FEATURE SELECTION METHOD OF WEB PAGE LANGUAGE IDENTIFICATION

NG CHOON CHING

UNIVERSITI TEKNOLOGI MALAYSIA

FEATURE SELECTION METHOD OF WEB PAGE LANGUAGE IDENTIFICATION

NG CHOON CHING

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

FEBRUARY 2010

To my beloved mother Mrs. Goh Ah Chew, family members and friends

ACKNOWLEDGEMENT

In preparation this thesis, there are many people have contributed towards my understanding and thoughts. I would like to acknowledge, and also I wish to express my deep gratitude to my supervisor, Assoc. Prof. Dr. Ali Selamat for guidance, valuable time, technical and friendly dealing through out my study. Without his continued support, guidance and interest, thesis would not have been the same as presented here.

I have received direct or indirect help and support from many personalities that motivated and enabled me to conduct this research. Thanks are due to Prof. Dr. Ahmad Zaki Abu Bakar, Assoc. Prof. Dr. Syed Malek Fakar Duani Syed Mustapha, Assoc. Prof. Daut Daman, Prof. Dr. Safaai Deris, Prof. Dr. Robert Michael Colomb, Assoc. Prof. Abdul Manan Ahmad, Dr. Mohd Zaidi Abd Rozan, Dr. Nor Zairah Abdul Rahim, Assoc. Prof. Dr. Siti Zaiton Mohd Hashim, Kak Lijah, Kak Amelia, Imam Much Ibnu Subroto, Ahmad Nadzri Muhammad Nasir, Muhammad Tarmizi Lockman, Siti Dianah Abdul Bujang, Muhammad Khairi Ismail, Siti Nurkhadijah Aishah Ibrahim, Morpheus Tey, Yiew-Siang Lee, Zhi-Sam Lee and Hui-Ming Teo for their help and valuable comments. In addition, I would like to present my sincere appreciation to Professor Richard L. Spear for the valuable suggestions in improving the thesis.

A great gratitude also goes to the Ministry of Science, Technology & Innovation (MOSTI), Malaysia and Research Management Center, Universiti Teknologi Malaysia (UTM), in providing scholarship of National Science Fellowship (NSF) for this work.

ABSTRACT

Globalization has led to a significant increase in the information flow between geographically remote locations with the realization of a common global market. When building a web site for use by various industries, developers need to deal with a wide range of users from different countries. Thus, a multilingual system must be implemented in order to provide the proper environment for those applications. Different languages can be produced by using the same script such as English, Malay, Spanish, etc., that uses Roman script. The issue is how to produce the reliable features of a web page that is to undergo language identification. Incorrectly identifying the language will results in garbled translations, faulty and incomplete analyses. The aim of this study is to enhance the effectiveness of feature selection method of web page language identification. A letter weighting method as feature selection embedded with fuzzy Adaptive Resonance Theory Map (ARTMAP) and simplified entropy embedded with decision tree are proposed to identify the language belonging to a web page. The methodology contains four major stages, namely; data preparation, data preprocessing, feature selection and identification. Data is collected from news website and then fed into preprocessing to filter out the noises. Feature selection reduces unnecessary attributes of the data in a proper feature representation. Language identification is to determine the predefined language of data. The scripts of languages such as Arabic, Hanzi, Roman, Indic and Cyrillic were used for the performance evaluation of web page language identification. Standard measurements such as T-test, f-fold cross validation, precision, recall and $F1$ measurements were used on results of the analysis. From the experimental analysis, it is observed that the simplified entropy outperforms the N-grams, entropy and letter weighting feature selection with an accuracy of 98.90%, 81.35%, 96.08% and 93.16%, respectively. The finding concludes that the proposed letter weighting and simplified entropy feature selection methods of web page language identification give promising results in terms of accuracy and retrieval performance at the letter representation level of web pages.

