CHAPTER 1

INTRODUCTION

1.1 Introduction

Language is a term used in this research to refer to a natural communication system used for humans either in spoken or written forms. There are 7000 languages that have been reported in *Ethnologue*, a widely cited reference work on the languages around the world (Gordon, 2005). Globalization has led to unlimited information sharing across the Internet, where the communication among people in a bilingual environment is a critical challenge to be faced. Abd Rozan et al. (2005) have noted the importance of monitoring the behaviour and activities of world languages in cyberspace. The information collected from such studies has implications for customized ubiquitous learning\(^1\), in which Information and Communication Technology (ICT) has to cope with the “digital divides”\(^2\) that exist both within countries and regions and between countries. In addition, Maclean (2006) has reasserted the status of language as a topic of major concern for researchers in the light of the rise in transnational corporations. Also, Redondo-Bellon (1999) has analyzed the effects of bilingualism on the consumers in Spain. All these examples reflect the significance of multi-languages in globalization. In the book *The World is Flat* by Friedman (2005), the author writes:

---

\(^1\)According to Abd Rozan et al. (2005), customized ubiquitous learning means that the learning is best conducted in the natural language of the student and present everywhere at once.

\(^2\)Digital divide refers to the disparity between those who have use of and access to ICT versus those who do not (Abd Rozan et al., 2005).
“The net result of this convergence was the creation of a global, Web-enabled playing field that allows for multiple forms of collaboration—the sharing of knowledge and work-in real time, without regard to geography, distance, or, in the near future, even language. No, not everyone has access yet to this platform, this playing field, but it is open today to more people in more places on more days in more ways than anything like it ever before in the history of the world. This is what I mean when I say the world has been flattened.”

(Friedman, 2005, pg. 119)

According to Internet World Stats, the Internet usage increased dramatically between 2000 and 2008 in the world, especially in Middle Eastern countries such as Iran, Syria, Saudi Arabia, Yemen, etc. There are many people such as the Japanese, Arabic, Chinese, etc., that do not use an international language like English, therefore language identification is needed to support a multilingual processing system (Miniwatts Marketing Group, 2008; Payack, 2007). Language identification is the process of determining the predefined language automatically from a given content (e.g., English, Malay, Chinese, Japanese, Arabic, etc.). Language is an indispensable tool for human communication, and presently the language dominating the Internet is English. A web page is a kind of digital document displayed on a web browser. The web page can be written using diverse languages or different encoding scripts such as Unicode (Allen, 2006).

Figure 1.1 shows an example of web pages that use diverse scripts to display the content. The languages used on these web pages are Indonesian, Spanish, Malay, English, Chinese, Hindi, Russian and Arabic. A computer system can identify the character set or encoding scheme that has been applied, but it is not able to discriminate the precise language of web page\(^3\). Therefore, an effective and automatic web page language identification method is needed to solve this problem. The following sections present the problem as it is dealt with in this work: the problem statement, hypothesis, aim, objective, scope, significant of this research and thesis organization.

\(^3\)The details of the character set can be found in Appendix A.
Figure 1.1: Example of different language web pages. a) Indonesian b) Spanish c) Malay d) English e) Chinese f) Hindi g) Russian and h) Arabic
1.2 Problem Background

Language identification is frequently the initial step in a text processing system that may involve machine translation, semantic understanding, categorization, searching, routing or storage for information retrieval (Chowdhury, 2003). In order to allow the correct dictionaries, sentence parsers, profiles, distribution lists and stop-word list to be used, the prerequisite is to know the language of the text. Incorrectly identifying the language results in garbled translations, faulty or no information analysis and poor precision or recall in searching (Lewandowski, 2008). Language identification has been typically performed by trained professionals (Jin and Wong, 2002). The manual language identification process is very time-consuming and costly if performed by diverse language experts, and thus is of limited applicability. To overcome the inefficiency of the manual process, learning-based language identification methods have emerged. While existing methods can produce reasonable results, they often do so at a large computational cost (in terms of both space and time) (Jin and Wong, 2002). Many methods require large lists of words and/or N-grams with associated frequency counts for each language. Others require matrices whose size is dependent on the number of unique words and the number of documents in the reference language set. Calculations on large lists and matrices make these methods expensive to use (Botha et al., 2006).

There are several important areas of concern for automatic language identification. As the global economic community expands, there is an increasing need for automatic language identification services (Constable and Simons, 2000; SantoshKumar and Ramasubramanian, 2005). For example, checking into a hotel, arranging a meeting or making travel arrangement can be difficult for a non-native speaker. Telephone companies will be better equipped to handle foreign language calls if an automatic language identification system can be used to route the call to an operator fluent in that language. Furthermore, rapid language identification and translation can even save lives. There are many reported cases of emergency response operators being unable to understand the language of a distressed caller. In response to these needs, an automatic language identification system would be able to serve as the front-end for a multi-language translation system (Levow et al., 2005; Xu et al., 2008) in which the input speech can be in one of several languages. The input language needs to be quickly identified before translation into the target language can begin (Xafopoulous et al., 2004).
There are several difficulties that arise when dealing with web pages. For example, the programming code used for visual appearance of the web page, the grammatical errors in the contents, the use of the character set in formatting, and the exceedingly frequent use abbreviated forms or terms that are applied throughout the Internet (Mikami and Suzuki, 2004). All these examples reflect the noises present on a web page that can interfere with the identification process (Xafopoulos et al., 2004). In the Section 2.4, the problem of web page language identification will be described in detail.

