information extractors and aggregators.

By developing these components as web services, we have provided the ability for other applications to use them to extract the information out of any text based document. To demonstrate this concept, we have developed an application that accepts a set of text documents, and extracts useful information. It uses "metadata files", which are dependent of the domain in which the ontology is created and populate the given ontology.

ACKNOWLEDGMENTS

I would like to express profound gratitude to my advisor, Dr. Daya Chinthana Wimalasuriya, for his invaluable support, encouragement, supervision and useful suggestions throughout this research work. His continuous guidance enabled me to complete my work successfully.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

IR	Information Retrieval
IE	Information Extraction
OBIE	Ontology Based Information Extraction
SOA	Service Oriented Architecture
NLP	Natural Language Processing
HTML	Hype-Text Markup Language
XML	Extensible Markup Language
OBCIE	Ontology Based Components for Information Extraction
WWW	World Wide Web
WSDL	Web Service Definition Language
UDDI	Universal Description Discovery and Integration
SOAP	Simple Object Access Protocol
REST	Representational State Transfer
RPC	Remote Procedure Call
HTTP	Hyper Text Transfer Protocol
EAI	Enterprise Application Integration
URI	Uniform Resource Identification
SOBA	SmartWeb Ontology Based Annotation
GATE	General Architecture for Text Engineering
JAX-RS	Java API for RESTFul Web Services
JAPE	Java Annotation Pattern Engine
GDM	GATE Document Manager
CREOLE	Collection of Re-usable Objects for Language Engineering
GGI	GATE Graphical Interface
WEKA	Waikato Environment for Knowledge Analysis
MALLET	Machine Learning for Language Toolkit

JSON	Java Script Object Notation
MIME	Multi-purpose Internet Mail Extension
CDDL	Common Development and Distribution License
API	Application Program Interface