ABSTRAK

Era globalisasi telah menyebabkan peningkatan yang bermakna dalam pembangunan maklumat di antara lokasi terpencil dengan pasaran global. Oleh itu, pembangun perisian perlu berurusan dengan para pengguna dari seluruh dunia ketika membina laman web yang akan digunakan oleh pelbagai industri. Dengan demikian, sistem pelbagai bahasa harus dilaksanakan untuk memudahkan para pengguna. Pelbagai bahasa seperti Bahasa Inggeris, Bahasa Melayu dan Sepanyol dapat dihasilkan dengan menggunakan script *Roman*. Persoalannya adalah bagaimana untuk menghasilkan ciri-ciri laman web yang dapat digunakan untuk pengenalan bahasa. Ini disebabkan salah mengenalpasti bahasa akan menghasilkan penterjemahan yang salah dan tidak lengkap. Maka, objektif kajian ini adalah untuk meningkatkan keberkesanan kaedah pilihan ciri laman web untuk pengenalan bahasa. Dua kaedah telah dicadangkan untuk mengenalpasti bahasa laman web iaitu pemberatan huruf dengan *fuzzy ARTMAP* dan entropi ringkas bersama dengan *decision tree*. Terdapat empat tahap utama dalam metodologi, iaitu: persediaan data, pra-pemprosesan data, pilihan ciri dan pengenalan. Persediaan data dilaksanakan dengan mengumpul data daripada laman berita dan menjalankan pra-pemprosesan data untuk membuat penapisan. Seterusnya, pilihan ciri digunakan untuk mengurangkan atribut daripada data dengan representasi ciri yang tepat. Akhirnya, pengenalan bahasa dilaksanakan untuk menentukan bahasa laman web. Script bahasa seperti *Arab*, *Hanzi*, *Roman*, *Cyrillic* dan *Indic* digunakan untuk penilaian prestasi pengenalan bahasa. Pengukuran standard seperti *T-test*, *f-fold cross validation*, presisi, *recall* dan $F1$ telah digunakan untuk menganalisis keputusan. Dari analisis percubaan, didapati bahawa kaedah entropi ringkas lebih baik berbanding dengan kaedah N-grams, entropi dan pemberatan huruf yang mempunyai ketepatan masing-masing 98.90%, 81.35%, 96.08% dan 93.16%. Kesimpulannya bahawa pengenalan bahasa laman web memberikan hasil yang menjanjikan ketepatan dan peningkatan prestasi pada peringkat penggunaan huruf dalam laman web.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	-	Artificial Neural Networks
ART	-	Adaptive Resonance Theory
ARTMAP	-	Adaptive Resonance Theory Map
ASCII	-	American Standard Code for Information Interchange
BBC	-	British Broadcasting Corporation
BPNN	-	Backpropagation Neural Networks
CLIR	-	Cross Language Information Retrieval
CNN	-	Cable News Network
CI	-	Confidence Interval
CPBF	-	Class Profile-Based Approach
DBI	-	Dictionary Based Identification
DNA	-	Deoxyribonucleic Acid
ECI	-	European Corpus Initiative
FESE	-	Feature Selection Method
GMM	-	Gaussian Mixture Model
HMMs	-	Hidden Markov Models
HTML	-	Hypertext Markup Language
ICA	-	Independent Component Analysis
ICT	-	Information and Communication Technology
ICU	-	International Components for Unicode
ISO	-	International Organization for Standardization

KNN	-	K-Nearest Neighbor
LAN	-	Local Area Network
LFDF	-	Letter Frequency Document Frequency
LID	-	Language Identification
ML	-	Maximum Likelihood
MMI	-	Maximum Mutual Information
OCR	-	Optical Character Recognition
PCA	-	Principle Component Analysis
PDF	-	Portable Document Format
RMSE	-	Root Mean Squared Error
RNA	-	Ribonucleic Acid
SMS	-	Short Message Services
SOM	-	Self Organizing Map
SVD	-	Singular Value Decomposition
SVM	-	Support Vector Machine
SWT	-	Small Word Technique
TFIDF	-	Term Frequency Inverse Document Frequency
TT	-	Trigram Technique
UDHR	-	Universal Declaration of Human Rights
URL	-	Uniform Resource Locator
UTF-8	-	8-bit UCS / Unicode Transformation Format
VQ	-	Vector Quantization
WAN	-	Wide Area Network
WTA	-	Winner Take All
WWW	-	World Wide Web
XHTML	-	Extensible Hypertext Markup Language

