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. Safaaui 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.
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 result 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

### TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DECLARATION</td>
<td>ii</td>
<td></td>
</tr>
<tr>
<td>DEDICATION</td>
<td>iii</td>
<td></td>
</tr>
<tr>
<td>ACKNOWLEDGEMENT</td>
<td>iv</td>
<td></td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>v</td>
<td></td>
</tr>
<tr>
<td>ABSTRAK</td>
<td>vi</td>
<td></td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>vii</td>
<td></td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>xi</td>
<td></td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xii</td>
<td></td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>xiv</td>
<td></td>
</tr>
<tr>
<td>LIST OF SYMBOLS</td>
<td>xvi</td>
<td></td>
</tr>
<tr>
<td>LIST OF APPENDICES</td>
<td>xxii</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>Problem Background</td>
<td>4</td>
</tr>
<tr>
<td>1.3</td>
<td>Problem Statement</td>
<td>5</td>
</tr>
<tr>
<td>1.4</td>
<td>Hypothesis</td>
<td>6</td>
</tr>
<tr>
<td>1.5</td>
<td>Aim</td>
<td>7</td>
</tr>
<tr>
<td>1.6</td>
<td>Objectives</td>
<td>7</td>
</tr>
<tr>
<td>1.7</td>
<td>Scope</td>
<td>8</td>
</tr>
<tr>
<td>1.8</td>
<td>Significance of the Research</td>
<td>9</td>
</tr>
<tr>
<td>1.9</td>
<td>Contribution of the Research</td>
<td>9</td>
</tr>
<tr>
<td>1.10</td>
<td>Thesis Organization</td>
<td>9</td>
</tr>
</tbody>
</table>
## 2 LITERATURE REVIEW

### 2.1 Introduction 11

### 2.2 Internet

- 2.2.1 Evolution of Computer Network 12
- 2.2.2 Web Pages 14
- 2.2.3 Text Encoding / Character Set (Charset) 15

### 2.3 Overview of Language Identification

- 2.3.1 Importance of Language Identification 18
- 2.3.2 Language Identification Applications 20
- 2.3.3 Minority Language Identification 22
- 2.3.4 Multilingual Identification 23
- 2.3.5 Supervised / Unsupervised Identification 24
- 2.3.6 Feature Processing 25

### 2.4 Problems of Web Page

- 2.4.1 Problem of Web Page Format 27
- 2.4.2 Problem of Grammatical / Morphological Error 28
- 2.4.3 Problem of Tremendous Abbreviations 28
- 2.4.4 Problem of Encoding Issue 29

### 2.5 Feature Selection Problem of Web Page Language Identification 31

### 2.6 Conventional Web Page Language Identification Process

- 2.6.1 Preprocessing Step 34
- 2.6.2 Representation Step 35
- 2.6.3 Induction Step 35

### 2.7 Feature Selection Method Review

- 2.7.1 Filter and Wrapper of Feature Selection 36
- 2.7.2 Statistical and Linguistic of Feature Selection 38
- 2.7.3 Statistical
  - 2.7.3.1 Entropy 42
  - 2.7.3.2 Principle Component Analysis (PCA) 44
  - 2.7.3.3 N-grams Approach 45
  - 2.7.3.4 Windowing Algorithm 48
- 2.7.4 Linguistic
  - 2.7.4.1 Small Word Technique 49
  - 2.7.4.2 Unicode Based Identification 50
2.7.4.3 Web Page Information 52
2.7.4.4 Hidden Markov Models (HMMs) 53

2.8 Identification Method Review 54
2.8.1 General Language Identification Methods 55
2.8.2 Artificial Neural Networks (ANN) 55
2.8.3 Fuzzy ARTMAP 59
2.8.4 Support Vector Machine (SVM) 62
2.8.5 Decision Trees 63
2.8.6 Vector Quantization (VQ) 64
   2.8.6.1 Vector Quantization (VQ) Training 65
   2.8.6.2 Vector Quantization (VQ) Testing 65
2.8.7 K-Nearest Neighbor (KNN) 66

2.9 Evaluation Approach 67
2.9.1 T-test 68
2.9.2 Precision, Recall and $F_1$ Measurements 70
2.9.3 Cross Validation and Accuracy 72

2.10 Summary 73

3 RESEARCH METHODOLOGY 74
3.1 Introduction 74
3.2 Operational Framework 75
3.3 Methodology Flow 78
   3.3.1 N-grams Feature Selection Method 79
   3.3.2 Entropy Feature Selection Method 81
   3.3.3 Letter Weighting Feature Selection Method 84
   3.3.4 Simplified Entropy Feature Selection Method 86
3.4 Data Preparation Design 88
3.5 Data Preprocessing Design 90
3.6 Feature Selection Design 91
   3.6.1 N-grams 92
   3.6.2 Entropy 93
   3.6.3 Letter Weighting 93
   3.6.4 Simplified Entropy 95
3.7 Language Identification Design 98
   3.7.1 Argument Minimum 99
   3.7.2 Fuzzy ARTMAP 99
   3.7.3 Decision Trees for Simplified Entropy 99
# EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction 103
4.2 Experimental Setting 103
4.3 Results of Corpora Preparation 105
  4.3.1 Letter Frequency Justification 106
  4.3.2 T-test Analysis of the Preprocessing Step 107
4.4 Results of Feature Selection Methods Evaluation 112
  4.4.1 Retrieval Performance 112
  4.4.2 Comparison of Identification Performances 117
  4.4.3 Letter Frequency Constraints 119
4.5 Discussion 120
  4.5.1 Character Versus Word 120
  4.5.2 The Impact of Dataset 121
  4.5.3 Noise Tolerance 122
  4.5.4 Accuracy of Language Identification Methods 122
  4.5.5 Shortcoming of the Proposed Method 124
4.6 Summary 124

# CONCLUSION

5.1 Introduction 125
5.2 Research Findings 126
5.3 Thesis Contributions 129
5.4 Future Work 129
5.5 Summary 130

REFERENCES 131

APPENDICES A - F 143 - 156
## LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Supervised and unsupervised methods</td>
<td>25</td>
</tr>
<tr>
<td>2.2</td>
<td>Comparison of feature selection methods</td>
<td>40</td>
</tr>
<tr>
<td>2.3</td>
<td>Example of entropy feature selection</td>
<td>43</td>
</tr>
<tr>
<td>2.4</td>
<td>Inlink and outlink</td>
<td>53</td>
</tr>
<tr>
<td>2.5</td>
<td>Orthogonal language codes</td>
<td>57</td>
</tr>
<tr>
<td>2.6</td>
<td>Evolution of the adaptive neural networks</td>
<td>60</td>
</tr>
<tr>
<td>2.7</td>
<td>The definitions of the parameters $\tilde{a}$, $\tilde{b}$ and $\tilde{c}$</td>
<td>71</td>
</tr>
<tr>
<td>2.8</td>
<td>Decision matrix for calculating the classification accuracies</td>
<td>71</td>
</tr>
<tr>
<td>3.1</td>
<td>Comparison of the proposed methods</td>
<td>78</td>
</tr>
<tr>
<td>3.2</td>
<td>Demonstration local letter weighting $\alpha_{jk}$</td>
<td>94</td>
</tr>
<tr>
<td>3.3</td>
<td>Demonstration of global letter weighting $\beta_{jk}$</td>
<td>95</td>
</tr>
<tr>
<td>3.4</td>
<td>Simplified entropy feature selection</td>
<td>97</td>
</tr>
<tr>
<td>4.1</td>
<td>News website data sets</td>
<td>104</td>
</tr>
<tr>
<td>4.2</td>
<td>The Unicode boundary in decimals</td>
<td>105</td>
</tr>
<tr>
<td>4.3</td>
<td>The different groups of T-test</td>
<td>109</td>
</tr>
<tr>
<td>4.4</td>
<td>The backpropagation neural networks structure</td>
<td>109</td>
</tr>
<tr>
<td>4.5</td>
<td>Analysis of the statistical hypothesis test</td>
<td>111</td>
</tr>
<tr>
<td>4.6</td>
<td>Average accuracy of identification according to Figure 4.11</td>
<td>118</td>
</tr>
<tr>
<td>5.1</td>
<td>Objective versus outcome</td>
<td>127</td>
</tr>
<tr>
<td>A.1</td>
<td>Detectable character sets</td>
<td>143</td>
</tr>
<tr>
<td>B.1</td>
<td>Comparison of the previous works</td>
<td>145</td>
</tr>
<tr>
<td>B.2</td>
<td>Comparison strength and weakness of previous works</td>
<td>146</td>
</tr>
<tr>
<td>C.1</td>
<td>Critical values of T-test</td>
<td>151</td>
</tr>
<tr>
<td>D.1</td>
<td>Features found based on different feature selection methods</td>
<td>152</td>
</tr>
</tbody>
</table>
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Example of different language web pages</td>
<td>3</td>
</tr>
<tr>
<td>2.1</td>
<td>Internet applications</td>
<td>12</td>
</tr>
<tr>
<td>2.2</td>
<td>A basic component of computer</td>
<td>13</td>
</tr>
<tr>
<td>2.3</td>
<td>Local area network and wide area network</td>
<td>13</td>
</tr>
<tr>
<td>2.4</td>
<td>Internet users in the world on year 2009</td>
<td>14</td>
</tr>
<tr>
<td>2.5</td>
<td>Elements of a web page</td>
<td>15</td>
</tr>
<tr>
<td>2.6</td>
<td>Encoding, scripts and language identification</td>
<td>16</td>
</tr>
<tr>
<td>2.7</td>
<td>Human language technology overviews</td>
<td>17</td>
</tr>
<tr>
<td>2.8</td>
<td>Bilingual processing system</td>
<td>19</td>
</tr>
<tr>
<td>2.9</td>
<td>Conventional web page language identification</td>
<td>21</td>
</tr>
<tr>
<td>2.10</td>
<td>Various types of language texts</td>
<td>23</td>
</tr>
<tr>
<td>2.11</td>
<td>Monolingual and multilingual web page</td>
<td>24</td>
</tr>
<tr>
<td>2.12</td>
<td>Different of programming code and consolidated text</td>
<td>27</td>
</tr>
<tr>
<td>2.13</td>
<td>The grammatical and morphological errors</td>
<td>28</td>
</tr>
<tr>
<td>2.14</td>
<td>Short forms and abbreviations</td>
<td>28</td>
</tr>
<tr>
<td>2.15</td>
<td>Various encodings / charsets</td>
<td>30</td>
</tr>
<tr>
<td>2.16</td>
<td>Flow of the automatic language identification</td>
<td>33</td>
</tr>
<tr>
<td>2.17</td>
<td>A representation step either in boolean or numeric</td>
<td>35</td>
</tr>
<tr>
<td>2.18</td>
<td>Filter and wrapper approach</td>
<td>37</td>
</tr>
<tr>
<td>2.19</td>
<td>Types of features found on a web page</td>
<td>39</td>
</tr>
<tr>
<td>2.20</td>
<td>A simulation of the N-gram approach</td>
<td>46</td>
</tr>
<tr>
<td>2.21</td>
<td>Overview of non-sliding window algorithm</td>
<td>48</td>
</tr>
<tr>
<td>2.22</td>
<td>Probability of small word technique</td>
<td>50</td>
</tr>
<tr>
<td>2.23</td>
<td>The scripts of Unicode</td>
<td>51</td>
</tr>
<tr>
<td>2.24</td>
<td>Markov chain model</td>
<td>54</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Networks</td>
<td></td>
</tr>
<tr>
<td>ART</td>
<td>Adaptive Resonance Theory</td>
<td></td>
</tr>
<tr>
<td>ARTMAP</td>
<td>Adaptive Resonance Theory Map</td>
<td></td>
</tr>
<tr>
<td>ASCII</td>
<td>American Standard Code for Information Interchange</td>
<td></td>
</tr>
<tr>
<td>BBC</td>
<td>British Broadcasting Corporation</td>
<td></td>
</tr>
<tr>
<td>BPNN</td>
<td>Backpropagation Neural Networks</td>
<td></td>
</tr>
<tr>
<td>CLIR</td>
<td>Cross Language Information Retrieval</td>
<td></td>
</tr>
<tr>
<td>CNN</td>
<td>Cable News Network</td>
<td></td>
</tr>
<tr>
<td>CI</td>
<td>Confidence Interval</td>
<td></td>
</tr>
<tr>
<td>CPBF</td>
<td>Class Profile-Based Approach</td>
<td></td>
</tr>
<tr>
<td>DBI</td>
<td>Dictionary Based Identification</td>
<td></td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
<td></td>
</tr>
<tr>
<td>ECI</td>
<td>European Corpus Initiative</td>
<td></td>
</tr>
<tr>
<td>FESE</td>
<td>Feature Selection Method</td>
<td></td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Model</td>
<td></td>
</tr>
<tr>
<td>HMMs</td>
<td>Hidden Markov Models</td>
<td></td>
</tr>
<tr>
<td>HTML</td>
<td>Hypertext Markup Language</td>
<td></td>
</tr>
<tr>
<td>ICA</td>
<td>Independent Component Analysis</td>
<td></td>
</tr>
<tr>
<td>ICT</td>
<td>Information and Communication Technology</td>
<td></td>
</tr>
<tr>
<td>ICU</td>
<td>International Components for Unicode</td>
<td></td>
</tr>
<tr>
<td>ISO</td>
<td>International Organization for Standardization</td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>$a$ is a feature vector ($0 \leq a \leq 1$) of fuzzy ARTMAP</td>
</tr>
<tr>
<td>$\tilde{a}$</td>
<td>The system and the expert agree with the assigned category</td>
</tr>
<tr>
<td>$a_c$</td>
<td>The complement of $a$ (e.g., if $a = 0.4$ then $a_c = -0.6$)</td>
</tr>
<tr>
<td>$A$</td>
<td>A matrix document $N$-grams</td>
</tr>
<tr>
<td>$\text{Acc}_D$</td>
<td>Accuracy of the experiment data set $D$</td>
</tr>
<tr>
<td>$A_{ik}$</td>
<td>Average accumulated frequency of particular letter $k$ in particular language $i$</td>
</tr>
<tr>
<td>$\alpha_{ik}$</td>
<td>Local letter weighting of particular letter $k$ of particular language $i$</td>
</tr>
<tr>
<td>$\hat{\alpha}$</td>
<td>A non-zero vector of PCA</td>
</tr>
<tr>
<td>$b$</td>
<td>Number of subsets $D_h$ in collection $D$</td>
</tr>
<tr>
<td>$\tilde{b}$</td>
<td>The system disagrees with the assigned category but the expert did</td>
</tr>
<tr>
<td>$\hat{b}$</td>
<td>The characters of word</td>
</tr>
<tr>
<td>$\beta_{ik}$</td>
<td>Global letter weighting of particular letter $k$ of particular language $i$</td>
</tr>
<tr>
<td>$\hat{\beta}_{\tilde{q}}$</td>
<td>A codebook of language $\tilde{q}$ for algorithm VQ</td>
</tr>
<tr>
<td>$B$</td>
<td>Complement coded input vector of fuzzy ARTMAP</td>
</tr>
<tr>
<td>$\tilde{c}$</td>
<td>The expert disagrees with the assigned category but the system did</td>
</tr>
<tr>
<td>$C$</td>
<td>Number of committed coding nodes of fuzzy ARTMAP</td>
</tr>
<tr>
<td>$\hat{C}$</td>
<td>A real symmetric matrix of PCA</td>
</tr>
<tr>
<td>Symbol</td>
<td>Definition</td>
</tr>
<tr>
<td>--------</td>
<td>------------</td>
</tr>
<tr>
<td>$\hat{C}$</td>
<td>The universe of languages of algorithm VQ</td>
</tr>
<tr>
<td>$\chi$</td>
<td>Convergence point of simplified entropy</td>
</tr>
<tr>
<td>$co$</td>
<td>Number of correct identifications</td>
</tr>
<tr>
<td>$d$</td>
<td>The desired value that appropriate to be feed into identifier of PCA</td>
</tr>
<tr>
<td>$\hat{d}$</td>
<td>The system and the expert disagree with the assigned category</td>
</tr>
<tr>
<td>$d_j$</td>
<td>Particular document $j$</td>
</tr>
<tr>
<td>$\delta_q$</td>
<td>The generalized error through a layer $q$ of the Backpropagation Neural Networks (BPNN)</td>
</tr>
<tr>
<td>$\delta_r$</td>
<td>The generalized error through a layer $q$ and $r$ of the BPNN</td>
</tr>
<tr>
<td>$D$</td>
<td>A collection of information (or web documents)</td>
</tr>
<tr>
<td>$D_h$</td>
<td>Subset $h$ of collection $D$</td>
</tr>
<tr>
<td>$DF_k$</td>
<td>Document frequency of particular letter $k$</td>
</tr>
<tr>
<td>$DoF$</td>
<td>Degree of freedom of T-test</td>
</tr>
<tr>
<td>$E_h$</td>
<td>Accuracy of subset $h$</td>
</tr>
<tr>
<td>$EN_k$</td>
<td>Entropy weighting of particular letter $k$ in collection, $D$</td>
</tr>
<tr>
<td>$EN_{jk}$</td>
<td>Entropy weighting of particular letter $k$ in particular document $j$</td>
</tr>
<tr>
<td>$\ell$</td>
<td>Sum of features of simplified entropy</td>
</tr>
<tr>
<td>$\hat{\ell}$</td>
<td>Iteration number of BPNN</td>
</tr>
<tr>
<td>$\ell_{new}$</td>
<td>Total of current feature frequency of simplified entropy</td>
</tr>
<tr>
<td>$\ell_{old}$</td>
<td>Total of previous feature frequency of simplified entropy</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Learning rate of the BPNN</td>
</tr>
<tr>
<td>$f$</td>
<td>Number of folds in cross validation</td>
</tr>
<tr>
<td>$\hat{f}$</td>
<td>The frequency of the word of Zipf’s Law</td>
</tr>
<tr>
<td>$F1$</td>
<td>The F1 measure is the average of precision, $\tilde{p}$ and recall, $\tilde{r}$</td>
</tr>
<tr>
<td>$F_k$</td>
<td>Frequency of letter $k$ in the collection $D$</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Prefix convergence point of simplified entropy</td>
</tr>
<tr>
<td>$\Gamma$</td>
<td>Momentum rate of the BPNN</td>
</tr>
</tbody>
</table>
$G_k$ - 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

$m$ - 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

$ngm_{\bar{r}}$ - $N$-grams of the testing document ordered descending

$ngm_{Lm}$ - $N$-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

$\omega_{ik}$ - Letter weighting of particular letter $k$ of particular language $i$

$p$ - Input layer of BPNN

$\hat{p}$ - The precision describes the probability that an desired document (randomly selected) retrieved document is relevant to a certain language

$\hat{\rho}$ - 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
\( \tilde{r} \) - The rank of the \( N \)-grams in the predicted text ordered descending by the frequency of Zipf’s Law

\( \hat{r} \) - The recall describes the probability of a relevant language being retrieved

\( R \) - Threshold or number of features of input patterns

\( \rho \) - Vigilance variable of fuzzy ARTMAP

\( s \) - Window size of windowing algorithm

\( \hat{S} \) - Particular script of languages

\( \hat{S}_{\text{begin}} \) - The begin codepoint of a particular script, \( \hat{S} \)

\( \hat{S}_{\text{end}} \) - The end codepoint of particular script, \( \hat{S} \)

\( S \) - Standard deviation of the mean

\( \sigma \) - Standard deviation is a measure of the dispersion of a set of values

\( t \) - Critical value of T-Test

\( \hat{i} \) - 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 \( j \)

\( \theta_q \) - A bias on hidden unit \( q \) of the BPNN

\( \theta_r \) - A bias on output unit \( r \) of the BPNN

\( \varepsilon \) - Match tracking (\( \varepsilon \in (-1, 1) \)) of fuzzy ARTMAP

\( \varphi_z \) - Original input patterns of machine learning

\( \varphi_z \) - Normalized input patterns of machine learning

\( w \) - Weight vector of fuzzy ARTMAP

\( \hat{w} \) - The elements of set \( \hat{W} \), words

\( w_p \) - The coding node weight vector \( \hat{p} \) of fuzzy ARTMAP

\( w_q \) - The output class weight vector \( \hat{q} \) of fuzzy ARTMAP

\( W_{qp} \) - The \( q^{th} \) weight to the unit \( p^{th} \) of the BPNN
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_{rq}$</td>
<td>The $r^{th}$ weight to the unit $q^{th}$ of the BPNN</td>
</tr>
<tr>
<td>$\hat{W}$</td>
<td>A set of words $\hat{w}$ for algorithm VQ</td>
</tr>
<tr>
<td>$x$</td>
<td>The element of $A$</td>
</tr>
<tr>
<td>$\hat{x}$</td>
<td>The samples of KNN</td>
</tr>
<tr>
<td>$\bar{x}$</td>
<td>The mean of $x$</td>
</tr>
<tr>
<td>$\hat{x}$</td>
<td>Mean is the arithmetic average of a set of values or distribution</td>
</tr>
<tr>
<td>$x_{\hat{n}m}$</td>
<td>All the $N$-grams exist in the collection, $\hat{n}$ is the number of $N$-grams and $\hat{m}$ is the number of documents</td>
</tr>
<tr>
<td>$y$</td>
<td>Coding field activation pattern of fuzzy ARTMAP</td>
</tr>
<tr>
<td>$z$</td>
<td>Dimension of input pattern of machine learning</td>
</tr>
<tr>
<td>$\zeta_{cur}$</td>
<td>Index position of current feature of simplified entropy with frequency changed</td>
</tr>
<tr>
<td>$\zeta_{pre}$</td>
<td>Index previous position of current feature of simplified entropy with frequency changed</td>
</tr>
</tbody>
</table>
# LIST OF APPENDICES

<table>
<thead>
<tr>
<th>APPENDIX NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Character Set Detection</td>
<td>143</td>
</tr>
<tr>
<td>B</td>
<td>Comparison Conventional Methods</td>
<td>144</td>
</tr>
<tr>
<td>C</td>
<td>Critical Values of T-test</td>
<td>151</td>
</tr>
<tr>
<td>D</td>
<td>Output of Feature Selection</td>
<td>152</td>
</tr>
<tr>
<td>E</td>
<td>Examples of the Arabic Script Letter Distribution</td>
<td>153</td>
</tr>
<tr>
<td>F</td>
<td>List of Publication and Recognition</td>
<td>154</td>
</tr>
</tbody>
</table>
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 page3. 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.

3The 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; Santosh Kumar 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 of 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 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.
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 $F_1$ 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.
CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this chapter, the Internet or World Wide Web (WWW) and their applications are briefly discussed. An overview of the web page language identification is then presented. This is followed by a discussion of the problems related to web page formats. Then, it is an elaboration of feature selection problem and offers a brief explanation of the conventional web page language identification processes. The chapter continues with a review of the literature, specifically that concerning the two significant aspects of the processes, called feature selection and identification. The measurements used, such as T-test, precision, recall, $F1$ and cross validation, are also discussed. The chapter ends with a summary.

2.2 Internet

Since the late 1960s, the Internet has grown from a single experimental network serving a dozen sites in the United States to a network of networks linking millions of
computers worldwide. The Internet is a global network of interconnected computers, enabling users to share information along multiple channels. Computer users typically send and receive information using web browsers (Berners-Lee and Cailliau, 1994; Lawrence and Giles, 1998). Other software for user interface with computer networks includes specialized programs for electronic mail, online chatting, file transfer and file sharing, as shown in the Figure 2.1.

Figure 2.1: Internet applications

2.2.1 Evolution of Computer Network

Development of the global Internet has been surprisingly. It has grown from a small, closed, text based computer network of a few thousand scientific and government users to 43 million Internet host that supporting an open global network of an estimated 150 million Internet users in early 1999 (Petrazzini and Kibati, 1999). In general, basic components of computer are input, processor, storage and output as shown in Figure
2.2. Input of computer is data such as image, audio, text, etc. Processor is the engine of the computer that responsible for data processing such as insert and delete. Storage is the space for saving data, it is divided into temporary and permanent memory. Output is information produced by the computer program and perceived by the user (Shelly et al., 2008).

Figure 2.3 illustrates the Local Area Network (LAN) and Wide Area Network (WAN). Local area networks, generally called LANs, are privately-owned networks within a single building or campus of up to a few kilometers in size (Tanenbaum, 2002). A WAN spans a large geographical area, often a country or continent. A collection of interconnected networks is called an internetwork or internet. A common form of internet is a collection of LANs connected by a WAN (Tanenbaum, 2002). With the rapid development on hardware, internet has now becomes part of the human life. People use it to send / receive email, on-line chatting, searching information, e-business, etc (Abu Bakar, 2003). Figure 2.4 presents the analysis of the internet users in the world by geographic on year 2009. The estimated internet users are more than 1 billion on June 30, 2009. Asia internet users are the highest regions which are
704.2 million users and the lowest is Oceania / Australia which is 20.8 million users (Miniwatts Marketing Group, 2008).

Figure 2.4: The internet users in the world by geographic regions on year 2009 (Miniwatts Marketing Group, 2008)

2.2.2 Web Pages

A web page is a page of resource information that has been placed on the WWW or Internet and can be accessed through a web browser and displayed on a computer screen. This information is usually in Hypertext Markup Language (HTML) or Extensible Hypertext Markup Language (XHTML) format and the hyperlink may provide navigation to other web pages (Berners-Lee and Hendler, 2001). Usually the web page has content such as textual information, images, multimedia (audio / video), hyperlinks, forms, comments, metadata, script, etc. This information is able to be seen, heard or interacted with by the end users. Usually, a web page is located on a specific Uniform Resource Locator (URL), with multiple web pages being located on same URL using certain scripts. There are various softwares that can be used to
create, edit, save and view web pages, such as TextPad (Helios Software Solutions, 2009) and Macromedia Dreamweaver (Adobe Systems, 2008). A web browser has a graphical user interface such as Internet Explorer, Mozilla Firefox and Google Chrome to view the content of the web page (Oreilly, 2007). Figure 2.5 presents a sample of Google Chrome web page. It consists of URL, HTML, web page, script / source code, encoding, metadata, etc.