With the rapid emergence of the Internet and the trend toward globalization, a tremendous number of web pages written in different languages are electronically accessible online. Efficiently and effectively managing these web pages is important to organizations and individuals. For this purpose, many studies have been carried out in order to identify automatically the language in which the information is written on a web page (Xafopoulos et al., 2004). A suitable method of feature selection or extraction of web pages is required to extract the usable features from web pages before the identification process is begun. One of the fundamental motivations for feature selection is the curse of dimensionality (Friedman et al., 2001). The number of features is a key factor that determines the size of the hypothesis space containing all hypotheses that can be learned from data (Mitchell, 1997). The more features, the larger the hypothesis space. Indirectly, the classification performance can be expedited if the features used are reliable and robust (Botha et al., 2006). With the increasing number of web pages on the Internet, it has become a necessity to provide some techniques to identify and retrieve effectively encoded information automatically.

1.3 Problem Statement

In this study, it is intends to come up with a method to provide insights into solving the feature selection and classification of web page language identification. The research question is:
How can one produce reliable features that are able to be used for identifying the language of web pages accurately?

In order to answer the main issue raised here, the following issues need to be addressed:

(i) How have previous works solved the problem of web page language identification?

(ii) It is well known that web pages consist of many noises, such as programming language and nonstandard encoding schemes. How can this be overcome?

(iii) What is the problem of existing methods like N-grams and entropy for selecting features from the web pages? How can this be overcome?

(iv) What is the most suitable classification method for web page language identification?

(v) How can one perform web page language identification based on finer granularity within a web page such as characters, words, sentences, etc.?

(vi) How can one test the bias of web page data set and the accuracy of web page language identification?

1.4 Hypothesis

In this research, the proposed feature selection method on the web page language identification is used to improve the performance in terms of accuracy. Therefore, several assumptions have been made:

(i) that the preprocessing method being applied will increase the effectiveness of web page language identification.
that the feature selection method is one of the impact factors in the performance of web page language identification, and that the feature selection method being used in web page language identification will enhance the identification results.

(ii) that the use of a suitable identifier from the machine learning methods will increase the identification results.

1.5 Aim

The aim of this study is to enhance the performance of web page language identification.

1.6 Objectives

In order to achieve this aim, the following objectives have been established:

(i) To review previous research related to web page language identification.

(ii) To propose an improved feature selection method for web page language identification.

(iii) To test the performance of the proposed method on web page language identification.
1.7 Scope

The scope of this research has been limited to the following:

(i) This research focuses only on web page language identification, and does not include web documents such as Portable Document Format (PDF), Word documents, Excel documents, etc.

(ii) The data set used is Roman, Arabic, Cyrillic, Indic and Hanzi script web pages only.

(iii) The machine learning methods involved are supervised neural networks such as a decision tree, a Backpropagation Neural Networks (BPNN) and the adaptive neural networks.

(iv) The data sets are obtained from news websites such as British Broadcasting Corporation (BBC), Cable News Network (CNN) or other available web repositories.

(v) The collection contains news articles concerning politic, sport and health in order to obtain a reasonable degree of diversity, but it does not include scientific web pages such as biology, chemical, etc.

(vi) The method involves process crawling of web pages using HTTPTrack crawler (Roche, 2008).

(vii) The $f$-fold cross validation procedure is used as a benchmark of evaluation.

(viii) The standard measurements such as accuracy, precision, recall and $F1$ measurements are used for evaluating performance of web page language identification.

(ix) This research does not involve character set or encoding scheme identification; it is assumes all the web pages are converted into Unicode.

(x) This work is based on Java programming.
1.8 Significance of the Research

(i) It improves the conventional method into two feature selection methods; letter weighting and simplified entropy.

(ii) It demonstrates the importance of the preprocessing step in web page language identification.

(iii) It reveals the actual performance procedures of various classification methods for web page language identification.

1.9 Contribution of the Research

(i) It supports the existing language identification technology in order to realize the natural language processing automatically on computer.

(ii) It promotes the ubiquitous learning environment based on one’s native language either is study or working.

(iii) It prevents the digital divide of minority languages on internet.

1.10 Thesis Organization

The thesis consists of 5 chapters, each of which is briefly described as follows:

(i) Chapter 1 describes the background, problem statement, hypothesis, aim, objectives, scope, significance of research and ends with an overview of the thesis organization.
(ii) Chapter 2 presents an introduction to the Internet, an overview of language identification, the problems of web page format, the problem of feature selection method, the conventional web page language identification processes, concluding with a review of the literature on the feature selection method, the classification method and the evaluation approach.

(iii) Chapter 3 describes the operational framework and also the methodological steps adopted to perform the web page language identification, such as data preparation, data preprocessing, feature selection and identification methods.

(iv) Chapter 4 compares the results and the discussion of each experiment.

(v) Chapter 5 concludes the study with finding of this research, thesis contributions, suggestions for future research and a summarizing conclusion.

1.11 Summary

The introduction to web page language identification has been discussed, including the problem background, objectives, scope, etc. In order to enhance the method of web page language identification, the following section describes the advantages and disadvantages of the previous work related to web page language identification. Following this, an operational framework has been proposed for improving the web page language identification based on the objectives that have been defined in this chapter.