Faculty of Computers and Information
Cairo University

# Knowledge Discovery from the Web

By
Maryam Mohy El Din Mohamed Hazman

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy
Computer Science

Supervised by
Prof. Dr. Ahmed Rafea
Prof. Dr. Salwa ElGamel
Dr. Samhaa El-Beltagy

2009

A thesis submitted to the faculty of Computers and Information, Cairo University, in partial fulfillment of requirements for the degree of Doctor of Philosophy of Computer Science in the department of Computer Science.

I certify that this work has not been accepted in substance for any academic degree and is not being concurrently submitted in candidature for any other degree.

Any portion of this thesis for which I am indebted to other sources are mentioned and explicit references are given.

Student: Maryam Mohy El Din Mohamed Hazman

# Approval Sheet

## Knowledge Discovery from the Web

By

Maryam Mohy El Din Mohamed Hazman

This Thesis for The Doctor Degree in Department of Computer Science, Faculty of Computers and Information, Cairo University, has been approved by:

| Name | Signature |
|---|---|
| 1- Prof. Dr. Ahmed Rafea | |
| 2- Prof. Dr. Salwa ElGamel | |
| 3- Prof. Dr. Mohammed Zaki | M. Zak |
| 4- Dr. Akram Salah | |
| 5- Dr. Samhaa El-Beltagy | Samhaa El Beltagy |

2009

# Acknowledgements

I would like to express my deepest appreciation and gratitude to the supervisors' committee Prof. Dr. Ahmad Rafea, Prof. Dr. Salwa ElGamal and Dr. Samhaa El-Beltagy for their contribution to the thesis.

I would like to express my deepest gratitude to Prof. Dr. Ahmed Rafea, who has enriched this work with his vast experience and knowledgeable guidance. It was only through his meticulous supervision, support, and constructive advice, that this work has been completed. Also, I would like to give my special thanks to Dr. Samhaa El-Beltagy for her cooperation, for her continuous support, her care, and for her vast experience.

Deep thanks to all my colleagues in Central Laboratory for Agricultural Expert System (CLAES) at Agricultural Research Center, Egyptian Ministry of Agriculture and Land Reclamation.

This work was partially supported by the Center of Excellence for Data Mining and Computer modeling within the Egyptian Ministry of Communication and Information.

This list of persons could be extended much further, therefore I would like to thank all those not mentioned above, and who have helped and supported me.

I cannot end without thanking my family for supporting me to accomplish this work.

# Abstract

As the web continues to grow, more and more information rich documents, such as books, manuals, and educational brochures are being availed through it. This amount of text is a valuable resource of information and knowledge, but to find useful information in this resource is not an easy nor a fast task. One of the main goals of the semantic Web is to extend the current Web by linking web content with semantic information thus making it easier and cheaper to locate specific information. The success of the Semantic Web depends on the existence of the annotations of Web pages with meta-data obtained from ontologies. Ontologies thus play a major role in the facilitation of communication among different software applications and users. Ontology building is lengthy, costly and controversial. Most of existing ontologies learning systems support construction of concepts and relations for English.

The aim of this thesis is to investigate the automation of the process of adding semantic annotations to web document segments using a domain ontology. In this way a simple search model can be used to retrieve self contained information entities at a level of abstraction that is easy to analyze and digest. Also, ontology learning approaches are investigated to use in the absence of a domain specific ontology. Specifically, the thesis investigates how to learn an ontology from Arabic semi-structured sources.

To achieve the thesis objectives, semantic annotation and ontology learning approaches were surveyed. A system was built to automatically annotate web pages and another was developed for semi-automatic ontology learning. The built systems have been applied and tested for tasks of using an ontology in semantic web annotation and for ontology learning. Results showed the success of the proposed methodologies and tools.

# Table of Contents

# Chapter 1

# Introduction

# 1. Introduction

## 1.1. Introduction

The web is a large and growing collection of texts. This amount of text is an invaluable resource of information and knowledge. However, finding useful information in this source is not an easy nor a fast task. A typical searcher would want to extract useful information from these texts quickly and at a low cost [Loh et al., 2000], but now that there are more than a billion pages indexed by search engines, finding the desired information within all these documents has proven to be a difficult task. The classical way for finding information on the Web is done using keywords entered to a search engine, which returns a list of pages ranked based on their similarity to the input query. However, finding relevant information using Web search engines often fails. One of the reasons for this is that users typically submit queries that are short and general, retrieving a large numbers of documents, the vast majority of which are of no interest to the user. The low precision of Web search engines coupled with the ranked list presentation forces users to sift through a large number of documents and makes it hard for them to find the information they are looking for [Zamir, 1999]. The Semantic Web aims to add semantics to Web content in order to make it easier to find desired information. This formal information tagging is done using ontologies, which are the backbone of the Semantic Web. Manual ontology authoring is very time consuming, and expensive. Automatically constructing ontologies from information existing in Web documents is still a challenge for researchers who want to turn the Web into more useful information utility. New and sophisticated techniques that have been developed in the area of Web mining (knowledge discovery), can aid in the extraction of useful information from the web [Garofalakis et al., 1999]. This information can be used as an available source for building an ontology. The ontology knowledge acquired can be reused for semantically annotating various portions of Web documents in order to facilitate finding and returning specific information that is relevant to a user's query instead of returning whole documents.

