# Contents

List of Figures xi-xii

List of Tables xiii

Acknowledgments xiv-xv

1. Introduction to Web Mining 1

1.1 Introduction 1

1.2 Web Mining 4

1.2.1 Types of Web Data 7

1.2.1.1 Web Content Data 8

1.2.1.2 Web Structure Data 8

1.2.1.3 Web Usage Data 8

1.2.2 Characteristics of Web Data 9

1.2.3 Taxonomy of Web Mining 10

1.2.3.1 Web Content Mining 11

1.2.3.2 Web Usage Mining 12

1.2.3.3 Web Structure Mining 12

1.3 Web Content Mining 14

1.4 Applications of Web Mining 15

1.4.1 Predicting the Future Using Web Data 15

1.4.2 User Activity Identification and Tracking in E-Commerce 16

1.4.3 Google Search 17

1.4.4 Detection of Fake Reviews and Reviewers 17

1.4.5 Fraud Detection 18

1.5 Identification of the Research Problem 18

1.6 Objectives of the Research 21

1.7 Organization of the Thesis 22
List of Figures

Figure 1.1 Overall process of web mining 7
Figure 1.2 Web data types 8
Figure 1.3 Web mining taxonomy 11
Figure 1.4 Web structure graph 14
Figure 1.5 A structured webpage containing four data records 20
Figure 2.1 Related areas of web mining 25
Figure 2.2 Data mining, the core of KDD process 26
Figure 2.3 Schematic diagram of an information retrieval system 28
Figure 2.4 IE on top of IR 29
Figure 2.5 Data extraction approaches based on wrapper generation 30
Figure 2.6 Data extraction approaches based on degree of automation 32
Figure 2.7 Overall comparisons of various approaches 43
Figure 3.1 A datablock and its visual signal 46
Figure 3.2 Snapshot of an amazon store containing three data records 47
Figure 3.3 Block diagram of ESD method 48
Figure 3.4 (a) Different blocks of a webpage 50
Figure 3.4 (b) Corresponding Specification of Vision-Based Content 50
Figure 3.5 Block tree of a webpage 53
Figure 3.6 Main module of eVIPS 54
Figure 3.7 Get blocks procedure 54
Figure 3.8 Get separators procedure 55
Figure 3.9 Hierarchical Vision-Based Content Structure Procedure 55
Figure 3.10 Three data records at different levels 59
### List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Classification of retrieval and mining techniques</td>
<td>30</td>
</tr>
<tr>
<td>2.2</td>
<td>List of reviewed algorithms</td>
<td>34</td>
</tr>
<tr>
<td>3.1</td>
<td>Seven heuristic rules</td>
<td>52</td>
</tr>
<tr>
<td>3.2</td>
<td>Experimental results of ESD method</td>
<td>64</td>
</tr>
<tr>
<td>4.1</td>
<td>Experimental results of the our proposed approach</td>
<td>83</td>
</tr>
<tr>
<td>5.1</td>
<td>Various Measures</td>
<td>90</td>
</tr>
<tr>
<td>5.2</td>
<td>Confusion Matrix for Decision Tree Analysis</td>
<td>91</td>
</tr>
<tr>
<td>5.3</td>
<td>Confusion Matrix for Artificial Neural Network Analysis</td>
<td>93</td>
</tr>
</tbody>
</table>