CONTENTS

<table>
<thead>
<tr>
<th>Title</th>
<th>PAGE NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>DECLARATION</td>
<td>ii</td>
</tr>
<tr>
<td>BONAFIDE CERTIFICATE</td>
<td>iii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>iv</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENT</td>
<td>vi</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>xiv</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xv</td>
</tr>
<tr>
<td>LIST OF SYMBOLS AND ABBREVIATIONS</td>
<td>xix</td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2. Research Background</td>
<td>2</td>
</tr>
<tr>
<td>1.2.1. Web Mining</td>
<td>4</td>
</tr>
<tr>
<td>1.2.2. Types of Web Mining</td>
<td>5</td>
</tr>
<tr>
<td>1.2.3. Preliminaries in Web Usage Mining</td>
<td>8</td>
</tr>
<tr>
<td>1.2.4. Forms of Web Usage Mining</td>
<td>9</td>
</tr>
<tr>
<td>Data Source</td>
<td></td>
</tr>
<tr>
<td>1.2.5. Other Level Formatting</td>
<td>9</td>
</tr>
<tr>
<td>1.2.6. Browser Level Collection</td>
<td>10</td>
</tr>
<tr>
<td>1.2.7. Server Level Collection</td>
<td>11</td>
</tr>
<tr>
<td>1.2.8. Proxy Level Collection</td>
<td>11</td>
</tr>
</tbody>
</table>
1.2.9. Web Usage Terms 13
1.2.10. Steps And Techniques In Web Usage Mining 15

1.3. Motivation of the Research 22
1.4. Problem Specification 23
1.5. Objective of the Research 24
1.6. Contribution of the Research 24
1.7. Thesis Organization 27
1.8. Summary 28

2. REVIEW OF LITERATURE 29

2.1. Introduction 29
2.2. Web Mining 29
2.3. Weblog File 32
2.4. Web Usage Mining 33
   2.4.1. Web Personalization 35
   2.4.2. General Site Improvement 37
   2.4.3. Business Intelligence 39
   2.4.4. Site Modification 40
   2.4.5. Usage Characteristics 40
2.5. Data Preprocessing 41
2.6 Navigation Oriented Heuristics 43
2.7. Identification of Potential User 44
2.8 User Navigation Pattern Mining 45
   2.8.1. Statistical Pattern Mining 45
   2.8.2. Association Mining 47
2.9. Classification and Clustering Techniques 50
2.10. Summary 60
3. **PREPROCESS THE SERVER WEBLOG FILES AND TO CLUSTER THE SESSION**

3.1. Introduction
3.2. Proposed Methodology
   3.2.1. Weblog Files
3.3. Preprocessing
   3.3.1. Usage Preprocessing
   3.3.2. Content Preprocessing
   3.3.3. Structure Preprocessing
3.4. Proposed Preprocessing Algorithm
   3.4.1. Cleaning of Web Data
   3.4.2. Incomplete Path Filling
3.5. Access History List
3.6. User and Session Identification
   3.6.1. User Identification
   3.6.2. Session Identification
3.7. Immediate Link Analysis (IMMLINK)
3.8. Updating AHL (UPDATEAHL)
3.9. Backward Reference And Direct Reference Length (BRFDL)
   3.9.1. Visited Pages
   3.9.2. USIDALG with Access History List (AHL) Examples
3.10. Clustering Session
   3.10.1. Example of Clustering Session
3.11. Experimental Result
3.12. Summary
4. CLASSIFICATION ALGORITHMS TO IDENTIFY THE POTENTIAL USERS IN THE WEBLOG FILES

4.1. Introduction 100
4.2. Proposed Methodology 100
4.3. Classification 102
   4.3.1. Decision Trees 103
   4.3.2. C4.5 Algorithm 103
   4.3.3. Improved C4.5 Algorithm 104
   4.3.4. Information Gain 106
   4.3.5. Gain Ratio Criterion 107
   4.3.6. Decision Tree Generation Algorithm 108
4.4. Interest Measure 111
4.5. Hybrid method of Improved C4.5 and Interest Measure 111
4.6. Summary 117

5. PREDICTION MODEL 118

5.1. Introduction 118
5.2. Proposed Ensemble Clustering Classification 119
5.3. Ensemble Selection 120
5.4. Ensemble Clustering 122
   5.4.1. Ant Based Algorithm 123
   5.4.2. Graph Partitioned Algorithm 125
   5.4.3. Pairwise Nearest Neighbor 128
   5.4.4. Improved Pair Wise Nearest Neighbor Method 129
5.5. Ensemble Classification 130
5.5.1. Maximum Likelihood Classification 132
5.5.2. Longest Common Sequence Classification
   Algorithm 133
5.5.3. Markov Model 136
5.6. Majority Voting 139
5.7. Experimental Result 140
   5.7.1. Performance of Clustering Based 141
5.8. Summary 143

6 RESULTS AND DISCUSSION 144
6.1 Introduction 144
6.2 Experimental Results 145
   6.2.1. Preprocessing Results 145
   6.2.2. Identification of Potential Users 146
   6.2.3. Performance Measure 147
      6.2.3.1. Performance of Clustering Based
               Maximum Likelihood Classifier 148
      6.2.3.2. Performance of Clustering Based
               Longest Common Sequence Algorithm 150
      6.2.3.3. Performance of Clustering Based
               Markov Model 152
      6.2.3.4. Performance of Clustering Based
               Ensemble Clustering Based
               Ensemble Prediction 154
6.3. Summary 157
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONCLUSION AND FUTURE ENHANCEMENT</td>
<td>158</td>
</tr>
<tr>
<td>7.1. Conclusion</td>
<td>158</td>
</tr>
<tr>
<td>7.2. Future Enhancement</td>
<td>161</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>163</td>
</tr>
<tr>
<td>LIST OF PUBLICATIONS</td>
<td>180</td>
</tr>
<tr>
<td>CURRICULUM VITAE</td>
<td>181</td>
</tr>
</tbody>
</table>