LIST OF SYMBOLS

a	-	a is a feature vector $(0 \leq a_o \leq 1)$ of fuzzy ARTMAP
$\tilde{a}$	-	The system and the expert agree with the assigned category
a_c	-	The complement of a (e.g., if a = 0.4 then a_c = -0.6)
A	-	A matrix document N-grams
Acc_D	-	Accuracy of the experiment data set *D*
A_{ik}	-	Average accumulated frequency of particular letter *k* in particular language *i*
α_{ik}	-	Local letter weighting of particular letter *k* of particular language *i*
$\hat{\alpha}$	-	A non-zero vector of PCA
b	-	Number of subsets D_h in collection D
$\tilde{b}$	-	The system disagrees with the assigned category but the expert did
$\hat{b}$	-	The characters of word
β_{ik}	-	Global letter weighting of particular letter *k* of particular language *i*
$\hat{\beta}^{\tilde{q}}$	-	A codebook of language $\tilde{q}$ for algorithm VQ
B	-	Complement coded input vector of fuzzy ARTMAP
$\tilde{c}$	-	The expert disagrees with the assigned category but the system did
C	-	Number of committed coding nodes of fuzzy ARTMAP
$\hat{C}$	-	A real symmetric matrix of PCA

$\tilde{C}$	-	The universe of languages of algorithm VQ
χ	-	Convergence point of simplified entropy
co	-	Number of correct identifications
d	-	The desired value that appropriate to be feed into identifier of PCA
$\tilde{d}$	-	The system and the expert disagree with the assigned category
d_j	-	Particular document *j*
δ_q	-	The generalized error through a layer q of the Backpropagation Neural Networks (BPNN)
δ_r	-	The generalized error through a layer q and r of the BPNN
D	-	A collection of information (or web documents)
D_h	-	Subset *h* of collection *D*
DF_k	-	Document frequency of particular letter *k*
DoF	-	Degree of freedom of T-test
E_h	-	Accuracy of subset h
EN_k	-	Entropy weighting of particular letter *k* in collection, *D*
EN_{jk}	-	Entropy weighting of particular letter *k* in particular document *j*
ℓ	-	Sum of features of simplified entropy
$\hat{\ell}$	-	Iteration number of BPNN
ℓ_{new}	-	Total of current feature frequency of simplified entropy
ℓ_{old}	-	Total of previous feature frequency of simplified entropy
η	-	Learning rate of the BPNN
f	-	Number of folds in cross validation
$\hat{f}$	-	The frequency of the word of Zipf's Law
$F1$	-	The F1 measure is the average of precision, $\tilde{p}$ and recall, $\tilde{r}$
F_k	-	Frequency of letter *k* in the collection *D*
γ	-	Prefix convergence point of simplified entropy
Γ	-	Momentum rate of the BPNN

G_k	-	Global entropy of particular letter *k*
h	-	The hyperplane of SVM classifier
i	-	The number i^{th} language
I	-	Number of input features of fuzzy ARTMAP
in	-	The input values to the BPNN where $in \in [1, s]$
j	-	The number j^{th} document
J	-	During the training process, the fuzzy ARTMAP approach searches for a chosen coding node J that meets the matching criterion
k	-	Particular letter
K	-	Number of cluster of KNN
$\hat{\lambda}$	-	A positive real number of PCA
$\hat{L}$	-	A textual document of VQ
L_i	-	Particular language, *i*
L_{jk}	-	Local entropy of particular letter *k* in particular document *j*
L_m	-	Particular passive language, *m*
$\hat{L}_q$	-	A document $\hat{L}$ of language $\tilde{q}$ of VQ
LF_{jk}	-	Letter frequency of particular letter *k* in particular document *j*
$LFDF_{jk}$	-	Letter frequency document frequency of particular letter *k* in particular document *j*, so called simplified entropy
$\hat{m}$	-	Number of documents in PCA
M	-	A vector of number of fixed size M
$\hat{M}$	-	Total passive languages
min_z	-	Minimum value in the dimension z of input patterns
max_z	-	Maximum value in the dimension z of input patterns
n	-	Number of observations of T-test
$\hat{n}$	-	Number of N-grams of PCA
net_q	-	The first transfer function at hidden layer q of the BPNN
net_r	-	The second transfer function at output layer r of the BPNN

