CHAPTER ONE

INTRODUCTION

Over the last decade, the Internet or World Wide Web (the Web) has become a real necessity, an indispensable and dynamic resource through which public users and researchers can obtain desired information. The explosive growth is due to the effect of the Internet and communication technologies evolution on all aspects of our life. The Internet is the most important and largest repository of multilingual documents and Web sites. One of the most powerful properties of the Internet is its multilingualism. The profusion of resources has prompted the need for developing automatic mining techniques on the Web.

Recently, several methods to effectively utilize the Web infrastructure have been reportedly known as Web intelligence (WI). The concept of Web intelligence was introduced by Zhong et al [Zhong et al., 2000]. According to Zhong et al, who coined the term Web intelligence (WI), “WI exploits artificial intelligence (AI) and advanced information technology (IT) on the Web and Internet”. WI was first explicitly introduced in the year 2000 as a joint research effort in developing the next generation of Web-based intelligent systems, through combining expertise in intelligent agents, Web mining, data mining, and information retrieval.

Broadly speaking, WI encompasses the scientific research and development that explores the fundamental roles as well as practical impacts of Artificial Intelligence, such as autonomous agents and multi-agent systems, machine learning, Web mining, data-mining, and soft computing, as well as advanced information technology, such as wireless networks, grid computing, ubiquitous agents, and social networks, on Web-empowered systems, services, and activities. WI aims at producing new theories and technologies that will enable us to optimally utilize the global connectivity, as offered by the Web infrastructure, in life, work, and play.
As more detailed blueprints and issues of WI evolved and specified in recent years, numerous WI research studies and business enterprises have been established around the world. WI companies and research centers/labs have been launched in the USA, Europe, Japan, India, etc. Each one of them focuses on certain specific WI issues or products/services.

A knowledge-based subsystem can enhance the capabilities of decision support not only by providing the subject matter knowledge, but also by providing expertise in data management and modeling. All technologies use knowledge, which is organized in a knowledge base, to provide the needed support of information search and retrieval. These are considered as applications of soft computing.

The idea of soft computing (SC) is still in its initial stage of crystallization. According to Lotfi Zadeh, who coined the term soft computing “Soft computing is an emerging approach to computing which parallels the remarkable ability of human mind to reason and learn in an environment of uncertainty and imprecision”. In other words, soft computing combines techniques taken from fuzzy logic, neural network, genetic algorithm, probabilistic reasoning and signal processing tools such as wavelet transforms to obtain robust solutions at low cost for problems which would be intractable by conventional means. Soft computing is a partner in its domain. The principal constituent methodologies in SC are complementary rather than competitive [Mitra & Acharya, 2003; Thuillard, 2001; Thuillard, 2003, Thuillard, 2004].

Soft computing is the basis of computational intelligence (CI) which encompasses swarm intelligence (SI), artificial immune systems (AIS). Soft computing usually refers to a collective set of CI paradigms and probabilities methods [Eberhart & Shi, 2007; Engelbrecht, 2007]. Using hybrid intelligent system is a promising research field of soft computing focusing on synergistic combination of multiple approaches to develop and solve the complexity problem of e-learning. Further details about these methods may be found in [Castro et al., 2007; Romero & Ventura, 2006; Abonyi et al., 2005; Dr’a’zdilov’a et al., 2010].

Artificial Neural Networks (ANNs) have been developed as a generalization of mathematical models of biological nervous systems which is designed to mimic the decision–making ability of the brain by providing a mathematical model of combination of numerous neurons connected in a network [Qian-jin et al., 2005; Oropesa et al., 1999]. It possesses a good learning capability that learns from given
input/output data pairs and adjusts the design parameters through minimization of error function using a suitable learning algorithm. Back propagation (BP) is one of the most widely used training algorithms for multi-layer neural networks [Wang, 1998]. The multi-layer neural networks have one or more layers of nodes (neurons) between the input and the output nodes, which are called the hidden layer [Lippmann, 1987]. BP is a gradient descend method and was first described in detail in 1986 by Hinton, Rumelhart and Williams, and is sometimes called Generalized Delta Rule (GDR).

