Abstract

In a time of wide availability of communication technologies, language barriers are a serious bottleneck to multilingual Internet user's integration and to economic and cultural exchanges in general. 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.

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. 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.

Recently, several methods to effectively utilize the Web infrastructure have been reported known as Web intelligence (WI). The concept of Web intelligence was introduced by Zhong et al. 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, and 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 of 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 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.

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 to human-like decision making. The guiding principle is to devise methods of computation that lead to an acceptable solution by seeking for an approximate solution to an imprecisely/precisely formulated problem. 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."

This challenge has led to a huge demand for information management technologies which can access it in an effective way while supporting comprehension and make better use of the knowledge that the electronic documents hold. A Web-based search engine works by matching user queries against a previous constructed index of Web content. The formation and makeup of this index is vital to the quality of the results returned by the search engine. More effective tools to overcome such barriers, in the form
of software for indexing, query processing, compression and other cross-lingual textual information access tasks, are in strong demand.

This work is an endeavor to investigate the use of soft computing techniques to enhance the current usage of the Web, and make it available to Internet users, regardless of the language(s) they speak and understand. Moreover, the current work is concerned with how or whether those techniques can contribute to a more intelligent use of the Web.

Our proposed model for the query information retrieval is presented which deals with the query/document like a signal depending on Unicode standard and the wavelet transform as indexing. Unicode is the international standard for representing the characters used into a plurality of languages. Also, Unicode provides a unique numeric character, regardless of language, platform, and program in the world. Moreover, it is popularly used in the Internet and in any operating system of computers, as a device for the text visual representation and for writing systems in the whole world. The wavelet transform provides a way to index and query information retrieval model is introduced. The query of this technique can be a word(s), small and large sentences, snippets, or document. These signal queries do not exclude anything from the query such as the spaces, commas, question marks, and prepositions or applying parsing or stemming on the query. One of the reasons is that we need to keep the position of every word in the query or document in order to get full meaning of the query (document). In addition, two novel methods have been proposed and tested for multilingual query processing namely, the wavelet transform and multi-wavelet transform. Perhaps even more importantly, the query information retrieval model presented here by the wavelet transform or multi-wavelet provided a framework for future research into search engine construction.
The thesis investigates the application of multi-wavelets for multilingual query processing method to get better results and demonstrate their advantages in contrast to the simple wavelet. Multi-wavelets transform is able to solve the problem of selecting optimum wavelet filters for query entered with these languages by Internet user. The associated results and performance are given in the same conditions with the scalar wavelet case to elaborate a comparison.

The Ph.D thesis also proposed a model to query language identification based on ANN and the wavelet transform. The proposed cross-language identification model is applied to English-Arabic and English-Urdu cross-language identification experiments. The proposed model has been divided into three phases: preprocessing phase, feature extraction phase, and classification phase. In the preprocessing phase, series of steps are applied to the signal query. This is done to select a good filter from wavelet filters and to speed up the next step, i.e., feature extraction. In the feature extraction phase, the FWT maps the signal into a unique vector for every language that is used by the classification phase. Here, the main tool used in the feature extraction phase is the fast wavelet transform decomposition with three filters, namely, Haar, Bior 2.2 and Bior 3.1 and Power Deviation (PD) method. In the classification phase, the signal query is recognized by ANN from the unique vector obtained as an output of the feature extraction phase. We use an approach that is a combination of wavelet transform and artificial neural network. Also, the PD method for signal classification which had proved its previous ability for identifying any change on the signal is used. In that, after the native language users enter his/her query into a search engine, the algorithm converts the query to Unicode and deal with the query as a signal query without removing anything from it. One example of
converting the query to signal is shown. After that, algorithm first checks whether the length of the signal query is greater or less than $2^n$. The reason of that the wavelet transform requires a signal length that is a power of $2 (2^n)$. Depending upon the outcome, either zeros are inserted (zero padding) or/and the signal query is divided into frames. Then, the FWT is computed to select a good filter which can be applied depending on its length and quality of construction (decomposition) and reconstruction. Thereafter, we first compute power deviation of each level and apply Parseval's theorem to the power levels, and then deviation and the normalization of power are carried out. Finally, the signal query is extracted from the unique vector obtained. The proposed model has proven the aptitude of the wavelet transform in detecting the desired points in a signal query and depicting its features in the time-scale (time-frequency) plane and combining them with artificial neural network for the purpose of language identification.

The Ph.D thesis has also investigated the use of fourier transform and the wavelet transform for lossy text compression. The influence of compression size and threshold of wavelet filters and the fourier transform as well as two parameters (i.e. families of wavelet filters and decomposition levels on compression factor of text files) were investigated. The experimental results have shown that the wavelet and the fourier transforms are suitable for lossy text compression with non-stationary text signal files. In addition, the fourier transform is the most suitable with files which have same characters such as aaa.txt and aaaa.txt files.

By the inspiration of valuable features of wavelet transform, the present work has demonstrated aptitude of the wavelet transform to the multilingual lossy text compression. The influence of two parameters (i.e. families of wavelet filters and
decomposition levels on compression factor of text files) are investigated. The experimental results showed that the proposed method gives satisfactory performance for multilingual text compression using the wavelet transform. This work also represented a step forwards dealing with both images and text compression i.e. multimedia compression. In other words, 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.
List of Publication out of Ph.D Thesis

Journal:


Book Chapter:


Conferences:


International Conference on Information and Knowledge Engineering (IKE'09), a track at The 2009 World Congress in Computer Science, Computer Engineering, and Applied Computing (WORLDCOMP'09), pp.716-722, July 13–16, Las Vegas, Nevada, USA.