Knowledge discovery can be defined as "the non-trivial extraction of implicit, unknown, and potentially useful information from data" [Frawley et al., 1991]. The work carried out in this thesis explores knowledge discovery on two different levels both of which can help in the process of information finding. As stated before, ontologies are key to the success of the semantic web which in turn is vital for any future intelligent search. So the first knowledge discovery (KD) task addressed by this work is that of learning taxonomic ontologies from a set of domain specific documents. The second KD task is that of establishing a relationship between portions of web documents and concepts in the learned ontology via semantic annotation as this can greatly enhance search within these documents. Specifically, the second task relates to exploring the use of the acquired ontology for semantically annotating various portions of Web documents in order to facilitate finding and returning specific information that is relevant to a user's query instead of returning entire documents relevant to that query and wasting the user's time as s/he tries to find portions that correspond to their interests within these potentially long documents.

This chapter provides an introduction to this thesis. Section 2 describes the problem, and the fact that these changes should be reflected in the annotation of web page segments by their proper concepts. These concepts are selected from an ontology that can be learnt from an input set of web documents. In section 3 the objective of this thesis is outlined the methodology followed during the thesis is discussed. Finally, the structure of the thesis is presented in section 4. This gives a clear overview of the way the thesis is organized.

## 1.2. Problem Definition

The World Wide Web contains a humongous collection of texts, which is constantly growing. This amount of text is a valuable resource of information and knowledge. Finding useful information in this resource is not an easy task. People want to extract useful information from these texts quickly and at a low cost [Loh et al., 2000]. People either browse or use a search service when they want to find specific information in the web. Often, the desired information is contained within parts or paragraphs in a document rather than an entire document. In such cases, retrieving the paragraphs of interest would save the reader the time needed to go through the entire document in order to find the items of interest. However, all currently available search engines point to whole documents rather than to document fragments. In addition, when searching for an item of interest using a traditional search engine, numerous results are returned which requires the user to manually try to filter through these results. This places an overhead on the user in terms of time and effort.

Research in Web mining is moving the Web towards a more useful environment in which users can quickly and easily find information they need [Scime, 2004]. Also, semantic Web facilities aim at retrieving pages with high precision and recall by adding meta-data to web documents. The main idea of semantic annotation is to assign domain concepts in the form of semantic tags to documents, paragraphs, text experts, word segments, or phrases. These tags have to be well defined within a domain ontology and/or thesaurus. The tags can then act as descriptive meta-data that could facilitate retrieval based on their content [Soo et al., 2003] [Blythe & Gil, 2004]. Semantic annotations allow the use of concept search instead of keyword search and pave the way for more advanced search strategies. For example, users can specialize or generalize a query with the help of a concept hierarchy when too many or too few hits are found [Hollink et al., 2003]. Also, a concept can be used to define a context in which to focus an input query. The process of semantic annotation largely depends on the availability of ontologies [Cimiano et al., 2004].

Ontologies provide a shared and a common understanding of a domain that can be communicated between people and heterogeneous and distributed systems [Karoui et al., 2004]. In other words, ontologies are meta-data schemas, providing a controlled vocabulary of concepts, each with an explicitly defined and machine process-able semantics. By defining shared and common domain theories, ontologies help both people and machines to communicate and support the exchange of semantics and not only syntax.

Using ontologies in annotations can be said to serve two purposes. First, the user is immediately provided with the right context for finding an adequate index term. This ensures quicker and more precise indexing. Second, the hierarchical presentation of concepts helps to disambiguate terms [Hollink et al., 2003]. By using ontologies during search and information retrieval, it is possible to reduce the amount of non-relevant information in the returned results and retrieve only the relevant information [Lambrix, 2005]. However, for this task to be achieved we must first have an ontology. So, the cheap and fast construction of domain-specific ontologies is essential for the success and the proliferation of the Semantic Web [Maedche & Staab, 2001]. Building such ontologies is expensive, tedious, error-prone, biased towards their developer, inflexible and specific to the purpose that motivated their construction [Maedche & Staab, 2001] [Gomez-Perez & Manzano-Macho, 2003] [Shamsfard & Barforoush, 2003] [Sabou et al., 2005]. Ontology learning aims to accelerate the time and reduce the effort of building an ontology by acquiring concepts and relations semi-automatically or automatically from different information sources such as databases, documents, and/or web pages. Many ontology learning systems have been developed to speed up the learning ontology process using different sources of information and different techniques.

## 1.3. Thesis Research Objectives and Approaches

The main objective of this thesis is to devise means for annotating document segments with well defined meta-data. To achieve the objective of this work means for annotating document segments using an existing ontology as well as methods for learning an ontology when one does not exist, are investigated.

So, the thesis introduces a comprehensive methodology for building or using a domain-specific ontology for automatically annotating segments within web documents. This work takes advantage of information in existing organization web documents. First it investigates using an existing ontology for annotating segments that exist in Web document by their descriptive concepts to facilitate searching and browsing. Automatically finding the proper concepts for segments can be done by making use of segment heading titles which offer informal representations of segment content. The used ontology cannot contain all needed concepts for annotating the segments within the documents. Adding missing concepts in their right place in the ontology is done with confirmation from the user. To test the developed approach, it was applied to a set of agricultural extension documents and results compared to annotations provided by a human expert for the same set. The result of carrying out this experiment demonstrated that the proposed approach is capable of automatically annotating segments with concepts that describe a segment's content with a high degree of accuracy.

For cases when there is no available ontology, means of learning the required ontology were investigated. The presented learning approach aims to automate the process of constructing a taxonomic ontology using semi-structured domain specific web documents. To select the most representative candidate concepts from the documents, segment headings titles are used. Our system has been tested on building ontology in the agricultural domain using a set of Arabic extension documents.

## 1.4. Thesis Contributions

This thesis presents a methodology for the annotation of domain specific web document segments with minimal human effort. The methodology facilitates the retrieval of self contained information segments that are related to a user's query (focused search), as well as the browsing of annotated segments.

The main contributions of the thesis include:

- An approach for automatically annotating document segments (rather than an entire document) using their headings in conjunction with a domain ontology to