1.1 MOTIVATION

During the first decade of the search engines, most of the search engines were developed for the English language because English was a lingua franca. Currently, a report published by Internet World Stats (www.internetworldstats.com/stats7.htm) at the end of 2009 showed that more than 65% of all Internet users are from non-English speaking areas as shown in figure 1. It is also reported that the population of non-English speaking Internet users is growing much faster than that of English speaking users. Asia, Africa, the Middle East and Latin America are the areas with the fastest growing online population.

Nevertheless, the unequal use of the language and the users’ lack to access the relevant materials in a language that they speak and understand seem to reinforce the growing knowledge gap between information and cognition processing on the Internet. In other words, the main problem is information access on the Web which is also called Web information retrieval due to several reasons: the dynamic, heterogeneous, and unlabeled resources. A comprehensive coverage of the entire important topic is impossible because of the high dimensional and time varying nature of the Web.

This problem has motivated UNESCO to consider one of its long-term and important goals, "to achieve worldwide access to e-contents in all languages, improve the linguistic capabilities of users and create and develop tools for multilingual access to the Internet." In [Chowdhury, 2003; Peters & Picchi, 1997], the authors have reported that the multilingual information retrieval has become a major challenge to access the prolific information on the Web and digital libraries. In addition, they divided the problem of Web multilingual information retrieval into two sets:

1- Identification, manipulation, and display of multiple languages.
Most of popular search engines such as Google (www.google.com) and Yahoo! (www.yahoo.com) are available in more than 50 languages. A search engine user obtains relevant information in another language based on the queries entered for one language (English into Arabic or Arabic into English) which is called Cross language information retrieval (CLIR). CLIR is a subfield of information retrieval which is an important field of language identification as well.

Therefore, the basic knowledge of the language identification of user queries plays an important role in order to increase the accuracy and performance of search engines. The major difficulties of language identification may be summarized into two questions: (a) How can the search engine identify the language for the required information and get the relationships among these topics? (b) How can the search engines go in depth of content and make classification of this huge amount of information in the web? One of the answers to these questions is the automatic query language identification. The study of language identification is a critical problem, especially in identifying the language of user queries in the Web search engines.

1.2 LANGUAGE AND UNICODE ENCODING

Computers process and store the information in several natural languages by dealing with numbers. In 1987, the UNICODE encoding was introduced (http://unicode.org/history/earlyyears.html) to solve the problems of many encoding systems. Some of these problems were that these encoding systems could not represent sufficiently many characters as the number and diversity of languages have grown. Also, they often conflict with one another.

According to the definition of UNICODE from Unicode consortium (http://unicode.org/), “Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language”. The Unicode consortium is an organization that provides a standard for representing different characters in different languages with a unique encoding number. One of the properties of Unicode is that all ASCII characters still exist in Unicode, either prefixed or suffixed with a null.

Unicode is an international standard for representing the characters used into plurality of languages. It also provides a unique numeric character, regardless of language, platform, and program in the world. Furthermore, it is popularly used in
2- Multilingual or cross-language information search and retrieval.

The first problem is based on finding technology solution to enable users to use and access information in whatever language it is stored and to facilitate efficient support of prior language identification. The second problem implies allowing the users to get relevant information in another language based on the information needed for one language by users.

Due to the aforementioned fact, we should have an adequate solution of the problem of multilingualism on the Internet. The foremost reason is that we are in a multilingual and multicultural world.

![Top 10 languages in the Internet millions of users](image)

Figure 1.1: Top 10 languages in the Internet.

Everyone in the world should have the opportunity to utilize multilingualism on the Internet and universal access to cyberspace. The search engines (or search services) and Web directories are the most general approaches to facilitate searching for huge amount of information on the Web.
Internet and any operating system of computers, device for text and visual representation and writing system in the whole world. Many software companies often use it.

Tables 1.1 and 1.2 show an example of the Unicode of two questions of the English and Arabic languages.

Table 1.1:
Decimal Unicode character of "What is your name?" for the English language.

<table>
<thead>
<tr>
<th>Query</th>
<th>Decimal Unicode</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ASCII</td>
<td>87 104 97 116 32 105 115 32 121 111 114 32 110 97 109 101 63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1.2:
Decimal Unicode character of "ما إسمك؟" for the Arabic language.

