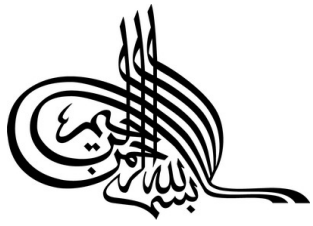
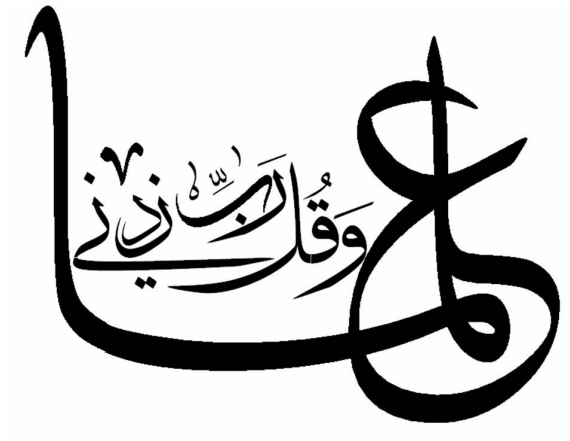

بسم الله الرحمن الرحيم


وقل رب زدني علما

Computer Science Department
Faculty of Computer and Information Sciences
Ain Shams University

Implementation of Enhanced Ontological Techniques for Information Retrieval Purposes

A thesis submitted as a partial fulfillment of the requirements for the degree of Master of Science in Computer and Information Sciences.

By

Amr Aly Amin Aly ElSehemy

B.Sc. in Computer & Information Sciences
Demonstrator at Computer Sciences Department
Faculty of Computer and Information Sciences, Ain Shams University

Under the supervision of

Prof. Dr. Taymoor Nazmy

Vice Dean for Research and Higher Studies
Faculty of Computer and Information Sciences, Ain Shams University

Dr. Mohamed Abdeen

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

Cairo, Egypt May 2015

Acknowledgement

All praise to Allah, the gracious and merciful, for helping me complete this work.

No words can describe my gratitude towards my parents, pushing me forward every step of the way. I could not have asked for a more supportive mother, doing everything she can to help. My father put no limits to the sacrifices he was willing to make to see me succeed. My sister and brother are the greatest gift my parents gave me, always putting me first. To them I owe the accomplishment of this work, and to them I put it forward as the fruit of their sacrifices.

I also present my sincerest thanks to my thesis advisors: Prof. Dr. Taymoor Nazmy, and Dr. Mohamed Abdeen. I thank them for the hours and hours spent brainstorming our way over obstacles. I'm lucky to have had such supportive advisors.

I sincerely appreciate the help of my friends, supporting me with great and insightful advice. I thank all my colleagues for their limitless support and encouragement.

Abstract

Keyword-based search engines were used to solve such problem, however, in many cases; keyword-based search engines either miss some documents or retrieve non-relevant documents. New technologies were presented to enhance the search results. One of these technologies is based on semantic web. Many semantic search techniques and models were made to enhance the traditional keyword-based search. Including conceptual knowledge such as ontologies in the information retrieval process contributes to the solution of major problems found in keyword-based search. Advances were made in languages such as English, German, French and Spanish. Although Arabic language is spoken by as many as 422 million native speakers, the literature does not fully cover it yet.

In this work, an architecture for an ontology-based information retrieval for Arabic language is presented. The architecture presented consists of four main modules, the query parser, the indexer, the search and the ranking modules. This work included building a semantic index; providing weighted links between the ontology concepts and the documents. Furthermore, a document categorizer to the architecture was built. The document categorizer was used as an additional process to enhance the overall document ranking.

To test our work, three Arabic domain ontologies were built. These ontologies are Sports, Economics and Politics. A knowledge base was built that consisted of 79 classes and 1456 instances. The document categorizer was evaluated using the WATAN-2004 corpus. We introduced some modifications merging the categories;

to be relevant to the work in here. The corpus contains around 20,000 articles in five categories (Culture, Religion, Economy, News and Sports). Twenty retrieval operations were manually chosen to assess different cases. The operations were applied on a sample of 40,316 documents with a size 320 Mega Bytes of pure text. Each operation was applied three times, once using the keyword-based search, and the other two using the presented ontology-based search with and without the classification module. The documents were downloaded from www.aljazeera.net news website.

This work was evaluated using the precision and recall metrics. The presented ontology-based search was compared against the traditional keyword-based search. The results showed that the ontology-based system performs better than the keyword-based system whenever ontology instances existed. The classification module was evaluated using the 10-fold evaluation technique. It achieved an average accuracy 78%, with a minimum error rate of 4%.

Further evaluation to the system was made by comparing between the ontology-based search with and without integrating the automatic document categorizer. The results showed an enhancement by 5% after integrating the categorizer module.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

IR	Information Retrieval
SPARQL	SPARQL Protocol and RDF Query Language
SW	Semantic Web
PWN	Princeton Word Net
AWN	Arabic Word Net
SSE	Semantic Search Engine
RDF	Resource Description Framework
OWL	Web Ontology Language
RDFS	Resource Description Framework Schema
NLP	Natural Language Processing
NN	Neural Networks
DCMI	Dublin Core Metadata Initiative
NER	Name Entity Recognizer
VSM	Vector Space Model
W3C	World Wide Web Consortium
KB	Knowledge Base
DARPA	Defense Advanced Research Projects Agency
DAML	DARPA Agent Markup Language
NB	Naïve Bayesian
KNN	K-Nearest Neighbor

Chapter 1

Introduction

Chapter 1: Introduction

1.1 Overview

Information Retrieval (IR) is one of the well-established research areas in Information Science. The goal of the IR discipline is to search and retrieve the most relevant documents to the information needs of the user. Therefore a good IR system should retrieve only those documents that satisfy the user needs, not other unnecessary data.

The increase in the amount and complexity of reachable information in the World Wide Web caused an excessive demand for tools and techniques that can handle data semantically. The current practice in information retrieval mostly relies on keyword-based search over full-text data. The keyword-based search is modeled with bag-of-words. Such a model introduces an issue, it does not present the actual semantic information embodied in text. To deal with this issue, ontologies are proposed for knowledge representation [1]. Ontologies are considered the backbone of semantic web applications. Both the information extraction and retrieval processes can benefit from such metadata, which gives semantics to plain text. It is also used to enhance the final ranking of the retrieved documents.

Having obtained the semantic knowledge and represented them via ontologies, the next step is querying the semantic data, also known as semantic search. There are several query languages designed for semantic querying. Currently, SPARQL Protocol and