ngm	-	A particular N-grams
$ngm_{\breve{r}}$	-	N-grams of the testing document ordered descending
$ngm_{\hat{r}}^{L_m}$	-	N-grams of the particular passive language model ordered descending
N	-	Number of document in the collection *D* of particular language *i*
NF	-	Particular N-grams frequency in a document
$\hat{N}$	-	The codewords of codebook for algorithm VQ
o	-	The input component index of fuzzy ARTMAP
out	-	The output values to the BPNN where $out \in [1, 2]$
O_p	-	Output on unit p of the BPNN
O_q	-	Output on unit q of the BPNN
O_r	-	Output on unit r of the BPNN
ω_{ik}	-	Letter weighting of particular letter *k* of particular language *i*
p	-	Input layer of BPNN
$\tilde{p}$	-	The precision describes the probability that an desired document (randomly selected) retrieved document is relevant to a certain language
$\hat{p}$	-	The coding node index of fuzzy ARTMAP
pa	-	Number of patterns / samples
P_{ik}^{m}	-	Average accumulated frequency of particular letter *k* in passive language *m*
q	-	Hidden layer of BPNN
$\hat{q}$	-	The output class index of fuzzy ARTMAP
$\tilde{q}$	-	A document's language of algorithm VQ
$\tilde{Q}$	-	Number of languages of $\tilde{q}$
r	-	Output layer of BPNN
$\hat{r}$	-	The rank of the word in the list ordered descending by the frequency of Zipf's Law

$\breve{r}$	-	The rank of the N-grams in the predicted text ordered descending by the frequency of Zipf's Law
$\tilde{r}$	-	The recall describes the probability of a relevant language being retrieved
R	-	Threshold or number of features of input patterns
ρ	-	Vigilance variable of fuzzy ARTMAP
s	-	Window size of windowing algorithm
$\hat{S}$	-	Particular script of languages
$\hat{S}_{begin}$	-	The begin codepoint of a particular script, $\hat{S}$
$\hat{S}_{end}$	-	The end codepoint of particular script, $\hat{S}$
S	-	Standard deviation of the mean
σ	-	Standard deviation is a measure of the dispersion of a set of values
t	-	Critical value of T-Test
$\hat{t}$	-	Number of letters in a document
t_k	-	Particular letter *k*
T_d	-	Total N-grams in a document
TF_j	-	Term frequency of all letters *k* in document *j*
TF_{jk}	-	Term frequency of particular letter *k* in document *j*
θ_q	-	A bias on hidden unit q of the BPNN
θ_r	-	A bias on output unit r of the BPNN
ε	-	Match tracking ($\varepsilon \in (-1, 1)$) of fuzzy ARTMAP
ϖ_z	-	Original input patterns of machine learning
φ_z	-	Normalized input patterns of machine learning
w	-	Weight vector of fuzzy ARTMAP
$\hat{w}$	-	The elements of set $\hat{W}$, words
$w_{\hat{p}}$	-	The coding node weight vector $\hat{p}$ of fuzzy ARTMAP
$w_{\hat{q}}$	-	The output class weight vector $\hat{q}$ of fuzzy ARTMAP
W_{qp}	-	The q^{th} weight to the unit p^{th} of the BPNN

W_{rq}	-	The r^{th} weight to the unit q^{th} of the BPNN
$\hat{W}$	-	A set of words $\hat{w}$ for algorithm VQ
x	-	The element of A
$\hat{x}$	-	The samples of KNN
$\bar{x}$	-	The mean of x
$\tilde{x}$	-	Mean is the arithmetic average of a set of values or distribution
$x_{\hat{n}m}$	-	All the N-grams exist in the collection, $\hat{n}$ is the number of N-grams and $\hat{m}$ is the number of documents
y	-	Coding field activation pattern of fuzzy ARTMAP
z	-	Dimension of input pattern of machine learning
ζ_{cur}	-	Index position of current feature of simplified entropy with frequency changed
ζ_{pre}	-	Index previous position of current feature of simplified entropy with frequency changed

LIST OF APPENDICES

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.