Figure 2.5: Elements of a web page

2.2.3 Text Encoding / Character Set (Charset)

Encoding is a process that transforms information from one format to another (Allen, 2006). The opposite operation is called decoding. This is used in many digital devices. Character encoding is a code that pairs a set of natural language characters (such as an alphabet or syllabary) with some other set, such as numbers or electrical pulses (Allen, 2006). The common encoding used in computer systems include Western (International Organization for Standardization (ISO)-8859-1), Unicode (8-bit UCS / Unicode Transformation Format (UTF-8)), Chinese Traditional (Big5) and Central European (Window-1250). Unicode and its parallel standard, ISO-10646 Universal Character Set, which together constitute the most modern character encoding systems,
their numbering, how those numbers are encoded as a series of “code units” (limited-size numbers) and finally how those units are encoded as a stream of octets (bytes). The traditional systems allow for a far simpler way to handle texts, but the latter allows any letter/diacritic combination to be used in a text (Allen, 2006). Other writing systems, such as Arabic and Hebrew, are represented with more complex language morphologies due to the need to accommodate such things as bidirectional text and glyphs\(^1\) that are joined together in different ways for different reasons (Allen, 2006). There are a number of scripts inside Unicode encoding or other encodings as well; for example; Arabic, Latin, Hanzi, Cyrillic and Indic scripts. Those scripts can be used to produce various languages. For this reason, language identification involves three steps, encoding detection, script determination and language identification. Due to the fact that each character within a particular encoding is unique, a standard encoding is recommended for a language identification process. An example is Unicode (UTF-8) (Allen, 2006). Script determination can be done easily by segmenting the characters of document based on their code point in Unicode, as shown in the Figure 2.23 in Section 2.7.4.2 (Allen, 2006). This script segmentation can increase the effectiveness of language identification indirectly but significantly. Figure 2.6 shows the relationships that exist among encoding, script and language. The arrow indicates the number of possibilities can be found. Therefore, it can be seen that there are a number of complex issues related to web page language identification and in this work it is focuses on the bottom line, which is web page language identification.

\[\text{Figure 2.6: Encoding, scripts and language identification}\]

\(^1\)Glyph refers to a symbolic figure or inscribed symbol, it may be an ideogram like Hanzi character
2.3 Overview of Language Identification

The field of human language technology covers a number of research activities, including the coding, identification, interpretation, translation and generation of language. The aim of such research is to enable humans to communicate with machines using natural language skills. Language technology research involves many disciplines, such as linguistics, psychology, electrical engineering and computer science. Cooperation among these disciplines is needed to create multimodal and

![Diagram of Human Language Technology](image)

Figure 2.7: Human language technology overviews and automatic language identification are part of the multilingualism (Muthusamy and Spitz, 1997)
multimedia systems that use the combination of text, speech, facial cues and gestures, both to improve language understanding and to produce more natural language processing by animated characters. Figure 2.7 illustrates the whole idea of human language technology and automatic language identification as a significant part of the multilingualism (Muthusamy and Spitz, 1997).

Language Identification (LID) is the means by which a language that has been used to write various types of documents is identified. For web page language identification, humans are the most accurate language identifiers. Within seconds of reading a passage, one is able to determine whether it is a language they can or cannot understand. If it is a language that they are not familiar with, they often can make a subjective judgement as to its similarity to a language that they know, e.g., “That sounds like Japanese or Korean?”

2.3.1 Importance of Language Identification

Language technologies play a key role in the age of information (Constable and Simons, 2000). Today, almost all device systems combine language understanding and generation that allow people to interact with computers using text or speech to obtain information, to study, to do business and to communicate with each other effectively (Kralisch and Koeppen, 2005).

The technology convergence in the processing of text, speech and images has lead to particular ability to make sense of the massive amounts of information now available via computer networks (Tanenbaum, 2002). For example, if a student wants to gather information about the art of getting things done, he or she can set in motion a set of procedures that locate, organize and summarize all available information related to the topic from books, periodicals, newspapers, etc. Translation of texts or speech from one language to another is needed to access and interpret all available material and present it to the student in his or her native language. As a result, it will increase academic interests of the student (Abd Rozan et al., 2005; McNamee, 2004).
Language is a basic requirement for any type of communication either between persons or organizations. Multilingual environments are spaces in which communication among people from different races or countries occurs. Therefore, language identification is a core technology to support various multilingual applications, such as Optical Character Recognition (OCR) (Amara and Bouslama, 2003; Hochberg et al., 1999; Joshi et al., 2006; Sibun and Spitz, 1994), automatic transliteration system (Jung et al., 2000), multilingual speech recognition (Li et al., 2007; Qian et al., 2009), text categorization and spoken document retrieval (Bian and Chen, 1998; Levow et al., 2005; Macherey et al., 2009; Sagiroglu et al., 2007; Sibun and Spitz, 1994) as shown in Figure 2.8. For example, the phonetic transcription of terms must be obtained online from text using either rules based or some other kind of pronunciation model (Hakkinen and Tian, 2001). Most pronunciation models depend on clearly expressed use of the language and therefore it must be learned by the system in order to enable the model.

Lewandowski (2008) has done research on the problems related to the use of web search engines to find material in foreign languages. In other words, the author wants to identify how the search engines can provide an accurate language-restricted search functionality. Moreover, Al-Salman et al. (2007) have developed a system namely Mubser, which translates Arabic and English Braille into normal text. The system can automatically detect the source languages and the Braille grade. Levow et al. (2005) have proposed a Cross Language Information Retrieval (CLIR) system that allows users to find documents written in various languages, while Kralisch and Mandl (2006) have investigated the barriers to information access across languages.
on the Internet and finds that language identification is extremely important. Those users who are not native speakers of one of the website languages pay a high price for accessing and understanding the information and services being offered on that website.

Even with criminal activity, languages play an important role. With the rapid growth of computer and network systems in recent years, there has been a corresponding explosion in cyber-crime activity (Nykodym et al., 2005; Wang, 2007). The most commonly occurring crimes involve hacking into computer systems, computer viruses, groundless allegations (rumor), threats, etc. Digital forensics can be defined as the practice of scientifically derived and proven technical methods and tools for the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of after-the-fact digital information derived from digital sources for the purpose of facilitating the reconstruction of events as forensic evidence (Enck et al., 2005). Moreover, it can be extended to the forensic examination of mobile devices to deal with the proliferation of mobile devices in society that has led to an incidental increase in their use in criminal activity (Mellars, 2004). For example, the mobile security includes protection against mobile Short Message Services (SMS) spam, spoofing and other SMS-related scams. It is important to raise the awareness of security threats among the mobile services in a multilingual environment (Nas, 2008).

2.3.2 Language Identification Applications

A multilingual environment can be rigidly defined as being one in which the speakers are native-like in two or more languages. Currently, there are number of tools applied in the multilingual environments, especially in web application. For example, Rosette Language Identifier (Toman, 2008), TextCat (Cavnar and Trenkle, 2008), Xeror Language Identifier (Xerox Corporation, 2008), etc., as shown in Figure 2.9. The most popular system, the TextCat Language Guesser makes use of the language-specific letter of $N$-grams distribution and can determine 69 different natural languages. Dunning (1994) claims that letter trigrams can identify the language almost error-free from a text-length of 500 bytes on. However, it has proved in this work that $N$-grams
Figure 2.9: Conventional web page language identification applications have limitations on identifying the language and also it focuses most tightly on common languages such as European.
2.3.3 Minority Language Identification

Various methods have been tested on European languages such as English, German, French, etc. (Martins and Silva, 2005; Sibun and Reynar, 1996). On the other hand, documents that use non-Roman script languages, such as Arabic, Indic and Chinese, are difficult to represent in encoded form. Available resources for technology development among the major languages are scarce (Murthy, 2003). According to Abd Rozan et al. (2005), this situation can be described as a digital divide among languages and termed as the “language digital divide” (Suzuki et al., 2002). Therefore, language identification is important for many applications in multi-lingual countries (Abd Rozan et al., 2005). Arabic script web page language identification has been given less attention by researchers, even though the number of Arabic script websites on the WWW has been reportedly increasing dramatically in the 21st century (Abd Rozan et al., 2005). The Arabic script is cursive, caseless, written right-to-left and belongs to the Semitic family of languages. It is used for languages such as Arabic, Farsi (the official language of Iran), Jawi, Kurdish, Pashto (the official language of Afghanistan), Sindhi and Urdu (the official language of Pakistan) (Mohamed Ould, 2007). Arabic itself is a language that is likely to present challenges in a traditional information retrieval environment and in popular search engines, because its morphology and word-formation rules are radically different from other languages. Although different Arabic dialects exist throughout the Arab world, there is only one form of the web page language found in printed works and it is known as فصحى or Standard Arabic (Mohamed Ould, 2007). Similar issues exist for India script as well. According to Murthy (2003), there are about 150 different languages spoken in India, of which 18 have been given a kind of constitutional recognition and are considered to be the major languages. Indian languages encompass four language families, namely Indo-Aryan, Dravidian, Tibeto-Burman and Austro-Asiatic. Multilingualism in India is very common, with many documents required to be in more than one language. With so many languages in use, identifying a specific language is a critical step in many applications (Murthy and Kumar, 2006). For Hanzi script languages like Traditional Chinese and Japanese, each character is a graphic symbol that represents an idea and is known as an ideograph or ideogram. Therefore, tokenization of Hanzi script character is difficult due to the fact that the characters do not have the exact word boundaries, as seen in European languages (Kranig, 2005). This increases the difficulty of web page language identification based on words or N-grams. Figure 2.10 illustrates the text example of different languages.
2.3.4 Multilingual Identification

For monolingual identification, the particular document or web page will usually consist of a single language. However, the multilingual identification may consist of two or more languages in a single document. For example, a web page may consist of 70% English and 30% Arabic or a mixture of both Arabic and Urdu languages (Kay et al., 1997). Figure 2.11 shows an example of monolingual and multilingual web page. Multilingual web page consists of Chinese, English and Japanese in one page, but monolingual only has one language. Therefore, the complexity of the multilingual identification is difficult than monolingual identification. It is assumed that the basic unit of any text is the sentence, with its construction based on a particular language that produces with some meaning. The words and phrases are not suitable because those objects might exist in more than one language. There are certain scripts in encoding that refer to one language only, such as Hangul Jamo, which in Unicode refers to the Korean language (Abd Rozan et al., 2005). However, one document may consists of mixed languages from a particular script (١) or same scripts such as Arabic or Latin.
script, which can be used to produce a number of languages. This fact required sentence based segmentation before proceed to language identification (Artemenko et al., 2006; Gigué, 1996).

Figure 2.11: Monolingual and multilingual web page

2.3.5 Supervised / Unsupervised Identification

For any supervised identification, the model is trained for a desired target with particular features. In other words, a model only can recognize the object being taught. For example, a particular model can only predict Latin script characters in its optical character recognition, and the process recognition will fail if a Chinese character is given to that model. This suggests that supervised language identification is limited to the trained languages only, not to all the languages in the world.
Table 2.1: Supervised and unsupervised methods

<table>
<thead>
<tr>
<th>Supervised Method</th>
<th>Unsupervised Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial Neural Networks (ANN)</td>
<td>Self Organizing Map (SOM)</td>
</tr>
<tr>
<td>Fuzzy ARTMAP</td>
<td>Adaptive Resonance Theory (ART)</td>
</tr>
<tr>
<td>Support Vector Machine (SVM)</td>
<td>Independent Component Analysis (ICA)</td>
</tr>
<tr>
<td>Gaussian Mixture Model (GMM)</td>
<td>Principle Component Analysis (PCA)</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>Hidden Markov Models (HMMs)</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>Vector Quantization (VQ)</td>
</tr>
</tbody>
</table>

Unsupervised identification is model learning without a desired target (Biemann and Teresiak, 2005). It is based on the features of particular object and finds the closest cluster according to an algorithm design such as KNN and SOM. The learning process is based on the features of a particular object, so the identification performance depends on the robustness features. Although, the unsupervised model can find the cluster itself, the problem is how to match this with the desired target. Therefore, the supervised identification is chosen as the experiment model in this study. Table 2.1 presents the supervised and unsupervised methods.

2.3.6 Feature Processing

As stated in Section 2.7.2 and shown in Figure 2.19, there are many types of feature that can be used in web page language identification. The problem is how to produce the most reliable features for a web page that is to undergo language identification. With respect to the linguistic method, it is based on the structure of the word or sentence and determines the semantic rules of particular language for language identification like a string kernel classifiers (Kruengkrai et al., 2006, 2005). As for the statistical method, it is based on the analysis of particular statistical data and searches out the unique patterns of each language as the features. Usually, this is the frequency or position of particular letter or words. There are number of methods that
have been proposed such as N-grams method (Cavnar and Trenkle, 1994), HMMs (Tran and Sharma, 2005) and web page information (Gustafson et al., 2005) to accomplish this task.

There are number of issues yet to be resolved in web page language identification. One fundamental problem is how to deal with an extremely large number of classes (all the languages in the world). Can the proposed model effectively maintain the identification performance on such huge scale? Another is how to detect a language that has not appeared in the training data. For these reasons, the feature selection methods are closely investigated in this work to justify their impact on web page language identification.

2.4 Problems of Web Page

A language identifier usually can produce higher identification accuracy with lower computational memory and shorter processing time. Mislabeled training documents will also affect the results of language identification (Zou et al., 2005). Sibun and Reynar (1996) have stated that the language identification factors need to be taken into consideration include the type of features to be used, the dimension of the data sets and the type of analysis to be used in validating the language identification results. Botha et al. (2006) state that the accuracy of web page language identification depends on a number of factors, including the size of the textual fragment, the amount and variety of training data, the classification algorithm employed and the similarity of the languages to be discriminated. In general, problems existing in web pages include irrelevant information, unstructured information, spelling or syntax errors and an overabundance of international terms (Benedetto et al., 2002; McNamee, 2005; Windisch and Csink, 2005). For example, when a word main is encountered, it can be either an English word (referring to what is most important) or a Malay word (referring to a word “play”) or other meanings as well. Biemann and Teresniak (2005) argue that supervised training has a major drawback. The language identifier will fail on languages that are not contained in its training, and it will for the most part have no way to cope with that and may well assign the data to some arbitrary or unknown language.
Web page language identification has received less focus than spoken language identification, as it is generally argued that is a straightforward task (Sibun and Reynar, 1996). However, Xafopoulos et al. (2004) and Hughes et al. (2006) argue that web page language identification presents a number of questions that remain open and ripe for further investigation, for example; how to deal with dynamic web page content, different encoding issue, grammatical error and tremendous abbreviations.

### 2.4.1 Problem of Web Page Format

The basic element of a web page is the Hypertext Markup Language (HTML). It provides a means to describe the structure of text-based information on a web page. For example, the elements of HTML like `<html> ... </html>`, `<head> ... </head>`, `<body> ... </body>`, etc., are written in the form of “tags” consisting of various elements surrounding by diamond brackets. Also, web documents may have textual information in a form that is useful when displaying the page but is disorganized when the documents are considered as consolidated text, for example, data formatted as lists or in tables. It is presented in Figure 2.12. All these examples shown that the web pages contain a significant amount of additional information for the visual appearance of a web page, all of which may interface with and confuse the main content, especially in the case of a faulty page composition (Xafopoulos et al., 2004).

![HTML Code](image)