<table>
<thead>
<tr>
<th>Query</th>
<th>Decimal Unicode</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Unicode</td>
<td>1605 1575 32 1573 1587 1605 1603 1567</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1.3 RESEARCH CHALLENGES

We have witnessed an explosive growth in the information available on the Web which is designed for humans to read and understand, not for machines and computer programs to manipulate meaningfully. Therefore, we need to incorporate and embed artificial intelligence into Web tools. This is precisely the objective of our work to mimic the brain structure. The necessity of creating server-side and client-side intelligent systems that can effectively mine for knowledge both across the Internet and in particular Web localities is drawing the attention of researchers from the domains of information retrieval, knowledge discovery, machine learning, and artificial intelligence among others.

However, the problem of developing automated tools in order to find, extract, filter, and evaluate the users desired information from unlabeled, distributed, and heterogeneous Web data is far from being solved.

In fact, we need to handle these characteristics and overcome some of the limitations of existing methodologies. Soft computing is a consortium of methodologies that works synergistically and provides, in one form or another flexible information processing capability for handling real-life ambiguous situations. Its aim is to exploit the tolerance for imprecision, uncertainty, approximate reasoning, and partial truth in order to achieve tractability, robustness, low-cost solutions, and closely resembles human-like decision making. The guiding principle is to devise methods of
computation that lead to an acceptable solution at low cost by seeking an approximate solution to an imprecisely/precisely formulated problem.

1.4 CONTRIBUTIONS

The work presented in the thesis is an attempt to investigate the use of soft computing technique to enhance the current usage of the Web, and is concerned with how or whether those techniques can contribute to a more intelligent use of the Web.

The main contributions of the Ph.D thesis are the following:

- The Ph.D thesis proposes a new model of query processing for multilingual Web information retrieval. The proposed method deals with the multilingual query processing as one language or domain using Unicode and signal processing techniques.
- The thesis investigates the effect of signal processing techniques; namely, wavelet transform and multi-wavelet transform on multilingual query processing and proposes wavelet-based indexing of information.
- The present thesis proposes a solution to the problem of selecting optimum wavelet filters for multilingual query information retrieval through multi-wavelets transform.
- The Ph.D thesis proposes a model to query language identification based on ANN and the wavelet transform. The proposed model divides the identification into three phases: preprocessing phase, feature extraction phase, and classification phase.
- Finally, the thesis presents a novel method for lossy text compression. It investigates the proposed method with the fourier transform and the wavelet transform. The effectiveness of the proposed method is investigated through its application to multilingual text compression.

1.5 OUTLINE OF THE THESIS

The rest of the thesis is organized in 7 chapters as follows:

**Chapter 2: Soft Computing Approaches In Query Language Processing and Identification.** This chapter contains an overview of the theoretical background of the mathematical material used. It begins with information retrieval systems and deals
with two major categories of information retrieval models: statistical analysis and semantic analysis. The chapter also presents the preliminaries of the wavelet transform, and extension to the multi-wavelet transform as well as the artificial neural network.

**Chapter 3: Query Information Retrieval Model.** In this chapter, the wavelet transform as a way to index, and query information retrieval model is introduced. One illustrative example for wavelet-based indexing and the proposed method of wavelet based searching are shown.

**Chapter 4: Multilingual Query Processing.** It describes the methods that were developed to highlight this work which uses the Unicode processing format. The Unicode processing format aims to represent the characters of any language. Indeed, this use of Unicode will help us to process several languages, like Arabic, Japanese, and Chinese, etc. One demonstrative example for the multilingual method of wavelet based indexing such as the Arabic language is presented. In addition, this chapter points out the way to apply multi-wavelets, to get better results and show their advantages in contrast to the simple wavelet. The associated results and performance comparison are made with the scalar wavelet case to elaborate the work.

**Chapter 5: Cross-Language Identification.** This chapter, we propose a model based on wavelet and artificial neural network. The main objective of the proposed model is to make the Internet user easily access it using one’s language or any language for the required information. In addition, the benefit of the proposed model is to use a smart way for automatic language identification to reduce human efforts in translation and to depend on machine translation by the search engine.

**Chapter 6: Multilingual Lossy Text Compression.** In general, this chapter addresses a new strategy to make lossy text compression and multilingual lossy text compression. In particular, it investigates using the fourier transform and the wavelet transform.

**Chapter 7: Conclusions and Future Directions.** It gives the conclusions of the current work and proposes the future work in the areas of Web intelligence.