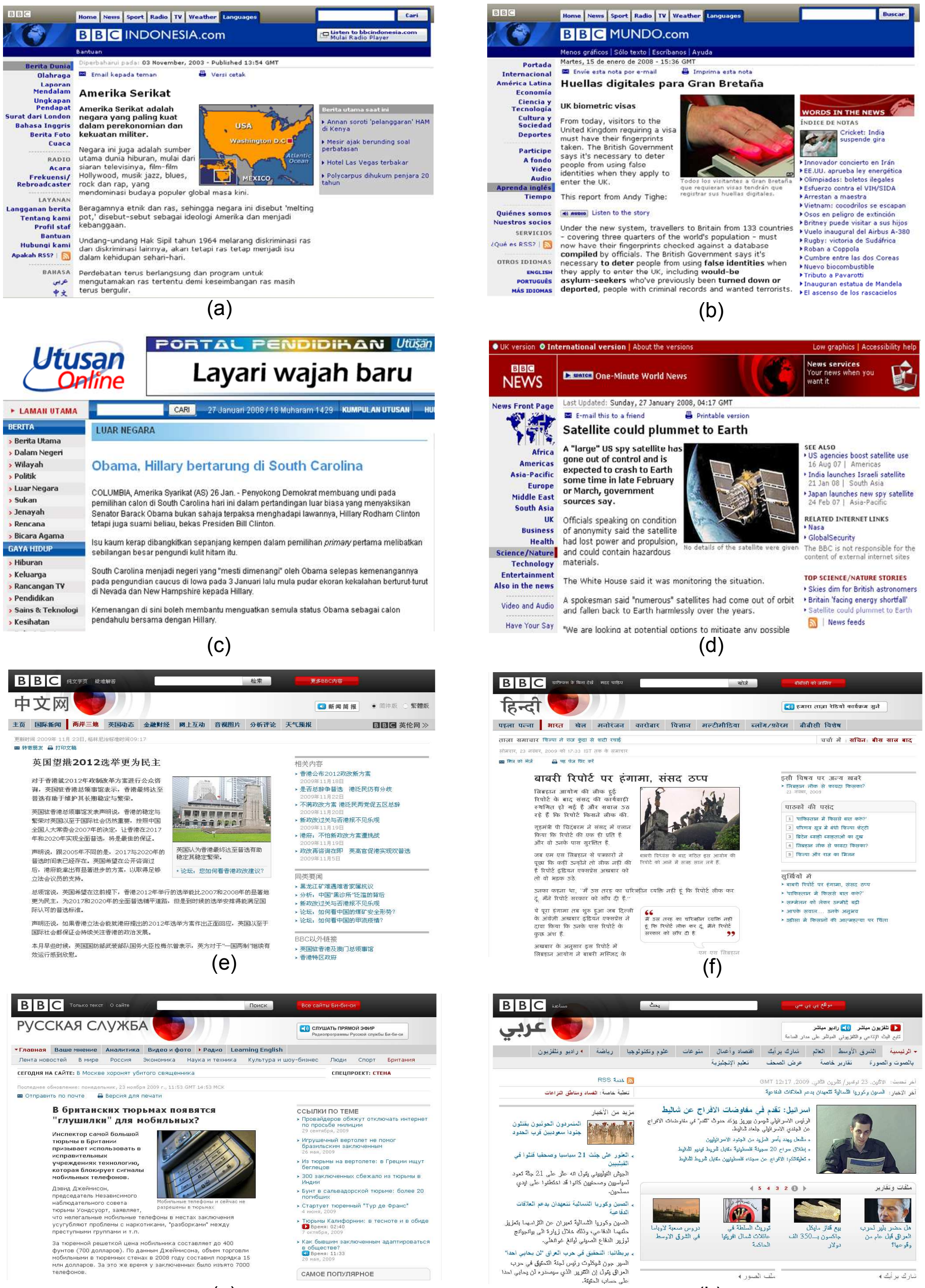

Figure 1.1: Example of different language web pages. a) Indonesian b) Spanish c) Malay d) English e) Chinese f) Hindi g) Russian and h) Arabic

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 (Xafopoulos *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 know that web pages consist of many noises, such as programming language and non standard 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.

(ii) 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.

(iii) 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.