**Figure 2.12:** Different of programming code and consolidated text

```html
<ul>
  <li type=square>First item</li>
  <li type=square>Second item</li>
  <li type=square>Third item</li>
</ul>

Consolidated Text
(Note that the `<ul>`, `<li>`, ```/ul>``` and many others HTML code is meaningless to a layman without knowledge of HTML programming)
2.4.2 Problem of Grammatical / Morphological Error

The main content of a web page may contain grammatical or morphological errors (Hughes et al., 2006) as shown in Figure 2.13. A grammatical error may be caused by careless typing wrong character selection or a misunderstanding of formal English usage. For example, the periods and commas should always go inside quotation marks, “Write like this,” for instance. With respect to morphological errors, there are frequently caused by the common sentence structural mistakes. For example, “somebody taked my shoe off” instead of “somebody took my shoe off.”

<table>
<thead>
<tr>
<th>Grammatical Error</th>
<th>Morphological Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Your vs. You’re</td>
<td>• I cutted my finger and was taking to</td>
</tr>
<tr>
<td>• It’s vs. Its</td>
<td>the hospital (cut)</td>
</tr>
<tr>
<td>• There vs. Their</td>
<td>• They have six childrens (children)</td>
</tr>
<tr>
<td>• Affect vs. Effect</td>
<td>• It is unsignificant (insignificant)</td>
</tr>
<tr>
<td>• The Dangling Participle</td>
<td>• The girl do not agreed because she</td>
</tr>
<tr>
<td>(e.g. Uhh… keep your decomposing</td>
<td>is innmatured (immatre)</td>
</tr>
<tr>
<td>brother away from me!)</td>
<td>• They did the work fastly (fast).</td>
</tr>
</tbody>
</table>

Figure 2.13: The grammatical and morphological errors

2.4.3 Problem of Tremendous Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>asap</td>
<td>(as soon as possible)</td>
</tr>
<tr>
<td>lol</td>
<td>(laughing out loud)</td>
</tr>
<tr>
<td>nvm</td>
<td>(never mind)</td>
</tr>
<tr>
<td>g9</td>
<td>(good night)</td>
</tr>
<tr>
<td>gtg</td>
<td>(got to go)</td>
</tr>
<tr>
<td>ic</td>
<td>(I see)</td>
</tr>
<tr>
<td>fama</td>
<td>(farther and mother)</td>
</tr>
<tr>
<td>foc</td>
<td>(free of charge)</td>
</tr>
<tr>
<td>etc</td>
<td>(et cetera or so on)</td>
</tr>
<tr>
<td>ibid</td>
<td>(Ion beam induced deposition)</td>
</tr>
<tr>
<td>et al</td>
<td>(and others)</td>
</tr>
<tr>
<td>DNA</td>
<td>(Deoxyribonucleic Acid)</td>
</tr>
<tr>
<td>RNA</td>
<td>(Ribonucleic Acid)</td>
</tr>
<tr>
<td>H1N1</td>
<td>(Influenza A (H1N1) Virus)</td>
</tr>
<tr>
<td>AIDS</td>
<td>(Acquired immune deficiency syndrome)</td>
</tr>
<tr>
<td>UTM</td>
<td>(Universiti Teknologi Malaysia)</td>
</tr>
<tr>
<td>R&amp;D</td>
<td>(Research and Development)</td>
</tr>
<tr>
<td>HR</td>
<td>(Human Resources)</td>
</tr>
<tr>
<td>CC</td>
<td>(Carbon Copy in Email)</td>
</tr>
<tr>
<td>IC</td>
<td>(Identity Card)</td>
</tr>
</tbody>
</table>

Figure 2.14: Short forms and abbreviations
The information shared across the Internet makes use of tremendous number of international terms and abbreviations as illustrated in Figure 2.14. These can be found in number of field such as business, politics, cuisine, economics, etc. For example, the abbreviation “AAA” is means American Automobile Association or The American Accounting Association or others. Different users can place diverse meanings on a particular abbreviation, not to mention users of various languages. Therefore, the abbreviation is considered as noise (Xafopoulos et al., 2004).

In the twenty first century, anyone can access Internet easily; even to post their own writing into the blog. At the same time, many short forms have been created from particular users (Xafopoulos et al., 2004). For example, the phrase “I see” becomes “ic” or “good night” becomes “g9”, etc. In normal language processing, these words are also considered as noise. Such short forms can hardly to be interpreted as a formal language word because different users may have different perspectives on them.

Besides, there are many abbreviations used in the sciences, such as chemistry and biology. Those are related to the particular objects in particular science. For example, the abbreviations like Deoxyribonucleic Acid (DNA) and Ribonucleic Acid (RNA) that used on biology science. In the field of automatic language identification, if the analysis was done on the statistical analysis of letter or word distribution, it should separate the scientific from the non scientific data for the analysis.

### 2.4.4 Problem of Encoding Issue

Character encoding used in a web page is another problem arising in language identification as the same language character may exist on a web page in different byte values because of the existence of quite a few different textual character encodings. Unicode encoding (Allen, 2006) is one of the solutions to this problem. However some users are still using different character encodings for their web based documents as shown in Figure 2.15. It will cause the browser cannot correctly display the desired characters. As a result, the web application related encoding issue cannot be functioned
Figure 2.15: Various encodings / charsets

properly to the failure of encoding detection. The details of the encoding have been discussed on Section 2.2.3.
2.5 Feature Selection Problem of Web Page Language Identification

Feature selection is essentially a task to remove irrelevant and/or redundant features. Irrelevant features can be removed without affecting learning performance (John et al., 1994). Redundant features are a type of irrelevant feature (Yu and Liu, 2004). The distinction is that a redundant feature implies the co-presence of another feature; individually, each feature is relevant, but the removal of one of them will not affect learning performance. The selection of features can be achieved in two ways. One is to rank features according to some criterion and select the top features, and the other is to select a minimum subset of features without learning performance deterioration. In other words, subset selection algorithms can automatically determine the number of selected features, while feature ranking algorithms need to rely on some given threshold to select features (Liu and Motoda, 2006).

Feature selection method of web page language identification is not much attention in the natural language processing as the $N$-grams is widely considered as a good feature selection method (McNamee, 2004). Actually, $N$-grams have the dimension issue which is not yet a completely solved problem. For one particular language like Pashto of Arabic script, the probability selected as actual language is higher than other languages of Arabic script due to the dimension of Pashto is smaller than others. If the dimension is smaller than others then distance calculated is the smallest in argument minimum. Therefore, accuracy of web page language identification will be affected. Another feature selection method is to choose the most important keywords of each class as the features, which is so called entropy (Selamat and Omatu, 2004). It depends on a particular threshold in selection the keywords of each class, for example 50 features of each class. The class is refers to one particular script of language. The idea is workable on different script languages but not same script languages. For example, the word “main” is exists on English and Malay, and the word “表示” (pronouns as “biaoshi” in Chinese) is exists on Chinese and Japanese, but the actual meaning is different on each language. These features are useful on text categorization but in web page language identification is a problem to be investigated in this research.

According to Liu and Motoda (2006), they experienced the fast data evolution in which extremely high-dimensional data, such as high-throughput data of bioinformatics
and web / text data became increasingly important. They stretch the capabilities of conventional data processing techniques, pose new challenges, and stimulate accelerated development of feature selection research in two major ways. One trend is to improve and expand the existing techniques to meet the new challenges. The other is to develop brand new algorithm directly targeting the arising challenges (Liu and Motoda, 2006). However, there are not many researches on feature selection method of web page language identification. Therefore, this work is presented an enhancement of existing feature selection method.

The evaluation of feature selection often entails two tasks. One is to compare two cases: before and after feature selection. The goal of this task is to observe if feature selection achieves its intended objectives (recall that feature selection does not confine it to improving classification performance). The aspects of evaluation can include the number of selected features, time, scalability, and learning model’s performance. The second task is to compare two feature selection algorithms to see if one is better than the other for a certain task (Liu and Motoda, 2006). As we know, there is no universally superior feature selection, and different feature selection algorithms have their special edges for various application. Hence, it is wise to find a suitable algorithm for a given application. An initial attempt to address the problem of selecting feature selection algorithms is presented in (Liu and Yu, 2005), aiming to mitigate the increasingly complexity of finding a suitable algorithm from many feature selection algorithms. This research is more specifically focus on learning model’s performance.

Another issue arising from feature selection evaluation is feature selection bias. Using the same training data in both feature selection and classification learning can result in this selection bias. According to statistical theory based on regression research, this bias can exacerbate data over-fitting and negatively affect classification performance (Liu and Motoda, 2006). A recommended practice is to use separate data for feature selection and for learning. In reality, however, separate datasets are rarely used in the selection and learning steps. This is because we want to use as much data as possible in both selection and learning. It is against this intuition to divide the training data into two datasets leading to the reduced data in both tasks. Feature selection bias in studied in (Singhi and Liu, 2006) to seek answers if there is discrepancy between the current practice and the statistical theory. The findings are that the statistical theory is correct, but feature selection bias has limited effect on feature selection for
classification. Therefore, the data set preparation is another issue to be investigated in this research.

2.6 Conventional Web Page Language Identification Process

Commonly, language identification task can be divided into two classes: they are spoken and written language identification. The spoken language identification method has to choose signal processing, even though language identification of texts is a particular symbolic or term processing task (Hakkinen and Tian, 2001). In this work, it is focusing on language identification based on written words rather than speech. The main reason for the apparent lack of activity in web page language identification is probably that it is not considered as a difficult problem (Schultz and Waibel, 1998). This might be true if the size of the texts available during the identification stage is large enough, more than 500 bytes, and a computational range of very unique words that are present in all embedded systems (Schultz and Waibel, 1998).

![Flow of the automatic language identification](image)

Figure 2.16: Flow of the automatic language identification

As mentioned previously, written language identification refers to the assignment of a textual document to one or more predefined languages based on its content. The challenging research issue for language identification is the development of a statistical classification or inductive learning method that can automatically
discover language identification patterns, based on a training set of manually
categorized documents. In general, automatic language identification patterns
encompass three main steps. They are the preprocessing step, the representation step
and the induction step, as shown in Figure 2.16. The preprocessing step determines
the set of features that will be used for representing the individual documents within
a collection. This is essentially a dictionary creation process. The representation step
maps each individual document into a training vector using the dictionary generated
in the previous step, and associates it with a label that is the target language. The
induction step is to find patterns based on identifiers (e.g., decision trees, support vector
machines, etc.) that distinguish categories one from another.

2.6.1 Preprocessing Step

The first step is to produce a set of attributes (or features), called a dictionary,
from a training set of manually categorized documents, and to do so in such a way that
a document text can be represented by a set of features in the representation step. The
text portion of the training documents is scanned to produce a list of nouns or noun
phrases in which neither noun or noun phrase belongs to a predefined list of stop words
or is a number or part of a proper name. To reduce the number of features, those words
(e.g., nouns or noun phrases) and those features with low frequency counts are removed
from the dictionary since features recurring only a few times are not statistically reliable
indicators. Two possible ways to establish features include a universal dictionary and
a local dictionary. A universal dictionary is created for all topic documents in text
databases with feature selection used to select words and phrases from this dictionary to
solve a specific language identification problem. However, this is very time consuming
in language identification, since there are too many words and phrases existing in the
world. A local dictionary is defined as the words found in documents on the given
topic. These are entered in the local dictionary. The authors have created a dictionary
based on letter frequency for each language to represent the documents.
2.6.2 Representation Step

![Diagram of representation step]

Figure 2.17: A representation step either in boolean or numeric

The next step is representation. Given a dictionary (local or universal), as created in the previous step, each training document is then represented in terms of the features in the dictionary. The document is labeled to indicate its category membership. This document consists of a value for each feature in the dictionary, where the value can be either Boolean (e.g., indicating whether or not the feature appears in the document), or numerical (e.g., the frequency of occurrence in the document being processed). Figure 2.17 illustrates a representation step either in boolean or numeric that will be used by the classifier.

2.6.3 Induction Step

The induction step automatically discovers the language identification patterns that differentiate languages from one another, based on a training set of manually determined documents. This, it should be noted, is much the same as the data classification process. Formally speaking, data classification is the process which finds the common properties among a set of objects in a database and classifies them into different classes according to a predetermined classification model. To construct such a classification model, a sample database is treated as the training set, in which
each sample consists of the same set of features (or attributes) and, traditionally, each sample has a known class identity (or label) associated with it. The objective of the classification is to analyze the training data and develop an accurate description or a model for each class, using the features available in the data. Such class descriptions are then used to classify future test data or to develop a better description (or a so-called classification rule) for each class in the database. Some well-known methods for data classification include decision trees, ANN and SVM.

2.7 Feature Selection Method Review

The language identification problem can be seen as an instance of a more general problem; that of classifying objects using attributes. For this purpose different kinds of attributes have been used: for example; character (Newman, 1987), word (Giguet, 1996; Ingle, 1976; Souter et al., 1994), word classes (Lins and Gonçalves, 2004), particular N-grams (Beesley, 1988; Cavnar and Trenkle, 1994; Dunning, 1994) and sentences (Biemann and Teresniak, 2005).

2.7.1 Filter and Wrapper of Feature Selection

Feature selection approach can also be classified into two categories based on whether or not feature selection is done independently of the learning algorithm used to construct the classifier. If feature selection is performed independently of the learning algorithm, the technique is said to follow a filter approach. Otherwise, it is said to follow a wrapper approach (Kohavi and John, 1997). While the filter approach is generally computationally more efficient than the wrapper approach, its major drawback is that an optimal selection of features may not be independent of the inductive and representational biases of the learning algorithm that is used to construct the classifier. The wrapper approach on the other hand, involved the computational
overhead of evaluating candidate feature by executing a selected learning algorithm on the dataset represented using each feature under consideration. This is feasible only if the learning algorithm used to train the classifier is relatively fast. Figure 2.18 summarizes the filter and wrapper approaches. Most of the approaches in this work are belong to filter approach such as N-grams, windowing algorithm, small word technique, etc (Yang and Honavar, 1998).

Figure 2.18: Two approaches to feature selection based on the incorporation of the learning algorithm. Features are selected independently of the learning algorithm in filter approach, while features are generated and evaluated by a learning algorithm in wrapper approach (Yang and Honavar, 1998)

Guyon and Elisseeff (2003) also have given similar definition of wrapper and filter approach in feature selection. They defined wrapper as utilize the learning machine of interest as a black box to score subsets of variable according to their predictive power. Filters select subsets of variables as a preprocessing step, independently of the chosen predictor. Wrappers are often criticized because they seem to be a method requiring massive amounts of computation, but it is not necessarily so. Efficient search strategies may be devised. Using such strategies does not necessarily mean sacrificing prediction performance. In fact, it appears to be the converse in some cases: coarse search strategies may alleviate the problem of over fitting. Several justifications for the use of filters for feature selection have been put forward in this special issue and elsewhere. It is argued that, compared to wrappers, filters are faster. Still recently proposed efficient embedded methods are competitive in that respect.
Another argument is that some filters (e.g., those based on mutual information criteria) provide a generic selection of features, not tuned for / by a given learning machine. Another compelling justification is that filtering can be used as a preprocessing step to reduce space dimensionality and overcome over fitting.

According to Liu and Motoda (1998), other important aspects of feature selection include models, search strategies, feature quality measures, and evaluation. The three typical models are filter, wrapper, and embedded. An embedded model of feature selection integrates the selection of features in model building. An example of such a model is the decision tree induction algorithm, in which at each branching node, a feature has to be selected. The research shows that even for such a learning algorithm, feature selection can result in improved learning performance. In a wrapper model, one employs a learning algorithm and uses its performance to determine the quality of selected features. In this work, the wrapper model is implemented for evaluating the feature selection methods of web page language identification.

2.7.2 Statistical and Linguistic of Feature Selection

Feature selection, also known as variable selection, feature reduction, attribute selection or variable subset selection, is the technique commonly used in machine learning and the selection subsets of relevant features for building robust learning models. There are many types of features that have been applied in web page language identification, such as characters, words, sequence characters (or N-grams), sentences, prefixes or suffixes and special characters or words. Figure 2.19 illustrates these web page features. The features to be used in web page language identification are harder than other text processing applications because of the different languages that have various morphologies. Some of the works are concerning with certain language features only, and this can lead to false results being produced on particular applications, since there are so many languages in the world. Table B.2 on Appendix B describes these features and feature selection method in details.
Feature selection has as its function the reduction of unnecessary attributes in the original content. It not only cuts down the load of the learning algorithm, but also reduces biases in the raw data and increases the effect of learned results. Several feature selection methods have been proposed in the literature: for example; entropy, the small word technique, Unicode based identification, web page information and PCA (Hakkinen and Tian, 2001; Ljubesic et al., 2007; Martins and Silva, 2005; Prager, 1999; Zhai et al., 2006). The following sections describe these in detail and the summary is given on Table 2.2.

Linguistic is the scientific study of natural language, encompassing a number of sub-fields. An important topical division must be between the study of language structure like grammar and the study of meaning as semantic (Aizawa, 2001). Based on such studies, linguistic can now be applied effective on web page language identification. However, there is a high computational cost for constructing the desired linguistic rules for each of the languages in the world. The WWW keeps changing, and so the maintenance of linguistic rules becomes a significant problem. Further, statistical is a mathematical science pertaining to the collection, analysis, interpretation or explanation and presentation of data. Statistical language modeling has been successfully developed for speech recognition and information retrieval (Chien and Wu, 2009). The statistical methods are entropy, PCA, N-grams approach
and windowing algorithm. The linguistic methods are small word technique, Unicode based identification, web page information and HMMs.

Table 2.2: Comparison of feature selection methods

<table>
<thead>
<tr>
<th>Category</th>
<th>Method</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical</td>
<td>Entropy (Selamat and Lee, 2008; Selamat and Omatu, 2004)</td>
<td>This method takes into consideration both local and global frequency in order to find out the most important keywords of data set. It is useful in text categorization.</td>
<td>Document of any languages having possibilities appears the most important keywords, it may not suitable be used on language identification.</td>
</tr>
<tr>
<td></td>
<td>PCA (James and Sankaran, 2005; Selamat and Lee, 2008; Selamat and Omatu, 2004)</td>
<td>The strength of PCA is to reduce the dimensionality of a data set through feature reduction</td>
<td>It is useful on large dimension features but not on characters due to certain scripts’ characters are limited.</td>
</tr>
<tr>
<td></td>
<td>$N$-grams (Botha et al., 2006; Cavnar and Trenkle, 1994; Grefenstette, 1995; Martins and Silva, 2005; Robertson and Willet, 1998)</td>
<td>$N$-gram method performs well against the grammatical or morphological error</td>
<td>Feature dimension of this method is a problem, for example implementing on Hanzi script documents may produce up to million possibilities.</td>
</tr>
<tr>
<td>Windowing</td>
<td>Windowing Algorithm (Adams and Resnik, 1997; Mandl et al., 2006)</td>
<td>Extensive features can be generated by windowing algorithm</td>
<td>This method has the feature dimension and feature representation problem.</td>
</tr>
<tr>
<td>Category</td>
<td>Method</td>
<td>Advantages</td>
<td>Disadvantages</td>
</tr>
<tr>
<td>-------------------</td>
<td>---------------------------------------------</td>
<td>----------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Linguistic</td>
<td>Small Word Technique (Gavrilov, 2003; Hanif et al., 2007; Hedlund et al., 2001)</td>
<td>It is based on the assumption that function words, short and highly frequent</td>
<td>The dynamic content of web pages is a challenge to this method.</td>
</tr>
<tr>
<td></td>
<td>Unicode Based (Hanif et al., 2007; Palmer, 2009)</td>
<td>This method is based on the defined boundary in Unicode</td>
<td>It is useful to determine the script of language but cannot do precise language identification.</td>
</tr>
<tr>
<td></td>
<td>Web Page Information (Baykan et al., 2008; Gustafson et al., 2005; Hayati, 2004; Simons, 2000)</td>
<td>Features such as URL, metadata, inlinks and outlinks are used on feature selection</td>
<td>This method having reliability issue as the actual content might be different with the desired language feature.</td>
</tr>
<tr>
<td></td>
<td>HMMs (Dunning, 1994; Tran and Sharma, 2005; Xafopoulos et al., 2004)</td>
<td>This method is to compute reference Markov chains for the training sets</td>
<td>The basic idea is too complicated to be implemented especially on large numbers of languages.</td>
</tr>
</tbody>
</table>

2.7.3 Statistical

In this Section, four statistical feature selection methods including entropy, PCA, N-grams approach and windowing algorithm are discussed. Statistical method is the study of taking sample documents from each language and subject them to a
statistical analysis based on some characteristic of the text. There are extra statistical feature selection methods as shown in Figure B.1 in Appendix B.

2.7.3.1 Entropy

Originally, the entropy feature selection is so called the class profile-based approach, as proposed by Selamat and Lee (2008); Selamat and Omatu (2004). The authors use it for web pages classification. It focuses on the most regular words that exist in each category. Then, those features are combined with features of PCA in order to be classified using neural networks. In this work, the method has been enhanced to adopt it to a problem of web page language identification, in which the entropy weighting $EN_{jk}$ on particular letter $t_k$ in particular document $d_j$ is calculated as $L_{jk} \times G_k$ where $L_{jk}$ is the local entropy of particular letter $t_k$ in particular document $d_j$ and $G_k$ is the global entropy of particular letter $t_k$. The $L_{jk}$ and $G_k$ are given by

$$L_{jk} = \begin{cases} 1 + \log_{10} TF_{jk} & (TF_{jk} > 0) \\ 0 & (TF_{jk} = 0) \end{cases}$$

(2.1)

and

$$G_k = \frac{1 + \sum_{j=1}^{N} \frac{TF_{jk}}{F_k} \log_{10} \frac{TF_{jk}}{F_k}}{\log_{10} N}$$

(2.2)

$$EN_{jk} = L_{jk} \times G_k$$

(2.3)
where \( N \) is the number of documents in a collection of a particular language \( L_i \), \( TF_{jk} \) is the letter frequency of a particular letter \( t_k \) in document \( d_j \), the \( F_k \) is number of documents where the letter \( t_k \) appears in the entire document collection \( D \) and \( EN_k \) is the entropy weighting of particular letter in entire collection. Table 2.3 shows an example of letters that have been selected during the experiment as a result of their entropy weighting \( EN_k \) being among the highest in the collection. It is noticed that the weighting of letter “e” in English and French are 100.8 and 80.33, respectively, so the entropy algorithm selects “e” as one feature only. \( R = 50 \) letters that have the highest entropy value are selected as such features (Selamat and Omatu, 2004). Duplicate features are discarded as well as those having entropy lower than zero. This technique is mainly designed for text categorization, so it focuses on the weighting of the most important words among the training data set. Therefore, it may produce bad results during the web page language identification process.

Table 2.3: Example of entropy feature selection

<table>
<thead>
<tr>
<th>Language, ( L_i )</th>
<th>Letter, ( t_k )</th>
<th>Entropy, ( EN_k )</th>
<th>Selected?</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>e</td>
<td>100.8</td>
<td>Yes</td>
</tr>
<tr>
<td>Indonesia</td>
<td>a</td>
<td>90.99</td>
<td>Yes</td>
</tr>
<tr>
<td>English</td>
<td>i</td>
<td>90.01</td>
<td>Yes</td>
</tr>
<tr>
<td>French</td>
<td>e</td>
<td>80.33</td>
<td>No</td>
</tr>
<tr>
<td>Spanish</td>
<td>c</td>
<td>77.34</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Legend:
- Particular language, \( i (L_i) \)
- Particular letter \( k (t_k) \)
- Entropy weighting of particular letter \( k \) in collection, \( D (EN_k) \)
2.7.3.2 Principle Component Analysis (PCA)

PCA is a multivariate procedure that rotates the data in such a way that the maximum variability is projected onto the axes. Sets of correlated variables get transformed into uncorrelated variables, ordered by the reduction variability. These are linear combinations of the original variables, the last of which can be removed with minimum loss of real data. PCA is primarily used to reduce the dimensionality of a data set through feature reduction, while retaining as much information as possible. In earlier methods, Bayesian classification was performed on languages, with the sum of the probabilities in the spectrum of the random variables (in the \( N \)-grams) summed up to one. PCA captures information from them all and discriminates better by classifying them with respect to their principle components (James and Sankaran, 2005; Selamat and Lee, 2008; Selamat and Omatu, 2004).

According to James and Sankaran (2005), PCA is used primarily to reduce the dimensionality of a data set, while retaining as much information as possible. In principal component analysis there is the need to find the directions in the data having the most variation, for instance the eigenvectors corresponding to the largest eigenvalues of the covariance matrix, and then project the data onto these directions. Assume \( A \) is a matrix document of \( N \)-grams in the Equation 2.5.

\[
A = \begin{bmatrix}
    x_{11} & x_{12} & \cdots & x_{1\hat{n}} \\
    \vdots & \vdots & \ddots & \vdots \\
    x_{\hat{n}1} & x_{n2} & \cdots & x_{\hat{n}\hat{m}}
\end{bmatrix}
\]  

(2.5)

where, \( x_{nm} \) is the \( N \)-grams that exist in the collection, \( \hat{n} \) is the number of \( N \)-grams and \( \hat{m} \) is the number of documents. Then, the mean \( \bar{x} \) is calculated and subtracted from each data point \( x_i - \bar{x} \). Next, the variance-covariance matrix \( A \) is calculated, where the new value of \( x_{ij} = (x_j - \bar{x})(x_i - \bar{x}) \). Then, the eigenvalues and eigenvectors of the matrix determine which \( \hat{C} \) is a real symmetric matrix, so that a positive real number \( \hat{\lambda} \) and a nonzero vector \( \hat{\alpha} \) can be found such that \( \hat{C}\hat{\alpha} = \hat{\lambda}\hat{\alpha} \), where \( \hat{\lambda} \) is called an eigenvalue and \( \hat{\alpha} \) is an eigenvector of \( \hat{C} \). In order to find a nonzero vector \( \hat{\alpha} \), the characteristic
equation \( |\hat{C} - \hat{\lambda}I| = 0 \) must be solved. By using \((\hat{C} - \hat{\lambda}I) \hat{\alpha} = 0\), all corresponding eigenvectors can be found. The eigenvalues (and corresponding eigenvectors) will be sorted, so that \( \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \hat{\lambda}_3 \cdots \geq \hat{\lambda}_n \). Then, it is selected the first \( d \leq n \) eigenvectors (where \( d \) is the appropriate value to be fed into the identifier) and generate the data set in the new representation (usually compressed). Alternatively, it is possible to find the desired values based on the Singular Value Decomposition (SVD) method.

### 2.7.3.3 N-grams Approach

N-grams approach is the sequencing of \( N \) items from a given sequence. The \( N \)-gram approach performs well in learning lexical structure information of the text, and it is suitable for short words such as names, titles and bits of grammar (Hirsimaki et al., 2009; Sethy et al., 2009). It has been widely used in various applications such as spelling error detection and correction, query expansion, information retrieval with serials, inverted and signature files, dictionary look-up, text compression and language identification (Robertson and Willett, 1998). Although this method functions rapidly and accurately in most cases, it is limited when noise is exist in the documents.

The unigram approach (Martins and Silva, 2005) has been used on each of the letters in respective documents. The decision tree approach will figure out the most likely language for each letter on each web page. The language is determined by asking a series of questions about the context of the current letter. For instance, a string “Hello World” would be composed to number of unigram and bigram as shown in Figure 2.20. The fine dotted line means to shift the token to next targeted gram according to the size of \( N \)-grams. Therefore, the tokens of unigram after processing are م، ل، م، ل، م، ل، م، ل، م، ل، م، ل، م، ل، م، ل، م، ل، م، ل، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، م، ل، م، M (the meaning of “العالممرحبا” is “helloworld”).

The Trigram Technique (TT) calculates the frequency of the strings of three letters in a large sample. TT captures the intuition that, for example, a word ending in \(-ck\) is more likely to be an English word than a French, while a word ending
in -ez is more likely to be French. In the training phase for the implementation of TT, trigram statistics were collected from the European Corpus Initiative (ECI) corpora using the first million characters, including spaces, in each language from a group of 10 (Grefenstette, 1995). Each text was tokenized. For each language, all trigrams appearing more than 100 times were retained as attributes; each language was characterized by a number ranging between 2550 and 3560 of trigrams. The probability of each trigram is calculated for each language, and a minimal probability is assigned to each unseen trigram. The probability of the sequence of trigrams is calculated for each language. The most probable language is chosen as the identity. Both TT and Small Word Technique (SWT) reach 100% precision in most common situations (Grefenstette, 1995). All of these approaches have the problem of data sparseness. Most of them need to apply some smoothing techniques to deal with the probability of potentially missing grams (Mohamed Ould, 2007). Moreover, the use of higher N-gram frequency decreases the classification performance (Pham and Tran, 2003) and additional computation costs arise when using a large value for N (Botha et al., 2006).

---

2The European Corpus Initiative (ECI) was founded to oversee the acquisition and preparation of a large multilingual corpus (ECI/MCI) to be made available in digital form for scientific research at a low cost as possible. The corpus has been available on CD-ROM since 1994 and is being distributed by ELSNET, retrievable at http://www.elsnet.org/incs/framer.html?/ecilisting.html.
The experience with natural languages that some words occur more frequently than others is formally expressed by what is known as Zipf’s Law (Benedetto et al., 2002). Cavnar and Trenkle (1994) summarized Zipf’s Law as the $N^{th}$ most common words in a human language text when the frequency is inversely proportional to $N$. In other words,

$$\hat{f} \propto \frac{1}{\hat{r}}$$  \hspace{1cm} (2.6)

where $\hat{f}$ is the frequency of the word and $\hat{r}$ is the rank of the word in the list ordered descending by the frequency (Cavnar and Trenkle, 1994). For $N$-grams frequency $NF$, it is based on the occurrences of the particular $N$-grams in a document, not the whole data set. The number of a particular $N$-grams contributed to the document is an important factor in a language identification. For example, the $N$-grams “de” appears in Spanish more frequently than in English, so a Spanish document has higher occurrences of that N-grams. The formula of $N$-grams frequency ($NF$) is given by Equation 2.7, where $T_d$ is the total $N$-grams in the document, $d_j$. Each unigram, bigrams and trigrams calculation is done separately and the highest $NF$ is selected as features.

$$NF_L (ngm) = \sum_{d=1}^{D} \left( \frac{\sum ngm_d}{T_d} \right)$$  \hspace{1cm} (2.7)

It is assumed that the $N$-grams of particular passive language ($L_m$ is given by $ngm_{Lm}$ in the training. Then, the N-grams of predicted text are given by $ngm_{\hat{r}}$ where $\hat{r}$ is the rank of the $N$-grams in the predicted text ordered descending by the frequency of Zipf’s Law. Therefore, the output of prediction (winner) is given by the argument minimum,

$$winner = \arg \min_{1..m} \left( \sum_{1..\hat{r}} \sqrt{(ngm_{\hat{r}} - ngm_{Lm})^2} \right)$$  \hspace{1cm} (2.8)
where the winner is selected with the minimum distance to particular language.

2.7.3.4 Windowing Algorithm

Windowing algorithms have been used for capturing the letters inside each document. It is slightly different with N-grams that windowing algorithms consist of a sliding window and a non-sliding window algorithm. Figure 2.21 shows how the non-sliding window algorithm works. On the left side of figure is the original Arabic script document and the right side is the encoded document. If the window size is 3, then the window will capture the first token that consist of the first three letters, رون، which correspond to the document value of decimal, 1585, 1608 and 1606. Then it will shift to next three letters، ل، for second token. The processes will be repeated until the end of the document (Adams and Resnik, 1997; Mandl et al., 2006).
Windowing algorithms include sliding window and non-sliding window. These algorithms are similar to \(N\)-gram algorithms. Windowing algorithms lack of reliability if applied to language identification. The feature dimension is too large since windowing algorithms do not have a proper way to reduce the dimension. If the process of document preprocessing is not sophisticated, then it will greatly affect the performance of a windowing algorithm. However, this algorithm can be further developed by integrating algorithms such as PCA or ICA to increase the reliability of input features (Mandl et al., 2006).

### 2.7.4 Linguistic

In this Section, linguistic methods such as small word technique, Unicode based identification, web page information and HMMs are presented. Linguistic method uses certain aspects of the language to identify it, for example a set of words in each language. There are extra linguistic methods as shown in Figure B.1 in Appendix B.

#### 2.7.4.1 Small Word Technique

The SWT is representative of a group of approaches using particular words (Gavrilov, 2003). SWT is based on the assumption that function words, short and highly frequent, such as determiners, conjunctions and prepositions are good clues for determining a language. SWT uses a corpus for training and for each language. In this procedure words that appear more often than a frequency threshold and having five characters or less are retained. So, given a new document to classify, it is tokenized. Tokens appearing in the short word list are assigned their probabilities; tokens not in the list are assigned a minimum probability. The probability of a web page being in a given language is taken as the product of the probabilities of each token belonging
to that language (Ingle, 1976). Figure 2.22 illustrates a flow chart for calculating the probability of small word technique.

Figure 2.22: Probability of small word technique

The small word technique is similar to the so-called Dictionary Based Identification (DBI) (Hanif et al., 2007; Hedlund et al., 2001). The DBI uses unigrams rather than compounds. Unigrams cover all the linguistics features. They are generated and stored in a dictionary. One of the biggest advantages is its performance is that it gives good results when comparing languages belonging to different scripts, $\hat{S}$. However, the drawback in such dictionary based implementations is that they fail to perform well when used to compare languages using the same script. Also a document containing too few characters gives too little information to DBI for it to identify a language effectively. Also, in cases of multilingual web pages, the results are not able to distinguish a language consistently (Hanif et al., 2007; Hedlund et al., 2001).

2.7.4.2 Unicode Based Identification

Unicode is a standard encoding or character set (charset) that has been widely used for any text encoding, including web page encoding. Each script ($\hat{S}$) / language ($L_i$) in Unicode has defined by code ranges as shown in Figure 2.23 (Palmer, 2009). This information is provided in the form of Unicode Database which gives type, language, code point and other information for every character of every language
represented in Unicode (Allen, 2006). Any textual documents based on Unicode can easily be segmented or determined according to the particular script or language. However, for multiple languages using the same Unicode script, the possibility of false identification of a particular language is very high. For example, there are a number of possible languages that can be generated from Arabic script of Unicode such as Arabic, Persian, Urdu, Pashto, etc. Therefore, Unicode based identification can serve as the initial processing for web page language identification but not for multilingual script language identification (Hanif et al., 2007). The issue of encoding, script and language identification can be seen in Section 2.2.3.
2.7.4.3 Web Page Information

There are a number of elements in a web page that can be used as features for web page language identification. These include, for example; URL, metadata, inlinks and outlinks. Although the process of web page language identification can be done very quickly, there are certain problems that need to be taken into consideration, for example; the wrong encoding is specified, the usage of non-European language without specifying the encoding, the loss of characters or scrambling of texts when performing a translation, false identification of multilingual page and little or no text in one particular web page (Gustafson et al., 2005).

In general, a URL is referred to a web page address. Every URL is made up of some combination of the following: the scheme name or resource type; a registered domain name or Internet protocol address, a port number, the path name of the file to be fetched or the program to be run, the query string with html files, an anchor for where the page should be displayed. Usually, the top-level domains are used as features for language prediction. For example, www.google.com.tw and www.google.co.jp are two different domains; one could simply justify www.google.com.tw as being a Chinese web page and www.google.co.jp as a Japanese. However, this assumption does not work if the domain belongs to a generic one. For instance, the general domains are .com, .net, .org, etc. Moreover, the top-level domain is an Internet top-level domain and generally used or reserved for a country or a dependent territory. In other words, the top-level domain is based on a country code and not a language. Therefore, a top level domain or URL is not suitable for use as a feature (Baykan et al., 2008; Hayati, 2004).

Usually, a web page will consist of metadata that describes the characteristics of a resource. For example, a simple metadata consists of few elements such as; title, creator, publisher, identifier, format, relation, charset and language. Therefore, one can justify the language of a web page by using metadata (Simons, 2000). However, it is noticed that a document with a UTF-8 charset is not only used for writing one language, but has multiple languages on same page. So, the metadata is not reliable for multilingual web page language identification.
Table 2.4: Inlink and outlink (Yi and Jin, 2008)

<table>
<thead>
<tr>
<th></th>
<th>Inlink</th>
<th>Outlink</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
<td>A link pointing into a site</td>
<td>A link pointing out of a website</td>
</tr>
<tr>
<td><strong>Synonyms</strong></td>
<td>Incoming link, inbound link, inward link, back link</td>
<td>Outgoing link, outbound link, outward link, forward link</td>
</tr>
</tbody>
</table>

Inlink or outlink within a page have also been used as the features for web page language identification (Hayati, 2004). The term hyperlink refers to a clickable navigation element imbedded in a web page that leads to another web page (external linking or outlink) or to another portion within the same page (internal linking or inlink) (Yi and Jin, 2008). Table 2.4 shows the different between inlink and outlink in terms of definition and synonyms. However, these features are not reliable, since the link content might not be the same language as the source web page. For example, the BBC website provides different links to web pages of different languages. For this reason, the targeted link cannot be used reliably for predicting the language of a particular web page.

### 2.7.4.4 Hidden Markov Models (HMMs)

The HMMs are another line of research (Xafopoulos et al., 2004). They have been found to be slightly less efficient then the $N$-gram method. They are also not easy to be implemented. The basic idea behind HMMs is to compute reference Markov chains for the training sets. Then when a text comes in, it must compute the probability that a certain model produced the output. The model with the highest probability is chosen as an answer (Dunning, 1994; Tran and Sharma, 2005).

Figure 2.24 presents Markov chains proposed by Tran and Sharma (2005), in which having 3 states $t$, $o$ and $n$ represent English words such as *no*, *on*, *to*, *not*, *ton*,...
too, tot, toot and noon, French word non, German words tot and toon and Spanish word no and tono.

Figure 2.24: Markov chains having 3 states t, o and n represent English words such as no, on, to, not, ton, too, tot, toot and noon, French word non, German words tot and toon and Spanish word no and tono (Tran and Sharma, 2005).

In general, HMMs consist of three problems to be resolved:

(i) Given the parameters of the model, compute the probability of a particular output sequence.

(ii) Given the parameters of the model and a particular output sequence, find the state sequence that is most likely to have generated that output sequence.

(iii) Given an output sequence or a set of such sequences, find the most likely set of state transition and output probabilities. In other words, derive the maximum likelihood estimate of the parameters of the HMM given a dataset of output sequences.

2.8 Identification Method Review

Web page language identification is a multi-classes problem. For one particular script like Arabic, it can be used to write Arabic, Persian, Urdu, Jawi and Pashto
Figure 2.25: Summary of language identification methods

languages. Therefore, an Arabic script model that has been trained by Arabic script data is needed to do Arabic script language prediction or identification, same goes on other scripts of languages as well. In this Section, different models are presented to illustrate the advantages and disadvantages of each algorithm. Figure 2.25 shows the language identification methods including supervised and unsupervised methods. Supervised methods are ANN, fuzzy ARTMAP, SVM and decision tree while unsupervised methods are VQ and KNN.

2.8.1 General Language Identification Methods

There are many things in the world in need of identification such as library books, documents, web data, films, etc. However in this work, it is focuses on the web page language identification due to the explosive expansion of data on the web. Language identification is defined as the task of assigning an electronic document to one or more languages, based on its contents (Xafopoulos et al., 2004). It is usually involved with machine learning methods such as ANN or SVM for generalizing predictive models (Dunning, 1994). Table B.2 in Appendix B describes the identification method in detail.

2.8.2 Artificial Neural Networks (ANN)
ANN is fundamentally a parallel processor that each node performs the weight calculation simultaneously. ANNs are computer programs that are biologically inspired to simulate the way in which the human brain processes information. It has been applied to many applications including language identification because of their fascinating features, such as learning, generalizing, fast real-time computation and classification capabilities (Sagiroglu et al., 2007). For example, MacNamara et al. (1998) used ANN in combination with Roman letters in the identification of the language of the entries in a library catalogue; Sagiroglu et al. (2007) applied ANN to frequency analysis of letters in the identification of languages in multilingual documents.

Figure 2.26: Backpropagation neural networks architecture

Figure 2.26 shows the example of architecture of BPNN identification (Selamat and Omatu, 2004). This BPNN consists of one input layer (p), one hidden layer (q) and one output layer (r). The total nodes of input layer depend upon the feature size, s, used for capturing input patterns. If the feature size, s, is 15 then the number of input layers will be set correspondingly. The number of one hidden layer is eight units. The number of an output layer consists of 2 units, which are based on the corresponding output. Table 2.5 shows the corresponding orthogonal language codes used in the backpropagation neural network output layer. The output are binary forms that (0 0) represent Arabic language, (0 1) represent Persian language, (1 0) represent Urdu language and (1 1) represent Jawi language, respectively. However, at times the
orthogonal language codes might be different due to the design of BPNN architecture. Usually, the actual output produced by the model is compared to the desired output in order to insure the accuracy of the model. It is a justification of the model performance.

Table 2.5: Orthogonal language codes

<table>
<thead>
<tr>
<th>Language</th>
<th>Corresponding vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>1 0</td>
</tr>
<tr>
<td>Persian</td>
<td>0 1</td>
</tr>
</tbody>
</table>

The neural networks parameters are defined as $\hat{\ell}$ for the iteration number, $\hat{t}$ for the number of letter in a document, $\eta$ for the learning rate, $\Gamma$ for the momentum rate, $O_p$ for the output on unit $p$, $O_q$ for the output on unit $q$, $O_r$ for the output on unit $r$, $W_{qp}$ for the $q^{th}$ weight to the unit $p^{th}$, $W_{rq}$ for the $r^{th}$ weight to the unit $q^{th}$, $net_q$ for the first transfer function at hidden layer $q$, $net_r$ for the second transfer function at output layer $r$, $\theta_q$ is for the bias on hidden unit $q$, $\theta_r$ is for the bias on output unit $r$, $\delta_q$ is for the generalized error through a layer $q$ and $\delta_r$ is for the generalized error through a layer $q$ and $r$. The input values of the backpropagation neural network are represented by $in$ where $in$ is between 1 and $s$ ($in \in [1, s]$), where $s$ is for the window size of the windowing algorithm. The output values to the backpropagation neural network are represented by $out$ where $out \in [1, 2]$, which correspond to Table 2.5. Adaptation of the weight between hidden ($q$) and input ($p$) layers is given by,

$$W_{qp}(\hat{\ell} + 1) = W_{qp}(\hat{\ell}) + \Delta W_{qp}(\hat{\ell} + 1) \quad (2.9)$$

where

$$\Delta W_{qp}(\hat{\ell} + 1) = \eta \delta_q O_p + \Gamma \Delta W_{qp}(\hat{\ell}) \quad (2.10)$$
and

\[ \delta_q = O_q(1 - O_q) \sum_r \delta_r W_{rq} \]  \hspace{1cm} (2.11)

Note that the first transfer function at the hidden layer \((q)\) is given by,

\[ \text{net}_q = \sum_q W_{qp}O_p + \theta_q \]  \hspace{1cm} (2.12)

and

\[ O_q = f(\text{net}_q) = \frac{1}{1 + e^{-\text{net}_q}} \]  \hspace{1cm} (2.13)

Adaptation of the weights between output \((r)\) and hidden \((q)\) layers is given by,

\[ W_{rq}\left(\hat{\ell} + 1\right) = W_{rq}\left(\hat{\ell}\right) + \Delta W_{rq}\left(\hat{\ell} + 1\right) \]  \hspace{1cm} (2.14)

where

\[ \Delta W_{rq}\left(\hat{\ell} + 1\right) = \eta \delta_r O_q + \Gamma \Delta W_{rq}\left(\hat{\ell}\right) \]  \hspace{1cm} (2.15)
and

$$\delta_r = O_r (1 - O_r) \left( \hat{\ell}_r - O_r \right) \quad (2.16)$$

Then the output function at the output layer \((r)\) is given by,

$$net_r = \sum_r W_{rq} O_q + \theta_r \quad (2.17)$$

and

$$O_r = f (net_r) = 1 / \left( 1 + e^{-net_r} \right) \quad (2.18)$$

### 2.8.3 Fuzzy ARTMAP

Table 2.6 shows the development of Adaptive Resonance Theory (ART) neural networks or adaptive neural network. The ART-1 is an approach introduced by Carpenter and Grossberg (Carpenter and Grossberg, 1987a). It was an unsupervised learning system to cluster binary \((0, 1)\) input patterns. ART-2 (Carpenter and Grossberg, 1987b) and fuzzy ART (Carpenter et al., 1991a) are extended from the ART-1 domain to categorize analog inputs that are represented by continuously variable as well as binary input patterns. ART-2A is a faster version evolved from ART-2
Table 2.6: Evolution of the adaptive neural networks

<table>
<thead>
<tr>
<th>No.</th>
<th>Approach</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>ART-1 (Carpenter and Grossberg, 1987a)</td>
<td>Binary version</td>
</tr>
<tr>
<td>2.</td>
<td>ART-2 (Carpenter and Grossberg, 1987b)</td>
<td>Analog version</td>
</tr>
<tr>
<td>3.</td>
<td>ART-2A (Carpenter et al., 1991c)</td>
<td>Fast version of ART-2</td>
</tr>
<tr>
<td>4.</td>
<td>ART-3 (Carpenter and Grossberg, 1990)</td>
<td>Hierarchical ART structure</td>
</tr>
<tr>
<td>5.</td>
<td>ARTMAP (Carpenter et al., 1991b)</td>
<td>Supervised version</td>
</tr>
<tr>
<td>6.</td>
<td>Fuzzy-ART (Carpenter et al., 1991a)</td>
<td>Hybrid variants, analog version</td>
</tr>
<tr>
<td>7.</td>
<td>Fuzzy-ARTMAP (Carpenter et al., 1992)</td>
<td>Hybrid variants, supervised fuzzy-ART</td>
</tr>
<tr>
<td>8.</td>
<td>Default ARTMAP (Carpenter, 2003)</td>
<td>Hybrid variants, default parameter values</td>
</tr>
</tbody>
</table>

(Carpenter et al., 1991c). ART-3 is an adaptive neural networks hierarchy approach that implements distributed pattern recognition codes (Carpenter and Grossberg, 1990).

The supervised ART architectures of neural network are called ARTMAP systems (Carpenter et al., 1991b). It is based on self-organizing an arbitrary map from input vectors, representing features such as spectral values and terrain variables. Furthermore, it produces output vectors, representing prediction such as face recognition in image processing (Carpenter et al., 2005). It creates stable recognition categories by maximizing code compression while minimizing predictive error in a real-time prediction. When fuzzy ART replaces ART-1 in an ARTMAP system (Carpenter et al., 1991b), the resulting fuzzy ARTMAP (Carpenter et al., 1992) architecture rapidly learns stable mappings between analog or binary input and output vectors. The default ARTMAP algorithm summarizes a series of steps taken to classify an arbitrary number of output classes in a supervised learning problem.

Fuzzy ARTMAP is the machine learning method used for classification. The fuzzy ARTMAP approach (Carpenter et al., 1992) is a supervised learning version of
ART neural network for representing discrete categories using real-valued features. It implements the Winner Take All (WTA) scheme in both training and testing processes. During the training process, the ARTMAP approach searches for a coding node $J$ that meets the matching criterion, $\alpha$ is feature vector ($0 \leq \alpha \leq 1$), $\alpha_c$ is the complement of $\alpha$ (e.g., if $\alpha = 0.4$ then $\alpha_c = -0.6$), $B$ is the complement coded input vector, $I$ is number of input features, $\rho$ is the vigilance variable, $\varepsilon$ is match tracking ($\varepsilon \in (-1,1)$), $y$ is the coding field activation pattern, $C$ is the number of committed coding nodes, $w$ is the weight vector of fuzzy ARTMAP, $o$ is the input component index, $\hat{p}$ is the coding node index, $\hat{q}$ is the output class index, $w_{\hat{p}}$ is the coding node weight vector $\hat{p}$, $w_{\hat{q}}$ is the output class weight vector $\hat{q}$ and $J$ is the chosen coding node (WTA) (Carpenter, 2003). This approach predicts the correct output class $K$, as follows:

Step 1. For the next sorted coding node ($\hat{p} = J$) that meets the matching criterion,

$$\left| \frac{B \land w_J}{I} \right| \geq \rho, \quad \text{set } y_J = 1 \text{ (WTA)} \quad (2.19)$$

Step 2. Output class prediction,

$$\sigma_{\hat{q}} = \sum_{\hat{p}=1}^{C} w_{\hat{p}\hat{q}} y_{\hat{p}} = w_{J\hat{q}} \quad (2.20)$$

Step 3. If the active code $J$ predicts the actual output class $K$ ($\sigma_{\hat{q}} = w_{J\hat{q}} = 1$), go to Step 5

Step 4. If the active code $J$ fails to predict the correct output class ($\sigma_{\hat{q}} = 0$), raise vigilance and continue searching,
\[ \rho = \frac{|B \land w_J|}{I} + \varepsilon \]  \hspace{1cm} (2.21)

Step 5. The process of learning occurs by updating the coding weights,

\[ w_{J}^{new} = \lambda (B \land w_{J}^{old}) + (1 - \lambda) w_{J}^{old} \]  \hspace{1cm} (2.22)

\[ B = (a, a^c), \text{ where } a \in [0, 1] \text{ and } |B| = I \]  \hspace{1cm} (2.23)

2.8.4 Support Vector Machine (SVM)

SVM is a relatively new statistical classification method proposed by Vapnik in 1995 (Cortes and Vapnik, 1995). Based on the principle of Structural Risk...
Minimization principle, SVM tries to find a separating hyper-plane with maximum margin to separate positives examples and negative examples from the training set. It makes decisions based on support vectors that are selected as the only effective elements from the training set (Selamat and Ibnu Subroto, 2007; Zou et al., 2005). Figure 2.27 shows the SVM finds the hyperplane $h$, which is separated from the positive and negative training examples with a maximum margin. The examples that close to the hyperplane are called support vectors which are marked with a circle. SVM classifier is design for binary classification that declared to positive and negative class. For example, the Arabic script LID actually is a multi-class problem for the Arabic, Persian, Urdu and Jawi class. These classes are divided into two groups, the positive and negative class. If Arabic is (+) and the others are (-), and so on for the other languages. So that these LID need 4 classifiers for four languages.

Mohamed Ould (2007) used SVM to perform Arabic script language identification, but the classification result is doubtful due to the stopping and stemming process. The question is how to perform stopping or stemming process for the document without knowing the predefined language because the corresponding process might be different (Kranig, 2005). For example, it cannot use the English stopping and stemming technique on a Malay document.

2.8.5 Decision Trees

The decision tree method has been applied to language identification of words and can learn lexical structure information very well, which makes it suitable for short words such as names, titles, grammars or characters. Intuitively, common words such as determiners, conjunctions and prepositions are good clues for identifying a language. The advantage of decision tree is that it is simple to understand and interpret. It performs well with large data in a short time and is a white box model.\footnote{If a given situation is observable in a model the explanation for the condition is easily explained by boolean logic, it is so called white box model. An example of a black box model is an artificial neural network since the explanation for the results is excessively complex to be comprehended.}
Hakkinen and Tian (2001) have proposed language identification using decision tree to determine the most likely language for each letter in the input word. The language is obtained by asking a series of questions about the context of the current letter, as defined by the corresponding decision tree. The most relevant context for a given input letter is found by splitting the data set in such a way that entropy is minimized. They use letter context both on the left and right hand side of the current letter. Since only the letter context is used and no frequency information is stored in the tree, a very compact presentation is obtained. A separate tree is trained for each letter contains an attribute and a tag representing the most likely language in the current context. Figure 2.28 depicts a simple decision tree. The tree is composed of a root node and internal nodes, each containing an attribute and a language tag and leaves that contain only language tags (Hakkinen and Tian, 2001).

![Decision Tree Example](image)

**Figure 2.28**: Example of decision tree showing the nodes and leaves with attribute \( a_i \) and language tags \( l_j \) (Hakkinen and Tian, 2001)

### 2.8.6 Vector Quantization (VQ)

VQ is an effective classification method for data compression on speech and image processing. According to Pham and Tran (2003), vector quantization can be efficient for web page language identification because it can process large volumes of training data and its solution is based on optimality criteria of the nearest neighbor and centroid conditions. Furthermore, it can speed up the classification tasks based on
reduced template matching, by which the noisy signal can also be tolerable. However, the work does not clearly show the comparison with state-of-art methods in web page language identification. Moreover, it does not justify the source of data, the preprocessing step and the measurement of time processing for supporting the work.

Given a language document $\hat{L}_q$, it is initiated with preprocessing the document by removing all common characters and punctuation marks such as commas, columns, semi-columns, quotes, stops, exclamation marks, question marks, signs, etc. The next step is to convert all the characters into lower case. These lower-case characters are translated into the corresponding American Standard Code for Information Interchange (ASCII) values. For example, $(a, b, ..., z) = (97, 98, ..., 122)$. The training and classification procedures are summarized as Section 2.8.6.1 and 2.8.6.2.

2.8.6.1 Vector Quantization (VQ) Training

(i) Given a textual document $\hat{L}$ of language $\hat{q}$: $\hat{L} \hat{q} \in \hat{C}$, where $\hat{C}$ is the universe of languages.

(ii) Remove all common characters from $\hat{L}_q$ to obtain a set of word $\hat{W}^\hat{q} = \{\hat{w}^\hat{q}_1, \hat{w}^\hat{q}_2, ..., \hat{w}^\hat{q}_G\}$.

(iii) Map each word of $\hat{b}$ characters, to a vector of numbers of fixed size $M$.

(iv) Build a codebook $\hat{\beta}^{\hat{q}}$ of $\hat{N}$ codewords.

2.8.6.2 Vector Quantization (VQ) Testing

(i) Given a textual document of an unknown language $\hat{L}$.
(ii) Do steps 2 and 3 for ̂L as described in the training phase to obtain a set of words ̂W ∈ ̂Z.

(iii) Calculate the average minimum distance between ̂W and ̂β̃q, ̃q = 1, 2, · · · , ̃Q, where ̃Q is the number of languages.

\[
\hat{d}_q = \frac{1}{N} \sum_{y=1}^{N} \min \left[ \hat{d} \left( \hat{W}, \hat{C}_q^y \right) \right]
\]  

(2.24)

(iv) Assign ̂L to the language ̃q that has the minimum distance:

\[
\hat{q}^* = \arg \min_{\hat{q}} \left( \hat{d}_q \right)
\]  

(2.25)

2.8.7 K-Nearest Neighbor (KNN)

![Figure 2.29: KNN classifier (Selamat and Ibnu Subroto, 2007)](image)

Selamat and Ibnu Subroto (2007) have proposed a KNN classifier to determine the class label of language identification. Any test examples are classified into the class which has dominant examples among its K closest neighbors. Figure 2.29 shows the KNN-classifier with K=3 (Selamat and Ibnu Subroto, 2007). The classifier calculates the three nearest data samples and predicts the class of the document closest to it. The distance for this experiment is determined by using the Euclidean distance formula.

From the three examples of documents ̂x1, ̂x2 and ̂x3 will predict the language using
KNN classifier, it is easy to classify $\hat{x}_1$ as predicted to (+) class, $\hat{x}_2$ predicted to (-) class and $\hat{x}_3$ predicted to (+) class.

2.9 Evaluation Approach

Forman (2003) have summarized various approaches used on the feature selection of text. The standard evaluation approaches used including T-test, $F_1$ measurement and $f$-fold cross validation as shown in Figure 2.30. The T-test is used to justify the significant different between the means of two group. As is well known, a classifier achieving good classification performance on training data does not necessarily generalize well on new test data, or the classifier might overfit the training data. Hence, $f$-fold cross validation or its variants are widely adopted to separate training and test data. Accuracy, precision, recall and $F_1$ are used to validate the identification accuracy and retrieval performance of web page language identification (Dougherty, 2005; Liu and Motoda, 2006; Singhi and Liu, 2006). $F_1$ measurement is the harmonic mean of precision and recall (Refaeilzadeh et al., 2007).
2.9.1 T-test

A T-test is applied when the population is assumed to be normally distributed but the sample sizes are small enough that the statistics upon which an inference is based is not normally distributed because it relies on an uncertain estimate of standard deviation rather than on a precisely known value. For example, a T-test compares the means of two groups, compare whether systolic blood pressure differs between control and treated group, between men and women, or any other two groups.

The T-test is a hypothesis test for answering questions about the mean where the data are collected from two random samples of independent observations, each from an underlying normal distribution. It also is a method for supporting stochastic models of the population viability. It is based on assessing the mean and variance of the predicted population size (Lee, 2008; Lee et al., 2008).

Usually, when using the T-test functions, it has to identify the ranges that contain two sets of sample data. A two-tailed test is a test in which both ends of the sampling distribution are taken into account when setting critical regions. One-tailed test is concerned with only one end of the sampling distribution (Lee, 2008; Lee et al., 2008).

When carrying out a T-test, it will formally entertain two hypotheses:

\[ H_0: \text{Population means are the same, } \mu_1 = \mu_2. \]
\[ H_1: \text{Population means are not the same, } \mu_1 \neq \mu_2. \]

Mean (\( \bar{x} \)) is the arithmetic average of a set of values or distribution and is given by Equation 2.26.
\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad (2.26) \]

Standard deviation is a measure of the dispersion of a set of values. It can apply to a probability distribution, a random variable or a population. The standard deviation is usually denoted with the letter \( \sigma \). It is defined as the square root of the variance given by Equation 2.27.

\[ \sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2} \quad (2.27) \]

The standard deviation of the mean \((S)\) is given by Equation 2.28.

\[ S = \frac{\sigma}{\sqrt{n}} \quad (2.28) \]

where \( \sigma \) is the standard deviation of the variable and \( n \) is the number of observations.

The degree of freedom \((DoF)\) is the number of parameters that may be independently varied and may be used in a problem much as population and distribution. If there are two means to be estimated, then \( DoF \) is given by Equation 2.29.

\[ DoF = n_1 + n_2 - 2 \quad (2.29) \]
A T-test is conducted to determine whether the means are different and it can be calculated as shown in Equation 2.30.

\[
t = \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1 - \bar{x}_2}}
\]  

(2.30)

\[
S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{(n_1 - 1) S^2_1 + (n_2 - 1) S^2_2}{n_1 + n_2 - 2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}
\]  

(2.31)

where \(x_1\) and \(x_2\) are the mean of first and second group, \(n_1\) and \(n_2\) are the number of observations of first and second group, \(S^2_1\) and \(S^2_2\) are the variance of first and second group, respectively. Note that in this case, \(S_{\bar{x}_1 - \bar{x}_2}\) is the pooled variance of the two samples. The \(t\)-value will be positive if the first mean is larger than the second and negative if it is smaller. The difference is not statistically significant if the \(t\)-value is within the critical region. On the contrary, the difference is statistically significant when the \(t\)-value is outside the critical region (Lee, 2008; Lee et al., 2008).

### 2.9.2 Precision, Recall and F1 Measurements

The proposed methods are evaluated using the standard of information retrieval measurements that are precision (\(\tilde{p}\)), recall (\(\tilde{r}\)) and \(F1\). They are defined as follows:

\[
\tilde{p} = \frac{\tilde{a}}{\tilde{a} + \tilde{b}}
\]  

(2.32)
\[
\tilde{r} = \frac{\tilde{a}}{\tilde{a} + \tilde{c}} \tag{2.33}
\]

\[
F_1 = \frac{2}{\frac{1}{p} + \frac{1}{\tilde{r}}} \tag{2.34}
\]

where the values of \(\tilde{a}, \tilde{b}\) and \(\tilde{c}\) are defined in Table 2.7. The relationship between the classifier and the expert adjustment is expressed using four values as shown in Table 2.8. The \(F_1\) measure is a kind of average of precision and recall.

Table 2.7: The definitions of the parameters \(\tilde{a}, \tilde{b}\) and \(\tilde{c}\) which are used in Table 2.8

<table>
<thead>
<tr>
<th>Value</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\tilde{a})</td>
<td>The system and the expert agree with the assigned category</td>
</tr>
<tr>
<td>(\tilde{b})</td>
<td>The system disagrees with the assigned category but the expert did</td>
</tr>
<tr>
<td>(\tilde{c})</td>
<td>The expert disagrees with the assigned category but the system did</td>
</tr>
<tr>
<td>(\tilde{d})</td>
<td>The system and the expert disagree with the assigned category</td>
</tr>
</tbody>
</table>

Table 2.8: Decision matrix for calculating the classification accuracies

<table>
<thead>
<tr>
<th>Expert</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>(\tilde{a})</td>
<td>(\tilde{b})</td>
</tr>
<tr>
<td>No</td>
<td>(\tilde{c})</td>
</tr>
</tbody>
</table>

The precision describes the probability that an retrieved Arabic document (randomly selected) is relevant to a certain language. The recall describes the probability of a relevant Arabic document being retrieved. \(F_1\) describes the average between precision and recall which is often used in the field of information retrieval for measuring search, document classification, and query classification performance (Selamat and Omatu, 2004).
2.9.3 Cross Validation and Accuracy

$f$-fold cross validation has been applied, which is one of the accuracy estimation methods for classification method that has been used to validate the experiments (Geisser, 1993; Kohavi, 1995). The $f$-fold cross validation is the statistical practice of partitioning a sample of data into subsets such that the analysis is initially performed on a single subset, while the other subset(s) are retained for subsequent use in confirming and validating the initial analysis. The initial subset of data is called the testing set; the other subset(s) are training sets. Data set ($D \in [d_1, d_N]$) has been used in that method by randomly split it into $f$ mutually exclusive subsets (namely the folds), $D_1, D_2, D_h, ...$ of approximately equal size. The cross-validation for estimating the accuracy, $E_1, E_2, E_h, ...$ is the overall number of correct identifications $co$, divided by the number of patterns $pa$ in the data set. Therefore, overall accuracy $Acc_D$ is given by

$$Acc_D = \frac{1}{b} \sum_{h=1}^{b} E_h$$

(2.35)

where $b$ is the number of subsets, $E_h$ is the accuracy of subset $h$ and is given by

$$E_h = \frac{co}{pa} * 100$$

(2.36)

Figure 2.31 shows the example of 5-cross validation and accuracy data set. The data set $D$ is divided into 5 equally size of subsets ($D_1, D_2, D_3, D_4$ and $D_5$). The evaluation is done in 5 loopings. For first loop, the subset 1 ($D_1$) is an unknown data set that used for testing the trained identifier. Other subsets are training data sets that are used for training the identifier. Only subset 1 will be used for calculating the accuracy of evaluation in first loop. The process will be repeated by using subset 2 ($D_2$) as testing subset and other subsets used as training data set. Accuracy is evaluated on subset 2. The process of switching the testing subset is repeated until the fifth subset ($D_5$).
Finally, accuracy of each looping will be averaged and produced the actual accuracy of this data set $Acc_D$.

### 2.10 Summary

Based on the studies that have been carried out, it is to find that most of the research is focusing on the common languages of the world. Many methods lack research on minority languages such as languages using Arabic script. This may lead to the raising of the digital divide in the WWW. Therefore, problem of the web page language identification has been reviewed by expanding the limitations of existing methods. Appendix B shows the comparison of feature selection and identification method related to web page language identification. Appendix F shows the related works have been done and published in the journal or conference. The following chapter discusses in detail the methods that have been applied and also the steps involved in the methodology.
CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Web pages are digitalized using a different encoding schemes or character sets for computer display. Each letter is represented by a code point inside the character set. Therefore, since the letters in each language are unique, it is best to deal with letters in web page language identification. In this work, it is assumed the character set of web page has been identified (Appendix A contains the details related to character set detection). The script within a character set also can be identified because each letter has a unique code point. Suzuki et al. (2002) provide an explanation of the relationships among languages, characters, scripts and encoding schemes. Therefore, this work only focusing on how to improve web page language identification in terms of feature selection methods.

This chapter highlights the four feature selection methods adopted for this study. The first method is N-grams method proposed by Cavnar and Trenkle (1994). The second conventional method is entropy feature selection proposed by Selamat and Omatu (2004). The third and fourth methods are the proposed methods, namely letter weighting and simplified entropy. Letter weighting is an enhancement of the entropy feature selection method in terms of selecting the most frequently appearing letters.
instead of the most relevant. Simplified entropy takes into consideration both letter frequency and feature position by simplifying both the $N$-grams and the entropy feature selection method. The objective of this work is to increase the classification accuracy for web page language identification. The following sections of this chapter describe each phase involved in web page language identification and the differences among each method. This is followed by a description of the standard measurements that have been applied in the experiments.

### 3.2 Operational Framework

Figure 3.1 depicts the operational framework utilized in this work. The overall research is divided to initial study and literature review, methodology, findings evaluation and discussion and conclusion. The methodology is divided into two phases: corpora preparation and feature selection methods evaluation. First phase corpora preparation is including data preparation (labelled as A) and data preprocessing (B). Second phase feature selection methods evaluation is including feature selection using $N$-grams (C), feature selection using entropy (D), feature selection using letter weighting (E), feature selection using simplified entropy (F), identification using argument minimum (G), identification using fuzzy ARTMAP (H) and identification using decision tree (I). The details of methodology flow are discussed on Section 3.3.

Initial study and literature review is to understand the existing methods and problems occurred in web page language identification. It is including the study of internet development, overview of web page language identification, web page format issues, feature selection problems, conventional web page language identification process, feature selection methods overview, identification methods overview and evaluation methods. Feature selection problem is the main focus in this research. Based on the study of feature selection methods, the strength and weakness of each method has been identified. Then, the weakness of particular method has been improved, for example $N$-grams and entropy as presented in this research. The summary of literature review is given on Chapter 2.
Methodology is the way to carry out the web page language identification experiments. It is divided into two phases as stated above which is corpora preparation and feature selection methods evaluation. Corpora preparation is to justify the data of interest to be used on web page language identification. The data has been identified from the web repository due to the lack of language expertises for manually developing the language data. There are two experiments have been simulated for determining the validity of collected data. It is presented in Section 4.3.1 and Section 4.3.2. The second phase feature selection methods evaluation is to justify the robustness and effectiveness of methods used. The conventional methods like $N$-grams and entropy have been revised and improved to letter weighting and simplified entropy as shown in the Section
4.4.1, 4.4.2 and 4.4.3. The following Section 3.3 will explain the methodology flow in details.

Findings evaluation is to validate the outcome of experiments and justify advantages and disadvantages of each method. There are number of feature selection issues have been discussed on Section 2.5 such as irrelevant and redundant features, bias occured and effectiveness of feature selection methods. The data set has been divided into training and testing data. It is to avoid the redundant features to be used again. Cross validation method has been utilized to reduce the bias occured. Accuracy, precision, recall and $F_1$ are used to validate the identification accuracy and retrieval performance of web page language identification (Dougherty, 2005; Liu and Motoda, 2006; Refaeilzadeh et al., 2007; Singhi and Liu, 2006). The discussion of each experiment is given afterward result analysis and conclusion of this research work is summarized on Chapter 5.

As shown in Figure 3.1, there are four common steps to be followed: data preparation (A), data preprocessing (B), Feature Selection (C, D, E and F) and Identification (G, H and I). Data preparation is employed to collect and identify the data of interest. Since this research is focusing on web pages language identification, most of the web pages have been collected from news websites. The detail of data preparation is given on Section 3.4. Data preprocessing filters and normalizes irrelevant contents from the original sources, such as the programming codes and unrecognized characters. Data preprocessing details are illustrated on Section 3.5. Feature selection identifies the appropriate features or attributes to be used for language identification. It can be divided into two categories, statistical and linguistic. The function of feature selection in text categorization is slightly different than in language identification. For example, the Term Frequency Inverse Document Frequency (TFIDF) used in text categorization may not be suitable for use in language identification due to the fact that word frequency is not reliable and varies in the same language, while frequency provides the significant features for text categorization. Feature selection algorithm is presented on Section 3.6. Identification is the process that determines the predefined language based on the features derived from the data and it is discussed on Section 3.7.
3.3 Methodology Flow

In this section, the second phase of operational framework as shown in Figure 3.1 is elaborated. It is related to the feature selection methods evaluation of web page language identification.

Table 3.1: Comparison of the proposed methods

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Preparation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Data Preprocessing</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Feature Selection Methods</td>
<td>Entropy</td>
<td>Letter Weighting</td>
<td>Simplified Entropy</td>
<td></td>
</tr>
<tr>
<td>Identification Method</td>
<td>Argument Minimum</td>
<td>Fuzzy ARTMAP</td>
<td>Fuzzy ARTMAP</td>
<td>Decision Tree</td>
</tr>
</tbody>
</table>

Table 3.1 summarizes the differences among each feature selection methods. As stated above, four feature selection methods are investigated in this work. Data preparation and data preprocessing are same for all methods. Entropy and letter weighting are utilize the fuzzy ARTMAP as the identification method. However, the \( N \)-grams and simplified entropy are applied to the argument minimum and decision tree, respectively.

There are four feature selection methods of language identification that are studied and whose results are compared to others in terms of identification accuracy. The first conventional method is \( N \)-grams feature selection. This method was proposed by Cavnar and Trenkle (1994) and used mainly in text categorization and language identification. A second method is used in web mining, which is proposed by Selamat and Omatu (2004), namely entropy feature selection. Letter weighting and simplified entropy is the proposed feature selection methods in this work. The first method \( N \)-grams utilizes \( N \)-gram frequency in producing a feature position for identification. The
second and third methods focus on letter frequency. The last method, which is feature selection using simplified entropy, is an enhancement of the $N$-grams, entropy and letter weighting feature selection methods.

### 3.3.1 $N$-grams Feature Selection Method

Figure 3.2: The $N$-grams feature selection method (note: the sequence is ordered by the label A, B, C and G which is data preparation, data preprocessing, feature selection using $N$-grams and identification using argument minimum, respectively)

Figure 3.2 shows the $N$-grams feature selection method has been used on web page language identification. This method was proposed by Cavnar and Trenkle (1994) and used mainly in text categorization and language identification. It is ordered by A (data preparation), B (data preprocessing), C (feature selection using $N$-grams) and G (identification using argument minimum).
Figure 3.3 shows the algorithm has been applied on the N-grams feature selection method. The input is a data set of experiments, \( D \); the parameters are \( d_j \), the particular document \( j \), \( L_i \), is the particular language \( i \), \( \hat{S} \), is the particular script of languages, \( \hat{S}_{\text{begin}} \), is the begin codepoint of a particular script, \( \hat{S} \), and \( \hat{S}_{\text{end}} \), is the end codepoint of particular script, \( \hat{S} \); and the outputs are \( NF \), is the particular N-grams frequency in a document and \( ngm^L_m \), is N-grams of the particular passive language model ordered descending. The details of N-grams algorithm is presented on Section 3.6.1 and 2.7.3.3.

1. INPUT: 
2. \( D \), is a data set of experiments 
3. PARAMETERS: 
4. \( d_j \), is the particular document \( j \). \( L_i \), is the particular language \( i \). \( \hat{S} \), is the particular script. \( \hat{S}_{\text{begin}} \), is the begin codepoint of a particular script. \( \hat{S}_{\text{end}} \), is the end codepoint of particular script. 
5. OUTPUT: 
6. \( NF \), Particular N-grams frequency in a document. \( ngm^L_m \), N-grams of the particular passive language model ordered descending. 
7. BEGIN 
8. \( D \leftarrow \hat{S} \); //set the experimental data set to particular script. 
9. set \( \hat{S}_{\text{begin}} \) and \( \hat{S}_{\text{end}} \); //based on the selected script 
10. for each \( L_i \in \hat{S} \) do 
11. for each \( d_j \in D \) do 
12. if \( \hat{S}_{\text{begin}} \leq ngm \leq \hat{S}_{\text{end}} \) then 
13. \( \text{find } \sum ngm_d \); 
14. \( d_j + 1; \) 
15. else 
16. set \( d_j \) as error document; 
17. \( d_j + 1; \) continue; //proceed to next document 
18. \( \text{find } \sum ngm_d/T_d; \) 
19. \( NF_L(ngm) = \sum_{d=1}^{D} ngm_d/T_d; \) //refer to N-grams approach 
20. sort \( NF_L(ngm) \) \( \rightarrow ngm^L_m \); 
21. set features = \( ngm^L_m \); //a particular language feature is set to \( ngm^L_m \) 
22. \( L_i + 1; \) //set to other languages of that particular script 
23. repeat the process from line 11 to 22 until the end of the languages of script; 
24. \( D \leftarrow \hat{S}; \) //set to other scripts of data set 
25. repeat the process from line 9 to 24 until the end of data set; 
26. END 

Figure 3.3: Algorithm of N-grams feature selection method

At the beginning of experiment, the experimental data set is set to particular script, \( D \leftarrow \hat{S} \) and the boundary of that script is set to \( \hat{S}_{\text{begin}} \) and \( \hat{S}_{\text{end}} \) as shown on line 8 and 9 in Figure 3.3. For each particular language, \( L_i \) which is a subset of a particular script, \( \hat{S} \), do the looping from line 11 to 22 until the end of the languages of that script. Then, for each particular document, \( d_j \), do the looping from line 12 to 18 until the end of
the document in the data set. If the particular $N$-grams, $ngm$ is between the boundary of $\hat{S}_{\text{begin}}$ and $\hat{S}_{\text{end}}$, then find the total of that particular $N$-grams, $\sum_{d_j} ngm$, in the particular document, $d_j$; find the $N$-grams weighting as line 18 and proceed to next document. Otherwise, set the particular document, $d_j$, as problematic document and proceed to next document. For line 19 until 21, it is to find out the $N$-grams sequence of particular language and then set the features as the $N$-grams ordered in descending. The process of $N$-grams feature selection is repeated until the end of the other scripts in data set as line 25. The algorithm is ended when all the features of each particular language are found. Those features position are used to compare with the predicted document feature position which is using the similar process on producing the $N$-grams features. The language is identified by using the argument minimum and the winner is the one with smallest distance as presented in Section 3.7.1.

3.3.2 Entropy Feature Selection Method

Figure 3.4 shows the second feature selection method using entropy. This method was proposed by Selamat and Omatu (2004) specifically used on text categorization. It is ordered by A (data preparation), B (data preprocessing), D (feature selection using entropy) and H (identification using fuzzy ARTMAP).

Figure 3.5 illustrates the algorithm of entropy feature selection method. The input is a data set of experiments, $D$; the parameters are $d_j$, the particular document $j$, $L_i$, is the particular language $i$, $\hat{S}$, is the particular script of languages, $\hat{S}_{\text{begin}}$, is the begin codepoint of a particular script, $\hat{S}$, $\hat{S}_{\text{end}}$, is the end codepoint of particular script, $\hat{S}$, and $R$, is the threshold or number of features of input patterns; the outputs are $L_{jk}$, is the local entropy of particular letter $k$ in particular document $j$, $G_k$, is the global entropy of particular letter $k$ in particular document $j$, $EN_{jk}$, is the entropy weighting of particular letter $k$ in particular document $j$, and $EN_k$, is the entropy weighting of particular letter $k$ in data set, $D$. The details algorithm of entropy feature selection is discussed on Section 3.6.2 and 2.7.3.1.
At the beginning of experiment, the experimental data set is set to particular script, $D \leftarrow \hat{S}$; the boundary of that script is set to $\hat{S}_{\text{begin}}$ and $\hat{S}_{\text{end}}$; and the threshold, $R$ is set to 50 as shown on line 8, 9 and 10 in Figure 3.5. For each particular language, $L$, which is a subset of a particular script, $\hat{S}$, do the looping from line 12 to 27 until the end of the languages of that script. Then, for each particular document, $d_j$, do the looping from line 13 to 23 until the end of the document in the data set. If the particular letter, $t_k$ is between the boundary of $\hat{S}_{\text{begin}}$ and $\hat{S}_{\text{end}}$, then find the term frequency, $TF_{jk}$, if $TF_{jk}$ is larger than 0 then calculate the local weighting, $L_{jk}$, else $L_{jk}$ is equal to zero. Otherwise, set the particular document, $d_j$ as problematic document and proceed to next document. For line 22 that if $t_k$ is a subset of $d_j$ then increment the $F_k$ and $d_j$. After that, find the global entropy, $G_k$ as shown in line 24. Line 25 and 26 is to find entropy weighting, $EN_{jk}$ and average entropy weighting, $EN_k$. The process is repeated on other languages as shown in line 28. Then, sort the $EN_k$ from the highest to the lowest. For each particular $EN_k$ which is a subset of $t_k$, do the looping until the end of the $k$. If $EN_k$ is lower or not equal to zero and threshold, $R$ is lower than 50 then check again if the existing features is not a subset of $EN_k$ then features are including $EN_k$. The process of entropy feature selection is repeated until the end of
INPUT: 
$D$, is a data set of experiments

PARAMETERS: 
- $d_j$, is the particular document $j$. $L_c$, is the particular language $i$. $S$, is the particular script. $S_{begin}$, is the begin codepoint of a particular script. $S_{end}$, is the end codepoint of a particular script. $R$, is the threshold.

OUTPUT: 
- $L_{jk}$, is the local entropy of particular letter $k$ in particular document $j$. $G_k$, is the global entropy of particular letter $k$. $EN_{jk}$, is the entropy weighting of particular letter $k$ in particular document $j$. $EN_k$, is the entropy weighting of particular letter $k$ in data set, $D$.

BEGIN: 
$D \leftarrow S$; //set the experimental data set to particular script.

set $S_{begin}$ and $S_{end}$; //based on the selected script

set $R = 50$;

for each $L_i \in S$ do
  for each $d_j \in D$ do
    if $S_{begin} \leq t_k \leq S_{end}$ then
      find $TF_{jk}$;
      if $TF_{jk} > 0$ then
        $L_{jk} = 1 + \log_{10} TF_{jk}$; //local weighting
      else
        $L_{jk} = 0$;
    else
      set $d_j$ as error document;
      $d_j + 1$; continue; //proceed to next document
    if $t_k \in d_j$ then
      $F_k + 1$; $d_j + 1$;
      $\sum_{j=1}^{N} \frac{TF_{jk}}{TF_{ik}} \times \log_{10} \left( \frac{TF_{jk}}{TF_{ik}} \right)$; //global weighting
      $G_k = \frac{1}{\log_{10} k}$
      $EN_{jk} = L_{jk} \times G_k$;
      $EN_k = \frac{\sum_{j=1}^{N} EN_{jk}}{N}$; //entropy weighting
      $L_i + 1$; //set to other languages of that particular script
    repeat the process from line 11 to 27 until the end of the languages of script;
    sort $EN_k$ from the highest to the lowest;
    for each $EN_k \in t_k$ do
      if $EN_k \leq 0$ && $R \leq 50$ then
        if $features \notin EN_k$ then
          $features \leftarrow EN_k$; //feature is set to $t_k$
        $D \leftarrow S$; //set to other scripts of data set
      repeat the process from line 9 to 34 until the end of data set;
  END

Figure 3.5: Algorithm of entropy feature selection method

the other scripts in data set as line 35. The algorithm is ended when all the features of each particular script are found. Those features are used to extract the frequency of particular document in order to do language identifier training and predicting the testing documents by using fuzzy ARTMAP as discussed in Section 3.7.2.
3.3.3 Letter Weighting Feature Selection Method

Figure 3.6: The letter weighting feature selection method (note: the sequence is ordered by the label A, B, E and H which is data preparation, data preprocessing, feature selection using letter weighting and identification using fuzzy ARTMAP, respectively)

Figure 3.6 shows the third feature selection method namely letter weighting. It is an enhancement of entropy method and the process order is A (data preparation), B (data preprocessing), E (feature selection using letter weighting) and H (identification using fuzzy ARTMAP).

Figure 3.7 depicts the algorithm of letter weighting feature selection method. The input is a data set of experiments, $D$; the parameters are $d_j$, the particular document $j$, $L_i$, is the particular language $i$, $\hat{S}$, is the particular script of languages, $\hat{S}_{begin}$, is the begin codepoint of a particular script, $\hat{S}$, $\hat{S}_{end}$, is the end codepoint of particular script, $\hat{S}$, and $R$, is the threshold or number of features of input patterns; the outputs are $\alpha_{ik}$, is the local letter weighting of particular letter $k$ of particular language $i$, $\beta_{ik}$, is the global letter weighting of particular letter $k$ of particular language $i$, and $\omega_{ik}$, is the
INPUT:

$D$, is a data set of experiments

PARAMETERS:

- $d_j$, is the particular document $j$.
- $L_i$, is the particular script.
- $\hat{S}_{\text{begin}}$, is the begin codepoint of a particular script.
- $\hat{S}_{\text{end}}$, is the end codepoint of particular script.
- $R$, is the threshold.

OUTPUT:

- $\alpha_{ik}$, is the local letter weighting of particular letter $k$ of particular language $i$.
- $\beta_{ik}$, is the global letter weighting of particular letter $k$ of particular language $i$.
- $\omega_{ik}$, is the letter weighting of particular letter $k$ of particular language $i$.

BEGIN:

1. $D \leftarrow \hat{S}$; //set the experimental data set to particular script.
2. set $\hat{S}_{\text{begin}}$ and $\hat{S}_{\text{end}}$; //based on the selected script
3. set $R = 5$;
4. for each $L_i \in \hat{S}$ do
   5.   for each $d_j \in D$ do
      6.     if $\hat{S}_{\text{begin}} \leq t_k \leq \hat{S}_{\text{end}}$ then
      7.         $TF_{jk} = TF_{jk} + 1$;
      8.     else
      9.         set $d_j$ as error document;
     10.        $d_j + 1$; continue; //proceed to next document
     11.    end
     12.   end
     13. end
     14. $N = D$;
     15. $A_{ik} = \sum_{j=1}^{N} TF_{jk} / N$ //average accumulated frequency of a particular letter $t_k$
     16. $\alpha_{ik} = A_{ik} / \max(A; A_{ik} \in A)$;
     17. $L_i + 1$; //set to other languages of that particular script
     18. repeat the process from line 11 to 21 until the end of the languages of script;
     19. $\hat{M} = \sum L_i - 1$;
     20. $\beta_{ik} = \sum_{m=1}^{\hat{M}} A_{ik} / P_{im}$,
     21. $\omega_{ik} = \alpha_{ik} + \beta_{ik}$;
     22. for each $L_i \in \hat{S}$ do
     23.     sort $\omega_{ik}$ from the highest to the lowest;
     24.     if $\omega_{ik} \neq 0 \& \& R \leq 5$ then
     25.         if features $\notin \omega_{ik}$ then
     26.             features $= \omega_{ik}$; //feature is set to $\omega_{ik}$
     27.         $D \leftarrow \hat{S}$; //set to other scripts of data set
     28.     end
     29. end
     30. repeat the process from line 9 to 31 until the end of data set;

END

Figure 3.7: Algorithm of letter weighting feature selection method

letter weighting of particular letter $k$ of particular language $i$. The details algorithm of letter weighting is presented on Section 3.6.3.

At the beginning of experiment, the experimental data set is set to particular script, $D \leftarrow \hat{S}$; the boundary of that script is set to $\hat{S}_{\text{begin}}$ and $\hat{S}_{\text{end}}$; and the threshold, $R$ is set to 5 as shown on line 8, 9 and 10 in Figure 3.7. For each particular language, $L_i$, which is a subset of a particular script, $\hat{S}$, do the looping from line 12 to 21 until the end of the languages of that script. Then, for each particular document, $d_j$, do the looping from line 13 to 17 until the end of the document in the data set. If the
particular letter, $t_k$ is between the boundary of $\hat{S}_{\text{begin}}$ and $\hat{S}_{\text{end}}$, then $TF_{jk}$ is increment one. Otherwise, set the particular document, $d_j$ as problematic document and proceed to next document. Set $N$ equal to $D$. Find the average accumulated frequency of a particular letter, $A_{ik}$ and local letter weighting, $\alpha_{ik}$. The process is repeated on other languages. The total passive language, $\hat{M}$, global letter weighting of particular letter, $\beta_{ik}$ and letter weighting of particular letter, $\omega_{ik}$ are calculated as shown in line 23, 24 and 25, respectively. For each particular language, $L_i$ which is a subset of a particular script, $\hat{S}$, do the looping from line 27 to 30 until the end of the languages of that script. Sort the $\omega_{ik}$ from the highest to the lowest; if $\omega_{ik}$ is lower or not equal to zero, threshold is lower or equal to five and if features is not including $\omega_{ik}$ then feature is set to $\omega_{ik}$.

The process of entropy feature selection is repeated until the end of the other scripts in data set as line 32. The algorithm is ended when all the features of each particular script are found. The features found of particular script are used to extract the letter frequency of web documents in order to do language identification training and testing by using fuzzy ARTMAP as described in Section 3.7.2.

### 3.3.4 Simplified Entropy Feature Selection Method

Figure 3.8 shows the fourth feature selection method, simplified entropy. It is an enhancement of $N$-grams and entropy method. The process order is A (data preparation), B (data preprocessing), F (feature selection using simplified entropy) and I (identification using decision tree).

Figure 3.9 shows the algorithm of simplified entropy feature selection method. The input is a data set of experiments, $D$; the parameters are $d_j$, the particular document $j$, $L_i$, is the particular language $i$, $\hat{S}$, is the particular script of languages, $\hat{S}_{\text{begin}}$, is the begin codepoint of a particular script, $\hat{S}$, and $\hat{S}_{\text{end}}$, is the end codepoint of particular script, $\hat{S}$; the outputs are $LF_{jk}$, is the letter frequency of particular letter $k$ in particular document $j$, $DF_k$, is the document frequency of particular letter $k$, and $LFDF_{jk}$, is the letter frequency document frequency of particular letter $k$ in particular document $j$, so called simplified entropy. The details algorithm of simplified entropy is shown on Section 3.6.4
At the beginning of experiment, the experimental data set is set to particular script, $D \leftarrow \hat{S}$ and the boundary of that script is set to $\hat{S}_{\text{begin}}$ and $\hat{S}_{\text{end}}$ as shown on line 8 and 9 in Figure 3.7, respectively. For each particular language, $L_i$ which is a subset of a particular script, $\hat{S}$, do the looping from line 11 to 21 until the end of the languages of that script. Then, for each particular document, $d_j$, do the looping from line 12 to 17 until the end of the document in the data set. If the particular letter, $t_k$ is between the boundary of $\hat{S}_{\text{begin}}$ and $\hat{S}_{\text{end}}$, then find the $TF_{jk}$ and $TF_j$. The local frequency is calculated by dividing $TF_{jk}$ with $TF_j$. Otherwise, set the particular document, $d_j$ as problematic document and proceed to next document. If one particular letter, $t_k$ is a subset of $d_j$ then $DF_k$ is the logarithm of $d_j$. The $LFDF_{jk}$ is calculated by multiplying $LF_{jk}$ and $DF_k$. After that, sort the $LFDF_{jk}$ from the highest to the lowest and features is set to $LFDF_{jk}$. The process is repeated on other languages and also other scripts as well. The algorithm is ended when all the features of each particular language are found. The features found are used to calculate letter frequency convergence and the fastest one is the winner based on decision tree as illustrated in Section 3.7.3.
1. INPUT: 
   D, is a data set of experiments 
2. PARAMETERS: 
   dj, is the particular document j. Li, is the particular script. Sbegin, is the begin codepoint of a particular script. Send, is the end codepoint of particular script. 
3. OUTPUT: 
   LFjk, is the letter frequency of particular letter k in particular document j. DFk, is the document frequency of particular letter k. LFDFjk, is the letter frequency document frequency of particular letter k in particular document j, so called simplified entropy. 
4. BEGIN: 
   D ← S; //set the experimental data set to particular script. 
   set Sbegin and Send; //based on the selected script 
   for each Li ∈ S do 
      for each tk ∈ D do 
         if Sbegin ≤ tk ≤ Send then 
            find TFjk,∑TFj; 
            LFjk = TFjk/∑TFj; 
         else 
            set dj as error document; 
            dj + 1; continue; //proceed to next document 
      DFk = log10(dj; tk ∈ dj); 
      LFDFjk = LFjk × DFk; 
      sort LFDFjk from the highest to the lowest; 
      set features = LFDFjk; //a particular language feature is set to LFDFjk 
      Li + 1; //set to other languages of that particular script 
   repeat the process from line 11 to 22 until the end of the languages of script; 
   D ← S; //set to other scripts of data set 
   repeat the process from line 9 to 24 until the end of data set; 
   END

Figure 3.9: Algorithm of simplified entropy feature selection method

3.4 Data Preparation Design

As seen in Figure 3.10, this phase consists of four steps: data requirement analysis, data collection, data labeling and data integration. For requirement analysis, data of the desired language obtained from a particular expert or source is used to justify the correctness of the language identification. For this reason, certain news websites such as BBC and CNN are preferable because the contents are written by native speakers. Moreover, this can reduce the bias content of the material to be used in the experiment analysis. Second, based on such requirements, a crawler (Roche, 2008) is implemented to collect the data from these news websites, World Language News, etc (Thompson, 2008; Turner, 2007), it is such a data collection method. The third step

1The particular seed URLs are randomly selected from the news domain for each language, so the content for particular web page is different with others, e.g., the first English web page is related to sport news but the first Indonesian web page is related to international news.
is data labeling. After collecting the data it is needed to identify the structure and its content. This is to ensure that the content of data corresponds to the needs of language identification. A tag corresponding to the predefined language of data source is applied to the data once the data has been downloaded. The data of same script’s language is gathered into the same folder manually in a process known as data integration.

There are two critical issues at this phase; the diversity of encoding schemes (or character sets) and multi-sources of data. Currently, the standard encoding scheme Unicode or UTF-8 is still lacks Internet user implementation (Abd Rozan et al., 2005). Different websites may have their own encoding schemes. For example, the language Tamil is written based on several character sets; namely, Iscii, Shree-Tam, Tab, Tam, Tscii, Vikatan, etc. (James and Sankaran, 2005) and some of the web pages in the Universal Declaration of Human Rights (UDHR) database have to put the content not able to be encoded in the form of PDF or images (Abd Rozan et al., 2005; Suzuki et al., 2002). Therefore, the encoding scheme of all the downloaded data has been converted into Unicode prior to processing. The web page character set can be identified by the header part of HTML documents that have been written in \texttt{charset=(encoding type)}, for example \texttt{charset=“UTF-8”} as shown in Figure 3.11a. After the preprocessing of the web pages, the preprocessed document that has been saved in the local computer is converted to UTF-8 as shown in Figure 3.11b. This is a standard encoding method for multilingual document processing.

For multi-sources data, it is to identify the unique seed URL for each data source to avoid data duplication and the inclusion of irrelevant data that may exist in databases by the Yahoo site explorer. The unrecognized data are filtered out of the database. This is to ensure the consistency of the data.
3.5 Data Preprocessing Design

After the data are collected and identified, they must be examined to identify any characteristics that may be unusual or indicative of more complex relationships. This is due to the fact that the data from a previous phase may be impure, divergent, untrustworthy, or even fraudulent. Therefore, data preprocessing is a transformation, or conditioning, of data designed to make modeling more robust. The goal of data inspection is to ensure data quantity and quality. The former is to check the size of datasets, the latter to determine any unusual dataset patterns. The data quantity check can be performed by observation. The data quality inspection can be performed by statistical methods including checking for missing data, unrecognized data, the over-fitting or under-fitting of data sizes, etc. There are five main issues usually occurring in data language identification. First, too many data have been collected in the database, a problem that can be solved by dividing the data into smaller portions for validation. This can increase processing time and also reduce the chances of tracking the faulty data. The second issue is that of having too few data for a certain language. It is needed to recollect data from other sources in order to increase the size. Some researchers simply multiply the data collected in order to overcome the insufficiency of data, but this has the problem of not representing the actual data. Third, the missing data may be simply replaced by using more data from the same language. Fourth, irrelevant data such as programming code and out of range data for certain scripts can be filtered out. Finally, multi-scale data appeared during entropy and letter weighting feature selection. This
can be solved by scoping the data to be extracted. To accomplish this, the first five hundred letters for each document has been utilized.

For any machine learning method, the input patterns should be normalized before being fed into the classification model. A min-max normalization method has been applied to all input patterns of fuzzy ARTMAP. The process takes the original data file and transforms each of non-class attributes, \( \varpi_z \) that is a value of a dimension \( z \) other than the class dimension of every data point in a real value \( \varphi_z \) between 0 and 1, using the following min-max normalization as shown in Equation 3.1.

\[
\varphi_z = \frac{\varpi_z - \min_z}{\max_z - \min_z}
\] (3.1)

where \( \min_z \) and \( \max_z \) are the minimum and maximum value of the dimension \( z \). Note that each normalized value is computed with respect to the minimum and maximum values in the corresponding column (dimension) rather than in the row (data point) (Al Shalabi et al., 2006; Saad, 2003). It is assumes here that \( \varpi_z, \min_z \) and \( \max_z \) are 30, 10 and 80, respectively. So, \( \varphi_z \) is calculated as \( (30 - 10)/(50 - 10) = 20/40 = 0.5 \).

### 3.6 Feature Selection Design

The following process of data preprocessing to be discussed is feature selection. The attributes found in this step are used to classifier training and prediction. At this phase, it is focuses on representing the data input patterns needed by the classifier. The preprocessed data will be divided into subsets for training and testing. It has used five cross validation methods as the benchmarks. This was done to insure that the results obtained correspond to all portions of the collected data. Feature selection is the process employed to establish the representative letters as features. For this reason four
feature selection methods have been compared; \textit{N}-grams, Entropy, Letter Weighting and Simplified Entropy.

At an earlier stage of this research, the letter weighting combination with the machine learning method has been made use for language identification. However, it was discovered that the letter weighting feature selection had limitations with respect to selecting the appropriate features. Therefore, the best way is to employ simplified entropy feature selection. It is also found that simplified entropy can maintain language identification in a stable manner. The following subsections will describe the proposed methods in detail.

### 3.6.1 \textit{N}-grams

As stated above, \textit{N}-grams feature selection has been used for text categorization and language identification by Cavnar and Trenkle (1994) and the inventor tool is the so-called TextCat (Cavnar and Trenkle, 2008). The algorithm is based on the sequencing of letters that slide across the text and derive the particular \textit{N}-grams. Then, each particular \textit{N}-gram’s frequency is analyzed through the data training in order to determine the features that correspond to each language. Finally, the position features are ordered from the highest to lowest weight and utilized in argument minimum in order to find out the closest distance from one particular language. In this work, only unigram is applied for a fair comparison as other feature selection methods are based on letter frequency only. The details of the algorithm can be seen in Section 2.7.3.3.
3.6.2 Entropy

Entropy feature selection is used for text categorization. This is to weight terms using the entropy weighting scheme before adding them to the feature vectors that have been selected from the PCA (Selamat and Lee, 2008; Selamat and Omatu, 2004). These feature vectors are then used as the input to a neural network for classification. The details of the algorithm can be seen in Section 2.7.3.1.

3.6.3 Letter Weighting

Letter weighting feature selection is an improvement over entropy feature selection. It is based on the letter frequency and takes into consideration other language weighting for the same letter. In other words, if one language has a higher global weighting for a particular letter, the more significant that letter is to that language. The formulation of letter weighting $\omega_{ik}$ is shown by,

$$\omega_{ik} = \alpha_{ik} + \beta_{ik}$$  \hspace{1cm} (3.2)

where $\alpha_{ik}$ is formulated as,

$$\alpha_{ik} = \frac{A_{ik}}{\text{Max} \{A : A_{ik} \in A\}}$$  \hspace{1cm} (3.3)
\[ A_{ik} = \frac{\sum_{j=1}^{N} TF_{jk}}{N} \]  \hspace{1cm} (3.4)

and \( \beta_{ik} \) is formulated as,

\[ \beta_{ik} = \sum_{m=1}^{\hat{M}} \frac{A_{ik}}{P_{ik}^m} \]  \hspace{1cm} (3.5)

where \( \alpha_{ik} \) is local letter weighting (as shown in Table 3.2), \( \beta_{ik} \) is global letter weighting (as shown in Table 3.3), \( A_{ik} \) is the average accumulated frequency of a particular letter \( t_k \) in particular language \( L_i \), \( P_{ik}^m \) is the average accumulated frequency of particular letter \( t_k \) in passive language \( L_m \) and total passive language is \( \hat{M} = \sum L_i - 1 \).

Table 3.2: Demonstration local letter weighting \( \alpha_{ik} \)

<table>
<thead>
<tr>
<th>Variable</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document 1</td>
<td>10</td>
<td>22</td>
<td>33</td>
<td>44</td>
<td>55</td>
<td>24</td>
</tr>
<tr>
<td>Document 2</td>
<td>12</td>
<td>23</td>
<td>34</td>
<td>45</td>
<td>56</td>
<td>56</td>
</tr>
<tr>
<td>Document 3</td>
<td>21</td>
<td>32</td>
<td>43</td>
<td>54</td>
<td>65</td>
<td>36</td>
</tr>
<tr>
<td>Document 4</td>
<td>12</td>
<td>23</td>
<td>34</td>
<td>54</td>
<td>65</td>
<td>23</td>
</tr>
<tr>
<td>Document 5</td>
<td>12</td>
<td>43</td>
<td>21</td>
<td>43</td>
<td>12</td>
<td>54</td>
</tr>
<tr>
<td>Document 6</td>
<td>32</td>
<td>43</td>
<td>54</td>
<td>67</td>
<td>87</td>
<td>54</td>
</tr>
<tr>
<td>Total</td>
<td>99</td>
<td>186</td>
<td>219</td>
<td>307</td>
<td>340</td>
<td>247</td>
</tr>
</tbody>
</table>

\[ A_{ik} \]

| \( A_{ik} \) | 16.50 | 31.00 | 36.50 | 51.17 | 56.67 | 41.17 |
| \( \alpha_{ik} \) | 0.29  | 0.55  | 0.64  | 0.90  | 1.00  | 0.73  |

Table 3.2 illustrates the location of letter weighting \( \alpha_{ik} \) of a particular letter \( k \). The frequency of letters a, b, c, d, e and f is calculated from the first document (Doc 1) to the sixth (Doc 6). The total is calculated as the frequency of a particular letter in all documents. For example, the total of “a” is \( 10 + 12 + 21 + 12 + 12 + 32 = 99 \). Then the average accumulated frequency \( A_{ik} \) is calculated as shown in Equation 3.4 and local letter weighting \( \alpha_{ik} \) is calculated as shown in Equation 3.3. For instance, \( A_{ik} \) of “a” is
99 ÷ 6 = 16.50 and \( \alpha_{ik} \) of “a” is 16.50 ÷ 56.67 = 0.29 were the maximum among \( \alpha_{ik} \) is 56.67.

### Table 3.3: Demonstration of global letter weighting \( \beta_{ik} \)

<table>
<thead>
<tr>
<th>Variable</th>
<th>English</th>
<th>Malay</th>
<th>French</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_{ik} )</td>
<td>16.50</td>
<td>6.50</td>
<td>1.50</td>
<td>55.50</td>
</tr>
<tr>
<td>( P_{ik}^{m} )</td>
<td>6.50,1.50,55.50</td>
<td>16.50,1.50,55.50</td>
<td>16.50,6.50,55.50</td>
<td>16.50,6.50,1.50</td>
</tr>
<tr>
<td>( \beta_{ik} )</td>
<td>13.84</td>
<td>4.84</td>
<td>0.35</td>
<td>48.90</td>
</tr>
</tbody>
</table>

Table 3.3 demonstrates the global letter weighting \( \beta_{ik} \) of a particular letter \( k \). The average accumulated frequency \( A_{ik} \) is calculated as shown in Equation 3.4. Global letter weighting \( \beta_{ik} \) is calculated as shown in Equation 3.5. For example, the \( A_{ik} \) of a particular letter “k” for English, Malay, French and Spanish are 16.50, 6.50, 1.50 and 55.50, respectively. Then, the \( P_{ik}^{m} \) for English is 6.50, 1.50 and 55.50, which is derived from the passive languages that are Malay, French and Spanish. Finally, the \( \beta_{ik} \) of the letter “k” of English is calculated as 16.50 ÷ 6.50 + 16.50 ÷ 1.50 + 55.50 ÷ 16.50 = 13.84. Letter weighting is used to determine the importance of a particular letter using local and global letter weighting. After this, the procedure sorts the letter weighting and ascertains the most important features to be selected for each language. The process of letter weighting feature selection is similar to entropy, as shown in Table 3.3. However, only the five highest weight features \( (R = 5) \) are selected from each language, duplicate features are discarded and zero letter weighting is not selected.

#### 3.6.4 Simplified Entropy

TFIDF is a method to weight terms often used in information retrieval or data mining in order to find the most important terms to represent each category (Chiang et al., 2008; Selamat and Omatu, 2004). However, simplified entropy is uses Letter Frequency Document Frequency (LFDF) to weight the letters and sort out the letters.
in order by sequence\(^2\). The sequence of the selected features is used in the proposed method to find the convergence point \(\chi\) and prefix converge point \(\gamma\).

The letter frequency in the given document is simply the number of times a given letter appears in that document. This measures the appearance of the letter \(t_k\) within the particular document \(d_j\) and is formulated as,

\[
LF_{jk} = \frac{TF_{jk}}{\sum TF_j}
\]

\[
DF_k = \log(d_j : t_k \in d_j)
\]

\[
LFDF_{jk} = LF_{jk} \times DF_k
\]

where \(TF_{jk}\) is the number of occurrences of the considered letter in document \(d_j\), the denominator is the number of occurrences of all letters \((TF_j)\) in document \(d_j\), and document frequency \(DF_k\) is the number of documents where the letter \(t_k\) appears. Then, the letter weight \(LFDF_{jk}\) is the product of letter frequency \(LF_{jk}\) and document frequency \(DF_k\).

A high weight of LFDF is reached by a high letter frequency (in the given document) and a high document frequency of the letter in the entire training set, hence the weight determines the most frequent appearance of a letter order by sequence. After this, the letters can be selected as features according to the sequence from highest to lowest weight.

\(^2\)LFDF is similar to concept term frequency inverse document frequency (TFIDF), in which it is focuses on letter instead of term.
Table 3.4 illustrates the selected letters as features and their correspondence weighting. For example, the first feature of the Pashto language is ږ, with the decimal Unicode point (code point) of 1608 and a weight of 162.14 units. This is followed by the second letter л with the weight of 134.73 units and so on. The sequence of features will contribute to the parameter of convergence point $\chi$ and prefix converge point $\gamma$. Although certain position features have the same letter, the sequence of feature is different for each language.

<table>
<thead>
<tr>
<th>Language</th>
<th>Legend</th>
<th>$1^{st}$ Feature</th>
<th>$2^{nd}$ Feature</th>
<th>$3^{rd}$ Feature</th>
<th>$4^{th}$ Feature</th>
<th>$5^{th}$ Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>Codepoint</td>
<td>1575</td>
<td>1604</td>
<td>1610</td>
<td>1605</td>
<td>1606</td>
</tr>
<tr>
<td></td>
<td>Character</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Weighting</td>
<td>211.36</td>
<td>147.52</td>
<td>107.65</td>
<td>75.8</td>
<td>71.42</td>
</tr>
<tr>
<td>Persian</td>
<td>Codepoint</td>
<td>1575</td>
<td>1585</td>
<td>1583</td>
<td>1606</td>
<td>1607</td>
</tr>
<tr>
<td></td>
<td>Character</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Weighting</td>
<td>197.52</td>
<td>121.75</td>
<td>92.92</td>
<td>92.34</td>
<td>82.44</td>
</tr>
<tr>
<td>Urdu</td>
<td>Codepoint</td>
<td>1575</td>
<td>1585</td>
<td>1608</td>
<td>1606</td>
<td>1605</td>
</tr>
<tr>
<td></td>
<td>Character</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Weighting</td>
<td>162.74</td>
<td>86.44</td>
<td>77.86</td>
<td>71.27</td>
<td>65.07</td>
</tr>
<tr>
<td>Pashto</td>
<td>Codepoint</td>
<td>1608</td>
<td>1575</td>
<td>1610</td>
<td>1607</td>
<td>1583</td>
</tr>
<tr>
<td></td>
<td>Character</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Weighting</td>
<td>162.14</td>
<td>134.73</td>
<td>103.95</td>
<td>97.59</td>
<td>83.3</td>
</tr>
</tbody>
</table>

The simplified entropy feature selection process selects features with a high letter weighting and the same total of features for each language, excluding those letter with a weighting of zero. A zero unit of letter weighting is useless in the classification because it never appears or appears only few times in the given document or collection. Finally, each selected group of feature for each language will be used in predicting the language of the document. For example, if a new, unknown language document is entered into the system, the sum of the features $\ell$ for one particular language will be calculated and then compared to others, with the highest the winner.
3.7 Language Identification Design

In this work, the argument maximum, fuzzy ARTMAP and decision tree are selected as the supervising identification method. Supervised learning is a machine learning technique for learning a function from training data. The training data consist of pairs of input objects (typically vectors) and their desired outputs. Most commonly, supervised learning generates a global model that maps input objects onto desired outputs. Fuzzy ARTMAP belongs to the supervised learning method. In data mining and machine learning, a decision tree is a predictive model; that is, a mapping from observations about an item to draw conclusions about its target value. In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. The machine learning technique for inducing a decision tree from data is called decision tree learning, or (colloquially) decision trees.

Figure 3.12: The argument minimum of N-grams
3.7.1 Argument Minimum

In general, the $N$-gram based method compares the successive $N$-grams derived from the text within a database of the $N$-gram sets generated for each language. Given an unknown document, the match for each language is obtained by comparing the $N$-gram frequency profile of the unknown document against the profile of each language using a distance measure (Cavnar and Trenkle, 1994). The unknown document is then classified as the language that is the closest distance as shown in Figure 3.12. The details of the algorithm can be found in Section 2.7.3.3.

3.7.2 Fuzzy ARTMAP

Based on the studies that have been done, the fuzzy ARTMAP has been chosen for this study as the machine learning method for identification. It has been done so because of its capability to map effectively different type of features. Moreover, it is able to produce promising results compared to other adaptive neural networks. The details of algorithms have been discussed in the Section 2.8.3.

3.7.3 Decision Trees for Simplified Entropy

In data mining and machine learning, a decision tree is the predictive model that maps observations about a sample onto conclusions about its target value. In this work, the parent of the decision tree is the input such as the sum of features $\ell$, convergence point $\chi$ and prefix convergence point $\gamma$. The leaves of the decision tree are the output of identification based on certain predefined rules. For example, if the highest sum of features $\ell$ is found in the first leaf, then the corresponding language is the winner. Figure 3.13 presents a simplified version of a decision tree being applied. However, it will be
noticed that it is not enough just applied the sum of features $\ell$ for a comparison. This is due to the fact that in certain cases features appear with the same high score. Therefore, also it was necessary to introduce another two parameters namely convergence point $\chi$ and prefix converge point $\gamma$. The sequence of letters for each language is important at this stage. It is noticed that the most likely language of one document have faster (or smaller) convergence than other languages in calculating the sum of features $\ell$. The formulation of $\chi$ and $\gamma$ is shown in Equation (3.9) and (3.10), respectively, where $\zeta_{cur}$ is the position of the current feature, $\ell_{new}$ is the total of the current feature frequency, $\ell_{old}$ is the total of a previous feature frequency and $\zeta_{pre}$ is the position of previous $\zeta_{cur}$ with frequency changed.

$$
\chi = \begin{cases} 
\zeta_{cur}, & \text{if } \ell_{new} > \ell_{old} \\
\text{continue} 
\end{cases}
$$

(3.9)

$$
\gamma = \begin{cases} 
\zeta_{pre}, & \text{if } \ell_{new} > \ell_{old} \\
\text{continue} 
\end{cases}
$$

(3.10)
Figure 3.14 illustrates the concepts of convergence points and prefix convergence points. An unknown document is evaluated from the point of view of each language, such as Arabic and Persian. Each language is organized as a group of features in order to count the sum of the features $\ell$ as shown in this figure. The index is the position of a feature according to the sequence letter weighting, while the code point is the decimal Unicode value. The letter is the actual character, while frequency is the occurrence of selected letters in a document and the total is the accumulated frequency of letters. Finally, the sum of features $\ell$ is obtained. The converge point and prefix converge point of Arabic are four and three while for Persian they are three and two. Figure 3.14 presents some of the highest frequencies found for Arabic and Persian. This will cause the sum of features $\ell$ to fail to identify the language. Therefore, the second condition has been checked by finding the lowest converge point. If the converge point is also same, then a third condition has been used by finding the lowest prefix convergence point. In other words, the proposed letter weighting feature selection has three conditions with respect to priorities. They are from the sum of features $\ell$, the convergence point $\chi$ and prefix convergence point $\gamma$. If any one of the conditions is satisfied, the language of document will be identified. The decision tree is derived as Figure 3.15. Notice that if none of the conditions are satisfied, the process is continued to next language model of same script.
3.8 Summary

There are four feature selection methods that have been applied in this work, namely; $N$-grams, entropy, letter weighting and simplified entropy. The steps involved in their employment and also the details of the algorithm have been discussed. The following chapter analyzes the experimental results of the proposed methods.
CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

In this chapter, the results of analysis will be described based on the experiments done on web page language identification. Firstly, the analysis of letter frequency is shown to justify the use of letter statistics as a feature. This is followed by a T-test analysis at the preprocessing step in order to differentiate the impact on the original data set and preprocessed data set. Next, the identification performance of the four methods used is compared. Then, the retrieval performance in terms of F1 measurements and noise tolerance performance on particular methods are discussed. Finally, the experimental results of letter constraints on simplified entropy method are highlighted.

4.2 Experimental Setting

This research has used the data sets of Arabic, Cyrillic, Hanzi, Indic and Roman script, as shown in Table 4.1. Each script is used by of several languages. For example, the Arabic script is used for Arabic, Persian, Urdu, Pashto; Cyrillic for
Russian, Macedonian and Uzbek; Hanzi for Traditional Chinese and Japanese; Indic for Hindi and Nepali; and the Roman script for Azeri, Indonesia, English, Serbian, Somali, Spanish, Turkish and Vietnamese. Each language comprises 1000 unit web pages. It has been collected the web pages from CNN and BBC news websites (Thompson, 2008) and assumed that a single web page is written in one language only. Then, those collected data are fed into preprocessing to filter out the noises. After that, feature selection methods such as entropy and letter weighting represent the data into appropriate form. Finally, language identification methods determine the language of web page. It is illustrated in the Figure 4.1. Standard encoding scheme Unicode (or UTF-8) is used for all experiments, so all the character sets of a web page are converted to UTF-8. The experiments were evaluated based on five cross validations. The total documents for each language is 1000 units, so if the number of testing documents is 200 units then the training document is 800 units, vice versa.

Table 4.2 shows the UTF-8 boundary for a particular script in decimal representation (or in bracket is hexadecimal). The script begin \((S_{begin})\) and end \((S_{end})\)
Table 4.2: The Unicode boundary in decimals

<table>
<thead>
<tr>
<th>Script</th>
<th>Unicode boundary in decimal (hexadecimal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>1536 – 1791 (U+0600 to U+06FF)</td>
</tr>
<tr>
<td>Cyrillic</td>
<td>1024 – 1280 (U+0400 to U+04FF)</td>
</tr>
<tr>
<td>Hanzi</td>
<td>19968 – 40959 (U+4E00 to U+9FFF)</td>
</tr>
<tr>
<td>Indic</td>
<td>2304 – 2435 (U+0900 to U+097F)</td>
</tr>
<tr>
<td>Roman</td>
<td>65 – 591 (U+0000 to U+024F)</td>
</tr>
</tbody>
</table>

of Arabic script is between 1536 and 1791; Cyrillic between 1024 and 1279; Hanzi between 19968 and 40959; Indic between 2304 and 2435 and the Roman between 65 and 591, respectively.

4.3 Results of Corpora Preparation

In this Section, two experiment results are presented: letter frequency justification and T-test analysis of the preprocessing step. The identification results are highly affected by the corpora has been used due to the bias included. Therefore, two experiments have been carried out in order to evaluate the data set has been collected.
The first experiment is to analyze the nature of letter distribution in each document and also the whole data set, while the second experiment is to justify the need of preprocessing step when utilizing the web page corpora.

4.3.1 Letter Frequency Justification

The objective of this experiment has been to explain why using letter frequency as the fundamental of analysis is advantageous. Figure 4.2a, 4.2b and 4.2c show the proportional results of letter distribution averages on 100 web pages including Hanzi, Indic and Cyrillic script web pages. It is based on the letter distribution of 100 web pages for each language. It is noticed that some letter distribution is significantly different, but others are not. For example, the Chinese and Japanese having similar frequencies for the character 年 (pronouns “NIAN” in Hanzi script) that is 223 and 244 units, respectively. However, the frequency of 有 (pronouns as “YOU”), 方 (pronouns as “FANG”) and 他 (pronouns as “TA”) in Chinese is 393, 162 and 480 units, respectively; in Japanese is 23, 99 and 3 units, respectively. Therefore, the variant of letter distribution is a significant attribute that can be used to discriminate languages (Windisch and Csík, 2005). Figure E.1 in Appendix E shows the letter distribution for 100 Arabic script web pages.

Another purpose of letter frequency distribution is to verify the content of the web pages have been collected. In other words, it is to check whether a particular content is spread throughout a given boundary of a particular script. For example, the Arabic web page should consist of Arabic script letters and Roman script letters (HTML codes) only, otherwise this web page is deselected as a data set. From the experiment observations, there are around 200 Uzbek web pages comprise Arabic script letters instead of Cyrillic script letters. Those web pages are discarded from the data set and replaced with correct Uzbek web pages. It is a critical corpora preparation stage in order to obtain the valid data set for experiments. As a result, the correct evaluation can be experimented on the feature selection methods of web page language identification.
Figure 4.2: Proportional results of letter distribution averages for 100 web pages in Chinese, Japanese, Hindi, Nepali, Russian, Uzbek and Ukrainian

4.3.2 T-test Analysis of the Preprocessing Step

The second experiment is to evaluate the impact of the preprocessing step in the web page language identification. In this work, the proposed methods are based on the letter frequency. Therefore, the letters of a document being used for feature selection affect the performance of the identification. From the experiment, it has been proven that data preprocessing is important for the original documents. The experiment was
done on Arabic script web pages that consist of 1000 web pages for each language. The languages included Arabic, Persian, Urdu and Pashto. The BPNN was used to justify the performance by using Root Mean Squared Error (RMSE) on a T-test analysis Lee (2008).

Firstly, the web pages are collected using crawler, as stated in the Section 3. The data are divided into two groups, preprocessed and original. The preprocessed group implements the data preprocessing on the original data. The programming codes

![Example of the web page before and after preprocessing](image)
in the original document are removed\(^1\). Then, those code free documents are further filtering in order to remove letters out of a range between 1536 and 1792\(^2\). Figure 4.3 offers an example of a web page before and after preprocessing. The original group retains the original web pages. Table 4.3 shows the difference between the preprocessed and original group data. Both groups are using same Arabic script data set, 1000 unit documents, the letter weighting feature selection method and the identifier BPNN.

<table>
<thead>
<tr>
<th>Description</th>
<th>Preprocessed Group</th>
<th>Original Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>Arabic Script</td>
<td>Arabic Script</td>
</tr>
<tr>
<td>Number of documents</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>Feature Selection</td>
<td>Letter Weighting</td>
<td>Letter Weighting</td>
</tr>
<tr>
<td>Identifier</td>
<td>BPNN</td>
<td>BPNN</td>
</tr>
<tr>
<td>Data Preprocessing</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

In order to do the analysis on T-test, RMSE of BPNN have been applied. Table 4.4 shows the parameters setting on the BPNN. The input node, hidden node and output node are 20, 8 and 1; the learning rate is 0.009, momentum rate is 0.0001, epochs is 1000, minimum RMSE is 0.01, features are normalized at between -1 and 1 and the output is normalized at between 0 and 1, respectively.

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function</td>
<td>Logistic</td>
</tr>
<tr>
<td>Input Node</td>
<td>20</td>
</tr>
<tr>
<td>Hidden Node</td>
<td>8</td>
</tr>
<tr>
<td>Output Node</td>
<td>1</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>0.009</td>
</tr>
<tr>
<td>Momentum Rate</td>
<td>0.0001</td>
</tr>
<tr>
<td>Epochs</td>
<td>1000</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.01</td>
</tr>
<tr>
<td>Features Normalized</td>
<td>-1 to 1</td>
</tr>
<tr>
<td>Output Normalized</td>
<td>0 to 1</td>
</tr>
</tbody>
</table>

\(^1\)The original document are parsing with standard java function (version 1.4.2) namely HTMLEditorKit.ParserCallback for removing the HTML code.

\(^2\)This is the decimal code point boundary of Arabic script letters in Unicode.
As noted that the T-Test analysis comprises the following hypotheses $H_0$ and $H_1$.

$H_0$: Population means are the same, $\mu_1 = \mu_2$.

$H_1$: Population means are not the same, $\mu_1 \neq \mu_2$.

The results of RMSE of BPNN simulation have been obtained based on a 5-fold cross validation, as shown in Figure 4.4. The RMSE of preprocessed group are $0.0384, 0.0387, 0.0348, 0.0390, 0.0365$ and until the last one is $0.0366$. The RMSE of the original group are $0.1053, 0.1017, 0.0971, 0.1127, 0.0950$ and until the last one is $0.1064$.

Based on the information collected for Figure 4.4, it is further validated the critical regions of the T-test analysis. Table 4.5 shows the analysis of statistical hypothesis test or T-test. It is noticed that the $\bar{x}$ of the preprocessed group and the original group are $0.0372$ and $0.1025$, respectively. The $\bar{x}$ of the preprocessed group is lower than that of original group. The standard deviation ($S$) of the preprocessed group
and the original group are 0.0011 and 0.0077, respectively. The number of observations 
(n) is 50, degree of freedom (DoF) is 98, the different in mean (\(\bar{x}_1 - \bar{x}_2\)) is -0.0653, the
different in standard deviation (\(S_{\bar{x}_1 - \bar{x}_2}\)) is 0.0133, so the critical value of t is -4.9098.

Table 4.5: Analysis of the statistical hypothesis test

<table>
<thead>
<tr>
<th>Description</th>
<th>Preprocessed Group</th>
<th>Original group</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\bar{x})</td>
<td>0.0372</td>
<td>0.1025</td>
</tr>
<tr>
<td>(S)</td>
<td>0.0011</td>
<td>0.0077</td>
</tr>
<tr>
<td>(n)</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>DoF</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>(\bar{x}_1 - \bar{x}_2)</td>
<td>-0.0653</td>
<td></td>
</tr>
<tr>
<td>(S_{\bar{x}_1 - \bar{x}_2})</td>
<td>0.0133</td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>-4.9098</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.5: Critical regions of T-test
Based on Table C.1 in Appendix C, it is observed that the critical region of the hypothesis is between -1.68 and 1.68, where the number of observations is 50, the significance level is 0.05 and the Confidence Interval (CI) is 95%. Through experiment, the t value (-4.9098) falls outside of the range of the critical region, as shown in Figure 4.5. This shows that the difference is statistically significant. Hence the null hypothesis \( H_0 \) is rejected. Therefore, it may be concluded that the preprocessed group with data preprocessing is significantly different from the original group of original documents.

### 4.4 Results of Feature Selection Methods Evaluation

Four feature selection methods are evaluated based on the data set that has been identified. There are three experiments that have been done on evaluating the feature selection methods. The first experiment is retrieval performance that each method is evaluated based on precision, recall and \( F_1 \) measurement. The second experiment is identification accuracy that four feature selection methods are compared based on the five cross-validation accuracy. The last experiment is to analyze the scalability of simplified entropy when increasing the number of features used.

#### 4.4.1 Retrieval Performance

The objective of this experiment has been to measure the retrieval performance of the proposed methods for each script web page language identification. The measurements used in this experiment are precision, recall and \( F_1 \) measurement. Precision describes the probability that a retrieved language’s web page (randomly selected) is relevant to a certain language. Recall describes the probability of a relevant language’s web page being retrieved. \( F_1 \) measurement is the average between precision and recall. In this experiment, 500 Indic script web documents were used for training, and another 500 documents for testing. There are 19 languages involved in this work.
which are Arabic, Pashto, Persian, Urdu, traditional Chinese, Japanese, Hindi, Nepali, Macedonian, Russian, Uzbek, Azeri, English, Indonesia, Serbian, Somali, Spanish, Turkish and Vietnamese. The formula of precision, recall and $F_1$ measurement is given by Equation 2.32, 2.33 and 2.34 in Section 2.9.

![Figure 4.6: Retrieval performance of $N$-grams](image)

Figure 4.6 illustrates the retrieval performance of $N$-grams feature selection method. $N$-grams failed to classify the languages like Pashto, Persian, Macedonian, Russian, Uzbek, Chinese, Spanish and Vietnamese that the precision is lower than 27%. In addition, the recall of Urdu, Macedonian, Russian, Uzbek, Chinese, Japanese, Azeri and English are 35.16%, 51.93%, 73.98%, 61.03%, 0%, 50%, 51.07% and 59.52%, respectively. The results have shown that $N$-grams is having feature dimension problem. This problem is found significantly on Chinese language that Hanzi script consists of more than 20000 characters and it will cause the bias of distance calculation in argument minimum. However, $N$-grams are useful on small size of script like Indic in which the precision, recall and $F_1$ of language identification are above 99%. The dimension issue of $N$-grams is related to irrelevant features that most of the features produced by $N$-grams are meaningless. It can be improved by including certain threshold to filter the irrelevant features or integrated with a particular ranking algorithm for discarding the useless features.
Figure 4.7 shows the precision, recall and $F_1$ of entropy feature selection method. All the results found are above 93% and it is out of expectation. Originally, entropy is a method on selecting the important keywords among the data set. In this work, the keywords are set to the particular characters only, not in words. Then, those characters above the threshold are selected as features and fed into machine learning method for language identification. Although, the idea of entropy feature selection is mainly used on text categorization, it is proved that this method is also workable in language identification.

Figure 4.8 depicts the retrieval performance of letter weighting feature selection method. The precision, recall and $F_1$ of Hindi and Nepali are 71.92%, 86.60%, 78.42%, 82.94%, 65.40% and 72.81%, respectively. For the others, the retrieval performances are above 90%. The low performance of Indic script is caused by the design of letter weighting algorithm. The letter weighting is an enhancement of $N$-grams that including the threshold and ranking algorithm. It is takes into consideration both internal and external weighting. External weighting is the letter weighting of other languages of same script. The penalty is given to those characters occurred in two or more languages. For Indic script, the total characters used are 131 units only and the threshold is set to 5 units in which only 10 characters are selected as features. It is noticed that the
penalty given has reduced the choice of selection. As a result, those characters of low level weighting are used as features and it will produce low performance of language identification due to those features occurred seldom in Indic text. This idea is applicable on large number units of script like Hanzi and many languages script like Roman but not Indic script.

Figure 4.9 shows the retrieval performance of simplified entropy feature selection method. It is suprisingly that all the results given are 100% except for Roman script languages. The precisions of Azeri, English, Indonesia, Serbian, Somali, Spanish, Turkish and Vietnamese are 100.00%, 82.80 %, 85.20%, 100.00%, 91.00%, 99.80%, 97.20% and 100.00%; the recalls are 97.28%, 84.15%, 90.83%, 100.00%, 84.26%, 100.00%, 100.00% and 100.00%; the $F_1$s are 98.62%, 83.47%, 87.93%, 100.00%, 87.50%, 99.90%, 98.58% and 100.00%, respectively. Simplified entropy is based on the letter frequency convergence, the fastest is the winner. Simplified entropy is an enhancement of $N$-grams and entropy. Initially, the entropy algorithm is calculates the letter frequency inverse document frequency. Then, this algorithm is simplified to letter frequency document frequency namely simplified entropy. Simplified entropy is assumes that a good feature should be sorted on the top ranking with high letter frequency and document frequency. With that, those relevant features of a particular
language are having higher priority in calculation of language identification. After that, simplified entropy is adopting the idea of \( N \)-grams by producing list of features for each language. It is used for calculating letter frequency convergence but not the distance difference between trained language model and to be predicted web page. The results have shown that simplified entropy is applicable on small and large dimension of script characters.

Figure 4.10 illustrates the average retrieval performance of feature selection methods. Precisions of \( N \)-grams, entropy, letter weighting and simplified entropy are 60.09\%, 98.83\%, 95.54\% and 97.68\%, respectively. Recalls of \( N \)-grams, entropy, letter weighting and simplified entropy are 76.54\%, 98.74\%, 95.35\% and 97.71\%, respectively. \( F1 \) of \( N \)-grams, entropy, letter weighting and simplified entropy are 55.23\%, 98.74\%, 95.30\% and 97.68\%, respectively. From these results, \( N \)-grams are considered as a weak feature selection method of web page language identification due to its dimension issue. In the following Section 4.4.2, the identification accuracy is evaluated on these four feature selection methods. It is based on five cross validation that each portion of data set is evaluated in order to reduce the bias of result.
4.4.2 Comparison of Identification Performances

The fourth experiment, which is the identification performance comparison on the proposed methods, has in part been discussed, such as entropy, letter weighting, \textit{N}-grams and simplified entropy. As stated in the discussion of the previous experiment, the identification performance might be different if more languages are selected. For this reason, the data has been expanded from Indic script documents to those in Arabic, Cyrillic, Hanzi and Roman script web pages. The total web pages used are 1000 units for each language. The details of the data set can be viewed in Table 4.1. It is also applied 5-fold cross validation in this experiment. This is a statistical practice of partitioning a sample of data into subsets such that the analysis is initially performed on a single subset, while the other subset(s) are retained for subsequent use in the validation of the initial analysis. This is to say, the testing results are produced from the whole data set instead of just a portion.

Figure 4.11 shows the identification performance of the proposed method in web page language identification. There are five scripts that have been selected for this experiment. They are; Arabic, Roman, Cyrillic, Hanzi and Indic. The accuracy
of $N$-grams method on Arabic, Roman, Cyrillic, Hanzi and Indic is 66.62%, 88.19%, 84.67%, 70.00% and 97.25%; for entropy, the accuracy is 91.52%, 98.94%, 90.60%, 99.35% and 100%; for letter weighting, the accuracy is 99.25%, 98.76%, 99.63%, 96.35% and 71.80%; for simplified entropy, the accuracy is 100%, 96.75%, 98.23%, 100% and 99.50%, respectively. Table 4.6 shows the average accuracy of the entropy, letter weighting and simplified entropy schemes on web page language identification based on Figure 4.11. Overall, the simplified entropy scheme is higher than the $N$-grams, entropy and letter weighting schemes, with an accuracy of 98.90%, 81.35%, 96.08% and 93.16%. Therefore, it can be concluded that the simplified entropy scheme performs significantly better than the $N$-grams, entropy and letter weighting schemes.

Table 4.6: Average accuracy of identification according to Figure 4.11

<table>
<thead>
<tr>
<th>Methods</th>
<th>$N$-grams</th>
<th>Entropy</th>
<th>Letter Weighting</th>
<th>Simplified Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average (%)</td>
<td>81.35</td>
<td>96.08</td>
<td>93.16</td>
<td>98.90</td>
</tr>
</tbody>
</table>
4.4.3 Letter Frequency Constraints on Simplified Entropy

Figure 4.12: Impact of letter frequency constraints on simplified entropy

Figure 4.12 shows the effect of features size on web page language identification performance using simplified entropy feature selection. The lowest and highest accuracy for language identification are 42.86% and 98.21% on 5 and 500 letters, respectively. As noted, the performance of language identification increases as the number of letters increases which is a linear trendline. Therefore, the number of letters in a document will significantly affect the accuracy of language identification. In addition, it was found that the accuracy of language identification is above 90% when the number of letters used is 150 units or 300 bytes, which is similar to the best performance reported by Cavnar and Trenkle (1994). However, this method may not be suitable for use on a too short text such as a single sentence. It requires at least 150 units for accurate calculation, therefore improvement is still needed to handle this shortcoming.
4.5 Discussion

There are a great number of languages written with many different scripts. When a document is acquired by a text processing application, its metadata often includes the statement that it is written in certain encoding, such as UTF-8, but not the language in which the document is written. This situation gives rise to the problem of identifying the language of a document from its contents.

4.5.1 Character Versus Word

A document’s content includes primarily the words in it and the characters that make up the words. Analyzing for words is not feasible because it is difficult to adopt the appropriate corresponding word library to a document in an unknown language. Such an approach also has difficulties in coping with inflected words, grammatical variations, or spelling errors (Grefenstette, 1995; Xafopoulos et al., 2004). Furthermore, the linguistic style will change according to the cultural traditions of a particular community especially on the WWW. For example, there are blogs using distinctive short forms to represent a particular word, such as when “you” is written as “u”. According to MacNamara et al. (1998), the inclusion of features that are peculiar to the linguistic characteristics of different languages did not increase performance substantially. Therefore, the most easily obtained content data from a web page are the characters used. Experiment on Section 4.3.1 shows the usage of characters in 100 web pages and it is noticed that character is suitable to be used as features of web page language identification.

Takcı and Soğukpınar (2005) has proven that the letter based text scoring method is easier and faster than short terms and $N$-gram methods and have also shown that accumulated score and position of the feature are factors that significantly affect the performance of identification (Li and Momoi, 2001; MacNamara et al., 1998). Moreover, Selamat and Ibu Subroto (2007) have also argued that the usage of characters as feature sets can reduce the dimension of the feature set as well as increase
the time consumed in web page language identification. However, it is commonly known that some languages such as Japanese have their own unique characters, those characters are not recommended to be used, as it might give false results derived from quotes (Windisch and Csink, 2005). Therefore, only the standard alphabets that exist in more than one language are applied in the web page language identification.

4.5.2 The Impact of Dataset

The domain of the data set in this work is another issue of concern. All the proposed methods are tested on general news websites such as international, local, entertainment and economy sites. The proposed methods might not function properly in other web domains such as those dealing with biology or chemistry, most of which contain specific characters or terminological terms. An effective and efficient language identification system should use light-weight features that can accrue significant statistics within a short span of time and text. Using letter features is an obvious choice given these desired results. The letter is the fundamental component of words and sentences, and so by using the code point of a letter, each word or sentence can be segmented to a particular language script of UTF-8. This will indirectly increase the accuracy of language identification. If the method used is based on words, some approach involving tokenization, the stemming or stopping process is required. This is not practical, since the preprocessing stage has not yet identifies the language and different languages have different morphological rules. So, the errors in language identification becomes higher (Giguet, 1996). Another problem, based on the vocabulary of each language, is how to cope with inflected words, that is, grammatical variations, as well as spelling errors (Grefenstette, 1995; Xafopoulos et al., 2004). Experiment on Section 4.3.2 presented a T-test analysis on two categories of data set: original web page and preprocessed web page. From this experiment, it is proved that preprocessed data is significant different from original data. In other words, preprocessed data is gives better result than original data.
4.5.3 Noise Tolerance

An algorithm in computer science is robust if it continues to operate despite noises in input. In this work, the measurements such as precision, recall and \( F1 \) measure are used to justify the robustness of particular method. In overall, entropy and simplified entropy perform more robust than others in web page language identification. In other words, these methods perform more reliable against the noises in the web page. However, this situation might be different if other domain data set is used.

Table D.1 in Appendix D shows the different features that have been found during the experiments on Roman script language identification. As observed from the experiment, the significant difference is that simplified entropy consists of more features than other feature selection methods. The particular language in the simplified entropy scheme contains particular letters to represent these features. These letters are arranged in descending order, from the most frequent feature to the lowest frequent. Based on such an order, the algorithm is able to identify the desired language with the fastest letter frequency convergence. This is one of the strength of simplified entropy that can produce such robust features compared to other feature selection methods.

4.5.4 Accuracy of Language Identification Methods

Effectiveness of methods is terms of accuracy. Accuracy is refers to how good a method is on prediction the desired language. In the experiment of identification performance, five different domain data sets including Arabic, Hanzi, Indic, Cyrillic and Roman scripts data are used for evaluation. It is purposely to test the effectiveness of the method against data diversity. From the experiment, it is noticed that \( N \)-grams have shortcoming if the particular script of language consists of many characters, for example Hanzi. \( N \)-grams depend on the closest distance of particular language model that cause almost one particular language always appears as predicted language. For
example, Japanese has 1000 N-grams but Chinese has 10000 N-grams, so Japanese will appear as winner in most of the time.

Another feature selection method is entropy. Entropy is used on text categorization and it performs very well on that field. When this method has been chosen and tried on language identification, the problems have been observed. It is used to find the most important key words on data mining. The particular key word found is meaningless to language identification due to the fact that such key word may appear in more than one language especially applied on character not word. Therefore, feature representation of entropy remains an unsolved problem in web page language identification.

The first proposed method, letter weighting, is an improvement of entropy method. Instead of find out the most important key word by entropy, letter weighting take into consideration of other languages weighting. For example, if both Arabic and Persian having similar global weighting on particular character, then that character is penalty from being chosen as the features. As a result, those characters considered as noises are filtered out from feature representation. However, a problem of letter weighting is occured on Indic script web pages. Many of the Indic script characters appear similar global weighting, so the number of choices as feature is limited. Therefore, the accuracy of identification is affected. It is noticed that the hybridization between letter weighting and N-grams can be a solution for this problem because N-grams produce many combination of characters.

The second proposed method is simplified entropy. Simplified entropy includes both letter frequency and feature position into the algorithm design. As discussed previously, the highest appearance letter frequency of particular language in a web page is assumed as winner. Based on this assumption, the first condition of decision tree is to determine the letter frequency. If sum of features appears the same in more than one language, then feature position will be used as second and third condition. This idea is derived from N-grams as the closest distance of particular language is considered as winner. Therefore, it is assumed that the desired language of a web page is having faster convergence of feature position than others. It is proven on the experiment of identification performance.
4.5.5 Shortcoming of the Proposed Method

Although simplified entropy performs the best compared to other feature selection methods, some problems of simplified entropy have been revealed. First problem is the size of a web page, simplified entropy only applicable to those web pages that has 200 or above letters. It is caused by the design of the algorithm that fundamentally based on letter frequency. Therefore, simplified entropy cannot be applied on too short web page or multilingual web page. Another problem is the predefined conditions of the decision tree. If one web page failed to match any conditions of all language models, then this algorithm cannot identify language of this web page. Therefore, improvement of this method is still needed. It is suggested to use the simplified entropy with \(N\)-grams and then fed into machine learning method like ANN for identification, the result might be outperformed on multilingual web page as well.

4.6 Summary

Overall, this research noticed that simplified entropy performs better than other feature selection methods. This is a significant improvement since web page language identification can be done with letter representation only. The number of letters for all languages is always much smaller than the number of words or \(N\)-grams, therefore the speed of identification can be improved. This work is concluded in the following chapter.
5.1 Introduction

Language identification is a core technology in various multilingual applications, especially in web services. This is a very challenging task due to the ever increasing number of internet users. There are a tremendous number of web pages emerging everyday. The problem is further aggravated by the complexity and diversity of web applications and services. In this study, improved feature selection methods have been proposed for web page language identification that has been determined to be accurate and effective on the experiments. Four methods are studied to achieve the desired results.

The first method is the conventional method of \(N\)-grams which comprises four steps; data preparation, data preprocessing, feature selection using \(N\)-grams and identification using argument maximum. \(N\)-grams utilize the position of a particular feature in argument maximum. However, the problem with \(N\)-grams is the dimension feature on various languages. Certain languages have a large number of features and thus cause that language to be misclassified. The second method is entropy that used in text categorization. It consists of four steps; data preparation, data preprocessing, feature selection using entropy and identification using fuzzy ARTMAP. It has
been discovered that the algorithm extracts the most important features instead of most frequent, and so is having issue of feature representation. The third method is an improvement of the second method, and comprises four steps: data preparation, data preprocessing, feature selection using letter weighting and identification using fuzzy ARTMAP. The letter weighting algorithm is limited in selecting appropriate features, where the number of letters in a certain script is limited with respect to number of possible selected as features. For this reason it may produce false language identification results. The last method is the enhancement of first and second method, in which the method takes into consideration both letter frequency and feature position. It consists also of four steps; data preparation, data preprocessing, feature selection using simplified entropy and identification using decision tree. This method is able to produce the most promising results in the process of language identification, as it was performed on thirteen languages, and is found in this work to be the best approach.

5.2 Research Findings

Table 5.1 shows the objective of this work and the corresponding outcome. First objective of this work is to review previous research related to web page language identification, second objective is to propose an improved feature selection method for web page language identification and third objective of this work is to test the performance of the proposed method on web page language identification.

The first objective of this work is to review previous research related to web page language identification. This objective is achieved on Chapter 2, literature review. First is the study of the Internet, followed by overview of language identification, problems of web page language identification, conventional web pages language identification process, feature selection method review, language identification method review and lastly is the evaluation method review. Initially, the study is conducted from the top view of the internet and the evolution of computer network. Then, it is followed by the description of the web pages, web browser and text encoding / character set of web pages. Finally, the problem of language identification is visualized at the bottom of the internet.
Table 5.1: Objective versus outcome

<table>
<thead>
<tr>
<th>No.</th>
<th>Objective</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>To review previous research related to web page language identification.</td>
<td>The study of the Internet, followed by overview of language identification, problems of web page, conventional web pages language identification process, feature selection method review, language identification method review and lastly is the evaluation method review.</td>
</tr>
<tr>
<td>3</td>
<td>To test the performance of the proposed method on web page language identification.</td>
<td>The experiments have been carried out are letter frequency justification, T-test analysis of the preprocessing step, retrieval performance, identification performance and letter constraints on simplified entropy. In overall, the simplified entropy scheme is higher than N-grams, entropy and letter weighting scheme, with accuracies of 98.90%, 81.35%, 96.08% and 93.16%, respectively.</td>
</tr>
</tbody>
</table>

In the overview of language identification, different issues of web page language identification are discussed including the importance of language identification, the applications related to web page language identification, minority language identification, multilingual identification, supervised/unsupervised identification and feature processing. Moreover, the problems of web page are further depicted such as web page format, grammatical/morphological error, encoding issue and tremendous abbreviations.

In the conventional web page language identification process, the preprocessing step, representation step and induction step are illustrated in order to provide an insight of web page language identification process. For feature selection method review, the general feature selection methods like wrapper, filter, statistical and linguistic are described. The statistical methods have been reported including entropy, PCA, N-grams approach and windowing algorithm. On the other hand, the linguistic methods
including small word technique, Unicode based technique, web page information and HMMs. In identification method review, the related language identification methods such as ANN, fuzzy ARTMAP, SVM, decision trees, VQ and KNN are further discussed. At the end of literature review, the evaluation approaches related to language identification are reviewed such as T-test, precision, recall, $F_1$ measure, cross validation and accuracy.

The second objective of this work is to propose an improved feature selection method for web page language identification. This objective is achieved as two improved methods namely letter weighting and simplified entropy are proposed to improve the performance of web page language identification. Initially, two conventional methods have been applied on web page language identification that is $N$-grams and entropy. $N$-grams is basically a well known approach in text categorization and language identification, while entropy is originally used on text categorization. $N$-grams have the problem on utilizing features been generated. The smallest dimension features of particular language mostly selected as winner that leads to high false positive. Entropy is capable in selecting the most important key words among the categories given on text mining. However, the problem occurred in language identification since that important key words might exist in two or more languages. Therefore, these two conventional methods have been improved to the proposed methods namely letter weighting and simplified entropy. Letter weighting take into consideration both local and global weighting when selecting the suitable features of web page language identification. Simplified entropy makes use of letter weighting and simplified entropy in web page language identification. Both proposed methods have slightly improved the conventional methods of web page language identification.

The third objective of this work is to test the performance of the proposed method on web page language identification. This objective is achieved in which five experiments have been carried out. The experiments are letter frequency justification, T-test analysis of the preprocessing step, retrieval performance, identification performance and letter constraints on simplified entropy. These experiments are reflected to the issues of language identification such as character versus word, the impact of dataset, robustness of feature selection methods, effectiveness of language identification methods and shortcoming of the proposed methods. In overall, the simplified entropy scheme is higher than $N$-grams, entropy and letter weighting scheme, with accuracies of 98.90%, 81.35%, 96.08% and 93.16%, respectively.
5.3 Thesis Contributions

Below are the deliverables that have been produced in the course of this research and it is intended to make available to the research community.

(i) Improvement of $N$-grams and entropy feature selection to two feature selection methods; letter weighting feature selection and simplified entropy feature selection.

(ii) Demonstration of the importance of preprocessing in web page language identification.

(iii) Clarification the actual performance of various identification methods for web page language identification.

5.4 Future Work

Language identification is an important technology in the Internet environment. Compared to identification methods, which are already known to work well, feature selection method needs to establish whether the proposed choice of features is appropriate. The issue of computational cost and scalability limitations is not addressed in this work. Many of the applications to be used require low computational expense in terms of time, speed and cost. Therefore, it is necessary to do analyses on such requirements in order to insure user satisfaction. There are thousands of languages that have been reported in this world, so how good of the proposed method is with respect noise tolerance when expanding to world languages remains a critical issue. Restriction to limited number of features may cause problems in a practical application. This work can also be extended to mixed-languages identification and problem of simultaneous detection of language and character sets.
Recent years have seen extensive efforts in feature selection research. The field of feature selection expands both in depth and in breadth, due to increasing demands for dimensionality reduction. The research expands from classic supervised feature selection to unsupervised and semi-supervised feature selection in order to meet the increasing demands of labeled, unlabeled, and partially labeled data. It is just one perspective of feature selection that encompasses many aspects. However, from this perspective, it can be clearly seen that as data evolve, feature selection research adapts and develops into new areas in various forms for emerging real-world applications especially in natural language processing.

5.5 Summary

The proposed method comes with some guarantees of optimality under certain limited conditions. Initially, the conventional $N$-grams and entropy feature selection have been applied on web page language identification. It is noticed that it is not suitable in language identification due to the required algorithm having dimensional and feature selectivity problems. Then too, these methods are expanded into two feature selection methods, namely letter weighting and simplified entropy. The letter weighting scheme is able to produce stable identification for Arabic, Roman, Cyrillic and Hanzi script but not Indic. This may be due to the features that have been identified are not sufficiently representative. The proposed letter weighting scheme in combination with fuzzy ARTMAP have trouble in discriminating two similar languages such as Hindi and Nepali. The experiments find, however, that this method is able to determine the ratio between active language weighting and passive language weighting. Certain cases appear with fuzzy global or local language weighting have also been selected as features. Therefore, the performance of language identification is affected. In overall, the simplified entropy scheme is higher than $N$-grams, entropy and letter weighting scheme, with accuracies of 98.90%, 81.35%, 96.08% and 93.16%, respectively. Therefore, the simplified entropy scheme performs better than other feature selection methods. In conclusion it can state that web page language identification gives promising results at the finer granularity level (or letter representation) of web documents.
REFERENCES


Character set detection is the process of determining the character set, or encoding, of character data in an unknown format. This is, at best, an imprecise operation using statistics and heuristics. Because of this, detection works best if one supplies at least a few hundred bytes of character data most of which is in a single language. In some cases, the language can be determined along with the encoding (IBM and Others, 2008). Several different techniques are used for character set detection. For multi-byte encodings, the sequence of bytes is checked for legal patterns. The detected characters are also checked against a list of the frequently used characters in that encoding. For single byte encodings, the data is checked against a list of the most commonly occurring three letter groups for each language that can be written using that encoding. The input data can either be a Java input stream, or an array of bytes. The output of the detection process is a list of possible character sets, with the most likely one first. Table A.1 shows all the encodings that can be detected (IBM and Others, 2008).

Table A.1: Detectable character sets provided by International Components for Unicode (ICU) (IBM and Others, 2008)

<table>
<thead>
<tr>
<th>Character Set</th>
<th>Specific Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTF-8</td>
<td>(included UTF-16BE / LE and UTF-32BE / LE are used for all languages)</td>
</tr>
<tr>
<td>Shift JIS</td>
<td>Japanese</td>
</tr>
<tr>
<td>ISO-2022-CN</td>
<td>Simplified Chinese</td>
</tr>
<tr>
<td>GB18030</td>
<td>-</td>
</tr>
<tr>
<td>Big5</td>
<td>Traditional Chinese</td>
</tr>
<tr>
<td>EUC-JP</td>
<td>Japanese</td>
</tr>
<tr>
<td>EUC-KR</td>
<td>Korean</td>
</tr>
<tr>
<td>ISO-8859-1</td>
<td>Danish, Dutch, English, French, German, Italian, Norwegian, Portuguese, Swedish</td>
</tr>
<tr>
<td>ISO-8859-2</td>
<td>Czech, Hungarian, Polish, Romanian</td>
</tr>
<tr>
<td>ISO-8859-5</td>
<td>Russian</td>
</tr>
<tr>
<td>ISO-8859-6</td>
<td>Arabic</td>
</tr>
<tr>
<td>ISO-8859-7</td>
<td>Greek</td>
</tr>
<tr>
<td>ISO-8859-8</td>
<td>Hebrew</td>
</tr>
<tr>
<td>windows-1251</td>
<td>Russian</td>
</tr>
<tr>
<td>windows-1256</td>
<td>Arabic</td>
</tr>
<tr>
<td>KOI8-R</td>
<td>Russian</td>
</tr>
<tr>
<td>ISO-8859-9</td>
<td>Turkish</td>
</tr>
</tbody>
</table>
Table B.1\(^1\) shows an attribute comparison of previous work on web page language identification. It is divided into the categories of linguistic (L) and statistic (S). It is a summary of feature selection methods and language identification methods that have been discussed in Section 2.7 and 2.8.

Statistical language identification takes sample documents from each language and subjects them to a statistical analysis based on some characteristic of the text. This is often \(N\)-gram character analysis or a counting of short words. Grefenstette (1995) compares the use of small words and tri-grams. Tri-grams work better on short sentences, since as the length of the text grows larger, they both perform about the same. To determine the language of a document, the same statistical procedure is done to the new document, and the document statistics are compared to those derived from the language trained documents to determine the closest match.

Linguistic methods use certain aspects of the language to identify it. These approaches are usually dictionary-based, with a set of words in each language. As large dictionaries take longer to process, the dictionaries are usually lists of the short, common words in the language such as pronouns, articles and prepositions. When trying to identify the language of an unknown document, each list is consulted and the document is declared to be of the language possessing the most matching words. Ingle (1976) creates a method using only one and two letter words. Having been designed for use by librarians it is easily automated. One starts by assuming that the unknown document could be in any language. Then the one and two letter words are examined to see which language uses them. Languages that do not use the word are eliminated. The process continues until all the one and two letter words have been examined or until all but a single language is identified.

A number of attributes are compared in Table B.1, which marks them as A to H. Then, the corresponding marks of either true (\(\checkmark\)) or false (X) is given. The first attribute, A, refers to whether this particular research is focusing on a non-European language such as Arabic, Chinese, Hindi, etc. rather than English, French, Spanish, etc. B indicates whether particular work handles those out of scope languages. In other

\(^1\)Due to the page limitation, this table use alphabet from A to H, L and S for particular annotation, it is not related to the symbol used on the mathematic equation of this work
Table B.1: A comparison of the attributes of previous work done on web page language identification (note: A - is utilizes a non Roman language? B - is manages out of scope languages? C - is utilizes sparse data? D - is manages multilingual documents? E - is utilizes standard corpora? F - is utilizes standard evaluations? G - is utilizes stemming & stopping? H - is utilize finer granularity only? L - linguistic examples; S - statistic examples)

<table>
<thead>
<tr>
<th>Category</th>
<th>Label</th>
<th>Author(s)</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical</td>
<td>S1</td>
<td>Hakkinen and Tian (2001)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>Ahmed et al. (2004)</td>
<td>X</td>
<td>X</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>Biemann and Teresniak (2005)</td>
<td>X</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>S4</td>
<td>James and Sankaran (2005)</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>S5</td>
<td>Martins and Silva (2005)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>S6</td>
<td>McNamee (2005)</td>
<td>X</td>
<td>X</td>
<td>√</td>
<td>X</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>S7</td>
<td>Takcı and Soğukpınar (2005)</td>
<td>X</td>
<td>X</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>√</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>S8</td>
<td>Windisch and Csink (2005)</td>
<td>X</td>
<td>X</td>
<td>√</td>
<td>X</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>S9</td>
<td>Artemenko et al. (2006)</td>
<td>√</td>
<td>X</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>S10</td>
<td>Botha et al. (2006)</td>
<td>√</td>
<td>X</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td>√</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>S12</td>
<td>Hammarstrom (2007)</td>
<td>X</td>
<td>X</td>
<td>√</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>S13</td>
<td>Mohamed Ould (2007)</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>√</td>
<td>√</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>S14</td>
<td>Sagiroglu et al. (2007)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>S15</td>
<td>Selamat and Ibnu Subroto (2007)</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>S16</td>
<td>Selamat and Lee (2008)</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>√</td>
<td>√</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Linguistics</td>
<td>L1</td>
<td>Benedetto et al. (2002)</td>
<td>√</td>
<td>X</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>L3</td>
<td>Xafopoulos et al. (2004)</td>
<td>X</td>
<td>X</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>L4</td>
<td>Tran and Sharma (2005)</td>
<td>X</td>
<td>X</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>L5</td>
<td>Gorkem Ozbek and Yeh (2006)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>√</td>
<td>X</td>
<td>√</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>L6</td>
<td>Ljubesic et al. (2007)</td>
<td>X</td>
<td>X</td>
<td>√</td>
<td>X</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

words, it determines if the undefined languages are beyond the predefined language in the learning process. C is designed to identify a particular language when only a sample of 50 / 100 / 200 characters or words is accessible. D is to check if the particular research can handle multi-language documents. E and F are use to verify the validity of particular research by utilizing available evaluation corpora and standard
performance evaluations such as precision, recall, $F_1$ measurement, cross validation instead of accuracy only. $G$ is used to determine whether the processes stemming or stopping have been applied or not on particular research. Such processes are not applicable to the preprocessing stage of language identification due to the fact that a predefined language for a particular document has not yet been identified. This will directly affect the following processes of language identification and the results of the identification. The final attribute $H$ is employed to determine if the particular work is utilizing a finer granularity level (or letter representation) of a web document instead of $N$-grams or words.

Table B.2: A comparison of the strength and weakness of previous work done on web page language identification (note: label in this table is referring to Table B.1; L - linguistic example; S - statistic example; Feature Selection Method (FESE)

<table>
<thead>
<tr>
<th>Label</th>
<th>Features</th>
<th>FESE</th>
<th>Identifier</th>
<th>Strength &amp; Weakness</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>$N$-grams, lexicon, letter</td>
<td>Maximum Likelihood (ML), Bayes’s rule</td>
<td>Decision tree</td>
<td>The lexicon source is not clear and the flexibility of decision tree on text-based language identification is another problem.</td>
</tr>
<tr>
<td>S2</td>
<td>$N$-grams</td>
<td>$N$-grams profile</td>
<td>Rank-order statistics, cumulative frequency addition, naïve Bayesian classifier</td>
<td>Data set is too small, lack of explanation on the preprocessing data set.</td>
</tr>
<tr>
<td>S3</td>
<td>Words, sentences</td>
<td>Sentence-based co-occurrence graphs (Poison distribution)</td>
<td>Babylonian confusion - unsupervised learning</td>
<td>Lack of explanation on the source of data set and how to produce bilingual corpus from monolingual corpus.</td>
</tr>
</tbody>
</table>

Continued on next page
Table B.2 -- continued from previous page

<table>
<thead>
<tr>
<th>Label</th>
<th>Features</th>
<th>FESE</th>
<th>Identifier</th>
<th>Strength &amp; Weakness</th>
</tr>
</thead>
<tbody>
<tr>
<td>S4</td>
<td><em>N</em>-grams</td>
<td>PCA</td>
<td>Multivariate analysis, compression, n-grams profile</td>
<td>The proposed method lack of explanation on how to apply PCA in the feature reduction and source of data, this is a critical part of getting promising result.</td>
</tr>
<tr>
<td>S5</td>
<td><em>N</em>-grams</td>
<td><em>N</em>-grams profile</td>
<td>Rank-order statistics</td>
<td>No significant contribution in this work especially the algorithm used.</td>
</tr>
<tr>
<td>S6</td>
<td>Words</td>
<td>-</td>
<td>-</td>
<td>The elaboration of feature selection method and identification method is not enough.</td>
</tr>
<tr>
<td>S7</td>
<td>Letters</td>
<td>Centroid values and text score</td>
<td>Maximum language score</td>
<td>The data source is not clear and problem occurred in classification if same or zero language score found.</td>
</tr>
<tr>
<td>S8</td>
<td>Word length, character ratios</td>
<td>-</td>
<td>-</td>
<td>This work provides an in depth view on the features and size of documents, but it lacks an explanation of feature selection and identification algorithms.</td>
</tr>
<tr>
<td>S9</td>
<td>Words</td>
<td>-</td>
<td>Vector space, out-of-space, Bayesian classification, word based method</td>
<td>The explanation is not extensive enough for the whole process of web page language identification.</td>
</tr>
<tr>
<td>S10</td>
<td><em>N</em>-grams</td>
<td>-</td>
<td>SVM and likelihood classifier</td>
<td>Lack of explanation on the algorithm of likelihood based classifiers and SVM.</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Label</th>
<th>Features</th>
<th>FESE</th>
<th>Identifier</th>
<th>Strength &amp; Weakness</th>
</tr>
</thead>
<tbody>
<tr>
<td>S11</td>
<td>N-grams</td>
<td>Discriminant ability, principal component analysis</td>
<td>Quadratic discrimination score</td>
<td>The training phase is time-consuming due to the use of $N$-grams, with $N$ ranging from 2 to 8.</td>
</tr>
<tr>
<td>S12</td>
<td>Words</td>
<td>Word emission probability (frequency dictionary)</td>
<td>Fine-grained model</td>
<td>The proposed method is time consuming compared to state-of-the-art methods because of its keyword repository. Furthermore, the proposed method does not show how to overcome the issue of similar words that are found in both languages.</td>
</tr>
<tr>
<td>S13</td>
<td>Keywords</td>
<td>SVM and rough set theory</td>
<td>Presents problems on stopping and stemming, it is very strange to apply such process to an undefined language document.</td>
<td></td>
</tr>
<tr>
<td>S14</td>
<td>Letters</td>
<td>Letter frequency</td>
<td>Neural networks</td>
<td>The result might be different if more languages were included in the system, due to the use of letters only.</td>
</tr>
<tr>
<td>S15</td>
<td>Letters</td>
<td>Vector space model and KNN</td>
<td>SVM</td>
<td>This work lacks an explanation on feature selection and the data sources are not clear.</td>
</tr>
<tr>
<td>S16</td>
<td>Words</td>
<td>Entropy and Class Profile-Based Approach (CPBF)</td>
<td>Neural networks</td>
<td>The experiment does not explain how to do a stopping process for language identification and the usability of feature selection being applied.</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Label</th>
<th>Features</th>
<th>FESE</th>
<th>Identifier</th>
<th>Strength &amp; Weakness</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>Character sequence, word frequency</td>
<td>Relative entropy</td>
<td>Zipf’s law</td>
<td>The explanation is not extensive enough for data sources and experiments.</td>
</tr>
<tr>
<td>L2</td>
<td>Closed word classes</td>
<td>Percentage of words</td>
<td>Decision tree</td>
<td>Lack of explanations of the algorithm and measurements used, flexibility of the closed word classes when domain or time change.</td>
</tr>
<tr>
<td>L3</td>
<td>Characters sequence, N-grams</td>
<td>ML and Maximum Mutual Information (MMI)</td>
<td>Discrete HMMs</td>
<td>The method lacks of implementation on standard measurements but the method does not require any linguistic or priori knowledge of the corpus under investigation.</td>
</tr>
<tr>
<td>L4</td>
<td>Letters, words</td>
<td>Markov chain representation</td>
<td>Maximum probability</td>
<td>Lack of explanation of the preprocessing stage of source data from the Internet.</td>
</tr>
<tr>
<td>L5</td>
<td>Morpheme</td>
<td>Creutz algorithm</td>
<td>Naïve-Bayes classifier</td>
<td>The explanation of algorithm and methodology is not extensive.</td>
</tr>
<tr>
<td>L6</td>
<td>Words</td>
<td>Most frequent words and characters elimination rule</td>
<td>Second order Markov model</td>
<td>The method lacks comparability with a state-of-the-art method in web page language identification and construction of a specific library requiring high computation cost.</td>
</tr>
</tbody>
</table>
Table B.2 shows a comparison of features, feature selection and identification method on web page language identification, including the strength and weakness of each proposed method. In general, most of the works lack an explanation of data sources, a preprocessing stage and a comparison with a state of the art web page language identification method. Based on the comparison, many of the works argue that tri-grams is the best feature selection method on web page language identification due to the reliability of tri-grams on dealing with grammatical errors in text documents.
## APPENDIX C

### Critical Values of T-test

Table C.1: Critical values of T-test

<table>
<thead>
<tr>
<th>DoF</th>
<th>0.1</th>
<th>0.05</th>
<th>0.03</th>
<th>0.01</th>
<th>$\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.08</td>
<td>6.31</td>
<td>12.71</td>
<td>31.82</td>
<td>63.66</td>
</tr>
<tr>
<td>2</td>
<td>1.89</td>
<td>2.92</td>
<td>4.3</td>
<td>6.97</td>
<td>9.93</td>
</tr>
<tr>
<td>3</td>
<td>1.64</td>
<td>2.35</td>
<td>3.18</td>
<td>4.54</td>
<td>5.84</td>
</tr>
<tr>
<td>4</td>
<td>1.53</td>
<td>2.13</td>
<td>2.78</td>
<td>3.75</td>
<td>4.6</td>
</tr>
<tr>
<td>5</td>
<td>1.48</td>
<td>2.02</td>
<td>2.57</td>
<td>3.37</td>
<td>4.03</td>
</tr>
<tr>
<td>6</td>
<td>1.44</td>
<td>1.94</td>
<td>2.45</td>
<td>3.14</td>
<td>3.71</td>
</tr>
<tr>
<td>7</td>
<td>1.42</td>
<td>1.9</td>
<td>2.37</td>
<td>3</td>
<td>3.5</td>
</tr>
<tr>
<td>8</td>
<td>1.4</td>
<td>1.86</td>
<td>2.31</td>
<td>2.9</td>
<td>3.36</td>
</tr>
<tr>
<td>9</td>
<td>1.38</td>
<td>1.83</td>
<td>2.26</td>
<td>2.82</td>
<td>3.25</td>
</tr>
<tr>
<td>10</td>
<td>1.37</td>
<td>1.81</td>
<td>2.23</td>
<td>2.76</td>
<td>3.17</td>
</tr>
<tr>
<td>11</td>
<td>1.36</td>
<td>1.8</td>
<td>2.2</td>
<td>2.72</td>
<td>3.11</td>
</tr>
<tr>
<td>12</td>
<td>1.36</td>
<td>1.78</td>
<td>2.18</td>
<td>2.68</td>
<td>3.06</td>
</tr>
<tr>
<td>13</td>
<td>1.35</td>
<td>1.77</td>
<td>2.16</td>
<td>2.65</td>
<td>3.01</td>
</tr>
<tr>
<td>14</td>
<td>1.35</td>
<td>1.76</td>
<td>2.15</td>
<td>2.62</td>
<td>2.98</td>
</tr>
<tr>
<td>15</td>
<td>1.34</td>
<td>1.75</td>
<td>2.13</td>
<td>2.6</td>
<td>2.95</td>
</tr>
<tr>
<td>16</td>
<td>1.34</td>
<td>1.75</td>
<td>2.12</td>
<td>2.58</td>
<td>2.92</td>
</tr>
<tr>
<td>17</td>
<td>1.33</td>
<td>1.74</td>
<td>2.11</td>
<td>2.57</td>
<td>2.9</td>
</tr>
<tr>
<td>18</td>
<td>1.33</td>
<td>1.73</td>
<td>2.1</td>
<td>2.55</td>
<td>2.88</td>
</tr>
<tr>
<td>19</td>
<td>1.33</td>
<td>1.73</td>
<td>2.09</td>
<td>2.54</td>
<td>2.86</td>
</tr>
<tr>
<td>20</td>
<td>1.33</td>
<td>1.73</td>
<td>2.09</td>
<td>2.53</td>
<td>2.85</td>
</tr>
<tr>
<td>21</td>
<td>1.32</td>
<td>1.72</td>
<td>2.08</td>
<td>2.52</td>
<td>2.83</td>
</tr>
<tr>
<td>22</td>
<td>1.32</td>
<td>1.72</td>
<td>2.07</td>
<td>2.51</td>
<td>2.82</td>
</tr>
<tr>
<td>23</td>
<td>1.32</td>
<td>1.71</td>
<td>2.07</td>
<td>2.5</td>
<td>2.81</td>
</tr>
<tr>
<td>24</td>
<td>1.32</td>
<td>1.71</td>
<td>2.06</td>
<td>2.49</td>
<td>2.8</td>
</tr>
<tr>
<td>25</td>
<td>1.32</td>
<td>1.71</td>
<td>2.06</td>
<td>2.49</td>
<td>2.79</td>
</tr>
<tr>
<td>26</td>
<td>1.32</td>
<td>1.71</td>
<td>2.06</td>
<td>2.48</td>
<td>2.78</td>
</tr>
<tr>
<td>27</td>
<td>1.31</td>
<td>1.7</td>
<td>2.05</td>
<td>2.47</td>
<td>2.77</td>
</tr>
<tr>
<td>28</td>
<td>1.31</td>
<td>1.7</td>
<td>2.05</td>
<td>2.47</td>
<td>2.76</td>
</tr>
<tr>
<td>29</td>
<td>1.31</td>
<td>1.7</td>
<td>2.05</td>
<td>2.46</td>
<td>2.76</td>
</tr>
<tr>
<td>30</td>
<td>1.31</td>
<td>1.7</td>
<td>2.04</td>
<td>2.46</td>
<td>2.75</td>
</tr>
<tr>
<td>40</td>
<td>1.3</td>
<td>1.68</td>
<td>2.02</td>
<td>2.42</td>
<td>2.7</td>
</tr>
<tr>
<td>60</td>
<td>1.3</td>
<td>1.67</td>
<td>2</td>
<td>2.39</td>
<td>2.66</td>
</tr>
<tr>
<td>120</td>
<td>1.29</td>
<td>1.66</td>
<td>1.98</td>
<td>2.54</td>
<td>2.62</td>
</tr>
<tr>
<td>$\infty$</td>
<td>1.28</td>
<td>1.65</td>
<td>1.96</td>
<td>2.33</td>
<td>2.76</td>
</tr>
</tbody>
</table>
Table D.1: Features produced by different feature selection methods on Roman script web pages

<table>
<thead>
<tr>
<th>Method</th>
<th>Features Found</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N-grams</strong></td>
<td></td>
</tr>
<tr>
<td>Azeri: _ , i, a, n, r, l, d, s, i, m, t, e, y, n_, b, ...</td>
<td></td>
</tr>
<tr>
<td>English: _ e, a, i, n, r, t, s, o, l, d, h, m, u, e_, ...</td>
<td></td>
</tr>
<tr>
<td>Indonesia: _ , a, e, i, n, r, t, s, o, l, d, m, u, h, n_, ...</td>
<td></td>
</tr>
<tr>
<td>Serbian: _ , a, e, i, n, r, t, o, s, d, l, u, m, k, h, ...</td>
<td></td>
</tr>
<tr>
<td>Somali: _ , a, e, i, n, o, r, t, s, d, l, u, m, k, a_, ...</td>
<td></td>
</tr>
<tr>
<td>Spanish: _ , a, e, i, n, o, r, s, t, d, l, u, a_, m, e_, ...</td>
<td></td>
</tr>
<tr>
<td>Turkish: _ , a, e, i, n, r, o, s, t, l, d, u, m, a_, e_, ...</td>
<td></td>
</tr>
<tr>
<td>Vietnamese: _ , a, e, n, i, r, o, s, t, l, d, u, a_, m, h, ...</td>
<td></td>
</tr>
<tr>
<td><strong>Entropy</strong></td>
<td></td>
</tr>
<tr>
<td>e, a, o, s, n, r, i, l, d, t, c, u, h, m, g, p, _ , Z, v, x, {, }, W, X, _ , £, w, o, ū, Ā, ā, Ň, ā, Q, è, ë, y, q, Ō, Ū, Ū, Ū, Є, V, [], Y, i, İ</td>
<td></td>
</tr>
<tr>
<td><strong>Letter Weighting</strong></td>
<td>q, d, r, b, T, c, e, p, g, a, j, K, o, v, S, D, w, G, f, h, c, s, A, L, t, l, k, i, y, M, n, Q, B, u, d</td>
</tr>
<tr>
<td><strong>Simplified Entropy</strong></td>
<td>Azeri: i, a, n, r, l, d, s, m, t, e, y, b, k, u, o, ü, q, z, v, ş, h, c, ç, ö, x, p, f, g, B, A, ğ, M, T, P, C, G, İ, S, O, K, R, H, D, N, Q, Ə, L, Y, Ş, E, F, V, j, X, Ō, Ç;</td>
</tr>
<tr>
<td>English: e, t, a, o, i, n, r, s, h, d, l, c, u, m, f, g, p, w, y, b, v, k, B, M, S, T, C, I, W, N, E, P, A, H, G, D, x,</td>
<td>, L, j, K, U, F, O, _ , q, R, X, ©, J, z, V, Y, £, Q, Z, [;</td>
</tr>
<tr>
<td>Indonesia: a, n, e, i, r, t, u, k, m, s, d, g, l, p, b, h, o, y, j, c, B, P, w, S, M, I, A, T, D, C, K, G,</td>
<td>, f, J, N, R, v, H, L, W, O, z, F, U, E, Y, V, Z, Q, q, x, X, [; ] , £;</td>
</tr>
<tr>
<td>Serbian: a, i, e, o, n, r, s, j, u, t, d, k, v, l, m, p, z, b, g, c, ş, č, ž, h, č, S, B, M, N, P, T, O, K, f, G,</td>
<td>, C, U, d, D, A, I, R, w, V, E, L, H, J, Z, F, Ć, Š, Ž, D, W, X;</td>
</tr>
<tr>
<td>Somali: a, i, o, d, e, y, n, u, l, h, s, k, r, m, g, b, t, w, x, c, q, M, S, B, f, C, G, A, j, W, D,</td>
<td>, T, K, I, H, p, X, J, E, Q, N, F, L, R, Y, P, O, U, z, v, Z, V, é, ê, i,</td>
</tr>
<tr>
<td>Spanish: e, a, o, s, n, r, i, l, t, d, c, u, p, m, g, b, v, h, S, A, E, y, C, f, i, ó, i, q, M, á, T, P, O, D, R, B, L, N, é, j, U, , z, x, V,</td>
<td>, w, G, H, Ň, F, Ľ, ú, k, Q, W, Y;</td>
</tr>
<tr>
<td>Turkish: a, e, c, i, n, l, r, r, d, k, t, s, y, u, m, o, b, ü, ş, g, z, ğ, h, č, c, v, ö, p, B, f, A, T, S, Ī, C, M, H, D, K,</td>
<td>, G, P, O, E, R, N, Y, F, Ş, I, j, L, Č, U, J, Ô, w, V;</td>
</tr>
<tr>
<td>Vietnamese: n, h, i, a, t, c, g, u, o, d, r, m, ŏ, v, l, ê, b, s, y, e, p, ā, σ, k, ā, d, T, à, C, q, ā, B, M, N, ŏ, H, P, G, x, V, A, K, ā, S, L, ĩ, i, l, D, Ō, Q, D, Z, ū, ŏ, w</td>
<td></td>
</tr>
</tbody>
</table>
Examples of the Arabic Script Letter Distribution

(a) Arabic patterns based on character frequency

(b) The average of Arabic pattern based on characters frequency (extension to Figure E.1a)

Figure E.1: Arabic script letter distribution on web pages
APPENDIX F
List of Publication and Recognition

Journal


Conference Paper


**Book Chapter**


**Patent**

The portion of this research has been certified under Intellectual Property Corporation of Malaysia with the application number PI 20084793 on 26 November 2008.

**Award**

The portion of this research has been developed as a commercial product namely “ReadMe: Multilingual Web Page Identification” and it is awarded silver medal on the 11th Industrial Art & Technology Exhibition (INATEX), 2009 at UTM.