

**WEB SERVICES FOR
ONTOLOGY BASED INFORMATION EXTRACTION**

L. C. T. Silva



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk 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



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk 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



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

.....

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.



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

TABLE OF CONTENTS

| | |
|---|------|
| DECLARATION | iii |
| ABSTRACT..... | iv |
| ACKNOWLEDGMENTS | v |
| TABLE OF CONTENTS..... | vi |
| LIST OF FIGURES | viii |
| LIST OF ABBREVIATIONS | ix |
| Chapter 1 Introduction | 1 |
| Chapter 2 Literature Review | 4 |
| 2.1 Ontology | 5 |
| 2.1.1 Ontology creation | 6 |
| 2.1.2 Ontology Population..... | 6 |
| 2.2 Information Extraction..... | 7 |
| 2.2.1 Ontology Based Information Extraction..... | 9 |
| 2.3 Ontology Based Components for Information Extraction | 11 |
| 2.3.1 Web Services for Ontology Based Information Extraction..... | 12 |
| 2.4 Web Services | 12 |
| 2.4.1 Service Oriented Architecture..... | 14 |
| 2.4.2 SOAP..... | 15 |
| 2.4.3 REST | 16 |
| 2.5 Related Work | 18 |
| 2.5.1 The OwlExporter | 18 |
| 2.5.2 SOBA | 19 |
| Chapter 3 Methodology | 21 |
| 3.1 High Level Architecture | 22 |
| 3.2 Extraction Methods..... | 22 |
| 3.2.1 Extraction Rules | 22 |
| 3.2.2 Two Phase Classification..... | 23 |
| 3.2.3 Tools used..... | 24 |
| 3.3 Framework for Building RESTful Web Services | 26 |
| 3.3.1 Apache Wink | 26 |
| 3.3.2 Project Jersey | 27 |
| 3.3.3 JBoss RESTEasy | 27 |
| 3.3.4 Restlet Framework..... | 28 |
| 3.3.5 Comparison..... | 28 |

| | |
|---|----|
| 3.4 Implementation | 29 |
| 3.4.1 Front End | 30 |
| 3.4.2 Preprocessor | 31 |
| 3.4.3 Extractor | 32 |
| Chapter 4 Results and Conclusion | 37 |
| 4.1 Implementation | 38 |
| 4.2 Conclusion and Future Work..... | 40 |
| REFERENCES | 41 |
| APPENDIX A: Sample Metadata Files – Information Extractor | 44 |
| APPENDIX B: Sample Metadata File – Two-Phase Classifier..... | 46 |



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

LIST OF FIGURES

| | |
|---|----|
| Figure 1 - Relationship between two classes | 5 |
| Figure 2 - Part of generalization/specialization hierarchy | 5 |
| Figure 3 - General Architecture of an OBIE System [3] | 10 |
| Figure 4 – Pipeline for Ontology Population in OBCIE | 12 |
| Figure 5 – General Process of Engaging Web Services [13] | 14 |
| Figure 6 – Obtaining REST architectural style from WWW | 17 |
| Figure 7 - General workflow of the OwlExporter [23] | 19 |
| Figure 8 – Pipeline used in SOBA | 20 |
| Figure 9 – High-level architecture | 22 |
| Figure 10 – Performance Metrics for different JAX-RS implementation [32] | 29 |
| Figure 11 – Message passing between web services | 31 |
| Figure 12 – A sample request json object sent to preprocessor | 32 |
| Figure 13 – Information Extractor method: a sample response json object from preprocessor | 32 |
| Figure 14 – Two-phase classifier method: a sample response json object from preprocessor. | 33 |
| Figure 15 – Information Extractor method: request json object for processing the metadata file | 34 |
| Figure 16 – Two-phase classifier method: request json object for processing the metadata file | 34 |
| Figure 17 – Information Extractor method: response json object for processing the metadata file | 35 |
| Figure 18 - Two-phase classifier method: response json object for processing the metadata file | 35 |
| Figure 19 – Information Extractor Method: A sample request json object to extractor | 35 |
| Figure 20 – Two-phase classifier Method: A sample request json object to extractor | 36 |
| Figure 21 - Information Extractor Method: a sample response json object to extractor | 36 |
| Figure 22 – UI developed to populate ontology using information extractor method | 38 |
| Figure 23 – UI developed to populate ontology using two-phase classifier method | 39 |
| Figure 24 - UI developed to view a given ontology | 39 |

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 |



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mru.ac.lk

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



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk